MINISTÈRE DE L'**A**GRICULTURE
ECOLE **N**ATIONALE **S**UPÉRIEURE **A**GRONOMIQUE DE **M**ONTPELLIER

THESE

présentée à l'Ecole Nationale Supérieure Agronomique de Montpellier
Pour obtenir le DIPLOME DE DOCTORAT

DISCIPLINE : *Biologie de l'Evolution*
FORMATION DOCTORALE : *Biologie de l'évolution et écologie*
ECOLE DOCTORALE : *Biologie des systèmes intégrés, Agronomie – Environnement*
LABORATOIRE : *CEFE-CNRS, Génétique et dynamique des populations*

DYNAMIQUE DE L'ADAPTATION :
Théorie et confrontation aux données expérimentales

par

Guillaume MARTIN

Soutenue le 9 Novembre 2005 devant le Jury composé de:

Mme Marie-Laure NAVAS, Professeur ENSAM	Examinateur
M. David WAXMAN, Professeur, Université du Sussex (Angleterre)	Rapporteur
M. Santiago F. ELENA, Directeur de Recherche CSIC Valencia (Espagne)	Rapporteur
M. Claude RISPE, Directeur de Recherche INRA Rennes	Examinateur
M. Philippe JARNE, Directeur de recherche CNRS Montpellier	Directeur de thèse
M. Thomas LENORMAND, Chargé de Recherche CNRS Montpellier	Co-directeur de thèse

Résumé :
L'un des buts ultimes de la biologie évolutive est de comprendre la dynamique de l'adaptation et les facteurs qui l'affectent. Toutefois, malgré de récents développements, les théories de l'adaptation restent cantonnées à des prédictions qualitatives. L'une des raisons en est que l'on sait mal prédire la relation entre mutation et valeur sélective (fitness). Durant cette thèse, j'ai développé un modèle permettant une telle prédiction. J'ai ensuite validé ce modèle à partir de données empiriques existantes chez des espèces modèles. Cette approche a permis de dégager des tendances générales pour la distribution de l'effet des mutations sur la fitness, et pour sa variation entre espèces et entre environnements. Elle a également permis de proposer des prédictions testables pour la vitesse de l'adaptation dans les études d'évolution expérimentale en milieux contrôlés.

J'ai aussi développé des protocoles expérimentaux pour tester mon modèle sur le crustacé Artemia. Bien qu'encore à l'état de préliminaire, cette dernière approche pourrait permettre de tester les théories de l'adaptation au laboratoire mais aussi *in natura*.

Enfin, j'ai étudié l'impact de deux facteurs limitant l'adaptation : la dérive génétique et le système de reproduction, ainsi que l'effet de leur interaction sur la vitesse de l'adaptation. Cette approche a permis de montrer comment, dans des populations soumises à dérive génétique (y compris en populations structurées), la recombinaison permet d'augmenter significativement la vitesse d'adaptation. La reproduction sexuée peut en retour être favorisée dans ce contexte.

Mots-clé : distribution de l'effet des mutations – paysages adaptatifs – recombinaison – réponse à la sélection - Artemia

Summary :
One of the main goals of evolutionary biology is to understand the dynamics of adaptation and the factors that influence it. However, despite recent developments, theories of adaptation remain limited to qualitative predictions. One reason for this is a lack of a general analytic approach to model the relationship between mutation and fitness. During this PhD, I developed such an approach. I then validated the model using the existing empirical data in various model species. This approach allowed to detect some general trends in the distribution of mutation fitness effects, its variation across species and across environments. It also allowed to make testable predictions for the speed of adaptation to a new environment in evolution experiments in controlled conditions.

I also developed protocols to test the model in the crustacean Artemia. Although still in a preliminary phase, I suggest that this approach could eventually allow to test theories of adaptation both in the laboratory and *in natura*.

Finally, I studied the influence of two factors limiting adaptation: genetic drift, and the genetic system, including the effect of their interaction on the speed of adaptation. In particular, I showed, using an analytic model, that in subdivided populations, recombination allows to increase the speed of adaptation, and that sexual reproduction can then in turn be favoured by this mechanism.

Key words : mutation fitness effect – adaptive landscapes – recombination – response to selection - Artemia

Laboratoire d'accueil :
Equipe Génétique et dynamique des populations, Centre d'Ecologie Fonctionnelle et Evolutive CEFE-CNRS UMR 5175, 1919, Route de Mende 34 293 Montpellier Cedex 5

Remerciements :

5 ans à Montpellier (tudieu ça me rajeunit pas !), et beaucoup de plaisir, scientifique et humain. Cette thèse n'aurait pas pu se faire sans l'aide de pas mal de monde, j'espère ne pas (trop ?) en oublier:

Plusieurs étudiants m'ont beaucoup aidé au laboratoire. Ils ont gentiment accepté de se lancer dans des expériences parfois plus que préliminaires, des bricolages en tous genres, des discussions théoriques « complexes » qui ont bien fait avancer mes modèles, et ils ont mis l'ambiance au milieu des crevettes, ou dans les marais. Dans l'ordre de passage: Pascal Tillie et les joies du karyotypage, Emilie Lefevre et les niches écologiques, Raquel Calatayud qui a beaucoup contribué à produire la souche LAPX qui a été la base de mon travail expérimental, et nous a appris à utiliser LE carottier (gracias tambien por el calendario!), Luis Chevin qui a fait toute l'analyse des distributions de cystes d'Aigues-Mortes (et m'a fait découvrir Desmond Decker, yessai), Fanny Ramel (Madame Bricolage) qui a extrait une quantité impressionnante de cystes de sédiments pas trop ragoûtants, et m'a beaucoup aidé pour construire un système d'élevage à grande échelle… Merci aussi à Serge Muller (Equipe palynologie UM II) pour de joyeuses sorties terrain (sous la pluie), et ses conseils pour l'étude des carottes (pas celles qui se mangent, les autres).

Christian, David et Jérémie du TE : avant de venir au CEFE, je ne savais pas tenir un tournevis (ouais bon, je sais toujours pas trop..), l'ensemble des mises en place expérimentales sur *Artemia* n'ont été possibles que grâce à leur aide et à leurs bonnes idées.

Chantal, France et Marie-Pierre m'ont donné plein de conseils et de produits chimiques bizarres (en toute légalité) alors que je suis même pas béhème !

Plusieurs spécialistes (des mines d'or) m'ont mis un pied dans le monde passionnant de l'Artemia. Dino Facca (Salins du Midi) a accepté de prendre sur ses journées chargées au Salin d'Aigues-Mortes pour nous servir de guide dans les marais et nous a fait part (avec le sourire) de ses 40 ans d'expérience du fonctionnement des populations naturelles de la région. Francisco Amat (CSIC Torre la Sal, Espagne) m'a donné plein de cystes de sa collection, et de nombreux tuyaux pour reconnaître, élever et éclore les petites bêtes. Il s'est aussi occupé de la multiplication des cystes anciens. Gilbert Van Stappen (Artemia Reference Center, Ghent) a toujours été dispos pour répondre à mes questions ou pour m'envoyer des cystes de sa fabuleuse collection.

Karine et Anne-Marie font tourner (avec d'autres !) la partie administrative du CEFE avec plein de bonne humeur, et ça aide beaucoup ! Idem pour Jean-Marie à l'ED BSIAE.

Mon travail théorique (et aussi empirique) a été nourri par des discussions avec les chercheurs qui sont passés au CEFE et m'ont donné de précieux conseils, entre autres : Sally Otto (qui a co-encadré mon DEA), Mark Kirkpatrick, Ruth Shaw…sans oublier les montpellierains ! Ceux qui m'ont conseillé pendant et en dehors de mes comités de thèse (Thomas Bataillon, Michel Raymond, Yannis Michalakis) et Isabelle Olivieri qui m'a donné l'opportunité de m'essayer à l'enseignement.

L'équipe Génétique et Dynamique du CEFE a été un environnement très agréable et enrichissant, on y apprend plein de choses: c'est grâce en bonne partie à ceux qui la font tourner. Patrice est toujours là pour expliquer un concept (et le rendre clair comme par magie) et pour apporter de précieux conseils théoriques ou expérimentaux, tout ça avec le sourire. Merci aussi à Patricia pour la bonne humeur et les discussions politiques qui m'ont bien plu, et à Violette pour son aide au TE. Si l'équipe est un endroit où il fait bon vivre (voire même, travailler ??) c'est aussi grâce à ceux qui y passent (sans mauvais esprit, juré). Je pense à tous les thésards qui m'ont vu arriver (frais et innocent), m'ont accueilli dans la bande à coups de pauses café et de repas à la cantine, et que j'ai vus partir (moyennement frais), la larme à l'œil: Marie-France, Nathalie et Anne (les drôles de dames), Cyril (« Maître » ou « Tamanoir Astigmate », mais il préfère

« Maître »), Jesus, Pierre-Yves et Benoît (on se revoit bientôt !). Je pense aussi aux nouveaux (dans l'ordre de fraîcheur et d'innocence ?): Carole, Guillaume, François, Seb et Juan.

Evidement, il en manque deux qui contribuent pas mal à faire que l'Equipe Gendyn tourne et qu'elle est agréable à vivre. Philippe a été un directeur de thèse bien plus qu' « officiel »: il m'a suivi depuis le DEA, m'a conseillé pour les manips comme pour la théorie, et a corrigé mes manuscris et le présant documant (sauf les remaircimant). Il m'a aussi donné confiance en moi, et c'est un beau cadeau à faire à un étudiant.

Enfin, merci à Thomas pour… à peu près TOUT.
Après mon DEA, j'ai choisi de faire cette thèse en bonne partie parce que j'avais envie de continuer à travailler avec toi, et je l'ai jamais regretté. C'est difficile de résumer tout ce que tu m'as apporté (et tu connais bien ma concision naturelle). Je pourrais dire: pas mal de clopes que je te dois encore, mais ça résume un peu trop. Alors je vais faire une petite liste (avec des gros points noirs, là.. comme tu les aimes). Merci pour :

- Ta disponibilité quasi-permanente pour mes diverses questions de toutes sortes pendant ces 5 ans
- Ta façon d'y répondre qui rend l'autre plus intelligent après (j'ai pas dit *intelligent*, j'ai dit *plus* intelligent)
- Ta patience, notamment pendant cette fin de thèse, et pour avoir pas compté ton temps pour m'aider à avancer, dans les articles, la thèse, les manips
- tes intuitions géniales qui se révèlent tellement souvent justes que ça en serait louche
- m'avoir laissé me lâcher et suivi mes idées, tout en m'évitant de virer trop à l'ouest
- m'avoir fait profiter de ta mine intarrissable d'idées et de ta grande culture Evolutive
- m'avoir fait bien marrer (même à 2h du mat' sur des corrections de manuscrit, alors que franchement, y a pas de quoi rire)

Et je suis sûr que j'en oublie parce que tu te débrouilles pour former sans déformer si bien qu'on s'en rend même plus compte...Bref, tu m'as donné envie de faire de la science (alors que pourtant ça sert à rien !), donc merci très beaucoup.

Et puis pour finir, merci aux co-loques (huhuhu Old boys!) qui font qu'on passe 5 ans à Montpellier sans les voir passer (je rentre pas dans les détails, pas de traces, vous serez peut-être connus un jour): Franck, Naig, Mota, Gad, et Olive (solidaires bob!)

Les amis(ies), de Montpellier et d'ailleurs font qu'on se souvient pourquoi il faut finir cette thèse et être en vacances (merci Nath, Jérémie, Etienne, Mathieu et les autres..). Merci aussi à ma maman, mon papa, ma belle-maman, mes frères zé ma sœur, qui ne m'ont pas vu beaucoup ces derniers temps et ne m'en ont pas voulu (trop. j'espère.). Merci aussi à Glougloute d'avoir supporté une bonne partie de cette fin de thèse…

INTRODUCTION

L'une des questions centrales de la biologie évolutive est la compréhension de l'adaptation: sa dynamique avec le temps et les facteurs qui l'affectent. L'adaptation peut être définie de plusieurs manières. La première définition caractérise un état: un trait ou un ensemble de traits sont adaptés s'ils ont évolué en réponse à la sélection naturelle. Cela peut se traduire par la correspondance, à un moment donné, entre certaines caractéristiques phénotypiques d'un organisme et celles de son environnement. On peut aussi définir l'adaptation par le processus dynamique de réponse à la sélection naturelle, se traduisant par une augmentation de la valeur sélective (ou fitness) moyenne. Une population s'adapte à un environnement donné si sa fitness moyenne augmente avec le temps dans cet environnement (Fisher 1930). Il convient également de noter que l'adaptation peut être définie en fonction de différents niveaux de sélection (par ex., un seul sexe dans la population, un gène dans le génome). Par exemple, l'augmentation en fréquence d'un trait soumis à sélection sexuelle traduit une adaptation des mâles leur permettant d'attirer les femelles, bien qu'elle n'entraîne pas une augmentation de la fitness moyenne de la population dans son ensemble, elle peut même la réduire (coût de la sélection sexuelle). Il en va de même pour différents niveaux de sélection (par ex. gènes égoïstes etc.): une adaptation est définie à un niveau de sélection donné.

Quelles que soient les définitions considérées, une des questions centrales en Evolution est la dynamique de l'adaptation et les facteurs qui l'influencent. De nombreux facteurs peuvent affecter/limiter la réponse à la sélection naturelle et donc la dynamique de l'adaptation. On peut les classer en facteurs affectant le niveau de variation disponible à court terme (*par ex.* dérive, mutation, système de reproduction etc.), ou accessibles à plus long terme (contraintes physiques ou développementales), et l'influence du type de sélection sur la réponse attendue (sélection hétérogène, coût de la sélection, conflits entre différents niveaux de sélection), pour revue, voir (Barton and Partridge 2000). Comprendre l'influence des ces différents facteurs selon les espèces ou les environnements est un des buts fondamentaux de la biologie évolutive. Un autre aspect important des théories de l'adaptation est la question du lien entre micro-évolution (le changement évolutif dans un population à l'échelle de plusieurs générations) et macro-évolution (les grandes transitions évolutives et l'apparition de nouvelles espèces). Le débat sur la façon dont ces deux échelles de temps sont reliées n'est pas tranché : on ne sait pas dans quelle mesure les processus micro-évolutifs peuvent rendre compte des changements macro-évolutifs observés dans les séries fossiles (pour une revue de ces questions, voir le numéro spécial (N°112) de Genetica, notamment Arnold et al. 2001; Gingerich 2001; Hendry and Kinnison 2001; Kinnison and Hendry 2001). Pour résoudre ces questions, un pas nécessaire est la compréhension de la dynamique du changement adaptatif micro-évolutif, et donc la construction et la validation de modèles permettant de prédire l'adaptation à court mais aussi à moyen terme.

La compréhension de l'adaptation est aussi un problème appliqué: peut-on prédire l'adaptation des espèces d'importance économique ou médicale, face à une modification de leur environnement. Ces questions peuvent s'appliquer à la dynamique d'un ou plusieurs traits (d'importance agronomique par exemple), comme à l'évolution de la fitness elle-même en réponse à un changement environnemental (évolution de la virulence des pathogènes, rendement et survie des espèces domestiquées, survie des espèces menacées).

La dynamique de l'adaptation est un phénomène très complexe (même à l'échelle micro-évolutive), et son étude est passée par la réduction de cette complexité. Dans cette introduction, je présente un rapide aperçu historique des approches permettant une telle simplification, puis les résultats obtenus dans les cas simples. Ensuite, je présente l'état de nos connaissances empiriques et théoriques sur les processus affectant l'adaptation dans ces cas

simplifiés. Finalement, je présente comment cette thèse s'inscrit dans la recherche d'un modèle satisfaisant pour la dynamique du processus adaptatif.

Comment simplifier le processus adaptatif

L'adaptation implique un processus de sélection naturelle. Sacrifiant à la coutume, je citerai donc Darwin (1859) puis Fisher (1930) qui distinguent trois hypothèses nécessaires et suffisantes pour observer un phénomène de sélection naturelle: (i) variation: existence d'une variation du phénotype dans la population, (ii) sélection: corrélation entre ce phénotype et le nombre de descendants à l'âge de reproduction (fitness), et (iii) héritabilité: corrélation de la valeur phénotypique entre parents et descendants.

La prédiction du taux d'adaptation à un moment donné de l'histoire d'une population dépend d'une connaissance de chacun de ces aspects du changement évolutif :

(i) D'abord, quelle est la distribution des phénotypes sous sélection ? Celle-ci dépend de la variance génétique présente et de celle introduite par les mutations (selon le taux et les effets de celles-ci), et la migration (selon le nombre et la variance génétique des migrants). La distribution des phénotypes dépend aussi de la façon dont la variation génétique se traduit en phénotypes selon l'environnement (plasticité phénotypique). Ces divers effets varient entre différents traits, plus ou moins corrélés au niveau génétique ou phénotypique.

(ii) Ensuite, comment la variation génétique est-elle sélectionnée ? Chaque allèle peut avoir des effets variables sur la fitness selon le background génétique (interactions épistatiques entre loci, relations de dominance entre allèles à un même locus). Dans un environnement donné, la sélection peut être directionnelle, stabilisante autour d'un optimum phénotypique unique ou diversifiante (avec plusieurs optima). Elle peut en outre varier dans l'espace (milieux hétérogènes) ou dans le temps (fluctuations du milieu, coévolutions), et affecter les organismes différemment selon leur stade.

(iii) Enfin, comment cette variance en fitness est-elle héritée d'une génération à l'autre ? La transmission de combinaisons d'allèles dépend du système de reproduction (ségrégation, liaison) et de son effet variable selon les loci. La variation génétique peut être perdue aléatoirement par effet d'échantillonnage (dérive), qui dépend lui-même du fonctionnement démographique des populations et de leur système de reproduction. Au niveau phénotypique, les variations environnementales d'une génération à l'autre peuvent affecter l'héritabilité de chaque trait (lorsque ceux-ci sont plastiques).

Par ailleurs, la prédiction à plus long terme nécessite de savoir comment chacun de ces paramètres varie à une échelle de temps supérieure à la génération. On voit que pour proposer des prédictions quantitatives, il est nécessaire de réduire cette grande complexité inhérente au processus adaptatif. Comme souvent, ce sont les « pères » de la synthèse néo-darwinienne (notamment Fisher et Haldane) qui ont suggéré les principales simplifications possibles, du point de vue théorique. C'est beaucoup plus tard que des approches nouvelles ont permis la même réduction dans des systèmes empiriques, ainsi que de récents développements théoriques.

I.1 SIMPLIFICATION THEORIQUE

D'une manière générale, la plupart des modèles prédictifs se sont limités à des situations simples où la sélection est supposée directionnelle ou stabilisante avec une population non structurée et en environnement constant. Des modèles existent bien sûr pour prendre en compte des situations plus complexes (*par ex.* sélection hétérogène, disruptive, en populations structurées etc.). Nous verrons un exemple d'une telle approche en Chapitre III. Toutefois, ces modèles font d'autres simplifications quant à la base génétique de l'adaptation (se limitant souvent à un ou deux loci) de sorte qu'il est difficile de les utiliser pour prédire l'adaptation à un trait multigénique (voir toutefois les modèles généraux de Barton et Turelli: Barton and Turelli 1991; Barton and Turelli 2004; Turelli and Barton 1990).

La génétique quantitative a fourni le premier modèle prédictif de l'effet de la sélection naturelle, en réduisant le déterminisme génétique complexe des traits quantitatifs au modèle infinitésimal (Fisher 1930): un très grand nombre de loci génétiquement indépendants, chacun de très faible effet. Cette approximation permet de réduire la distribution des phénotypes sous sélection à une Gaussienne dont les paramètres sont mesurables empiriquement. Surtout, elle permet de prédire cette distribution à la génération suivante pour un niveau de sélection connu. Cette théorie a ensuite été étendue à des phénotypes multivariés (plusieurs traits sélectionnés simultanément) par Lande (1979). Ces modèles ont aussi été affinés pour incorporer l'effet de la mutation dans ce contexte (Lande 1980; Turelli 1985). Cette approche est un des plus grands succès de la biologie évolutive, et elle est couramment appliquée en agronomie pour prédire l'évolution des traits soumis à sélection artificielle.

Toutefois, elle présente certaines limites lorsqu'il s'agit de prédire l'évolution de la fitness elle-même, ou de traits de fitness (survie, fécondité etc.), particulièrement à long terme. D'abord les distributions de fitness ne sont en général pas gaussiennes, et sont très biaisées (vers les effets délétères) et asymétriques, de sorte qu'il est difficile de les « rendre gaussiennes » par une transformation. Cela est vrai que l'on considère la variation présente dans les populations naturelles (e.g. chez les virus, Bonhoeffer et al. 2004; Sanjuan et al. 2004b) ou créée par mutation (Garcia-Dorado et al. 1999; Lynch et al. 1999). Cette forte dissymétrie limite grandement l'application des approches de génétique quantitative basées sur la Gaussienne. Ensuite, l'un des buts des théories de l'adaptation est de prédire le changement de fitness sur de nombreuses générations, or la génétique quantitative est plutôt destinée à proposer des prédictions sur un nombre limité de générations à partir de la variance génétique existante, supposée approximativement constante au fil des générations (Turelli 1988). Cette approche est donc moins appropriée pour prédire l'adaptation à long terme, par des évènements rares, tels que la fixation de nouvelles mutations apparaissant en copies uniques.

Les théories dites « de l'adaptation » (Orr 1998) se basent donc sur une simplification alternative: se limiter à l'étude de l'adaptation d'une population initialement peu variable génétiquement (par opposition à la génétique quantitative qui s'intéresse surtout à la variance existante). Ainsi, seules des mutations nouvellement apparues participent au processus adaptatif. Haldane, puis Fisher (1930), ont fourni les bases de cette approche en calculant la probabilité qu'une mutation d'effet sélectif s, apparaissant initialement en une copie finissent par envahir une population (probabilité de fixation). Là aussi, une réduction de la complexité du phénomène est faite en supposant une population finie mais de taille assez grande, panmictique et sexuée. Toutefois, des modèles plus généraux ont été fournis depuis (détaillés plus loin). Par ailleurs, Fisher (1930) puis Kimura (1979) ont développé un modèle simplifié (trop simplifié, peut-être) de l'effet des mutations sur la fitness qui sera détaillé plus loin et en Chapitre I. L'adaptation à long terme peut alors être modélisée comme l'apparition puis la fixation de mutations avantageuses dont l'effet est variable (tiré dans une distribution donnée). Cette approche a reçu des développements récents, notamment par Orr (1998; 1999; 2000b; 2003) et Gerrish (2001; 1998), pour une revue voir (Orr 2005a). Mon travail de thèse

s'est surtout basé sur cette seconde approche. Je détaillerai dans la partie suivante les hypothèses faites pour modéliser les taux d'adaptation dans ce contexte.

I.2 SIMPLIFICATION EMPIRIQUE

Bien que le but à terme soit de comprendre le processus adaptatif *in natura*, il est souvent difficile d'y tester des modèles d'adaptation. L'étude de l'évolution de traits quantitatifs dans les populations naturelles permet d'estimer la force de la sélection sur un ensemble de traits (pour revue, voir Hereford et al. 2004; Kingsolver et al. 2001) et d'étudier leur évolution d'une génération à l'autre. Toutefois, de telles études permettent rarement de valider des prédictions à long terme, à cause du temps de génération relativement long des espèces étudiées, et de la variabilité non contrôlée des conditions environnementales. Un certain nombre d'études ont toutefois documenté des adaptations rapides à un changement de l'environnement ou lors de la colonisation d'un nouveau milieu : changement d'hôte, de milieu physique, changements anthropogéniques etc., pour revue voir (Reznick and Ghalambor 2001). Ces études montrent que l'adaptation d'un grand nombre de traits simultanément peut survenir rapidement lors d'un changement environnemental. Toutefois, rares sont celles qui fournissent la dynamique de l'adaptation avec le temps. Un exemple (connu !) est l'évolution de la résistance aux insecticides chez le moustique C*ulex pipiens*, sur plus de 30 ans, dans la région de Montpellier... De telles études permettent de suivre l'évolution, sur de nombreuses générations et dans un milieu hétérogène, d'un ou plusieurs gènes clairement identifiés comme étant soumis à sélection, ainsi que d'étudier la dynamique des coûts associés à cette adaptation (Lenormand et al. 1999). Toutefois, de tels suivis à long terme sont très rares, et traitent d'adaptation ayant une base génétique simple (un ou deux gènes de résistance), bien qu'elles impliquent des coûts au déterminisme potentiellement complexe.

Récemment, des approches expérimentales ont été proposées pour étudier le processus adaptatif au laboratoire (en environnement contrôlé), chez des organismes modèles à temps de génération court (souvent des micro-organismes). Ces approches ont permis (i) des études sur parfois plusieurs milliers de générations, (ii) en très grandes populations (limitant la stochasticité), et (iii) des mesures précises de la fitness (paramètre démographique malthusien) et non d'une de ses composantes seulement. Ces expériences, initiées par Lenski et Travisano sur *E. coli* (Lenski et al. 1991; Lenski and Travisano 1994) et Novella et al. (1995) sur le virus à ARN *VSV*, ont grandement amélioré nos connaissances sur la dynamique de l'adaptation a long terme : elles ont également permis de tester explicitement diverses prédictions théoriques (pour revue, voir Elena and Lenski 2003). Il existe aussi quelques rares expériences d'adaptation à relativement long terme (de l'ordre de 100 générations) chez la drosophile (Frankham 2005; Gilligan and Frankham 2003; Latter and Mulley 1995; Matos et al. 2002). Toutefois, en raison d'une contrainte évidente de temps, il y a un manque criant de validations des théories de l'adaptation dans des organismes autres qu'unicellulaires. Je résume maintenant un ensemble de caractéristiques de la dynamique de l'adaptation qui sont prédites par la deuxième approche de simplification théorique citée plus haut, et qui ont été validées par ces expériences d'adaptation au laboratoire. Je détaillerai plus loin les modèles dont découlent ces prédictions.

I.3 MUTATION ET ADAPTATION: RESULTATS THEORIQUES CONFIRMES EMPIRIQUEMENT

Une perspective historique et un résumé des résultats importants de ces théories de l'adaptation peut être trouvé dans deux revue récentes par Orr (2005a; 2005b). Je résume ici quelques prédictions théoriques qui ont été confirmées empiriquement.

Distribution de l'effet des mutations contribuant à l'adaptation: sous l'influence notamment de Fisher (1930), les mutations de très faible effet ont longtemps été considérées comme la source majeur de l'adaptation, les autres mutations ayant de grandes probabilités d'être délétères (micro-mutationnisme). Un des résultats majeurs des théories récentes, obtenu par Orr par deux approches différentes (Orr 2005b) ont remis en cause ce paradigme (du moins partiellement). Orr a étudié la distribution des mutations responsables de l'adaptation, *i.e.* fixées par la sélection durant une phase d'adaptation à un nouvel environnement. Il est apparu que cette distribution est approximativement exponentielle. Elle comprend donc une majorité d'effets faibles, mais aussi quelques rares mutations de fort effet avantageux, et donc contribuant pour une part importante au processus adaptatif dans son ensemble (voir détails plus bas). Ce résultat a été qualitativement confirmé empiriquement chez *E. coli* (Imhof and Schlotterer 2001). La logique amenant à cette prédiction est plus indirecte qu'il n'y paraît. Le distribution de l'effet des mutations avantageuses apparaissant au hasard peut prendre diverses formes, mais on s'attend (Fisher 1930) à ce qu'elle comprenne une majorité de mutations de très faible effet (voir l'Encadré 1.a). En effet, plus une mutation a un effet faible, plus elle a de chances d'être avantageuse. On peut aussi s'intéresser à la distribution attendue des effets d'une mutation fixée (pour de nombreux répliquas d'une unique substitution adaptative). Celle-ci diffère de la distribution des effets fixés lors une phase d'adaptation (plusieurs substitutions successives). La distribution des effets fixés par une substitution compte une majorité d'effets intermédiaires, avec quelques rares effets forts, et une proportion réduite d'effets très faibles (Kimura 1983) (voir Encadré 1.b). En effet, les mutations d'avantage très faible sont certes les plus fréquentes, mais elles sont facilement perdues par la dérive dès leur apparition, et ont donc de faibles chances de se fixer. Ce résultat a été confirmé empiriquement chez *E. coli* (Rozen et al. 2002) et chez un virus à ADN (Rokyta et al. 2005). Les mutations d'effet avantageux très faible ont donc peu de chances de se fixer, à chaque génération, mais elles représentent la classe la plus fréquente de mutations avantageuses, et ce, d'autant plus que la population s'approche d'un optimum. Au final, elles constituent donc une part importante de l'ensemble des mutations fixées lors une phase d'adaptation, de sorte que cette dernière distribution est à nouveau exponentielle, comme mentionné plus haut (Orr 1998) (voir Encadré 1.c).

Effet du système de reproduction : les taux d'adaptation ont été modélisés chez les sexués, puis chez les asexués (Gerrish and Lenski 1998; Orr 2000b). L'approche est détaillée dans la partie suivante de cette introduction. Il apparaît que l'impact du flux de mutants par génération $N\,U$ (taille de la population N * taux de mutation U) diffère entre ces deux systèmes de reproduction :

- Chez les sexués le taux d'adaptation croît linéairement avec $N\,U$,
- Chez les asexués cette croissance est limitée pour les grands $N\,U$ (« diminishing return »).

Curieusement, c'est la prédiction chez les asexués qui a été confirmée empiriquement (de Visser et al. 1999; Miralles et al. 2000), je n'ai pas connaissance d'un test de la prédiction pour les sexués.

A long terme, décroissance du taux d'adaptation à un environnement fixe : Enfin, une dernière prédiction est que le taux d'adaptation décroît au fur et à mesure que la population s'adapte à son nouvel environnement (Orr 1998; Orr 2000a). Comme pour la distribution des effets fixés lors d'une phase d'adaptation, cette prédiction est commune à deux approches théoriques différentes (que nous détaillerons plus bas). Le principe est que plus une population s'approche d'un optimum donné (fixe), moins les mutations ont de chances d'améliorer encore son niveau d'adaptation (*i.e.* d'être avantageuses). Une telle dynamique est confirmée par la plupart des expériences d'adaptation à long terme dans un environnement fixe (voir la revue de Elena & Lenski 2003). Toutefois, même s'il y a décroissance du taux

d'adaptation, un « plateau » de fitness n'est pas toujours observé, par exemple chez le virus *VSV* (Novella et al. 1995).

Ces prédictions n'ont été confirmées que qualitativement, et souvent avec des protocoles expérimentaux complexes. Il n'existe pas de modèle permettant directement de prédire le changement de la fitness moyenne au cours du temps, même pour une espèce connue dans un milieu contrôlé. Il existe toutefois des approches pour modéliser cette trajectoire de fitness, et de nombreux progrès ont été faits dans la connaissance des paramètres de tels modèles. Dans les deux section suivantes, je développe ces deux points.

Comment quantifier la dynamique de l'adaptation dans les cas simples

Des approches théoriques ont permis de prédire la dynamique temporelle de l'adaptation en réduisant le cadre d'étude à un système « simple ». Considérons une population génétiquement homogène, confrontée à un nouvel environnement fixé. En l'absence de variation pré-existante, la fitness moyenne de cette population augmente par l'apparition de mutations avantageuses qui augmentent en fréquence jusqu'à leur fixation, le tout constituant un balayage sélectif. A l'échelle des générations, ce processus est fortement stochastique puisque (i) l'augmentation en fréquence des mutations avantageuses peut être erratique en petite population et (ii) des mutations d'effets sélectifs différents apparaissent à divers instants et envahissent à des vitesses différentes. Le résultat est une augmentation partiellement aléatoire de la fitness avec le temps (mesuré en générations). En outre, l'effet net sur la fitness de la fixation successive de deux mutations avantageuses dépend des relations épistatiques entre ces mutations, qui sont elles-même variables. Il convient donc d'abord de clarifier les hypothèses qui sont généralement faites pour simplifier/quantifier ce processus.

En général, pour contourner le problème de la forte stochasticité de la fitness avec le temps à l'échelle des générations, le taux d'adaptation par génération est calculé en moyenne sur une échelle de temps supérieure, bien que cela soit rarement mentionné. Ainsi, on considère que la dynamique transitoire pendant laquelle une mutation avantageuse donnée augmente en fréquence peut être négligée. L'augmentation progressive de la fitness moyenne $\overline{W}$ due à l'augmentation en fréquence d'un allèle avantageux est donc résumée à un saut en fitness de $\overline{W}$ à $\overline{W}(1+s)$, où s est l'effet sélectif de l'allèle considéré. Le changement de fitness moyenne est approximé par une fonction continue du temps en considérant l'effet moyen (sur de nombreuses générations) d'un flux constant de mutations se fixant dans la population.

L'Encadré 2 illustre ce principe : la succession de balayages sélectifs d'allèles avantageux, d'effets sélectifs variables, entraîne une augmentation irrégulière de la fitness moyenne de la population. Comme le montre l'Encadré 2, à une échelle de temps limitée, la fitness semble augmenter par sauts successifs, souvent interprétés empiriquement comme la fixation d'une mutation avantageuse (Elena and Lenski 2003). Notons toutefois que le parallélisme attendu entre sauts de fitness et balayages sélectifs n'est pas parfait, particulièrement chez les asexués (Gerrish and Lenski 1998). A une échelle de temps suffisamment plus longue (de l'ordre du temps de fixation des mutations avantageuses), cette augmentation de la fitness est approximée par un fonction linéaire du temps. Ainsi, la dynamique moyenne de l'adaptation, qui dépend de processus discrets et stochastiques à faible échelle de temps, peut être traitée de façon continue à une échelle de temps plus longue. Le taux d'adaptation est alors modélisé comme la dérivée de la fitness en fonction du temps (approximation en temps continu).

Ensuite, on suppose que les mutations interagissent de façon multiplicative, ce qui revient à négliger l'effet de l'épistasie entre mutations avantageuses relativement à s. Notons que les modèles additifs et multiplicatifs sont équivalents si les effets s sont faibles. En considérant le logarithme de la fitness moyenne, on obtient alors une relation simple entre le coefficient de sélection des mutations fixées et le taux d'adaptation. Si une première substitution d'effet s_1 a lieu puis (après de nombreuses générations) une seconde d'effet s_2, la log-fitness est donnée par $\log(1 + s_1) \approx s_1$ après la première substitution puis $\log(1 + s_1 + s_2) \approx \log((1 + s_1)(1 + s_2)) \approx s_1 + s_2$, après la seconde etc. Ainsi, le taux d'adaptation est donc modélisé par l'effet sur la log-fitness d'un flux (*transitoirement* constant) de fixations de mutations avantageuses.

L'incrément de log-fitness par génération est alors donné par le nombre de mutations apparaissant dans la population à chaque génération (flux entrant de mutations) et l'espérance, sur l'ensemble des mutations, du coefficient de sélection de celles qui se fixent. Si $f(s)$ est la distribution de l'effet sélectif des mutations, cette espérance est donc pondérée par (i) la probabilité $f(s)$ qu'une mutation ait un effet s et (ii) la probabilité que celle-ci se fixe $p_{fix}(s)$. Pour une population de N individus avec un taux de mutation U par individu par génération, le flux entrant de mutants par génération est $N\,U$. On obtient donc l'expression

$$\frac{d\,\log(W(t))}{dt} \approx N\,U \int_{s>0} p_{fix}(s)\; s\; f(s)\; ds\,. \qquad (1)$$

Notons que cette expression néglige la fixation de mutations délétères: l'intégrale est sur $s>0$, on néglige $p_{fix}(s)$ si $s<0$. Elle n'est donc pas valide dans des populations de très petites taille, ou de taille plus grande mais asexuées. En effet, en l'absence de recombinaison, des mutations délétères peuvent être fixées dans des populations d'effectif efficace relativement grand (Charlesworth et al. 1993). Toutefois, dans les expériences d'évolution expérimentale sur des asexués, les effectifs sont souvent très grands ($>10^6$), de sorte que le problème est souvent négligé a priori. Une expression plus générale peut de toute façon être proposée en intégrant l'Eq. (1) sur tout le spectre des effets (*i.e.* sans négliger $p_{fix}(s)$ si $s<0$). Notons aussi que ce taux d'adaptation est une moyenne sur une échelle de temps supérieure à quelques générations, mais qu'il peut changer à une échelle encore plus longue, comme nous le discuterons ci-dessous (I.3).

D'une manière générale, cette approche permet de contourner la forte stochasticité des processus adaptatifs due à la stochasticité de la mutation et de la dérive. Elle donne une mesure moyenne du taux d'adaptation, sur une échelle de temps intermédiaire. Un tel taux peut être mesuré par la moyenne des taux d'adaptations observés dans plusieurs répliquas de populations, ou directement dans une seule population de grande taille. Notons enfin que l'échelle en question peut varier selon la stochasticité démographique (taille efficace), le nombre de populations répliquées, et le système de reproduction (sexués / asexués).

En résumé, dans la situation simplifiée décrite ci-dessus, il est nécessaire, pour prédire les taux d'adaptation, de connaître (Eq. (1)) :

(i) la probabilité qu'une mutation avantageuse d'effet s se fixe dans la population $p_{fix}(s)$
(ii) le taux d'apparition des mutations $N\,U$
(iii) la distribution de leurs effets sur la fitness $f(s)$ (notamment proportion et effets des mutations avantageuses)

Je détaille maintenant l'état actuel de nos connaissances sur chacun de ces éléments.

I.4 PROBABILITES DE FIXATION

Les deux facteurs majeurs affectant $p_{fix}(s)$ sont le système de reproduction (sexué, asexué, consanguinité) et la dérive génétique (taille efficace des populations, structuration des populations en dèmes ou isolement par la distance, dynamique métapopulationnelle). L'influence de la dérive sur $p_{fix}(s)$ a été bien étudiée sur le plan théorique, y compris pour des populations variant en taille dans l'espace ou le temps (Barton and Whitlock 1997; Otto and Whitlock 1997). Ces résultats ont été obtenus pour une population sexuée, en supposant que les mutations ségrègent de manière indépendante à plusieurs loci (en négligeant la liaison). A l'autre extrême, la probabilité de fixation $p_{fix}(s)$ dans une population purement asexuée et non structurée a également été obtenue pour une population panmictique (Gerrish and Lenski 1998 ; Orr 2000b). Dans de nombreux protocoles expérimentaux, les paramètres déterminant $p_{fix}(s)$ sont connus et contrôlés (tailles efficaces et variations temporelles des effectifs, système de reproduction etc.), de sorte que les prédictions théoriques peuvent être appliquées. Le problème est beaucoup plus complexe *in natura*, mais pas insoluble: des mesures de structuration et de taille efficace sont en principe possibles à partir de marqueurs neutres, particulièrement lorsque les populations sont suivies sur plusieurs générations (Beaumont 2001).

I.5 TAILLE DES POPULATIONS

La taille de la population est souvent contrôlée dans les systèmes expérimentaux. Là-aussi, sa mesure en nature est en revanche plus problématique, mais possible par les approches de marquage-recapture. A défaut, la taille efficace N_e peut être estimée (voir plus haut). Plus généralement, le problème de l'étude de l'adaptation en milieu naturel est aussi compliqué par la difficulté de *délimiter* une population, comme une unité évolutive claire. On peut caractériser des tailles de voisinages, dans un modèle continu d'isolement par la distance à partir de marqueurs neutres. Toutefois, l'équivalence entre ces tailles de voisinage et un nombre d'individus constituant une population panmictique équivalente (l'unité évolutive citée plus haut) est problématique (Rousset 2001).

I.6 TAUX DE MUTATION

Le taux de mutation par génome par génération peut aussi être estimé empiriquement chez les organismes modèles. Je donne un résumé rapide des différentes méthodes d'estimation dont une revue peut être trouvée dans (Bataillon 2000; Keightley 2004; Lynch et al. 1999). On peut distinguer deux types de données empiriques qui ont fournit une estimation de U dans de nombreuses espèces.

Mesures directes à partir du taux de mutation génique: le taux de mutation par gène peut être mesuré directement par les nombreuses analyses de lignées mutantes pour un gène connu. Dans les organismes modèles dont le nombre de gènes total est connu, le taux de mutation peut ensuite être inféré pour tout le génome, après une correction pour la proportion de mutations neutres. Une revue de ces estimations est donnée par Drake et al. (1998) pour plusieurs espèces modèles.

Expériences d'accumulation de mutations: une autre approche est basée sur des analyses d'expériences d'accumulation de mutations (MA). Cette méthode introduite par Bateman

(1959) et Mukai (1964) consiste à maintenir des lignées isolées pendant de nombreuses générations en limitant la sélection au maximum. Les lignées doivent être fondées à partir d'un génotype fixé génétiquement (génotype très homozygote, ou clone), de façon à ce que les seules sources de variance génétique entre lignées soient les mutations apparues pendant l'expérience. Dans la suite de ce manuscrit, j'appellerai « ancêtre » ou génotype/lignée ancestral(e) le génotype initial dans lequel les mutations apparaissent, *c.a.d.* ici, la lignée utilisée pour fonder les lignées d'accumulation. A chaque génération, un petit nombre de descendants (souvent un ou deux) est pris au hasard pour former la génération suivante, dans chaque lignée. Dans les organismes « supérieurs », on isole le descendant à un stade juvénile pour éviter toute compétition intra-lignée. Cette méthode consiste donc à pratiquer de forts goulots d'étranglements ($N_e \approx 1$) sur de nombreuses lignées indépendantes, dans des conditions les plus permissives (optimales) possibles. L'impact de la sélection est ainsi fortement réduit relativement à celui de la dérive, pendant un maximum de générations: les mutations qui apparaissent durant la période d'accumulation se fixent donc indépendamment de leur effet sélectif (en dehors des mutations létales à sublétales, *i.e.* réduisant la fertilité ou la survie à près de zéro). On a donc accès au spectre de fitness que produit la mutation sans le filtre de la sélection. Seules les mutations létales sont éliminées mais leur taux d'apparition peut être mesuré.

Si les mutations spontanées sont délétères en moyenne, on attend dans un tel protocole d'accumulation (i) une diminution de la fitness moyenne sur l'ensemble des lignées relativement à leur ancêtre et (ii) une augmentation de la variance de la fitness entre ces lignées (par l'accumulation de mutations d'effets différents dans chaque lignée). L'Encadré 3 donne une description d'un tel protocole, ainsi que le changement de la fitness moyenne et de la variance interlignée au fur et à mesure des générations d'accumulation. Pour plus de réalisme, j'ai simulé une expérience d'accumulation sur 50 générations avec 100 lignées d'accumulations, en considérant des paramètres mutationnels correspondant à la drosophile. On constate bien un changement à peu près linéaire de la variance ΔV et de la moyenne ΔM de la fitness (relative à la lignée ancestrale) avec le temps. Etant donnée la lourdeur de tels protocoles, des méthodes alternatives utilisent la mutagenèse aléatoire pour augmenter artificiellement la production de mutations en un plus petit nombre de générations (Halligan et al. 2003; Keightley et al. 2000; Keightley and Ohnishi 1998). On peut ainsi observer une grande variance mutationnelle en un nombre réduit de générations.

Etant donné le changement moyen par génération de la variance ΔV et de la moyenne ΔM (respectivement dV et dM), on peut estimer (i) le taux de mutation U et (ii) l'effet moyen de ces mutations $\bar{s}$. Ces estimateurs «de Bateman-Mukai » sont toutefois biaisés lorsque l'effet des mutations est variable (c'est à dire toujours !). Soit CV(s) le coefficient de variation de la distribution de l'effet de mutations spontanées aléatoires, les estimateurs de U et $\bar{s}$ sont alors :

$$U_{BM} = \frac{dM^2}{dV} = U\,(1 + CV(s)^2)$$
$$s_{BM} = \frac{dV}{dM} = \frac{\bar{s}}{(1 + CV(s)^2)} \qquad (2)$$

Le biais induit par la variation de s peut entraîner une erreur de plus d'un ordre de grandeur par la méthode de Bateman-Mukai (Kibota and Lynch 1996). Il est donc important de pouvoir estimer les paramètres de la distribution de s, pour pouvoir avoir une estimation fiable de U. Cette distribution affecte aussi de nombreuses autres prédictions en biologie évolutive, qui seront plus détaillées en Chapitre I. Cela a motivé l'étude empirique de la distribution des effets des mutations délétères.

I.7 DISTRIBUTION DE L'EFFET DES MUTATIONS DELETERES

Mesures indirectes: des approches ont été développées pour estimer U à partir de données MA sans recourir à l'hypothèse de constance des effets des mutations (CV(s) = 0). Ces approches permettent parallèlement d'estimer les paramètres de la distribution de l'effet des mutations $f(s)$. Elles supposent que la distribution de l'effet des mutations est une gamma négative (lorsque seules des mutations délétères sont observées). En utilisant la distribution de fitness des mutants, les deux paramètres de la gamma et le taux de mutation U peuvent alors être estimés par les méthodes de maximum de vraisemblance (Keightley 1994) ou de distance minimum (Garcia-Dorado and Marin 1998), voir (Keightley 2004) pour une comparaison des deux approches. Ces méthodes permettent d'éviter le biais d'estimation inhérent à la méthode de Bateman-Mukai. Elle peuvent aussi prendre en compte l'existence de mutations avantageuses, dont la distribution est souvent modélisée par une gamma positive (« gamma réfléchie », voir l'Encadré 4). Cependant, une telle distribution présente une discontinuité en zéro qui semble assez peu « naturelle ». Garcia-Dorado & Marin (1998) ont aussi utilisé un mélange d'une gamma réfléchie et d'une gaussienne pour prendre en compte les mutations avantageuses sans générer de distribution bimodale. Shaw et al. (2002) a proposé un modèle alternatif supposant que $f(s)$ est la distribution d'une constante positive moins une gamma (« gamma déplacée », voir l'Encadré 4). Ces méthodes sont prometteuses mais permettent rarement d'obtenir une estimation précise de U et des paramètres de $f(s)$ indépendamment. En général, le paramètre de forme de la gamma doit être fixé pour estimer le taux de mutation U et vice versa. Toutefois, Shaw et al. (2002) a obtenu des estimations relativement précises de U et des trois paramètres pour la gamma déplacée. Vassilieva et Lynch (2000) ont aussi utilisé le maximum de vraisemblance sur des données de MA pour estimer directement le coefficient de variation CV(s), ainsi que U et $\bar{s}$ et ont obtenu une estimation assez précise de ces paramètres.

Un des problèmes majeurs de ces approches est qu'elles supposent une distribution des effets des mutations *a priori*. Or, en l'absence d'un attendu théorique clair sur $f(s)$, il est difficile de justifier le choix d'une distribution a priori plutôt qu'une autre (Bataillon 2003).

Mesures directes: la difficulté à estimer $f(s)$ à partir des expériences MA vient du fait que la distribution des fitness des lignées d'accumulation reflète deux distributions sous-jacentes: celle du nombre de mutations par lignée (une loi de Poisson de paramètre U) et celle de l'effet des chaque mutation $f(s)$. Une alternative empirique consiste donc à produire un ensemble de lignées dont on sait qu'elles portent chacune une unique mutation. La distribution des fitness entre lignées reflète alors directement $f(s)$. Les protocoles permettant d'obtenir de telles lignées sont plus complexes que l'accumulation de mutations et dépendent de l'espèce considérée. Toutefois, ils ont été appliqués avec succès à la plupart des espèce modèles en évolution: le virus à ARN *VSV*, la bactérie *E.coli*, la levure *S. cerevisiae*, et la drosophile *D. melanogaster*. Des détails sur ces protocoles sont donnés en chapitre II et dans l'Article 1. Ces approches ont permis d'obtenir, chez des espèces modèles, une mesure précise et directe de la distribution de l'effet de mutations spontanées ou obtenues par mutagenèse. Leur seul défaut est qu'elles peuvent *a priori* entraîner un biais dans l'estimation de $\bar{s}$ lorsque la méthode utilisée a un effet direct sur la fitness.

En résumé, nous disposons donc de bonnes informations sur la mutation délétère chez les espèces modèle: nombreuses estimations du taux de mutation génomique et de l'effet moyen de ces mutations, et parfois paramètres de la distribution des effets.

I.8 PROPORTION ET EFFET DES MUTATIONS AVANTAGEUSES

L'un des résultats les plus surprenants qui émerge de l'analyse de l'effet de mutations aléatoires, est que la plupart des mutations ainsi générées (sinon toutes) se sont révélées être

délétères. L'analyse empirique des mutations avantageuses est donc plus difficile. Pourtant, il est clair que l'adaptation correspond à la fixation de mutations avantageuses.

Mesures indirectes: devant la difficulté d'observer directement les mutations avantageuses, des méthodes indirectes ont été développées pour estimer la proportion p_{av} et la distribution de l'effet sélectif des mutations avantageuses. Le principe général est d'estimer les paramètres de l'ensemble des mutations avantageuses via la mesure de celles qui sont fixées par la sélection (donc accessibles empiriquement). Le problème est donc de détecter les évènements de fixation. Gerrish et Lenski (1998) ont détecté indirectement les substitutions adaptatives, par les sauts de fitness observés au cours du temps (fitness en escalier comme représenté en Encadré 2). Plus récemment, des expériences chez *E. coli* et le virus *VSV*, ont utilisé des marqueurs associés à différents clones de la population pour identifier directement les clones portant une mutation avantageuse par l'augmentation en fréquence du marqueur qui lui est associé (Imhof and Schlotterer 2001 ; Miralles et al. 1999; Rozen et al. 2002). Ces approches restent cependant indirectes du fait qu'elles se fondent sur des prédictions théoriques entre les grandeurs mesurées et les paramètres à estimer. Elles sont donc partiellement dépendantes de la validité des prédictions théoriques (Rozen et al. 2002). Trois hypothèses sont communes à toutes ces approches :

(i) La distribution de l'effet des mutations avantageuses est exponentielle.

(ii) cette distribution et la proportion de mutations avantageuses sont constantes dans le temps.

(iii) l'influence des mutations délétères sur l'adaptation (« background selection ») est négligée.

Les limites de ces différentes hypothèses ont été mises en évidence dès le papier fondateur par Gerrish et Lenski (1998), qui proposèrent le premier modèle de taux d'adaptation chez les asexués et posa les bases de ces méthodes d'estimation.

Les hypothèses (i) et (ii) sont difficiles à tester empiriquement étant donné la difficulté d'estimer la distribution d'effet de l'ensemble des mutations avantageuses. En effet, celles-ci sont rares, et seules peuvent être observées celles qui échappent à la dérive dans les premières générations après leur apparition, et à l'interférence clonale ensuite, pour les asexués (Orr 2003). L'interférence clonale sera vue plus loin : elle correspond au fait qu'en l'absence de recombinaison, plusieurs mutations avantageuses ségrégeant simultanément dans des clones distincts entrent en compétition pour la fixation, ce qui réduit la probabilité qu'a chacune de se fixer (Gerrish and Lenski 1998). Les *a priori* théoriques quant à ces hypothèses seront discutés plus loin dans cette introduction et en Chapitre I. Toutefois, une observation fréquente dans les expériences d'adaptation à un environnement fixé est que le taux d'adaptation décroît avec le temps à long terme (Elena and Lenski 2003). Il est clair que sous l'hypothèse (ii), un tel comportement n'est pas prédit (voir l'Eq. (1)). Cependant, il est possible qu'elle soit valide à l'échelle de temps limitée de ces études. Reste le problème d'inférer le changement des paramètres de mutation avantageuse à plus long terme.

L'hypothèse (i) suppose que le taux de mutation (délétère) est suffisamment faible pour que la sélection de fonds (« background selection ») soit négligée. Elle semble *a priori* réaliste chez *E. coli* ($U \sim 0.0002$), mais peut être pas chez le virus *VSV* ($U \geq 1$), par exemple. Toutefois, Miralles et al. (Miralles et al. 1999) proposent une correction pour prendre en compte l'influence des mutations délétères sur leurs estimations et montrent qu'elle est faible.

Notons aussi que certaines de ces études constituent bien un test (au moins partiel) des prédictions théoriques. La proportion de mutations avantageuses, p_{av}, et le paramètre α de la distribution (exponentielle) de leurs effets sur la fitness sont estimés en maximisant la

vraisemblance des observations. Cette méthode repose sur un modèle explicite prédisant une relation donnée entre le taux d'adaptation et la taille de la population (Miralles et al. 1999), ou une distribution donnée des effets des mutations fixées (Rozen et al. 2002). Le bon ajustement du modèle avec les données permet donc de valider les prédictions théoriques. Cela a permis la validation des prédictions qui ont été citées en partie I.1 de cette introduction.

En revanche, comme le soulignent Rozen et al. (Rozen et al. 2002), ces expériences ne permettent malheureusement pas de déterminer précisément la distribution des effets avantageux. En effet, différentes $f(s)$ peuvent générer le même résultat pour la distribution des effets fixés par la sélection. Zeyl (2004) a récemment fait une revue des estimations de p_{av} obtenues par ces méthodes chez les micro-organismes. Dans toutes les espèces (*S. cerevisiae*, *E. coli*, *VSV*), la proportion de mutations avantageuses estimée est très faible ($<10^{-8}$).

Mesures directes: dans certaines expériences MA, des mutations avantageuses ont toutefois été observées *directement*, parfois dans des proportions bien supérieures aux estimations indirectes. Ces expériences suggèrent que la proportion de mutations avantageuses varie selon l'environnement dans lequel la fitness est mesurée ou la souche utilisée pour initier l'accumulation de mutations (l'ancêtre). Plusieurs auteurs (Remold and Lenski 2001; Sanjuan et al. 2004b) ont proposé que la proportion de mutations avantageuses augmente avec la maladaptation de la souche ancestrale aux conditions de mesure. Cet argument est assez intuitif, dans un génotype maladapté, de nombreuses mutations peuvent compenser les défauts : « plus on est mauvais, plus y a de l'espoir ». Cela pourrait expliquer pourquoi les mutations avantageuses sont très rarement détectées dans les MA classiques où la souche d'origine est souvent bien adaptée aux conditions du laboratoire dans lesquelles la fitness est mesurée. En outre, une maladaptation du génotype ancestral semble plausible dans trois des quatres études ayant justement détecté des mutations avantageuses :

(i) Chez *E. coli*, la proportion de mutations avantageuses dans des lignées mutantes descendant d'une souche adaptée au glucose augmente de 0 à 19% lorsqu'on mesure la fitness dans le milieu optimal (glucose) ou dans un milieu non-optimal (riche en maltose, Remold and Lenski 2001).

(ii) Chez le virus *VSV*, des mutants issus d'un génotype synthétisé à partir de deux génotypes naturels (« chimère ») présentent 4% de mutations avantageuses. Le génotype « chimère » ancestral pourrait être maladapté aux conditions du laboratoire (Sanjuan et al. 2004b).

(iii) Chez la levure *S. cerevisiae*, une proportion d'environ 6% de mutations avantageuses a été observée parmi des mutants issus d'une souche initialement diploïde et rendue entièrement homozygote en une génération (Joseph and Hall 2004). Il est possible qu'une forte dépression de consanguinité se soit exprimée dans la souche d'origine réduisant sa fitness en général. Notons que dans une MA classique, la souche ancestrale est également rendue fortement homozygote, mais en plusieurs générations, de sorte que les mutations récessives très délétères sont purgées avant la MA.

(iv) Chez *Arabidopsis thaliana*, une proportion de presque 50% de mutations avantageuses a été detectée (Shaw et al. 2002; Shaw et al. 2000). Cela pourrait s'expliquer si les traits de fitness utilisés (nombre de fruits, nombre de graines par fruit) ne sont pas les plus soumis à sélection en nature, comme suggéré par Keightley et Lynch (2003), mais cela reste hypothétique. L'argument semble donc s'appliquer de façon moins évidente à cette étude. L'effet ne semble pas inhérent à

l'espèce puisqu'une autre étude sur *A. thaliana* n'a identifié aucune mutation avantageuse (Schultz et al. 1999).

Les mesures directes suggèrent donc que les mutations avantageuses peuvent être relativement fréquentes dans des génotypes peu adaptés au milieu où la fitness est mesurée. Les grandes différences d'estimations entre les mesures directes et indirectes ($<10^{-8}$ contre plusieurs % en mesure directe!) suggèrent donc que l'on ne peut directement extrapoler les taux mesurés entre espèces et même entre environnements ou entre lignées pour une même espèce. Je n'ai pas trouvé d'analyses tentant de relier ces résultats apparemment contradictoires. En résumé, contrairement aux mutations délétères, les mutations avantageuses sont difficiles à détecter, et on observe de grandes différences d'estimations selon les méthodes utilisées, l'espèce, la lignée ou l'environnement. Ces différents problèmes rendent incertaines les estimations de p_{av} ou α (du moins les estimations indirectes) et leur extrapolation à d'autres espèces ou environnement.

Ce résumé montre les avancées qui ont été faites dans la connaissance empirique des paramètres qui affectent les taux d'adaptation. La principale inconnue reste en fait la distribution de l'effet des mutations sur la fitness $f(s)$, mais on ne peut prédire un taux d'adaptation sans elle. Il serait donc utile de proposer un modèle satisfaisant sur la distribution des effets des mutations, et ce pour des raisons

empiriques: pour déterminer quelle(s) distribution(s) sont « réalistes », sur des bases biologiques, et peuvent être utilisée dans l'analyse d'expériences MA, et les études indirecte de mutations avantageuses

théoriques: pour pouvoir prédire les propriétés de cette(ces) distribution(s) et les facteurs qui l'affectent (lignée ancestrale, environnement) qu'il s'agisse des mutations avantageuses ou délétères, et par conséquent, pour pouvoir ainsi prédire comment varient les taux d'adaptation.

Dans la section suivante, je détaille plusieurs approches qui ont permis de faire diverses prédictions sur $f(s)$.

Résultats théoriques sur la distribution de l'effet des mutations

Jusqu'à récemment, peu de travaux théoriques ont été proposés pour prédire la distribution de l'effet des mutations sur la fitness. Je détaille seulement les trois approches qui ont eu un fort impact théorique, voire plus récemment empirique (pour revue, voir Orr 2005a).

I.9 MODELE DE FISHER (MF):

Du point de vue théorique, la seule approche permettant de prédire la distribution des coefficients de sélection de l'ensemble des mutations (délétères *et* avantageuses) est due à Fisher (1930). Son modèle géométrique proposé en quelques lignes dans son livre de 1930 a eu un impact important sur notre façon d'appréhender la mutation et l'adaptation. Il fut initialement un plaidoyer pour le gradualisme (l'idée que l'évolution avance « à petits pas »). Il a été récemment remis au goût du jour par Orr (1998; 2000a) dans un ensemble de travaux fondateurs en théorie de l'adaptation (synthétisés dans Orr 2005b). Le modèle de Fisher (MF)

part du principe qu'un organisme peut être modélisé par un ensemble de n traits phénotypiques continus soumis à sélection stabilisante pour un optimum, et que la mutation correspond à un déplacement aléatoire dans cet espace phénotypique. Son intérêt est qu'il relie l'effet des mutations sur un ensemble de traits phénotypiques (produits enzymatiques, structure etc..) et la fitness. Son autre intérêt est qu'il englobe à la fois les mutations délétères et avantageuses. Enfin, par la définition d'une distance entre le phénotype ancestral et un optimum adaptatif donné, ce modèle permet d'expliciter l'effet de l'adaptation du génotype ancestral à son environnement. Celle ci est caractérisée par la distance entre le phénotype de l'ancêtre et un optimum.

Toutefois, ce modèle est souvent critiqué comme une vision trop idéalisée de ce qu'est un phénotype et comment il détermine la fitness (Clarke and Arthur 2000; Orr 2001). Notamment, il repose sur des hypothèses très fortes de symétrie: tous les traits phénotypiques ont des effets équivalents sur la fitness et varient de la même manière par mutation. Le chapitre I (et l'article 1) traitent plus en détail des hypothèses de ce modèle, et de leur validité, je ne les détaille donc pas ici. Je présente maintenant des approches basées sur un modèle génétique (et non phénotypique comme le MF).

I.10 PAYSAGES MUTATIONNELS ET THEORIE DES VALEURS EXTREMES:

Orr (2002 ; 2003), a proposé une approche pour s'affranchir du fait que nous ne connaissons que très peu la distribution des effets avantageux des mutations. Il se base sur les modèles de paysages mutationnels (PM) introduits par Maynard Smith puis développés par Gillespie (1984). Cette approche modélise de façon explicite le processus de mutation au niveau de la séquence d'un gène. On suppose que le taux de mutation est suffisamment faible pour qu'il n'y ait pas plus d'une seule mutation par individu à chaque génération, et que les mutations sont ponctuelles (modifient seulement un nucléotide). Considérons un gène ou un petit génome de L paires de bases, le nombre de séquences possibles produites par mutation est $3L$ (chaque nucléotide peut muter vers les 3 autres). Chaque mutation (ponctuelle) produit un génotype dont la fitness est tirée dans une distribution *arbitraire*. Nous noterons la distribution de l'effet dw des mutations sur la fitness *absolue* $f_w(dw)$: notons que $f_w(dw)$ est différente de $f(s)$ qui fait référence aux coefficients de sélection des mutations (donc aux effets sur la fitness *relative* des mutants). Il est supposé que le génotype ancestral est bien adapté: il a une fitness élevée relativement à l'ensemble des mutants possibles. Sa fitness est donc dans la queue (droite) de la distribution des fitness. Cette hypothèse-clé permet d'appliquer la théorie des valeurs extrêmes (extreme value theory: EVT), qui prédit la distribution des valeurs tirées dans la queue d'une distribution, quasi-indépendamment de la nature de celle-ci. Ainsi, on peut déterminer $f_w(dw^+)$: la distribution de l'effet des mutations avantageuses, sans connaître $f_w(dw)$ pour l'ensemble des mutations. La distribution ainsi obtenue est une exponentielle, elle est indépendante de $f_w(dw)$, de la fitness de l'ancêtre (tant que celle-ci est élevée) et de la longueur de la séquence L. Le seul paramètre du modèle est donc celui de l'exponentielle, $E[\Delta_1]$: l'écart moyen en fitness entre le meilleur allèle accessible par mutation (celui conférant la plus grande fitness) et le deuxième meilleur allèle.

L'intérêt majeur de l'EVT est qu'elle fait des prédictions indépendantes de la distribution exacte de la fitness des mutants (un paramètre peu accessible empiriquement, notamment pour les mutations avantageuses voir I.3). Le modèle fait cependant une hypothèse sur la classe de distribution à laquelle appartient $f_w(dw)$: elles doivent être du type Gumbel III, c'est à dire un grand nombre de distributions classiques, excluant certaines distributions bornées vers la droite (pour plus de détails voir Orr 2003). Toutefois, certaines distributions bornées à droite sont aussi du type Gumbel III, nous en verrons un exemple en Chapitre I.

Un autre avantage parfois invoqué (Rokyta et al. 2005), relativement au modèle de Fisher, est que le PM se fonde sur un modèle apparemment plus « réaliste » de l'effet de la

mutation puisque directement relié au processus de mutation ponctuelle dans des séquences codantes. Toutefois, les substitutions nucléotidiques ne sont pas les seules sources d'adaptation, vue l'importance des délétions, insertions, duplications et réarrangements chromosomiques dans les mutations avantageuses (Kidwell and Lisch 1997; Schneider and Lenski 2004; Zeyl 2004). Néanmoins, le modèle n'est peut-être pas trop sensible au mode de mutation supposé (bien que je n'ai pas trouvé de discussion de la question). Une autre hypothèse est que la sélection est assez forte, relativement à la taille de la population ($N_e\, s >$ 1). Cette hypothèse n'est que peu restrictive lorsqu'on considère des expériences sur des populations de micro-organismes qui sont typiquement de grandes taille, mais pourrait l'être pour les expériences sur des pluricellulaires (par ex. Drosophile).

Il est important de clarifier les prédictions indépendantes de la biologie du système étudié, et celles qui en dépendent. Pour cela, rappelons que $f_w(w)$ n'est pas la distribution de l'effet sélectif des mutations $f(s)$ (celle qui est utilisée pour calculer le taux d'adaptation en Eq (1)). Nous avons vu que $f_w(dw^+)$ est assez indépendante du système étudié mais $f(s^+)$ ($f(s)$ parmi les mutations avantageuses) ne l'est pas. Cette dernière est dépendante du génotype ancestral même dans les modèles basés sur l'EVT, puisque l'on standardise la fitness absolue w par la fitness de l'ancêtre pour calculer s. Ainsi, comme $f_w(dw^+)$ est une exponentielle de moyenne $E[\Delta_1]$, $f(s^+)$ est une exponentielle de moyenne $E[\Delta_1]/w$. Pour pouvoir prédire comment varie $f(s)$, avec le temps, il faut donc aussi prédire comment varie w dans le temps. Orr (2002) a montré que le saut de fitness absolue moyen par substitution adaptative est $2(i - 1)/i\ E[\Delta_1]$, où i est le rang de fitness du génotype ancestral, relativement au meilleur génotype (l'ancêtre est le $i^{\text{ème}}$ meilleur génotype). Cette quantité dépend très peu du génotype ancestral tant que celui-ci n'est pas parfaitement adapté (de rang i proche de 1): pour $i \geq 5$, $2(i - 1)/i \approx 2$. En revanche, ce saut de fitness moyen par substitution dépend de la distribution de fitness des mutants $f_w(dw)$ parce qu'il dépend de $E[\Delta_1]$. Notons aussi que ce résultat implique des probabilités de fixation qui sont calculées sous l'hypothèse d'une population sexuée de grande taille ($p_{fix}(s) = 2s$). De là, Orr donne aussi (sur la base de simulations) la distribution des sauts de fitness sur une série de substitutions adaptatives, sous l'hypothèse que $f_w(dw)$ ne change pas dans le temps (donc $E[\Delta_1]$ constant). Une hypothèse similaire est faite dans les modèles présentés ci-dessous.

I.11 PAYSAGES *NK* ET MODELE EN « CHATEAU DE CARTES »

Kauffman et Levin (1987) ont proposé un modèle qui correspond approximativement à une extension des paysages mutationnels pour prendre en compte les interactions épistatiques entre mutations. Ce modèle suppose que l'ensemble des effets des mutations à n loci (ou n sites nucléotidiques) est tiré dans une distribution, et que chaque locus interagit avec k autres loci (ou sites), d'où le terme « *nk* ». Ce type de modèle permet de prendre en compte l'existence de plusieurs optimums locaux déterminés par les interactions épistatiques. Il permet aussi de prédire $f(s)$ pour les mutations délétères comme avantageuses, mais suivant autant, voire plus d'hypothèses que le MF: grossièrement, on peut faire un parallèle entre les n traits quantitatifs dans le MF et les n loci du modèle *nk*, auquels s'ajoute le nombre d'interactions épistatiques k. Récemment, Welch et Waxman (2005) ont montré que lorsqu'on considère qu'un grand nombre d'allèles peut être produit par mutation à chaque locus (infinite allele approximation), les optimums locaux n'existent plus et le modèle converge vers un modèle plus simple sans épistasie: le modèle en « château de cartes » (« House of Cards » HC) (Tachida 1991). Comme le PM, ce modèle suppose que la distribution de l'effet des mutations sur la fitness *absolue* est constante dans le temps, mais il suppose en outre que c'est une gaussienne centrée en zéro. Sous ces hypothèses (relativement restrictives), il permet de prédire des taux d'adaptation. Les modèles *nk* et HC font à peu près les même hypothèses limitantes que le modèle PM dont ils sont proches, mais font en outre certaines hypothèses

supplémentaires. Je tente maintenant de résumer les points communs et les différences entre les modèles présentés jusqu'ici.

I.12 LES POINTS COMMUNS ET DIFFERENCES ENTRE CES MODELES

Orr (Orr 2005b) a récemment comparé les approches basées sur le modèle de Fisher et sur les paysages mutationnels. Il montre que ces deux modèles prédisent une dynamique de l'adaptation similaire:

(i) le taux d'adaptation décroît au fur et à mesure que la population s'adapte

(ii) la distribution des effets sélectifs s fixés durant une phase d'adaptation (des pas vers un optimum) est approximativement exponentielle

Ces conclusions sont qualitativement en accord avec les résultats expérimentaux cités en I.3. Notons également que les modèles basés sur les paysages *nk* et le modèle HC prédisent aussi (Welch and Waxman 2005) la décroissance du taux d'adaptation avec le temps (point (i)). Cette convergence pour la prédiction (i) découle cependant d'hypothèses différentes, entre le MF d'une part, et les autres modèles (PM, HC paysages *nk*) d'autre part. Cette différence réside dans ce qui est supposé constant au cours du temps. Le MF suppose que la distribution de l'effet des mutations sur les *traits phénotypiques* est constante, alors que les autres approches supposent une distribution constante des effets sur la *fitness absolue*. Dans tous ces modèles toutefois, la décroissance du taux d'adaptation avec le temps vient du fait que $f(s)$ est déterminée par les fitness absolues des mutants standardisées par celle du génotype fixé dans la population à un temps t. Comme ce génotype s'adapte, sa fitness augmente avec le temps, de sorte qu'une fraction décroissante des mutations sont avantageuses. Comme l'a suggéré Orr (2005b), la prédiction (i) est donc probablement commune à tout modèle d'adaptation dans un environnement fixe. Par ailleurs, étant donné que l'EVT fait un ensemble de prédictions indépendantes de $f(s)$, tout modèle d'adaptation basé sur une $f(s)$ de type Gumbel III génère les même conclusions. Nous verrons par exemple (Chapitre I et Annexe 1) que $f(s)$ prédit par une extension du modèle de Fisher est du type Gumbel III, de sorte que les prédictions de l'EVT s'y appliquent.

Toutefois, des prédictions communes ne permettent pas de discriminer les différents modèles sur la base de données expérimentales. Il est donc également intéressant de souligner les prédictions qui diffèrent entre ces deux approches. Je compare ici plus particulièrement le modèle de Fisher et l'approche basée sur l'EVT.

Une première différence réside dans le fait que l'approche EVT n'est valide que lorsque le génotype ancestral est bien adapté. Il semble que la théorie des valeurs extrêmes reste assez fiable même pour des génotypes assez mal adaptés, mais la limite reste importante (le génotype doit être dans les 5% les mieux adaptés environ (Orr 2002)). Le MF n'a pas ce genre de limite.

Une deuxième différence réside dans la façon de prédire le changement de $f(s)$ sur plusieurs générations. Le MF suppose constante la distribution des effets phénotypiques de la mutation. Dans ce cas, la distribution des phénotypes des mutants change en fonction du phénotype ancestral, puisque les *effets* (de distribution constante) s'*ajoutent* au phénotype ancestral. En revanche, les autres approches supposent que la distribution des fitness absolues des mutants est constante. Ainsi, quel que soit la fitness de l'ancêtre, il produit toujours une même distribution de fitness absolues par mutation. Cela correspond à appliquer les hypothèses des modèles en « château de cartes » de la génétique quantitative (Turelli 1985) au trait « fitness absolue » (d'où le lien avec les approches HC et *nk*). Il est difficile empiriquement de déterminer laquelle de ces deux hypothèses d'invariance est la plus valide. Toutefois, on peut tester la deuxième ($f_w(dw)$ constante) en mesurant la distribution des fitness

absolues de mutants obtenues à partir de plusieurs génotypes ancestraux, ou à partir d'un unique ancêtre en mesurant $f_w(dw)$ dans plusieurs environnements (auquel l'ancêtre est plus ou moins adapté). Une étude (non encore publiée Kassen and Bataillon 2005) a utilisé la seconde approche et a détecté des différences de $f_w(dw)$ entre environnements. Le plus important a priori serait toutefois de mesurer la variation du paramètre $E[\Delta_1]$ qui est la clé pour prédire le changement de fitness au cours du temps. Celui-ci pourrait de fait être invariant même si $f_w(dw)$ varie. De son côté, le MF suppose que $f_w(dw)$ varie et prédit son évolution, au fur et à mesure que la population s'adapte. En revanche, cette prédiction est faite à partir de quantités non-mesurables (par ex. le nombre de traits phénotypiques sous sélection) et sous des hypothèses très limitantes d'équivalence entre tous les traits phénotypiques. Une telle limite est absente de l'approche par l'EVT.

Une troisième différence est que l'approche EVT ne dit rien sur la *proportion* de mutations avantageuses, un paramètre pourtant nécessaire pour prédire le taux d'adaptation. On connaît le changement de fitness par substitution à partir de l'EVT, mais on ne peut prédire le taux de substitutions (Orr 2002). A l'opposé, le MF génère une prédictions sur la distribution de l'effet de l'ensemble des mutations $f(s)$, donc notamment sur la proportion de celles-ci qui sont avantageuses.

Enfin, les calculs nécessaires pour prédire comment $f(s)$ varie au cours du temps à partir de l'approche EVT impliquent des probabilités de fixations (pour calculer le changement moyen de fitness par substitution). Les résultats analytiques sont pour le moment limités au cas d'une population sexuée de très grande taille. Cela réduit donc leur application à des micro-organismes asexués, notamment si le taux de mutation est élevé. En effet, dans ce cas, les mutations avantageuses peuvent ségréger conjointement avec d'autres mutations avantageuses (interférence clonale) ou avec des mutations délétères (« background selection »). Ces deux processus peuvent modifier fortement les probabilités de fixation dans les espèces asexuées (De Visser and Rozen 2005; Gerrish and Lenski 1998; Johnson and Barton 2002; Orr 2000b). Bien que le MF n'ait pas non plus été étudié pour un système de reproduction asexué, cette limite est absente pour le MF, parce qu'il prédit directement $f(s)$ (et non $f_w(dw^+)$), donc on peut utiliser les résultats théoriques existants pour calculer $p_{fix}(s)$ selon le système de reproduction (comme nous le verrons en Chapitre II). Par ailleurs, le MF permet de prédire la distribution de l'effet des mutations délétères, donc l'effet de la background selection.

Deux confirmations empiriques des prédictions de l'EVT ont été proposées récemment (Kassen and Bataillon 2005; Rokyta et al. 2005), mais elles apportent une information réduite sur $f(s)$ puisqu'on attend les mêmes observations sous de nombreux modèles. En revanche Rokyta et al. (2005) ont montré que la qualité des prédictions était sensible aux hypothèses faites sur le type de mutations ponctuelles (excès ou non de transversions) et sur la façon de calculer les probabilités de fixation.

Toutes ces approches posent le problème qu'elles sont difficilement testables, et qu'elles ne peuvent être utilisées pour prédire un taux d'adaptation. Cela vient du fait que de nombreux paramètres de ces modèles ne sont pas mesurables empiriquement (Orr 2005a). Le modèle de Fisher est basé sur des « pas » dans un espace phénotypique abstrait, qu'on ne peut mesurer directement, et il fait des hypothèses fortes sur les bases phénotypiques de l'adaptation. Les modèles *nk* nécessitent de poser une distribution donnée de l'effet des mutations à différents loci, ainsi que le nombre et le type d'interactions épistatiques. Le modèle basé sur les paysages mutationnels et l'EVT fait des prédictions robustes sur la forme générale de la distribution des effets des mutations avantageuses. Toutefois, pour relier les prédictions à un taux d'adaptation, des quantités difficilement mesurables doivent être connues : $E[\Delta_1]$ et la proportion de mutations avantageuses. Les modèles MF et EVT sont tous les deux assez contraints dans leurs prédictions à long terme mais de façon différente, ce qui devrait permettre de proposer des tests entre les deux approches, bien que je n'en aie pas vu dans la littérature.

Les buts de cette thèse

Le but principal de ma thèse était de proposer et de valider expérimentalement un modèle le plus réaliste possible de la distribution de l'effet des mutations sur la fitness $f(s)$. Il s'agissait notamment de caractériser la variation de $f(s)$ avec l'environnement, le génotype ancestral dans lequel les mutations apparaissent et de pouvoir paramétrer et tester le modèle à partir de données expérimentales.

Pour aborder ces questions, je me suis basé sur le modèle de Fisher, pour plusieurs raisons :

(i) Il prend en compte de façon explicite l'adaptation du génotype ancestral à l'environnement de mesure en définissant une distance à un optimum phénotypique qui est déterminée par l'environnement et l'ancêtre.

(ii) Il permet de modéliser les mutations délétères et avantageuses, et de faire le lien entre elles. Or, notre connaissance empirique des mutations délétères est bien meilleure que celle des mutations avantageuses, comme nous l'avons vu. Par ailleurs, dans le cas des asexués au moins, les mutations délétères peuvent influer sur le taux d'adaptation.

(iii) Il permet de modéliser à la fois la proportion de mutations avantageuses et la distribution de leurs effets, or ces deux paramètres interviennent dans la dynamique de l'adaptation (voir Eq. 1).

Un but du mon travail théorique a donc été de proposer un modèle basé sur l'approche de Fisher-Orr, en se passant de certaines des hypothèses limitantes du modèle. Le chapitre I présente ce travail.

Le deuxième but de cette thèse était de valider les prédictions du modèle proposé, à partir de données empiriques existantes. Pour cela, j 'ai développé, pendant ma thèse, des protocoles d'élevage et d'accumulation de mutations sur le crustacé *Artemia*. Le but était de comparer la distribution de l'effet des mutations entre différents environnements, ici entre plusieurs salinités. Je n'ai malheureusement pu mener à bien cette expérience, (mais je présente en annexe les protocoles expérimentaux que j'ai mis en place). J'ai donc utilisé une autre approche pour tester nos prédictions. J'ai utilisé les résultats des mesures empiriques directes et indirectes de l'effet des mutations délétères (voir I.3). J'ai aussi tenté de valider les prédictions de notre modèle directement sur des trajectoires de fitness observées dans des expériences d'adaptation à long terme. Pour aller au delà de l'effet de mutations isolées, j'ai aussi utilisé le modèle pour prédire la distribution des interactions épistatiques entre paires de mutations. J'ai comparé ces prédictions aux données disponibles sur le virus *VSV*. L'ensemble de ces travaux est présenté en chapitre II.

Il existe une forte interaction entre adaptation, mutations, et système de reproduction. Dans une troisième partie (Chapitre III), je discute les implications de nos résultats pour l'évolution de la reproduction sexuée. Les modèles présentés dans le chapitre I et II supposent une reproduction purement asexuée ou sexuée en négligeant la liaison, et en populations non-structurées. J'ai étudié un modèle génétique explicite, permettant d'aborder le cas intermédiaire où des mutations avantageuses ségrègent à deux loci partiellement liés, dans une population subdivisée. J'ai déduit de ce modèle comment la recombinaison entre les deux loci sous sélection pouvait évoluer en réponse à ce processus.

Un des points limitant de toutes les études empiriques de l'adaptation à long terme est qu'elles se sont surtout focalisées sur les micro-organismes dans des conditions contrôlées. Il serait intéressant de pouvoir aussi étudier l'adaptation en nature, et sur d'autres types d'organismes. En Discussion, je tenterai de montrer comment des protocoles expérimentaux développés pendant cette thèse sur *Artemia* pourraient permettre l'étude de l'adaptation à long terme et en nature. Je discuterais également l'intérêt, les limites et les perspectives de nos résultats théoriques.

CHAPITRE I: UN MODELE DE DISTRIBUTION DES EFFETS SELECTIFS DES MUTATIONS

Article 1: Martin, G. & Lenormand, T. A multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution* 60, 893-907 (2006).

Article 3: Martin, G. & Lenormand, T. Predicting adaptation rates in model organisms. *in prep.*

Les mutations sont la source ultime de variabilité génétique, il est donc logique que la distribution de leur effet sélectif $f(s)$ soit un paramètre clé en Evolution. Une connaissance de la variation de $f(s)$ entre espèces et entre environnement est donc utile dans un cadre plus général que la seule étude de la dynamique de l'adaptation. Je ne ferai pas ici une revue précise des raisons théoriques pour lesquelles une bonne connaissance de $f(s)$ est necessaire dans de nombreuses prédictions en évolution. Une revue de l'impact des mutations délétères dans les processus évolutifs peut être trouvée dans (Bataillon 2000; Charlesworth and Charlesworth 1998; Lynch et al. 1999). $f(s)$ peut affecter aussi bien les paramètres de populations à l'équilibre (via les équilibres mutation-sélection) que la dynamique du processus d'adaptation (chez les sexués comme les asexués) et les coûts associés (pléiotropie, spécialisation écologique), ainsi que l'évolution du système génétique lui-même (taux de recombinaison et de mutation). Enfin, comme nous l'avons vu en Introduction, l'estimation de paramètres mutationnels a partir de données d'accumulation de mutations dépend elle-même d'hypothèses sur $f(s)$. Ce chapitre présente un modèle conçu pour prédire la forme générale de $f(s)$ et les facteurs affectant sa variation entre espèces ou entre environnements.

Parmi les approches utilisées pour modéliser $f(s)$, j'ai choisi d'utiliser les modèles de paysages adaptatifs tels que le modèle de Fisher (MF), qui a été présenté brièvement en Introduction, il est aussi illustré dans l'Encadré 5.a. Malgré son rôle fondateur dans la modélisation de l'interaction entre mutation et sélection, ce modèle fait plusieurs hypothèses fortes, ce qui a amené à le considérer comme une approche plus heuristique que quantitativement valable:

« *It is important to bear in mind that Fisher's model merely tries to capture the essence of Darwinian evolution* » (Orr 2005b).

Il peut être utile d'évaluer quelles hypothèses limitent le réalisme de ce modèle et celles dont on peut s'affranchir. Une approche pour évaluer la validité d'un modèle en général peut se diviser entre (i) le réalisme de ses prémisses (hypothèses initiales), et (ii) la validité des prédictions qu'il fait en les comparant à des résultats empiriques. Je discute brièvement ici des la validité des prémices du modèle de Fisher, puis les extensions que nous avons proposées pour le rendre plus « réaliste », et pour que ses paramètres soient mesurables. Je discute enfin certaines conséquences évolutives de ces résultats. Des tests de la validité de notre modèle seront proposés dans le chapitre suivant.

Validité des prémices du modèle de Fisher

On peut diviser les différentes hypothèses du MF entre celles ayant trait à l'effet de la mutation sur le phénotype, et celles ayant trait à l'effet du phénotype sur la fitness.

Phénotype et fitness: le mode de sélection supposé est la sélection stabilisante pour un ensemble de traits phénotypiques autour d'un optimum unique. Notons qu'un tel schéma n'interdit pas la sélection directionnelle lorsqu'on est loin de l'optimum. Les preuves de sélection stabilisante ou directionnelle sur des traits quantitatifs sont nombreuses (Kingsolver et al. 2001), mais des preuves de sélection disruptive pour des traits quantitatifs existent aussi (Barton and Keightley 2002). De même, des exemples d'évolution convergente dans les expériences d'adaptation à long terme suggèrent un optimum unique, mais il existe aussi des exemples d'évolution divergente dans des conditions identiques répliquées (Elena and Lenski 2003). Toutefois, ces divergences sont rarement importantes relativement à la trajectoire moyenne. En résumé, si l'hypothèse d'un optimum unique semble parfois valide, elle ne l'est pas toujours. Globalement, l'approche géométrique peut donc donner de bonnes approximations pour $f(s)$, lorsqu'on considère un phénotype ancestral relativement près d'un optimum. La notion de « proximité » ici est relative à l'effet des mutations sur le phénotype. Elle sous-entend donc que la domaine de variation des phénotypes produits par mutation est sous l'influence de la sélection stabilisante exercée par un seul optimum local.

Définition des traits: il convient à ce stade de faire la distinction entre les traits phénotypiques définis dans le modèle de Fisher (et que nous appellerons « traits d'adaptation ») et les traits de fitness. En effet, la fitness est un trait : empiriquement, elle est mesurée comme une composante, plus ou moins intégrée de la survie, du temps de développement, et de la fécondité. Toutefois, ces traits ne sont pas sous sélection pour un optimum, comme le sont les traits d'adaptation: la fitness est toujours sous sélection directionnelle. Par ailleurs, les expériences MA montrent clairement une différence qualitative entre les distributions d'effets des mutations sur les traits de fitness et celles affectant les traits morphologiques (le nombre de soies chez la drosophile par exemple). Les premières présentent souvent très peu voire aucune mutation avantageuse (donc des effets très biaisés sur la valeur du trait), alors que les dernières présentent en général des proportions équivalentes d'effets positifs et negatifs, même si la moyenne peut être biaisée dans un sens (Garcia-Dorado et al. 1999; Keightley et al. 2000; Keightley and Ohnishi 1998; Yang et al. 2001).

Mutation et phénotype: La distribution de l'effet des mutations sur le phénotype (les traits d'adaptation) est supposée continue et symétrique, souvent Gaussienne dans le MF. L'hypothèse de continuité ne semble pas trop choquante, même les traits discrets résultent parfois de traits sous-jacents continus avec un phénotype à seuil (Orr 2001). En revanche, la symétrie des effets ne va pas de soi. On pourrait par exemple supposer que la mutation tend à entraîner en moyenne la perte de certaines fonctions enzymatiques. Toutefois, dans un système comprenant de nombreuses boucles de régulation positives comme négatives, la perte d'une fonction enzymatique donnée peut *a priori* entraîner aussi bien l'augmentation que la diminution des produits métaboliques dont elles influencent la synthèse. Si la concentration de ces métabolites est vue comme un trait d'adaptation soumise à sélection stabilisante, la distribution de l'effet des mutations sur ce trait peut alors être symétrique. Globalement, l'hypothèse de symétrie des effets des mutations sur les traits phénotypiques semble au moins un bon modèle nul.

Finalement, l'hypothèse d'une distribution gaussienne des effets, qui permet nombre des traitements mathématiques, a de bonnes chances d'être fausse en général (Garcia-Dorado et al.). Cependant, la force du modèle de Fisher, comme l'a déjà souligné Orr (Orr) est qu'il

requiert seulement l'existence d'un système équivalent de traits pour lesquels le traitement mathématique soit possible et qui génère un effet global similaire sur $f(s)$. Or la façon de mesurer les traits d'adaptation est un choix arbitraire. L'hypothèse d'une distribution gaussienne de l'effet des mutations sur les traits d'adaptation se réduit donc en fait à l'hypothèse qu'il existe, pour chaque trait, une transformation par laquelle cette distribution devient gaussienne. Il me semble donc que le problème d'une distribution gaussienne est plus ou moins un faux problème lorsqu'il s'agit d'intégrer un ensemble de traits pour déterminer la fitness. Le problème reste en revanche entier lorsqu'on s'intéresse à un trait donné !

En résumé, il me semble que les hypothèses de base de l'approche par les paysages adaptatifs ne sont pas si irréalistes, du moins pour traiter de $f(s)$ lorsqu'on est près d'un optimum. Le problème de la validité du modèle loin de l'optimum est lié à la forme de la fonction de fitness: jusqu'à quelle distance de l'optimum l'approximation quadratique reste-t-elle valide ? Ce problème peut être abordé à partir des données empiriques (voir Chapitre II). Il n'y a en revanche pas de raisons de supposer que la validité des hypothèses sur l'effet phénotypique des mutations (continuité, et symétrie) dépendent de la distance à l'optimum, puisque celles-ci ont trait à la relation génotype-phénotype et non à la relation phénotype fitness.

On peut en revanche distinguer deux hypothèses fortement restrictives dans ce modèle: la pléiotropie universelle (chaque mutation affecte tous les traits d'adaptation) et les hypothèses de symétrie: l'effet des mutations sur les traits et l'effet de ceux-ci sur la fitness sont isotropes dans l'espace phénotypique. Biologiquement cela signifie qu'il n'y a (i) pas de différences entre traits pour la variance mutationnelle ou pour la force de la sélection stabilisante et (ii) pas de corrélations entre traits par mutation ou dans leur effet sur la fitness. Ces hypothèses sont clairement irréalistes et peuvent avoir un effet très important sur les prédictions quant à la forme de $f(s)$ et à ses conséquences évolutives. Je décris ici un modèle permettant de relâcher ces hypothèses.

Relâchement des hypothèses de symétrie dans le modèle de Fisher

I.1 PRINCIPE DU MODELE

Ce modèle est détaillé dans l'article 1, j'explique ici son principe, il est aussi illustré dans l'Encadré 5.b. Un phénotype est modélisé comme un vecteur $\mathbf{z}$ de n traits continus sous sélection stabilisante pour un optimum (arbitrairement le vecteur nul $\mathbf{0}$). La mutation crée une déviation $\mathbf{dz}$ sur ce vecteur phénotypique distribuée de façon gaussienne et symétrique pour chaque trait. L'apport par rapport aux hypothèses de base du modèle de Fisher, est de considérer que la sélection et la mutation ont des effets variables et potentiellement corrélés sur les traits, en utilisant les modèles multivariés empruntés à la génétiques quantitative, par ex. (Lande 1980). Ainsi on définit une matrice arbitraire $\mathbf{S}$ décrivant la force de la sélection stabilisante sur chaque trait et les interactions entre traits pour déterminer la fitness de $\mathbf{z}$: $W(\mathbf{z}) = Exp(-½\ \mathbf{z}^T\ \mathbf{S}\ \mathbf{z})$ (^{T} dénote la transposition). De même, on définit une matrice de covariances mutationnelles arbitraire $\mathbf{M}$ qui définit les variances mutationnelles pour chaque trait et les corrélations entre traits par mutation. Le principe est qu'un tel système peut être réduit par une transformation linéaire des traits (qui n'altère pas les hypothèses gaussiennes). Ainsi, il existe toujours un espace de Fisher équivalent dans lequel les traits (i) sont non corrélés pour la fitness et (ii) mutent de façon indépendante et équivalente entre traits (voir Annexe 1 de l'article). Le système transformé n'est pas pour autant isotrope, l'effet λ_i de chaque trait (transformé) sur la fitness diffère entre traits. Les λ_i sont simplement les valeurs propres de $\mathbf{S}$

M, le produit des matrices de covariances sélectives et mutationnelles. On suppose que le génotype ancestral a un phénotype $\mathbf{z}_o$ (la distance à l'optimum), $\mathbf{x}_o$ dans l'espace transformé. Ce modèle suppose une équivalence entre génotype et phénotype : dans la suite, j'utiliserai l'un ou l'autre terme de manière équivalente.

I.2 MOMENTS ET DENSITE DE PROBABILITE $f(s)$

On peut alors donner la distribution exacte de s (une forme quadratique de vecteurs gaussiens) mais sa densité $f(s)$ n'a pas d'expression analytique. On peut en revanche obtenir une expression simple pour tous les cumulants de $f(s)$ (les trois premiers cumulants sont les trois premiers moments centrés). Nous nous focaliserons ici sur la variance $V(s)$ et la moyenne E(s) de $f(s)$. Définissons d'abord $s_o = \mathbf{z}_o^T \mathbf{S} \mathbf{z}_o$, la distance à l'optimum exprimée en terme de fitness. Le point important est que cette grandeur est mesurable empiriquement: c'est le logarithme du ratio de fitness entre un génotype très bien adapté au milieu considéré (*i.e.* de phénotype optimal $\mathbf{z} \approx \mathbf{0}$) et le génotype ancestral (de phénotype $\mathbf{z}_o$). Elle est aussi interprétable biologiquement: s_o mesure le niveau d'adaptation de l'ancêtre à l'environnement où la fitness est mesurée. Définissons aussi $\bar{\lambda}$ et CV(λ) la moyenne et le coefficient de variation des valeurs propres λ_i de **S M**. Enfin, considérons que la sélection est faible de sorte que $\log(1+s) \approx s$ (ou à défaut définissons s comme l'effet sur la log fitness). On a alors

$$\begin{aligned} E(\mathrm{s}) &= -\bar{s} = -\frac{n}{2}\bar{\lambda} \\ V(s) &= \frac{n}{2}\bar{\lambda}^2(1+\mathrm{CV}(\lambda)^2)\left(1+2\,q\,\frac{s_o}{\bar{s}}\right), \end{aligned} \qquad (3)$$

Où q est un coefficient qui dépend de la direction de $\mathbf{x}_o$ (non de sa norme). Ainsi, différents phénotypes se trouvant à la même distance en fitness de l'optimum (même s_o), mais dans des directions différentes correspondent à différentes valeurs de q. Ce paramètre n'est pas mesurable, mais

(i) on peut prouver que sa moyenne est 1 (pour différentes valeurs de $\mathbf{x}_o$ prises au hasard)

(ii) sa variance autour de cette moyenne est réduite. Sous un modèle assez général, q oscille environ entre 0.5 et 1.5.

Ces arguments sont détaillés dans l'article 3, et sont confirmés par des simulations.

L'Eq. (3) permet déjà de tirer deux conclusions sur l'influence de la maladaptation du génotype ancestral à l'environnement:

(i) La maladaptation n'influence pas l'effet moyen des mutations ($\bar{s}$ ne dépend pas de s_o)

(ii) La variance des effets $V(s)$ augmente linéairement avec la maladaptation (proportionnellement à s_o)

On peut ensuite utiliser ces deux premiers moments pour obtenir une bonne approximation analytique de la densité de probabilité de s en termes d'une gamma déplacée (définie en Introduction, voir Figure 4). C'est la distribution de s_o - γ où γ suit une loi gamma de paramètres d'échelle α et de forme β choisis pour que la distribution ait la variance $V(s)$ et la

moyenne $E(s)$ données par L'Eq. (3). Cette distribution est une approximation intéressante car :

(i) Lorsque les valeurs propres sont égales (CV(λ) = 0: traits équivalents comme dans le MF original) et que $s_o = 0$ (mutations délétères uniquement) la gamma déplacée correspond à la distribution exacte (une gamma).

(ii) La distribution exacte de s est également bornée à droite par s_o.

(iii) La gamma déplacée peut être facilement utilisée dans les analyses par maximum de vraisemblance des données MA (voir Introduction et Shaw et al. 2002)

(iv) Cette approximation donne un très bon ajustement aux simulations exactes avec peu de paramètres

(v) Enfin, la qualité de l'approximation peut être calculée en comparant les moments supérieurs de l'approximation et de la distribution exacte. Cette qualité reste assez bonne quelle que soit la distance à l'optimum et la variance de la distribution des λ_i (si celle-ci est supposée gamma, par exemple).

L'autre intérêt de cette approximation est qu'elle définit explicitement un « nombre effectif de dimensions » n_e, qui ne dépend que des paramètres de mutation (**M**) et de sélection (**S**) de l'espèce dans un environnement donné (mais pas du phénotype ancestral $\mathbf{z}_o$). L'approximation consiste en effet à considérer la distribution d'un nombre n_e de traits qui auraient tous les même effets λ_e sur la fitness (correspondant à un espace de type Fisher isotrope) et génèreraient la même moyenne et variance de s. n_e correspond donc bien au nombre effectif de dimensions initialement proposé par Orr (1998; 2000a), puis repris par Barton & Keightley (2002).

En utilisant ces résultats et en considérant $s_o = 0$ (génotype ancestral très bien adapté à l'environnement), toutes les mutations sont d'effet délétère et leur distribution est une gamma négative de paramètres

$$\begin{aligned} \beta_o &= \frac{n_e}{2} = \frac{1}{2}\,\frac{n}{1+CV(\lambda)^2} \\ \alpha_o &= \frac{\bar{s}}{\beta_o} = \lambda_e = \bar{\lambda}\ (1+\mathrm{CV}(\lambda)^2) \end{aligned} \quad . \qquad (4)$$

Dans ce modèle assez général, $f(s)$ est déterminée par un ensemble d'interactions phénotypiques mutationnelles et sélectives (les matrices **S** et **M**), mais peut finalement être approximée en ne connaissant que le nombre, la moyenne et le coefficient de variation des valeurs propres λ_i de la matrice **S M**.

Lorsque le génotype ancestral n'est pas parfaitement adapté ($s_o > 0$), certaines mutations sont avantageuses et la distribution est une gamma déplacée de densité

$$f(s) = \frac{e^{-\frac{s_o - s}{\alpha}}(s_o - s)^{\beta-1}\alpha^{-\beta}}{\Gamma(\beta)}, \qquad (5)$$

où Γ(.) est la fonction gamma et où les paramètres de forme β et d'échelle α sont

$$\beta = \beta_o \frac{(1 + s_o / \bar{s})^2}{(1 + 2\,q\,s_o / \bar{s})} \text{ and } \alpha = \frac{\bar{s}}{\beta_o} \frac{(1 + 2\,q\,s_o / \bar{s})}{(1 + s_o / \bar{s})}, \qquad (6)$$

Où, comme on l'a vu, q = 1 en moyenne et varie peu autour de cette espérance. Cette approximation est très correcte (comme l'illustre l'**Erreur ! Source du renvoi introuvable.**) bien qu'elle ait peu de paramètres (s_o, β_o et $\bar{s}$). Lorsque l'on considère des génotypes ancestraux assez loin de l'optimum, c'est à dire pour $s_o >> \bar{s}$, on a, avec q = 1, $\beta \approx \beta_o \left(\frac{3}{4} + \frac{s_o}{2\,\bar{s}} \right)$ et $\alpha \approx 2\alpha_o$. On voit qu'alors que le paramètre de forme β augmente à peu près linéairement avec la maladaptation s_o, celui d'échelle reste à peu près constant (compris entre α_o et 2 α_o). Lorsque $s_o >> \bar{s}$, la distribution est proche d'une gaussienne de moyenne $\bar{s}$ et de variance $V(s)$ donnée en Eq. (3) (voir l'Encadré 6). Cela est dû au fait que β devient grand lorsque $s_o >> \bar{s}$, or une gamma de forme très grande se rapproche d'une gaussienne.

Enfin, on peut par exemple montrer la gamma déplacée proposée ici remplit la condition nécessaire et suffisante (Eq. A.2 dans (Orr 2003)), pour être de type Gumbel III (voir Appendice 1). Les prédictions de l'EVT s'appliquent donc à cette distribution (comme à nombre d'autres).

I.3 CONCLUSIONS

On peut tirer plusieurs conclusions de ce modèle quant à la distribution de l'effet des mutations (délétères comme avantageuses), notamment sur l'influence de divers facteurs sur $f(s)$.

Nombre de traits et leurs interactions mutationnelles et sélectives: L'influence de l'ensemble des variances et covariances mutationnelles et sélectives (**S** et **M**) se résume approximativement à leur influence sur la moyenne et la variance des valeurs propres de **S M** (*i.e.* sur $\bar{\lambda}$ et CV(λ)). L'Eq. (4) montre en outre que même si le nombre de traits n est très grand, le nombre de dimensions efficaces n_e peut être bien plus réduit si les interactions phénotypiques sont très hétérogènes entre traits (CV(λ) grand). Notons une conséquence empirique de ce point. On peut définir les organismes « complexes » comme ceux dans lesquels un grand nombre de fonctions différentes (de traits) peuvent être affectés par la mutation de façon pléiotrope, ce qui revient à n grand (par ex. organismes pluricellulaires différenciés). Ces résultats montrent que, même dans ces organismes, on s'attend à un n_e petit, si les traits phénotypiques sont hétérogènes et covarient pour la sélection et la mutation. On s'attend donc à observer des $f(s)$ fortement asymétriques (un paramètre de forme $\beta_o = n_e/2$ réduit), du moins lorsque les effets délétères dominent ($s_o \approx 0$, ancêtre bien adapté). Au contraire, en négligeant cette hétérogénéité (CV(λ) = 0), et en supposant un grand nombre de traits sous sélection, on obtient des distributions à peu près gaussiennes de s. Or, les distributions observées empiriquement pour des traits de fitness sont quasiment toujours nettement plus asymétriques que la gaussienne (Garcia-Dorado et al. 1999), même dans des espèces à plan d'organisation complexe comme la drosophile (Lyman et al. 1996).

Effet de l'adaptation du génotype ancestral à son environnement: le modèle permet de prédire comment change $f(s)$ avec s_o. Je dégagerai trois conclusions majeures :

(i) L'Eq. (3) montre que la variance de s augmente linéairement avec s_o et que sa moyenne est inchangée. Loin de l'optimum $f(s)$ tend à ressembler à une gaussienne (Encadré 6).

(ii) L'effet de la distance à l'optimum (s_o) est relatif à l'effet moyen des mutations: il est déterminé par $s_o/\bar{s}$. Le paramètre $\bar{s}$, l'effet moyen des mutations (constant entre environnements), est un paramètre d'échelle. Il donne une mesure du pas moyen parcouru par une mutation, mais ce pas est mesuré en termes de fitness, et non comme une distance dans l'espace phénotypique sous-jacent (modèle de Fisher classique). De façon similaire, s_o mesure la distance à parcourir pour atteindre l'optimum, là aussi en termes de fitness et non de phénotype. Il est logique que cette distance n'ait de sens que relativement à la taille d'un pas. Une telle mise à l'échelle de la distance à l'optimum a été mise en évidence par Orr, mais pour des distances mesurées dans l'espace phénotypique (Orr 1998; Orr 2000a).

(iii) Enfin, quand s_o augmente, la proportion et l'effet des mutations avantageuses augmente aussi. Cela semble assez logique, et correspond également à une conclusion classique du modèle de Fisher. Notre modèle permet seulement de quantifier cet effet dans un contexte plus « réaliste ». Notons que cette prédiction est en accord avec les observations empiriques (voir Introduction): lorsque le génotype ancestral est maladapté, on observe (i) une grande proportion de mutations avantageuses, et (ii) des taux d'adaptation plus élevés.

Notre approche montre donc que certaines des conclusions du modèle de Fisher classique sont valides dans un contexte plus réaliste qui ne néglige pas les différences et les interactions entre traits pour la mutation et la sélection. En revanche, la prise en compte des interactions phénotypiques, pour la mutation comme pour la sélection, change quantitativement les prédictions quant à l'influence du nombre de traits. La forme de $f(s)$ est en effet déterminée par n_e (et non n) qui peut être assez indépendant de n si les traits sont très hétérogènes. Nous discuterons de cet aspect plus loin. Au-delà de l'étude théorique des facteurs influençant $f(s)$, le modèle permet de faire des prédictions sur $f(s)$ à partir de mesures empiriques. Je détaille maintenant cette approche.

Estimation des paramètres et prédictions: Les paramètres du modèle sont mesurables, parce que nous opérons un changement de variable des distances phénotypiques classiquement utilisées (taille des pas, distance à l'optimum) vers des différences de fitness ($\bar{s}$, s_o, qui sont mesurables). On a vu en Introduction qu'il existe des méthodes pour estimer β_o et α_o (ou $\bar{s}$) d'après une distribution de fitness de lignées mutantes. Ces méthodes nécessitent des protocoles lourds, mais elles ont été appliquées avec succès chez plusieurs organismes modèles. Une mesure de s_o pour un environnement donné peut être obtenue en mesurant simultanément la fitness du génotype ancestral (W_o), et celle d'un génotype très bien adapté à cet environnement (W_{max}). Le log ratio de ces deux fitness mesure $s_o = \mathbf{z}_o\ \mathbf{S}\ \mathbf{z}_o = \log(W_{max}/W_o)$. Notons que cette formule correspond à une fonction de fitness gaussienne. Sous l'hypothèse d'une fonction de fitness quadratique on prendra $s_o = \mathbf{z}_o\ \mathbf{S}\ \mathbf{z}_o = W_{max} - W_o$. Ces deux expressions sont équivalentes si s_o est petit. En effet s_o est approximativement le coefficient de sélection du génotype parfaitement adapté relativement à l'ancêtre. Si on pose $W_{max}/W_o << 1$ alors on a $\log(W_{max}/W_o) \approx W_{max} - W_o = s_o$.

Notons aussi que l'on peut estimer ces paramètres dans un environnement auquel le génotype ancestral n'est pas parfaitement adapté ($s_o>0$), en ajustant la distribution empirique d'effets des mutations à une gamma déplacée au lieu d'une gamma. Dans ce cas on obtiendra

une estimation des trois paramètres: β_o, $\bar{s}$ et la distance à l'optimum s_o pour l'environnement considéré.

A partir de ces mesures, on peut théoriquement prédire $f(s)$ dans n'importe quel autre environnement pour lequel s_o est connu. On peut de là déduire la proportion de mutations avantageuses et la dynamique de l'adaptation (en prenant les expressions appropriées pour la probabilité de fixation des mutations). C'est l'approche que nous utilisons dans l'Article 3 (voir le chapitre suivant).

Cette prédiction repose toutefois sur une hypothèse clé: que β_o et $\bar{s}$ (donc $f(s)$ à l'optimum) sont approximativement les mêmes, quel que soit l'environnement considéré (quel que soit cet optimum).

Si cette hypothèse est valide, la mesure de ces paramètres dans un environnement standard du laboratoire, suffit pour prédire $f(s)$ dans un autre environnement, la seule mesure supplémentaire à faire dans ce nouvel environnement est celle de s_o. A première vue, cette hypothèse semble correspondre à l'idée qu'un changement de l'environnement détermine un changement de l'optimum pour les traits d'adaptation sous-jacents, mais sans affecter les paramètres mutationnels (**M**) et sélectifs (**S**) de chaque trait. Cela supposerait donc que l'environnement affecte peu l'expression du génotype (pas de plasticité sur les traits d'adaptation), et la force de la sélection. Notons que cette hypothèse est à la base du modèle classique de Via et Lande (1985) sur l'évolution des normes de réactions. L'hypothèse est en fait moins restrictive car $f(s)$ n'est déterminée que par le nombre (n), la moyenne ($\bar{\lambda}$) et le coefficient de variation (CV(λ)) des valeurs propres de **S M** (voir Eq. (4)), non par les termes exacts de chacune des matrices. Or de nombreuses valeurs de **S** et **M** peuvent générer la même moyenne et variance des valeurs propres de **S M**. Le sens et la force des interactions sélectives et mutationnelles peuvent donc changer d'un environnement à l'autre, il suffit que leur effet net sur la fitness ait la même moyenne et variance (même hétérogénéité entre traits). Cette hypothèse est donc moins restrictive. Supposons deux environnements e_1 et e_2, et supposons que les traits sont plastiques et que la force de la sélection sur chaque trait dépend de l'environnement (**M** et **S** changent avec l'environnement). Alors, β_o et $\bar{s}$ seront égaux dans les deux environnements si en moyenne, il y a autant de traits plus exprimés et plus sélectionnés dans e_1 que de traits plus exprimés/sélectionnés dans e_2 (même si ce ne sont pas les même traits). En effet dans ce cas l'effet net de la mutation et la sélection sur l'ensemble des traits donne à peu près la même moyenne et variance des λ_i dans les deux environnements.

Dans la partie suivante, je détaille un modèle nul (exposé dans l'Article 1) où les covariances **M** et **S** sont aléatoires. Avec ce modèle, on attend en fait une faible variation de la forme de $f(s)$ entre différentes espèces ou environnements.

Modèle d'interactions phénotypiques aléatoires

Comme nous venons de le voir, les paramètres de base de $f(s)$ (β_o et $\bar{s}$) peuvent être mesurés pour une espèce donnée dans un milieu donné à partir d'une distribution de mutations délétères. Toutefois, étant donné la lourdeur des protocoles expérimentaux permettant ces mesures, elles n'ont été réalisées que sur quelques lignées de quelques espèces modèle, et souvent dans l'environnement standard du laboratoire. Il serait donc utile de savoir dans quelle mesure ces estimations peuvent être extrapolées à d'autres environnements, et d'autres espèces. En outre, il serait intéressant de relier la distribution des λ_i à des caractéristiques (plus « biologiques ») des interactions phénotypiques entre traits.

Une approche pour aborder cette question est de considérer un modèle nul où les matrices de covariance (mutationnelles et sélectives), sont aléatoires (par exemple entre

environnements) et de grande dimension (n grand). On peut alors utiliser un outils puissant en théorie probabiliste: la théorie des matrices aléatoires. Celle-ci fournit plusieurs résultats généraux sur la distribution des valeurs propres de matrices tirées aléatoirement (voir une revue des résultats principaux dans (Bai 1999), et de leurs applications en physique et en finance dans (Forrester et al. 2003)). Les résultats sont asymptotiques (pour de grandes valeurs de n). Ils s'appliquent donc à notre problème si on suppose que le nombre de traits sous sélection (n) est toujours relativement grand, ce qui semble assez réaliste.

Des détails techniques peuvent être trouvés dans l'appendice théorique (section II) et l'Article 1, je ne donne ici que les résultats principaux. Les matrices **S** et **M** peuvent être exprimées comme un produit $\mathbf{S} = \mathbf{X_S}\ \mathbf{X_S}^T$ et $\mathbf{M} = \mathbf{X_M}\ \mathbf{X_M}^T$ car ce sont des matrices symétriques. Les matrices $\mathbf{X_S}$ et $\mathbf{X_M}$ sont de dimensions $n \times p_\mathbf{S}$ et $n \times p_\mathbf{M}$ respectivement, et leurs éléments sont tirés dans des distributions centrées de variances $\sigma_\mathbf{S}$ et $\sigma_\mathbf{M}$ respectivement (la nature des distributions peut différer entre **M** et **S**).

On peut alors dégager un résultat important. Les moments de la distribution des valeurs propres λ_i du produit **S M** (pour n'importe quelles **S** et **M** aléatoires) tendent vers des valeurs asymptotiques uniques (lorsque n est grand). En outre, ce résultat ne dépend pas des distributions exactes (gaussienne, uniforme etc..) dans lesquelles sont tirés les éléments de $\mathbf{X_S}$ et $\mathbf{X_M}$. Il dépend seulement de leur variance ($\sigma_\mathbf{S}$ et $\sigma_\mathbf{M}$) et du ratio des dimensions des matrices **X** ($y_\mathbf{S} = n/p_\mathbf{S}$ et $y_\mathbf{M} = n/p_\mathbf{M}$). Ainsi, tant que ces quatre paramètres, ainsi que n, sont constants, les paramètres $\bar{\lambda}$ et CV(λ) sont également constants, même si les matrices **S** et **M** varient aléatoirement (par exemple entre environnements ou entre lignées). On a alors, d'après les Eq. (3) et (4), les expressions suivantes:

$$\begin{aligned} \bar{s} &= \frac{n}{2}\,\bar{\lambda} \underset{n>>1}{\approx} \frac{n}{2}\,\sigma_\mathbf{S}^2\,\sigma_\mathbf{M}^2 \\ \beta_0 &= \frac{n_e}{2} = \frac{1}{2}\frac{n}{1+CV(\lambda)^2} \underset{n>>1}{\approx} \frac{1}{2}\frac{n}{1+y_\mathbf{S}+y_\mathbf{M}} \end{aligned} \qquad (7)$$

Ces résultats asymptotiques convergent rapidement, de sorte qu'ils sont assez précis même pour de faibles valeurs de n (par ex. $n \geq 30$ dans mes simulations) (Bai 1999). Ces résultats supposent aussi que $p_\mathbf{M}$ et $p_\mathbf{S}$ ne sont pas trop petits (même s'ils peuvent être nettement plus petits que n), là aussi l'asymptote est valable dès les faibles valeurs de $p_\mathbf{M}$ ou $p_\mathbf{S}$ (par ex. ≥ 7).

Il serait souhaitable d'obtenir une vision plus intuitive du sens biologique de ces paramètres : on sait déjà que $\sigma_\mathbf{S}$ et $\sigma_\mathbf{M}$ sont des paramètres d'échelle de la distribution des λ_i, mais que dire de $y_\mathbf{S}$ et $y_\mathbf{M}$? Pour cela, il faut faire l'hypothèse supplémentaire que les éléments des matrices **X** sont tirés dans des distributions gaussiennes : $N(0,\sigma_\mathbf{S})$ et $N(0,\sigma_\mathbf{M})$. Les matrices **S** et **M** sont alors tirées dans une distribution de Wishart, très utilisées en statistiques notamment. Sous cette hypothèse, on peut relier la distribution des corrélations phénotypiques dans les matrices **S** et de **M** aux paramètres $y_\mathbf{S}$ et $y_\mathbf{M}$. Cette distribution est donnée dans l'article 1, et illustrée dans l'Encadré 7. On peut mesurer la force des corrélations (en valeur absolue), comme la déviation standard de cette distribution, respectivement $\rho_\mathbf{S}$ et $\rho_\mathbf{M}$. On a alors $y_\mathbf{S} = n\rho_\mathbf{S}^2$ et $y_\mathbf{M} = n\rho_\mathbf{M}^2$ et on peut écrire

$$\beta_0 = \frac{n_e}{2} \underset{n>>1}{\approx} \frac{1}{2}\frac{n}{1+n(\rho_\mathbf{S}^2+\rho_\mathbf{M}^2)} \underset{n\to\infty}{\to} \frac{1}{2}\frac{1}{\rho_\mathbf{S}^2+\rho_\mathbf{M}^2}. \qquad (8)$$

Cette équation permet de relier la force des corrélations mutationnelles et sélectives et le nombre de traits sous sélection à la forme de $f(s)$ à l'optimum. Elle illustre une propriété

importante de ce modèle de covariances aléatoires: l'effet du nombre de traits sous sélection est limité. En effet, si $\rho_{\mathbf{S}}^2 + \rho_{\mathbf{M}}^2$ n'est pas trop petit, le paramètre de forme β_o converge assez vite vers une valeur indépendante du nombre de traits: $1/(2(\rho_{\mathbf{S}}^2 + \rho_{\mathbf{M}}^2))$. Pour illustration, en supposant des corrélations mutationnelles et sélectives moyenne de l'ordre de 0.3, β_o varie entre 1.78 et 2.77 pour n'importe quel nombre de traits (n) supérieur à 10. Ces valeurs sont comparables à celles observées empiriquement (voir Chapitre II).

En résumé, on peut tirer plusieurs conclusions sur la variation attendue des paramètres de $f(s)$ entre espèces ou entre environnements. Supposons qu'au sein d'une même espèce, différents environnements correspondent à des variations aléatoires des matrices **S** et **M**. On attend alors, entre différents environnements:

(i) une variation limitée de β_o si la force des corrélations reste approximativement la même dans ces environnements

(ii) cette prédictions devrait être assez robuste à un changement du nombre de traits sous sélection entre environnements, si ces corrélations sont relativement fortes

(iii) on prédit des valeurs assez faibles de β_o si ces corrélations sont raisonnablement fortes (donc des distributions $f(s)$ assez asymétriques).

Par ailleurs on voit aussi que la variation de β_o pourrait être assez limitée d'une espèce à l'autre (en supposant que celles-ci déterminent différentes valeur de n). En revanche, on ne peut attendre une variation limitée de $\bar{s}$ que si le nombre de traits sous sélection varie peu entre environnements et que la force de la sélection par trait varie peu en moyenne sur l'ensemble des traits (Eq. (7)).

En résumé, ce modèle nul suggère une variation limitée de la forme de la distribution (entre espèces et environnements) mais pas nécessairement de sa moyenne. Le débat, à terme, ne peut de toutes façons être tranché que par la comparaison de mesures empiriques (voir le Chapitre II). Dans la section suivante, je détaille certaines implications théoriques de notre modèle.

Implications évolutives

I.4 TAUX D'ADAPTATION ET COUT DE LA COMPLEXITE

Une conséquence importante de ce modèle a trait au coût de la complexité. Celui-ci, initialement identifié par Fisher (Fisher 1930), a reçu un traitement plus complet récemment avec les travaux de Orr (Orr 1998; Orr 2000a) qui résume ce coût ainsi: toutes choses étant égales par ailleurs, les mutations avantageuses sont moins probables et tendent à être de plus faible effet dans un organisme complexe (avec plus de traits sous sélection). Orr montre que cette contrainte peut constituer une forte limite à l'adaptation dans les organismes complexes. Cette conclusion découle d'une approche basée sur le modèle de Fisher, mais sans prendre en compte les corrélations phénotypiques. On peut la comprendre avec notre modèle en utilisant une approximation pour le taux d'adaptation

(i) dans une population sexuée de grande taille

(ii) lorsque le génotype ancestral est relativement loin de l'optimum ($s_o >> \bar{s}$)

Alors, on peut approximer $f(s)$ par une gaussienne (car $s_o >> \bar{s}$), et utiliser cette densité approximative pour calculer le taux d'adaptation (d'après l'Eq. (1)). En supposant une population d'individus diploïdes de taille efficace N_e, supposée grande on a $p_{fix}(s) = N_e/N\, s$ (où s est l'effet de la mutation à l'état *homozygote* en supposant la codominance) et le taux d'adaptation est

$$\frac{d\log(W)}{dt} \underset{s_o >> \bar{s}}{\approx} \frac{N_e U \bar{s}}{\beta_o} s_o , \qquad (9)$$

(en posant q = 1, voir calcul dans l'article 3). L'Encadré 8 illustre la valeur de cette approximation en fonction de la distance à l'optimum s_o. L'approximation n'est pas très bonne pour les faibles valeurs de s_o ou les grandes valeurs de β_o, mais elle donne une bonne intuition du comportement général du taux d'adaptation en fonction des paramètres du modèle.

A partir de cette approximation, on peut voir l'effet de la complexité (n) sur le taux d'adaptation. Supposons que tous les traits sont équivalents, de sorte que CV(λ) = 0 et $n_e = n$ (modèle de Fisher classique). Alors le paramètre de forme β_o vaut $n/2$, de sorte que toutes choses étant égales par ailleurs (pour une même valeur de $N_e U \bar{s}\, s_o$), le taux d'adaptation tend donc à décroître approximativement en $1/n$. En revanche, si on prend en compte les covariances sélectives ou mutationnelles, c'est le nombre *efficace* de traits (n_e) qui détermine le taux d'adaptation ($\beta_o = n_e/2$). Or nous avons vu dans la section précédente que pour des valeurs raisonnables des corrélations phénotypiques, la valeur de β_o est peu sensible au nombre de traits et reste toujours assez faible. Ainsi, le coût de la complexité est alors faible, car un grand nombre de traits hétérogènes (CV(λ)>>0) équivaut, en terme de distribution de s, à un nombre plus réduit de traits équivalents.

Par ailleurs, le taux d'adaptation est proportionnel à $\bar{s}$. Cela s'explique par le fait que $\bar{s}$ est une mesure de la taille moyenne des pas dans le paysage phénotypique. Il donne donc aussi l'échelle des pas vers l'optimum lors d'une phase d'adaptation, donc celle du taux d'adaptation. Comme $\bar{s}$ est lui-même proportionnel à n (Eq. (3)), on peut s'attendre à ce que cet effet puisse avantager les organismes plus complexes (ayant un n plus grand).

Finalement on voit que lorsqu'on modélise explicitement les interactions phénotypiques entre traits, le coût de la complexité peut se révéler quasiment nul. Là encore la prédiction est valable *toutes autres choses étant égales par ailleurs*. Or, il est possible par exemple que le niveau de sélection par trait $\bar{\lambda}$ soit plus élevé chez les organismes peu complexes, compensant ainsi les valeurs de $\bar{s}$. Enfin, il est maintenant établi que les organismes plus complexes ont des valeurs plus faibles de $N_e U$. Cette fois le terme « complexe » correspond à des espèces ayant des plans d'organisation plus ou moins complexes, par ex. micro-organismes *vs*. pluricellulaires (Lynch and Conery 2003).

Enfin, le modèle permet de caractériser un nombre de dimensions efficaces mesurable (n_e = 2 β_o). Notre approche par les matrices aléatoires suggère que β_o varie peu entre environnements ou entre génotypes. Dans ce cas, notre définition de n_e devrait être *a priori* inhérente à une espèce (*i.e.* constante pour cette espèce entre environnements et génotypes). Ce n_e est aussi peut-être plus facile à mesurer (à partir de mutations délétères) qu'un n_e défini à partir du taux d'adaptation (Orr 2000a), celui-ci étant plus variable intra-espèce.

I.5 REMARQUES SUR LA MODULARITE

Le modèle présenté ici suppose que la mutation affecte conjointement un ensemble de n traits. Or il est possible que le phénotype soit organisé en modules mutationnels, de sorte que chaque mutation n'affecte qu'un sous-ensemble des traits. Welch et Waxman (2003) ont

proposé un traitement général de l'effet de la modularité sur $f(s)$ et sur les taux d'adaptation, en supposant l'équivalence entre traits, mais pour une grande diversité de distributions de l'effet total des mutations sur le phénotype. Ils montrent que la modularité a un effet limité sur les taux d'adaptation et le coût de la complexité. Notons que si la modularité peut sembler être une caractéristique générale de la relation phénotype-génotype (par ex. modularité du développement), son effet sur la fitness lorsqu'on introduit des covariances sélectives peut être complexe (voir la discussion dans l'Article 1). Une modification du modèle présenté ici est nécessaire pour prendre en compte la modularité. En effet, un système modulaire général correspondrait à l'idée que selon le gène (ou le module) touché par une mutation, la distribution des effets sur l'ensemble des traits peut changer. Un modèle général de modularité peut être fait en définissant un ensemble de matrices de covariance mutationnelles $\mathbf{M}_m$, conditionnelles au module affecté par une mutation donnée. On peut décider que chaque module affecte un sous ensemble de traits, mais on peut aussi plus généralement supposer que les modules se superposent plus ou moins, ou que certains modules affectent les même traits mais en induisant des variances et des corrélations mutationnelles différentes etc. D'où une traduction en terme de matrices conditionnelles générales. La modularité n'affectant pas la sélection (donc la matrice **S**), on peut alors déduire la distribution de s avec notre modèle en calculant la distribution nette des effets phénotypiques de la mutation sur tous les modules, ce qui donne une multivariée gaussienne de matrice de covariance $\mathbf{M} = \sum_m p(m)\mathbf{M}_m$, où $p(m)$ est la probabilité qu'une mutation affecte le module m. On retrouve par cette approche un résultat important de Welch et Waxman (Welch and Waxman 2003) qui est que lorsqu'on considère des modules équivalents, la modularité n'a aucun effet sur la distribution. Dans notre cas, cela correspond à constater que toutes les partitions de m matrices équivalentes $\mathbf{M}_m$ donnent la même matrice **M** une fois sommées (quel que soit m). Je n'ai pas développé cette approche, mais il serait intéressant de prédire la distribution des valeurs propres λ_i de **S M** pour différentes formes de modularités en calculant la matrice **M** résultante.

I.6 CONCLUSION

Notre modèle fait diverses prédictions qui peuvent être testées à partir des données empiriques sur la distribution de l'effet des mutations délétères (i) entre espèces et (ii) entre environnements. On peut aussi tester plus indirectement ses prédictions à partir de la dynamique de l'adaptation à long terme. Pour certaines des questions posées par le modèle la théorie n'apporte aucun *a priori* net, et c'est alors l'analyse de données empiriques qui doit trancher. C'est le cas par exemple pour la validité du modèle loin de l'optimum, ou pour la variation de $\bar{s}$ avec n ou avec l'environnement. Des approches permettant de faire ces tests sont exposées dans le chapitre suivant.

CHAPITRE II: TESTS DU MODELE ET APPLICATIONS

Articles

Article 1: Martin, G. & Lenormand, T. A multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution* 60, 893-907 (2006).

Article 2: Martin, G. & Lenormand, T. The fitness effect of mutations in stressful environments: a survey in the light of fitness landscape models. *Evolution* (2006) in press.

Article 3: Martin, G. & Lenormand, T. Predicting adaptation rates in model organisms. *in prep.*

Pour pouvoir établir une théorie pertinente de l'adaptation, il est essentiel que les résultats théoriques soient validés (au moins partiellement) par des données expérimentales. L'un des intérêts du modèle de Fisher est qu'il traite à la fois des mutations délétères et avantageuses. Certains de nos résultats peuvent donc être testés (indirectement) à partir de l'effet des mutations délétères, pour lesquelles il existe maintenant un assez grande nombre d'études dans plusieurs espèces. Les deux premières sections de ce chapitre montrent comment une approche comparative entre plusieurs espèces (effet de la complexité n) ou plusieurs environnements (effet de la distance à l'optimum), permet de tester certaines des prédictions du modèle proposé. A terme, le but est de pouvoir prédire les taux d'adaptation et donc de valider également les prédictions sur les mutations avantageuses. La dernière section montre comment une telle prédiction peut être faite et testée sur des trajectoires de fitness observées. Les détails sur les données empiriques utilisées dans ces trois études sont donnés dans chacun des articles. Je me borne donc ici à donner le principe de l'approche, les résultats importants, et les conclusions et critiques que l'on peut tirer de ces études.

Revue : Distribution de l'effet des mutations délétères et complexité phénotypique

Article 1: A general multivariate extension of Fisher's geometric model and the distribution of mutation fitness effects across species

I.1 PRINCIPE DE L'APPROCHE

Notre modèle (Chapitre I) prédit comment $f(s)$ devrait varier avec n. Le paramètre n représente le nombre de traits affectés par la mutation et ayant un effet sur la fitness. On noteras que ces traits ne sont pas nécessairement affectés simultanément par chaque mutation (pléiotropie universelle), puisque la modularité de l'effet phénotypique des mutations affecte *a priori* peu la dimension du système (voir (Welch and Waxman 2003) et les remarques sur la modularité en Chapitre I). Une première approche pour tester le modèle est de comparer l'effet de n sur $f(s)$ à celui observé empiriquement. On peut pour cela comparer des distributions empiriques dans des situations correspondant à différentes valeurs de n.

Comme nous l'avons vu en Introduction, la distribution de l'effet, sur la fitness, de mutations aléatoires a été bien documentée empiriquement, grâce à des expériences MA ou à des mesures directes basées sur des lignées contenant exactement une mutation. Ces études font surtout état de mutations délétères, ce qui suggère que les génotypes ancestraux utilisés étaient proches de l'optimum ($s_o << \bar{s}$). Cela n'est pas gênant, au contraire. En effet, lorsque s_o n'est pas négligeable devant $\bar{s}$, les distributions observées dépendent de cette variable (Eq.(6)). Ainsi, des mesures de $f(s)$ obtenues à partir de génotypes maladaptés introduirait une variable confondante inutile (s_o différent selon l'étude). Reste à trouver une mesure a priori du nombre de traits n ou en d'autres termes de la « complexité » de l'organisme étudié. Des mesures de $f(s)$ existent (i) pour des traits de fitness plus ou moins intégrés (par ex. survie, fécondité, taux de croissance intégré) et (ii) pour des espèces très différentes dans leur plan d'organisation. Ces deux variations pourraient être reliées à une variation de n. Cependant, le deuxième facteur offre a priori un plus grand niveau de variation puisque les espèces pour lesquelles des paramètres mutationnels sont disponibles ont une très large distribution taxonomique (allant des virus aux plantes vasculaires). Nous avons donc choisi de comparer $f(s)$ entre espèces. Il convient ensuite de définir une mesure *a priori* de n pour chaque espèce. Nous avons choisi le nombre total de gènes du génome pour plusieurs raisons (détaillées dans l'article 1) :

(i) A priori plus une espèce a de gènes, plus la mutation peut affecter une grande diversité de produits enzymatiques et par extension, de traits

(ii) A une grande échelle phylogénétique, les différences de nombres de gènes correspondent bien à différents niveaux d'organisation intuitifs parmi les espèces considérées: virus, unicellulaires, invertébrés, plantes vasculaires

(iii) Il existe des estimations de ce nombre de gènes pour la plupart des espèces pour lesquelles des mesures de $f(s)$ sont documentées ou pour des espèces proches. Ce n'est pas le cas de mesures alternatives de n telles que le nombre de types cellulaires (Otto and Yong 2002) par exemple.

Nous avons donc réalisé une revue des mesures empiriques de paramètres de $f(s)$ proposés dans la littérature, pour différentes espèces. Nous avons obtenu une estimation du nombre de gènes pour ces espèces à partir du nombre d'ORF (Open Reading Frame) reporté dans la base de données online de la Kyoto Encyclopedia of Genes and Genomes (KEGG). Le paramètre de $f(s)$ le plus fréquemment estimé est l'effet moyen des mutations $\bar{s}$. Celui-ci est estimé soit de façon directe soit par l'estimateur de Bateman-Mukai s_{BM} (voir Introduction). Dans certains cas, une mesure de la variance et du 3ème moment centré était aussi disponible, à partir des mesures directes (distributions de fitness de mutants contenant une unique mutation). Nous avons donc pu tester trois prédictions du modèle

(i) L'effet moyen $\bar{s}$ augmente avec n

(ii) $f(s)$ est une gamma (quand $s_o = 0$)

(iii) Le paramètre de forme de cette gamma augmente avec n (si $s_o = 0$, $\beta = \beta_o = n_e/2$)

Je présente ci-dessous les résultats de cette revue.

I.2 RESULTATS

Effet moyen $\bar{s}$ et nombre de gènes: notre revue révèle une corrélation forte entre le nombre de gènes et l'effet moyen des mutations, avec une augmentation de un à deux ordres de grandeur des microorganismes procaryotes (*VSV* : $s_{BM} \approx 0.1\%$, *E. coli* : $s_{BM} \approx 1\%$) jusqu'aux invertébrés et aux angiospermes (*Blé, Arabidopsis* $s_{BM} \approx 20$ %). La figure *a* de l'Encadré 9 illustre cette corrélation. Cette relation repose majoritairement sur des estimations indirectes (Bateman-Mukai) mais est globalement confirmée par les estimations directes. Par ailleurs, les estimations indépendantes pour une même espèce sont assez proches entre elles (*E. coli*, *S. cerevisiae*, *D. melanogaster*, *C. elegans*) ce qui confirme un effet de l'espèce, comme on peut s'y attendre si elle est caractérisée par un nombre de traits donné. Le modèle prédit que $\bar{s} = -1/2\, n\, \bar{\lambda}$ (Eq. (3)).Une interprétation simple de ce résultat est donc que l'effet net de la mutation et de la sélection par trait ($\bar{\lambda}$) est peu variable entre espèces, ou en tous cas qu'il ne varie pas de façon suffisante pour masquer un effet du nombre de traits (n), mesuré indirectement par le nombre de gènes. Notons que la relation se manifeste à une grande échelle phylogénétique, mais probablement pas à une échelle plus fine (par exemple, en comparant *D. melanogaster* et *C. elegans* ou le Blé et Arabidopsis). Malgré une certaine cohérence des estimations au sein de chaque espèce, on constate en effet une variation non négligeable. Cette variation peut venir de nombreux facteurs. Nous avons pu prendre en compte certains d'entre eux, comme la méthode d'accumulation des mutations, la méthode d'estimation de $\bar{s}$, le trait de fitness mesuré (voir article 1 pour plus de détails) mais les diverses études peuvent aussi différer par par ex. la lignée ancestrale, les conditions d'élevage etc.

Type de distribution: Différentes distributions sont caractérisées par leurs moments. La valeur relative des 2nd et 3ème moments (variance $V(s)$ et asymétrie $\mu_3(s)$) peut ainsi indiquer qualitativement si $f(s)$ est bien distribuée comme une gamma (quand le génotype ancestral est bien adapté $s_o = 0$) comme le prédit le modèle. Si c'est le cas on s'attend à ce que $\mu_3(s) = 2\, V(s)^2/\bar{s}$. Les expériences mesurant directement la distribution des fitness de lignées contenant une mutation exactement permettent d'estimer ces moments. Bien que basé sur un nombre plus limité d'études, nous avons donc pu tester l'existence d'une relation linéaire entre $\mu_3(s)$ et $V(s)^2/\bar{s}$. A partir de huit estimations disponibles, chez le *VSV*, *E. coli*, *S. cerevisiae*, et *D. melanogaster* , on obtient une relation très nette ($\mu_3(s) \approx 3\, V(s)^2/\bar{s}$, $R^2 = 0.99$, figure *b* de l'Encadré 9). Proche de celle attendue. La pente de la relation linéaire ne vaut pas 2 comme attendu sous l'hypothèse que $f(s)$ suit une gamma, mais elle n'en est pas significativement différente. Cela suggère que la forme de $f(s)$ est proche de celle d'une gamma dans toutes ces espèces. Ce résultat est en accord avec des études antérieures qui ont mesuré l'ajustement d'une distribution gamma à des données d'accumulation de mutations délétères chez *D. melanogaster* (Keightley 1994), ou à des mesures directes de $f(s)$ chez *E. coli* (Elena et al. 1998) ou le *VSV* (Sanjuan et al. 2004b). Même si la gamma n'est pas toujours le meilleur modèle, elle a toujours donné un bon ajustement aux données expérimentales.

Variation de la forme de la distribution entre espèces: au delà de l'approximation de $f(s)$ par une gamma, le modèle prédit que le paramètre de forme de cette gamma est proportionnel à n_e. n_e est lui-même proportionnel à n et peut être estimé à partir du coefficient de variation de $f(s)$ par $2/\mathrm{CV}(s)^2$. Nous avons vérifié que les valeurs de n_e (estimées) augmentaient avec le nombre de gènes. Sur l'ensemble des 8 estimations la corrélation est vérifiée. Toutefois, les valeurs de n_e sont faibles (<3) et varient peu de *E. coli* à *D. melanogaster* ou *C. elegans*. J'ai présenté en chapitre I une approche basée sur les matrices aléatoires supposant que les corrélations entre traits phénotypiques sont aléatoirement positives ou négatives, avec un grand nombre de traits sous sélection. Sous ce modèle, on attend des valeurs réduites de n_e (pour des corrélations suffisamment fortes en moyenne par ex. $\rho > 0.1$) et une faible variation

de ce paramètre avec n (donc une corrélation faible voire indétectable entre ces grandeurs). Les valeurs observées pour des espèces de complexités a priori très variables sont donc en accord avec cette prédiction. La corrélation de n_e avec n pourrait même ne pas être détectée, en ajoutant des estimations provenant d'autres espèces ou avec d'autres conventions pour mesurer n_e (par ex. à partir de $\mu_3(s)$ et non CV(s) ou à partir du paramètre de forme d'une gamma ajustée aux données). En résumé, vu le faible nombre de données disponibles, la conclusion la plus robuste de cette comparaison est que la forme de la distribution varie peu entre espèces (même très différentes), et que les valeurs de nombre de traits efficaces (n_e) sont très petites. Ainsi, les distribution $f(s)$ sont très asymétriques ($\beta_o = n_e/2$ est réduit). Une telle observation ne peut s'expliquer, dans le modèle de Fisher, qu'en prenant en compte les corrélations phénotypiques.

Par ailleurs, notons que la valeur de n_e chez le *VSV* est de l'ordre de celle estimée pour *S. cerevisiae*¸ ce qui est supérieur à la prédiction, vu le faible nombre de gènes dans ce virus à ARN. Il est possible que notre modèle basé sur une approche phénotypique continue ne soit pas valide dans de tels organismes très « simples ». Il est aussi possible que le niveau de complexité d'un virus soit plus proche de celui de son hôte que de son propre génome. Les mutations dans un génome viral interagissent potentiellement avec de nombreux gènes de son hôte, et l'ensemble détermine la fitness d'une souche virale donnée. Cette interprétation, qui peut sembler « farfelue » est du moins testable, en comparant la distribution de s dans des hôtes présentant ue forte variation de n (par ex. procaryotes, eucaryotes et pluricellulaires).

I.3 CONCLUSIONS :

Caractéristiques de f(s) et test du modèle : En résumé, la comparaison des moments de $f(s)$ entre espèces de complexité différentes (et obtenus à partir d'un génotype ancestral bien adapté) révèle trois caractéristiques majeures de $f(s)$:

(i) L'effet moyen des mutations $\bar{s}$ augmente avec le nombre de gènes (avec n), variant de près de deux ordres de grandeur des procaryotes aux angiospermes

(ii) $f(s)$ a une forme générale proche de celle d'une gamma

(iii) Le paramètre de forme de $f(s)$ varie peu entre espèces et reste faible (distributions fortement asymétriques) et il tend à augmenter légèrement avec n.

L'ensemble de ces conclusions est en accord avec les prédictions de notre modèle, notamment celles basées sur les matrices aléatoires comme modèle des matrices de covariances phénotypiques. On peut certes critiquer l'utilisation du nombre de gènes estimé (estimation imparfaite qui plus est) comme mesure de n. Toutefois, c'est la meilleure mesure que nous ayons trouvée, et nous n'avons pas trouvé d'interprétation aux effets observés du nombre de gènes sur $f(s)$, autre qu'en termes d'un nombre de traits sous sélection (n).

Il reste cependant un facteur confondant possible pour l'augmentation de $\bar{s}$: le nombre de divisions germinales par génération. Lynch et al. (Lynch et al. 1999) ont montré que ce paramètre expliquait en partie la variation du taux de mutation (délétère) U entre espèces, une prédiction confirmée par l'analyse du polymorphisme de séquences (Keightley and Eyre-Walker 2000). Ce nombre de générations augmente des procaryotes aux organismes pluricellulaires. Cela pourrait *a priori* biaiser l'estimation de $\bar{s}$ par la méthode de Bateman-Mukai (s_{BM}) vers des valeurs supérieures chez les pluricellulaires. Toutefois, un tel biais n'est attendu que si les mutations interagissent majoritairement de façon épistatique synergique (de sorte que la fitness d'un double mutant est inférieure au produit des fitness de chaque simple mutant). En effet, s_{BM} est estimé en supposant des effets multiplicatifs entre mutations, de sorte qu'il sous-estime $\bar{s}$ si les épistasies sont synergiques, et ce biais augmente avec la

probabilité d'un double mutant (donc avec U). Cependant, un telle prépondérance de l'épistasie synergique ne semble pas clairement confirmée par les quelques mesures empiriques de distribution des interactions épistatiques (Bonhoeffer et al. 2004; Korona 2004; Sanjuan et al. 2004a). Par ailleurs, le biais devrait être important pour expliquer la variation de plus d'un ordre de grandeur observée entre procaryotes et pluricellulaires. Enfin, un tel biais n'est pas attendu pour les estimations directes de $\bar{s}$, or celles-ci tendent à confirmer l'augmentation de l'effet moyen avec le nombre de gènes, observée sur s_{BM}.

Coût de la complexité: Enfin, le fait que la forme de $f(s)$ (β_o dans notre modèle) varie peu entre espèces, notamment entre espèces plus ou moins « complexes », suggère que le coût de la complexité est assez réduit (voir section III du Chapitre I). Pour quantifier plus précisément comment les variations de $f(s)$ entre espèce se traduisent en variation du taux d'adaptation, nous avons calculé les taux prédits par espèce à partir des paramètres de $f(s)$ empiriques ($\beta_o = n_e/2$ et de $\bar{s}$). Pour ne prendre en compte que l'effet de n sur les paramètres de $f(s)$, nous avons calculé ce taux de façon homogène entre espèces, en supposant un mode de reproduction sexué, et en standardisant pour le nombre de mutants $N U$. On obtient, pour le taux d'adaptation, des tendances similaires à celles obtenues pour les valeurs de n_e : le taux d'adaptation augmente des pluricellulaires aux procaryotes, mais cette augmentation est faible. En résumé, notre analyse des données empiriques semble confirmer que le coût de la complexité est assez faible, un résultat cohérent avec les prédictions basées sur le modèle de Fisher mais prenant en compte les interactions phénotypiques.

Dans la partie suivante, je détaillerai les méthodes que j'ai utilisées pour tester les prédictions du modèles quant à l'effet de l'environnement sur $f(s)$.

Effet de l'environnement sur $F(S)$ et interactions G×E mutationnelles

Article 2. Effect of environmental stress on the distribution of mutation fitness effects: analysing the empirical data in the light of adaptive landscape models.

I.4 PRINCIPE D'UN TEST A PARTIR DES DONNEES EXISTANTES

Le modèle que j'ai présenté au chapitre I suppose que la fitness est une fonction quadratique d'une distance phénotypique à l'optimum (ou gaussienne, ce qui est équivalent en s faible). Cette approximation est a priori correcte lorsque le phénotype ancestral est assez proche de l'optimum (Lande 1980), de sorte que toute fonction de fitness $W(\mathbf{z})$ peut être approximée par un polynôme d'ordre deux en $\mathbf{z}$. Toutefois, lorsque le génotype est assez mal adapté à son environnement, il est possible que $W(\mathbf{z})$ ne soit plus bien approximée par une fonction quadratique. Un test de cette hypothèse peut être fait en comparant les effets de mutations accumulées dans différents génotypes ancestraux, plus ou moins adaptés. En effet, on peut montrer en considérant une fonction de fitness générale $W(\mathbf{z})$ que la variance de s est proportionnelle au carré de la dérivée première de $W(\mathbf{z})$ par rapport à $\mathbf{z}$, et que la moyenne de s est proportionnelle à la dérivée seconde de $W(\mathbf{z})$ (article 2). La variation de ces deux moments avec différents génotypes plus ou moins adaptés (correspondant à plusieurs valeurs de $\mathbf{z}$) renseigne donc sur la variation des dérivées de $W(\mathbf{z})$ avec $\mathbf{z}$, donc sur le type de fonction. Dans la suite, je dénoterai de façon équivalente, s_o et « maladaptation du génotype ancestral », la distance à l'optimum de ce génotype, mesurée en fitness. Avec une fonction quadratique (et uniquement avec une telle fonction) on s'attend à ce que (Eq. (3)) :

(i) la variance $V(s)$ augmente avec la maladaptation du génotype ancestral, parce que la dérivée première de $W(\mathbf{z})$ augmente (en valeur absolue) avec la distance à l'optimum $\mathbf{z}$.

(ii) à ce que l'effet moyen $\bar{s}$ ne soit pas affecté par la maladaptation du génotype ancestral, parce que la dérivée seconde de $W(\mathbf{z})$ est indépendante de $\mathbf{z}$.

La prédiction (i) est attendue avec n'importe quelle fonction concave en $\mathbf{z}$ (donc toute forme de sélection stabilisante autour d'un optimum), mais la prédiction (ii) n'est vraie que si $W(\mathbf{z})$ est un polynôme d'ordre 2 (pas plus) en $\mathbf{z}$ donc une fonction quadratique.

Toutefois, aucune étude n'a mesuré la variation de $f(s)$ pour plusieurs génotypes ancestraux, plus ou moins adaptés à un même environnement. En revanche, plusieurs expériences ont étudié comment varie la moyenne et la variance de fitness de lignées d'accumulation (obtenues à partir d'un même ancêtre), selon l'environnement dans lequel la fitness est mesurée. Comparer $f(s)$ pour un même ancêtre entre environnements revient à comparer $f(s)$ pour plusieurs ancêtres dans un seul environnement, si on considère que différents environnements déterminent différents optimums phénotypiques. Nous discuterons de cette hypothèse plus loin, mais on voit que dans ce cas, un même génotype ancestral est à différentes distances $\mathbf{z}$ de l'optimum selon son niveau d'adaptation à chaque environnement.

Nous avons donc réalisé une revue de l'ensemble des expériences mesurant l'effet de l'accumulation de mutations sur des lignées isogéniques (MA) dans différents environnements. A partir d'une expérience MA, on peut typiquement faire deux estimations: l'incrément de moyenne (ΔM) et de variance (ΔV) dû à la mutation (voir Introduction et Encadré 3). Nous avons étudié comment ces deux grandeurs changeaient lorsqu'on passait d'un environnement quasi-optimal pour le génotype ancestral (environnement « de référence ») à un environnement non-optimal (dans lequel la fitness de l'ancêtre est réduite, environnement « stressant »). Ces termes sont choisis pour la clarté de l'exposé, sachant que les définitions du « stress » n'impliquent pas toujours une réduction de fitness (Parsons 1991). Pour chaque environnement dans lequel les fitness sont mesurées, on a donc une estimation de ΔM et de ΔV qui peuvent être exprimées en fonction du taux de mutation U et de leurs effets dans l'environnement considéré (Mukai et al. 1972)

$$\begin{aligned} \Delta M &= U\,\bar{s} \\ \Delta V &= U\,\bar{s}^2\,(1 + CV(s)^2)\end{aligned}, \qquad (10)$$

Où CV(s) est le coefficient de variation de $f(s)$. On voit donc que la variation de ΔM et ΔV avec l'environnement nous renseigne sur celle de $\bar{s}$ et $V(s)$. On peut à partir de cette variation tester les prédictions (i) et (ii) citées plus haut.

D'après l'Eq. (10), on peut (arbitrairement) subdiviser l'effet possible de différents environnement en un effet sur U, sur $\bar{s}$ et sur CV(s) (ou de façon équivalente, sur $V(s)$). On peut tenter de clarifier les bases biologiques de chacun de ces effets, à partir du modèle général de paysage adaptatif basé sur une fonction $W(\mathbf{z})$ arbitraire (ci-dessus et article 2). On peut approximativement considérer que tout effet de l'environnement sur U dépend de la façon dont l'environnement affecte le phénotype ($\mathbf{z}$ + $\mathbf{dz}$) de chaque génotype mutant. Une influence de l'environnement sur U traduit donc l'expression environnement-dépendante de mutations, ou une plasticité de leur effet phénotypique entre environnements. Notons que dans les études considérées ici, les mutations ont été accumulées dans un environnement unique, donc les possibles effets de l'environnement sur le taux de mutation (moléculaire) n'influent pas sur notre analyse. En revanche, tout effet de l'environnement sur $\bar{s}$ et sur

CV(s) (ou $V(s)$) dépend de la façon dont l'environnement affecte la fonction de fitness et la distance à l'optimum (donc de comment les dérivées de $W(\mathbf{z})$ changent avec $\mathbf{z}$). En résumé, l'effet de l'environnement sur U a trait à l'influence de l'environnement sur la relation génotype-phénotype, tandis que son effet sur $\bar{s}$ et sur CV(s) a trait à son influence sur la relation phénotype-fitness.

Nous avons utilisé les données MA pour distinguer l'effet relatif de l'environnement sur ces trois variables (U, $\bar{s}$ et sur CV(s)), et évaluer la validité des prédictions (i) et (ii).

I.5 RESULTATS

Effet du stress sur la moyenne (ΔM) et la variance (ΔV) de W: Supposons que les prédictions (i) et (ii) soient correctes, même à de grandes distances à l'optimum (donc dans des environnements très stressants), et que les mutations n'aient pas d'effets plastiques. On s'attend alors à une augmentation de CV(s) loin de l'optimum (en environnement stressant), sans autre effet ni sur U ni sur $\bar{s}$. Dans ce cas, ΔV devrait augmenter dans les milieux stressants alors que ΔM devrait rester à peu près constant dans tous les milieux. Pour étudier les changements de ΔM et ΔV dans les milieux stressants nous avons étudié leur valeur relativement à la valeur dans l'environnement de référence (ΔM^* et ΔV^*) pour chaque étude, en définissant les ratios $\rho_M = \Delta M/\Delta M^*$ et $\rho_V = \Delta V/\Delta V^*$. Nos résultats montrent que les milieux stressants correspondent à des valeurs supérieures de ΔV ($\rho_V > 1$) tandis que les valeurs de ΔM tendent autant à décroître qu'à augmenter en milieu stressant, et montrent une variation plus réduite ($\rho_M \sim 1$). Globalement, ces résultats suggèrent donc que les milieux stressants entraînent une augmentation de la variance mais pas de la moyenne de fitness des lignées mutantes. Un tel effet est bien attendu dans un modèle de paysage adaptatif avec $W(\mathbf{z})$ quadratique et en supposant l'équivalence génotype-environnement. Toutefois, on ne peut rejeter la possibilité que cette observation résulte d'une interaction entre (U, $\bar{s}$ et CV(s)), de sorte que tous soient influencés par l'environnement mais que les différents effets s'annulent sur ΔM et augmentent ΔV en milieu stressant. Nous avons donc proposé un test de l'effet relatif sur U, $\bar{s}$ et CV(s) séparément.

Test d'hypothèses, relation entre U_{BM} et ΔV: On peut utiliser les estimations de ΔV et ΔM entre environnements pour obtenir une mesure plus quantitative de l'effet relatif de l'environnement sur celles-ci. L'estimateur de Bateman-Mukai du taux de mutations est

$$U_{BM} = \frac{\Delta M^2}{\Delta V} = \frac{U}{1 + CV(s)^2} \,. \qquad (11)$$

Comme pour ΔM et ΔV, définissons $\rho_U = U_{BM}/U_{BM}^*$ la valeur de U_{BM} standardisée par sa valeur dans l'environnement de référence. La pente de la relation entre $\log(\rho_V)$ et $\log(\rho_U)$ entre environnements permet de quantifier l'importance relative de la variation de U, $\bar{s}$ et CV(s) entre environnements. En effet, si l'environnement entraîne surtout une variation de U, on attend une pente de +1. Si c'est CV(s) qui est majoritairement affecté on attend une pente de –1. Enfin, si l'environnement entraîne surtout une variation de $\bar{s}$ on s'attend à n'observer aucune corrélation.

La pente observée est -0.69, qui est significativement inférieure (à $\alpha = 5\%$) à celle attendue si seul CV(s) est affecté (–1), mais ne l'est pas à $\alpha = 1\%$. Elle est fortement différente des relations attendues sous l'hypothèse que l'environnement affecte majoritairement U ou $\bar{s}$. Ces résultats suggèrent donc que l'effet quantitativement largement majoritaire est celui sur CV(s). Les prédictions (i) et (ii) semblent donc confirmées.

I.6 CONCLUSIONS

L'analyse de l'influence de différents environnements plus ou moins stressants sur la variance et la moyenne de la fitness de lignées d'accumulation suggère que le principal effet du milieu est d'augmenter la variance de $f(s)$ (ou de façon équivalente, CV(s)). Même s'il apparaît que l'environnement peut affecter aussi l'expression des mutations (U) ou leur effet moyen ($\bar{s}$), ces effets sont relativement faibles comparés à l'effet sur la variance.

Diverses implications de ces résultats sont discutées dans l'article 2. Je me borne ici à conclure quant à la validation des modèles de paysages adaptatifs. Il semble que l'on puisse raisonnablement approximer l'effet de l'environnement sur la distribution de l'effet des mutations comme un changement d'optimum dans un paysage gardant les mêmes propriétés d'échelle entre différents environnements:

(i) l'effet net de la mutation par trait ($\bar{\lambda}$) est constant (sinon $\bar{s}$ aurait fortement varié)

(ii) le nombre de traits sous sélection (n) est constant (sinon U et $\bar{s}$ auraient fortement varié)

Notons que d'autre fonctions $W(\mathbf{z})$ alternatives à la quadratique (par ex. quartique etc..) nous auraient amenés à détecter un effet de l'environnement sur $\bar{s}$, voire sur U pour certaines. Le test est donc bien une confrontation à des modèles alternatifs. Ce résultat ne garanti pas la validité du modèle, comme tout test, mais ne parvient pas à le rejeter. En revanche, ces résultats amènent à rejeter (quantitativement) plusieurs effets du stress proposés jusqu'ici :

(i) le stress ne tend apparemment pas à augmenter l'effet délétère des mutations ($\bar{s}$ semble varier peu entre environnements, et pas dans le sens d'une augmentation en milieu stressant).

(ii) le stress ne se traduit pas par une sur-expression des mutations (U semble varier peu entre environnements)

Ces conclusions sont à modérer. Elles sont limitées à l'analyse de données provenant de quelques espèces, dans quelques environnements. D'autre part, elles ne sont que relatives. Des variations significatives de l'effet moyen des mutations ont été observées dans plusieurs études prises en compte dans notre revue (Kishony and Leibler 2003; Korona 2004) mais pas dans certaines autres (Fry and Heinsohn 2002). Toutefois, ces effets semblent toujours faibles *relativement* à l'effet sur CV(s).

Par ailleurs, nous avons vu également (Chapitre I, partie III) que pour prédire $f(s)$ dans un nouvel environnement, il est nécessaire de supposer que ni $\bar{\lambda}$ ni CV(λ) ne changent trop entre environnements, de sorte que les paramètres $\bar{s}$ et β_o soient à peu près constants. Concrètement, cela correspond à supposer que même si les interactions sélectives (**S**) et mutationnelles (**M**) peuvent changer d'un environnement à l'autre, leur effet net par trait garde la même moyenne et variance. Les résultats de cette étude ne permettent pas de tester si β_o varie entre environnements. En effet, nous avons comparé ici des valeurs de CV(s) pour différentes distances à l'optimum, or d'après les Eqs. (3) et (4), $\mathrm{CV}(s) \approx 2\beta_o(1 + 2\, s_o / \bar{s})$. Nous avons interprété la variation de CV(s) par l'effet de l'environnement sur s_o, mais il est possible que β_o varie également entre environnements. Cependant, il semble raisonnable de supposer que ce paramètre est peu variable pour deux raisons. D'abord, nous avons vu qu'en supposant que la variation de **S** et **M** est aléatoire entre environnements, on prédit que le paramètre de forme ($\beta_o = n_e/2$) varie peu (Eq. (8)). Ensuite, nous avons vu dans la section précédente que ce paramètre était peu variable entre des groupes phylogénétiquement très distincts. Il est donc improbable que β_o varie de plusieurs ordres de grandeurs pour une même

espèce dans divers environnements. La forte variation observée ici a donc plus de chance d'être due à l'effet de l'environnement sur s_o.

En revanche, il n'y a pas de raisons théoriques fortes de supposer que $\bar{s}$ ne varie pas d'un environnement à l'autre (d'autant plus qu'on a vu en section I qu'il variait fortement entre espèces). Il est donc intéressant que l'analyse comparative présentée ici suggère que l'hypothèse qu'$\bar{s}$ est constant entre environnements est valide, même pour des environnements assez fortement stressants (par ex. diminuant la fitness de 50%). L'hypothèse d'une fonction $W(\mathbf{z})$ quadratique semble donc rester assez valide, même loin de l'optimum (en restant dans la limite où $W(\mathbf{z})>0$).

Enfin, les résultats présentés ici sont cohérents avec un modèle où U varie peu avec l'environnement. Cela ne veut pas dire qu'il n'y a pas d'expression environnement-dépendante des mutations ou de plasticité de leur effet. Cela signifie seulement que l'effet net des ces divers processus ne biaise pas les valeurs de U. Plus précisément, la variance phénotypique introduite par les mutations est à peu près constante, même si elle peut affecter des traits différents selon l'environnement.

Finalement, notons que notre modèle (comme tout modèle basé sur un paysage adaptatif) prédit aussi une augmentation de la proportion de mutations avantageuses lorsque le génotype ancestral est loin de l'optimum (milieux stressants). Une telle augmentation serait toutefois difficile à détecter *a priori* par des protocoles MA, qui ne mesurent que la variance et la moyenne des effets de l'ensemble des mutations (Bataillon 2000). Une telle augmentation de la proportion de mutations avantageuses a en revanche été observée par Remold & Lenski (Remold and Lenski 2001), qui ont analysé la distribution des fitness de lignées contenant exactement une mutation, dans des environnements contrastés (voir Introduction).

I.7 UN TEST PLUS DIRECT INITIE AVEC *ARTEMIA*

Les expériences citées ci-dessus se sont pour la plupart focalisées sur deux modalités d'une variable environnementale donnée (deux températures, deux niveaux de ressources nutritives etc.). Il semble plus direct de tester le modèle en comparant la distributions de des effets des mutations mesurés le long d'un gradient environnemental (par ex. 3 ou 4 températures etc.). Considérons une lignée ancestrale présentant une norme de réaction pour la fitness le long d'un gradient environnemental, par exemple la température. Cette norme de réaction définit sa niche pour la température. On peut supposer que différentes températures déterminent différents optimums phénotypiques, et que l'ancêtre est mieux adapté à l'une de ces températures (sa température optimale, notons la T*). Une telle niche montrant un optimum de température en T* correspond à dire que l'ancêtre est plus près de l'optimum déterminé par T* que de ceux déterminés par une température T>T* ou T<T*. Dans ce cas la distance à l'optimum s_o, pour l'ancêtre, varie de façon continue le long du gradient de température (et s_o est minimal en T*). Alors on peut vérifier de façon *quantitative* les relations prédites par l'Eq. (3) entre la variance $V(s)$ de $f(s)$ et s_o et entre l'effet moyen $\bar{s}$ et s_o. Plus précisément, pour une variation de s_o correspondant à un gradient d'une même variable environnementale on s'attend à ce que

(i) $V(s)$ augmente proportionnellement à 2 s_o.

(ii) $\bar{s}$ est constant avec s_o (même prédiction (ii) qu'en II.1 de ce chapitre).

La prédiction (i) est plus précise que son équivalent (prédiction (i)) présenté en partie II.1 de ce chapitre, puisqu'elle prédit non seulement une augmentation de $V(s)$ dans les environnements stressants (les températures autres que T*), mais elle la quantifie en fonction

de s_o. Une telle relation quantitative n'est pas attendue si on considère différentes valeurs de s_o correspondant à des changements environnementaux de natures différentes (par ex. température et salinité du milieu etc., au lieu de plusieurs températures), comme dans les expériences présentées précédemment (sections II.1-II.3). En effet, le paramètre q (voir Eq. (3)) est un facteur de correction correspondant à la direction du changement d'optimum imposé par un nouvel environnement, et il est donc spécifique à une variable environnementale (température *ou* salinité). Une même valeur de s_o imposée par une température plus basse ou par une salinité plus haute peut correspondre à différentes valeurs de $V(s)$, même dans notre modèle. Il serait donc préférable de tester le modèle à partir de comparaisons de $f(s)$, pour un même ancêtre, le long d'un gradient de température, ou de salinité par exemple, de sorte que q puisse être supposé constant pour toutes les valeurs de s_o.

L'un des buts initiaux de ma thèse était de faire un tel test expérimental. Celui-ci peut a priori être fait sur divers modèles biologiques, dès que l'on connaît une variable environnementale continue ayant un impact important et graduel sur leur fitness. J'ai pour ma part travaillé sur *Artemia*. Ce crustacé est adapté à la vie dans les milieux sursalés tels que les marais salants et les lacs salés. Une présentation générale de sa biologie peut être trouvée en Annexe 2. Les *Artemias* vivent dans des milieux extrêmes assez peu diversifiés (quasiment aucun prédateur aquatique, peu de compétiteurs). Leur niche écologique est donc simple et bien caractérisée : elle dépend majoritairement de facteurs abiotiques, notamment la température et la salinité du milieu, facilement reproductibles au laboratoire. Par ailleurs, *Artemia* présente les avantages classiques d'un « bon » modèle de laboratoire: temps de génération court (~15 à 20 jours), production de nombreux descendants (de l'ordre de 100 par semaines), possibilité d'élever un grand nombre d'individus dans un espace réduit. Enfin, elle présente des particularités utiles pour effectuer des protocoles d'accumulation de mutation: possibilité de garder des échantillons des générations passées sous forme de cystes (œufs de diapause), existence de lignées asexuées se reproduisant par automixie (produisant une descendance assez homozygote.

J'ai développé divers protocoles préliminaires à la réalisation d'une accumulation de mutations (ces protocoles sont détaillés en Annexe 2): création d'une lignée asexuée isogénique, développement d'un système d'élevage individuel, et de mesures de traits de fitness (survie, nombre d'œufs produits par ponte) et morphologiques (longueur totale du corps). Un préliminaire à cette expérience était de produire un ensemble de lignées mutantes. Etant donné le temps assez long qu'aurait pu prendre une accumulation de mutations spontanées (>20 générations), j'ai opté pour l'accumulation de mutations induites par mutagenèse aléatoire. Cette approche a déjà été utilisée sur *C. elegans* et *D . melanogaster* pour accélérer les expérience de MA (Halligan et al. 2003; Keightley et al. 2000; Keightley and Ohnishi 1998). J'ai utilisé le même mutagène que celui utilisé dans ces études l'EMS (ethyl méthane-sulfonate), qui génère principalement des mutations ponctuelles et est donc susceptible de produire des effets proches de ceux des mutations spontanées. J'ai aussi développé un protocole de mutagenèse avec l'ENU (ethyl nitro-sourea), un autre agent mutagène produisant des mutations ponctuelles.

Les résultats de ces mutagenèse sont présentées dans l'Annexe 2. J'ai réalisé 5 essais de mutagenèse à différents stades de développement. Le protocole final présenté en annexe me semble être le plus susceptible d'être efficace. Toutefois, j'ai obtenu des résultats pour le moins mitigés. La mutagenèse à l'ENU a eu l'effet inverse à celui attendu (les témoins ayant une survie moins élevée que les mutants). En revanche, la mutagenèse à l'EMS a produit des effets plus cohérent avec l'attendu : les lignées F1 issues de mutagenèse présentent, relativement au témoin, (i) une survie réduite et (ii) un nombre de descendants moins élevé et plus variable entre lignées. Toutefois, ces effets ne sont pas tous significatifs, notamment, l'effet sur la reproduction moyenne n'est significatif que sur la reproduction initiale, pas sur les pontes suivantes. Par ailleurs, l'augmentation de variance en fitness parmi les lignées de mutagenèse n'est pas significative. On peut calculer les estimations de Bateman-Mukai de U

et $\bar{s}$ (U_{BM} et s_{BM}) pour *Artemia* ainsi obtenues. Pour le nombre de descendants des premières pontes et le traitement 30mM d'EMS, on obtient s_{BM}~ -0.09 et U_{BM}~ 0.93, ce qui correspond assez bien aux estimations obtenues chez *C. elegans* et *D. melanogaster* pour des traitements similaires (30 mM d'EMS correspondant environ à 100 générations de mutations spontanées Keightley et al. 2000; Keightley and Ohnishi 1998). En revanche, pour la reproduction plus tardives (après le 35ème jour), on obtient des résultats moins cohérents avec des estimations précédentes dans d'autres espèces (s_{BM}~ -0.36 et U_{BM}~ 0.09), bien que celles-ci n'ont en général pas étudié la reproduction tardive en particulier. Enfin on obtient pour le nombre total de descendants sur 45 jours s_{BM}~ -0.194 et U_{BM}~ 0.36.

La principale difficulté rencontrée dans cette expérience a été que je n'ai pas pu, pour le moment, faire éclore les cystes qui ont été produits par les lignées F_1 (les F_2), qui auraient permis d'estimer plus clairement les paramètres mutationnelles en permettant d'estimer des effets lignées (on peut mesurer plusieurs répliquas par lignée, en F_2 pas en F_1). Les raisons de cet échec de l'éclosion (y compris parmi les témoins) restent assez peu claires (je les discute en Annexe 2). Il est possible que l'environnement d'élevage ait été stressant, bien que l'on ait obtenu une bonne survie et des pontes relativement nombreuses (de l'ordre de 80-100 individu par ponte). Cette possibilité est suggérée par le fait que la quasi-totalité des lignées F1 ont produit uniquement des cystes, ce qui peut traduire l'existence d'un stress nutritif par exemple (la production de cystes est favorisées par les conditions stressantes). Quoi qu'il en soit, je n'ai pu poursuivre l'expérience et mesurer l'effet des mutations le long d'un gradient de salinités et de températures. Je n'ai donc pas présenté mes résultats dans le texte principal de ma thèse mais en annexe, bien que ce travail ait représenté une part importante (en temps du moins) de mon travail de thèse. Toutefois, il me semble que le protocole de mutagenèse est maintenant assez bien en place, et que des modifications mineures du système d'élevage utilisé pourraient permettre de reprendre cette expérience, peut-être avec une lignée sexuée et la lignée asexuée que j'ai produite. Cela permettrait de comparer les paramètres mutationnels entre sexués et asexués dans deux espèces proches et avec le même système d'élevage. A ma connaissance, aucune comparaison de ce type n'a été entreprise jusqu'ici, puisque les organismes modèles classiquement utilisés ne présentent pas un choix de systèmes de reproduction comme on trouve chez *Artemia* (à l'exception de la levure, mais nous n'allons tout de même pas *tout* tester uniquement chez les microbes !). Une telle approche permettrait d'aborder empiriquement la question de l'évolution du taux de mutation selon le système de reproduction.

Au final, il me semble que malgré l'échec partiel de cette expérience, *Artemia* est un modèle biologique qui présente des particularités très intéressantes pour étudier l'interaction entre mutation, environnement, et système de reproduction. Les systèmes de mesures de traits et d'élevage sont maintenant assez fonctionnels et on peut (j'espère) envisager d'obtenir des résultats intéressants dans les années à venir.

Trajectoires de fitness dans un environnement fixe chez *E. coli* et la drosophile

Article 3 : Predicting fitness trajectories in fixed environments for sexual and asexual model organisms (*E. coli* and *D. melanogaster*)

Les revues des données empiriques présentées dans les parties I et II ne sont basées que sur la distribution d'effets de mutations délétères, car ce sont celles qui sont majoritairement observées dans les expériences d'accumulation de mutations. Toutefois, un des but principaux de notre modèle est de prédire la distribution de l'effet des mutations avantageuses, et les taux d'adaptation. Etant donnée la rareté des données empiriques sur les mutations avantageuses, il

est difficile de tester directement les prédictions du modèle quant aux mutations avantageuses. Par ailleurs, les tests des théories basées sur la théorie des valeurs extrêmes (par ex. Rokyta et al. 2005) ne permettent pas de distinguer plusieurs modèles de distribution de l'effet des mutations avantageuses (et de $f(s)$ en général), puisque les prédictions de l'EVT convergent pour un grand nombre de distributions $f(s)$ (y compris la gamma déplacée proposée en Chapitre I). Pourtant, à terme, les taux d'adaptation ne peuvent être prédits sans une connaissance de ces distributions (Orr 2002). Cette partie présente un test des prédictions de notre modèle par l'étude du changement à long terme de la fitness moyenne dans un nouvel environnement.

I.8 PRINCIPE DE L'APPROCHE

Comme nous l'avons vu en Introduction, le taux d'adaptation (mesuré comme le changement de la log fitness moyenne), peut être calculé dans un contexte donné si on connaît (Eq. (1))

(i) la probabilité de fixation d'une mutation d'effet s : $p_{fix}(s)$

(ii) le nombre de mutants $N\,U$

(iii) la distribution de l'effet des mutations $f(s)$

Dans une population sexuée, $p_{fix}(s)$ peut être calculé en connaissant la taille efficace N_e et la taille totale N, des grandeurs en général connues dans les protocoles d'adaptation au laboratoire. En revanche, pour des populations asexuées (comme de nombreux micro-organismes), les probabilités de fixations de mutations avantageuses sont affectées par (i) la présence d'autres mutations avantageuses (interférence clonale (Gerrish and Lenski 1998)) et (ii) de mutations délétères (sélection de fonds (Johnson and Barton 2002; Orr 2000b; Peck 1994)). Il est donc nécessaire de connaître la distribution de l'effet des mutations (y compris délétères) pour calculer $p_{fix}(s)$ dans ce cas.

Le taux de mutation U est connu à partir des estimations dont nous avons parlé en Introduction, pour certaines espèces modèles et souvent dans un seul environnement de laboratoire (assez optimal pour la souche utilisée). Le problème se pose alors de l'extension d'une estimation dans un environnement donné et sur une lignée donnée à d'autres situations. Des résultats empiriques suggèrent notamment que le taux de mutation chez les bactéries peut augmenter dans les environnements stressants, que ce soit le taux de mutation sur un ensemble de gènes marqueurs (Bjedov et al. 2003), ou le taux génomique U (Loewe et al. 2003). La sélection de génotypes mutateurs (ayant un taux de mutation très élévé) a été observée dans plusieurs expériences d'adaptation en milieu contrôlé chez des micro-organismes (Shaver et al. 2002), pour revue voir (Goho and Bell 2000)]. Des gènes augmentant le taux de mutation peuvent présenter un avantage adaptatif dans des populations asexuées, où un mutateur peut augmenter en fréquence par auto-stop sur les mutations avantageuses qu'il génère (de Visser). Ces différentes études ont en général mis en évidence une augmentation du taux de mutation lorsque les populations sont en phase stationnaire (pendant laquelle la démographie est stable). Ce processus de « mutagenèse en phase stationnaire » implique des mécanismes génétiques spécifiques qui augmentent le taux de mutation (erreur dans la synthèse d'ADN, perte des fonctions de réparation de l'ADN, activation de transposons, voir la revue de Kivisaar (2003)). Toutefois, la généralité de ce processus est encore en débat (Sniegowski 2004). Par ailleurs, dans plusieurs expériences, il semble que le taux de mutation ne soit augmenté qu'à certains gènes spécifiques impliqués dans la réponse au stress environnemental imposé (voir la revue de Massey (2002)). Au final, l'ensemble de ces résultats suggère qu'il n'est pas certain qu'on puisse extrapoler des estimations de U à différents environnements, particulièrement

(i) dans les environnements qui impliquent un stress physiologique rencontré dans les conditions naturelles. En effet des mécanismes de réponse à ce stress ont pu évoluer, tels que l'augmentation du taux de mutations à certains gènes.

(ii) dans des espèces asexuées

(iii) pour des populations de bactéries en phase stationnaire

Toutefois, étant donné que U doit être connu pour prédire un taux d'adaptation, et en l'absence d'attendus clairs quant à l'effet d'un environnement donné sur U, nous avons choisi d'utiliser les estimations obtenues dans les expériences MA (voir article 3). Par ailleurs, dans une des expériences sur lesquelles nous nous sommes basés, la population est maintenue en croissance exponentielle par des dilutions successives du milieu (Lenski and Travisano 1994), et n'est donc pas en phase stationnaire. En revanche, dans l'autre expérience sur *E. coli*, la population est maintenue à un niveau stable de sorte qu'il est possible *a priori* que des mécanismes de mutagenèse en phase stationnaire soient impliqués.

Comme nous l'avons vu en Chapitre I et II, $f(s)$ peut être approximée par une gamma déplacée dont la moyenne ($\bar{s}$) et le paramètre de forme (β) peuvent être obtenus à partir de leur estimation ($\bar{s}$ et β_o) dans un environnement standard, et d'une mesure de s_o. Nous avons discuté en Chapitre I de la validité d'une extension des estimations de $\bar{s}$ et β_o à un nouvel environnement. Les résultats théoriques basés sur les matrices aléatoires et les revues des données empiriques suggèrent qu'il est raisonnable de supposer qu'ils varient peu entre environnements. Le paramètre s_o doit, lui, être mesuré pour une lignée donnée (au moins pour une espèce) et un environnement donné. Nous avons donc basé nos tests sur les résultats d'études faisant un suivi de la fitness moyenne de populations en environnement fixe jusqu'à atteindre un plateau de fitness. Celui-ci donne alors une mesure de s_o. Nous avons trouvé une étude chez la drosophile (87 générations) et deux étude chez *E. coli* (adaptation sur 10 000 générations), dont les détails sont donnés dans l'article 3. Ces études constituent trois expériences indépendantes qui diffèrent par de nombreux points:

(i) Espèce et système de reproduction (sexués ou asexués)

(ii) environnement utilisé (captivité, milieu limitant en glucose ou milieu limitant en thymine)

(iii) régime démographique (petites populations constantes, grandes populations avec goulots d'étranglement réguliers, une grande population de taille constante)

Basé sur ces différentes mesures, nous avons pu comparer les trajectoires de log fitness observées à celles prédites par le modèle. Les expressions exactes pour les probabilités de fixation et le taux d'adaptation sont données dans l'article 3.

I.9 RESULTATS

Le modèle présenté en chapitre I (et article 1 et 3) prédit, pour $s_o >> \bar{s}$, un taux d'adaptation quasiment linéaire en fonction de s_o. Ceci est valable pour les sexués comme pour les asexués (voir l'expression analytique approximative pour les sexués: Eq. (9)). Si on définit a, la pente de cette relation linéaire, on obtient alors une expression approximative simple pour la log-fitness moyenne en fonction du temps :

$$\log(\overline{W}(t)) = \log(W_o) + s_o\,(1 - e^{-a\,t})\,, \qquad (12)$$

Où W_o est la fitness initiale.

Ajustement de la trajectoire prédite aux données expérimentales: Nous avons estimé l'ajustement d'une telle trajectoire de fitness aux données empiriques. Celui-ci est à peu près aussi bon (parfois meilleur) que les modèles log-hyperbolique ou hyperbolique utilisés fréquemment pour décrire de telles trajectoires. L'intérêt de la fonction proposée en Eq. (12) est que ses paramètres correspondent à des variables biologiques bien définies et distinctes

(i) le paramètre s_o est déterminé uniquement par l'adaptation du génotype ancestral (utilisé pour initier l'expérience) à l'environnement de l'expérience

(ii) le paramètre a est une mesure de la réponse à la sélection. Il dépend du système de reproduction et des paramètres mutationnels de l'espèce considérée (U, $\bar{s}$,β_o) et du régime démographique (N_e, N, goulots d'étranglements).

a est donc déterminé par le protocole d'élevage et le modèle biologique, tandis que l'environnement auquel la population s'adapte détermine s_o. Théoriquement, on prédit donc que a devrait varier peu d'une expérience à l'autre, si on garde un même protocole (espèce, mode d'élevage) et qu'on utilise un autre environnement (un autre s_o). Une confirmation de cette prédiction vient du fait que les valeurs de a estimées sur les données empiriques sont très proches entre les deux expériences menées chez *E. coli* et sont de deux ordres de grandeur inférieures à celles obtenues pour la drosophile (article 3). En revanche les valeurs de s_o varient fortement entre les trois études.

Prédiction de la valeur de a: On peut prédire la valeur de a à partir de la connaissance des paramètres mutationnels pour chaque espèce et du régime démographique. Pour cela, nous avons calculé le taux d'adaptation initial r_o prédit d'après ces paramètres et la valeur de s_o ajustée aux trajectoires empiriques. En supposant que le génotype ancestral est loin de l'optimum (on peut vérifier que $s_o >> \bar{s}$), le taux d'adaptation doit être linéaire de sorte que a est alors estimé par r_o/s_o. La prédiction sur a correspond donc au taux d'adaptation initial prédit sachant la fitness finalement atteinte et le protocole utilisé. La comparaison des valeurs observées (ajustées aux données) et prédites de a montre que le modèle explique assez bien la variation de réponse à la sélection, non seulement entre espèces (drosophile, vs. *E. coli*) mais aussi entre protocoles expérimentaux pour une même espèce (régimes démographiques). L'Encadré 10 illustre l'ajustement entre ces trajectoires attendues et observées pour les trois expériences que nous avons analysées.

I.10 CONCLUSIONS

Hypothèses inhérentes à l'approche : La bonne adéquation entre les trajectoires prédites (pour une valeur connue de s_o) et celles observées semble confirmer la validité de cette approche. Celle-ci repose pourtant sur un ensemble d'approximations inhérentes à la fois à notre modèle et aux autres résultats théoriques utilisés pour calculer les probabilités de fixation, chez les sexués et surtout chez les asexués. Les hypothèses inhérentes à un calcul du taux d'adaptation en temps continu ont été détaillées en Introduction.

Par ailleurs, pour le calcul des probabilités de fixation chez les asexués nous avons pris en compte la sélection de fonds, mais en supposant que toute mutation d'effet délétère $s < -1/N_e$ empêche la fixation des mutation avantageuses qui lui sont associées. Cela revient à considérer que les effets délétères sont plus forts que les effets avantageux (Orr 2000b).

Comme U est petit chez *E. coli*, l'effet prédit de la sélection de fonds est réduit (voir plus bas) de sorte que l'hypothèse a de toutes façons peu d'impact sur les prédictions. Elle est toutefois fausse a priori, lorsqu'on considère un génotype ancestral maladapté, pour lequel les effets délétères et avantageux sont a peu près égaux en force.

Enfin, nous avons supposé que l'on pouvait extrapoler les estimations des paramètres mutationnels (U, $\bar{s}$ et β_o) aux environnements utilisés dans ces expériences et qu'ils restaient constants au cours du temps. Il semble que dans le cadre des études que nous avons considérées, cette hypothèse soit valide. Nous avons vu précédemment (chapitre I et II section II) que $\bar{s}$ varie peu entre environnements pour une même espèce, et qu'on peut en attendre de même de β_o. En revanche, la validité de l'extrapolation du paramètre U n'est pas garantie. On peut s'attendre à ce que U varie peu avec l'environnement chez les espèces sexuées, dans lesquelles l'évolution de mutateurs par sélection indirecte est moins probable. Heureusement, le taux d'adaptation est assez insensible à $N\,U$ chez les asexués, de par l'interférence clonale (Gerrish and Lenski 1998), de sorte que la variation de U avec l'environnement pourrait avoir un impact limité.

Ajoutons que dans le cas de l'étude sur la Drosophile, la population était initialement non monogénique, de sorte qu'une variance pré-existante en fitness était présente. Nous avons supposé que celle-ci était limitée (voir la discussion de cette hypothèse en article 3), ce qui semble approximativement valable. En effet, dans le cas contraire, nous aurions dû sous-estimer la réponse à la sélection pour cette étude.

Impact des différentes limites à l'adaptation : Notre approche permet d'analyser l'impact de trois facteurs limitant potentiellement l'adaptation : la dérive, la sélection de fonds et l'interférence clonale. En effet, en prenant ou non en compte chaque facteur dans la prédiction de a, on peut estimer dans quelle mesure la prédiction est faussée. Les estimations suggèrent que chaque facteur doit être pris en compte pour obtenir la meilleure prédiction sur a. Cela suggère que le modèle est donc bien un test des différents résultats théoriques et empiriques puisque la prédiction converge vers la valeur observée lorsqu'on prend en compte chaque facteur. Par ailleurs, ces estimations relatives suggèrent que l'interférence clonale joue un rôle prépondérant dans la limite de la réponse à la sélection observée chez *E. coli.* Ce facteur explique quasi-entièrement la faible réponse observée (a petit) dans les deux populations d'*E. coli* relativement à la drosophile (avec pourtant une population beaucoup plus petite).

Conclusion générale : les résultats que nous avons obtenus par la confrontation des prédictions de notre modèle aux données empiriques sont encourageants et suggèrent que l'approche basée sur les paysages adaptatifs n'est pas si irréaliste qu'il est généralement supposé. Cette approche permet par ailleurs de faire des prédictions directement testables sur la dynamique de l'adaptation. Le test de la validité des prédictions sur les taux d'adaptation ne repose cependant que sur trois études empiriques, même si chacune permet un test des prédictions. Il est donc nécessaire de l'étendre à d'autres espèces et d'autres environnements. Si cette validité est confirmée, notre approche ouvre des perspectives empiriques intéressantes: on pourrait par exemple prédire la trajectoire évolutive complète d'une population à partir d'un taux d'adaptation initial et d'une bonne connaissance de l'espèce (système de reproduction, paramètres mutationnels, régime démographique dans un contexte donné).

Les prédictions quant aux taux d'adaptation sont sensibles aux paramètres mutationnels (β_o et $\bar{s}$) dont une estimation assez précise est nécessaire. Toutefois, nos résultats suggèrent que ces paramètres pourraient peut-être être inférés pour une espèce donnée, sans avoir à les mesurer. En effet des tendances semblent se dégager entre grands groupes phylogénétiques (faible variation de β_o, augmentation de $\bar{s}$) de sorte qu'une extrapolation d'une espèce à une autre espèce voisine ne semble pas absurde.

Les prédictions sont peu sensibles au taux de mutation chez les asexués, mais elles le sont dans le cas des espèces sexuées. Toutefois, lorsque la population n'est pas trop petite, le taux de mutation et la taille efficace déterminent le taux d'adaptation uniquement via le produit N_eU (voir Eq. (9)) qui peut être estimé à partir de données de séquence (voir par exemple (Lynch and Conery 2003)).

Il serait nécessaire de conduire une analyse plus quantitative de la sensibilité de nos prédictions à chaque paramètre. Il serait également intéressant de tester nos prédictions sur des micro-organismes sexués (par ex. *S. cerevisiae* diploïdes) et chez des asexués présentant un fort taux de mutation (par ex. *VSV*). En effet, il est possible que chez ces derniers, la sélection de fonds ait un fort impact, et une comparaison aux prédictions permettrait de tester la valeur de notre approche dans ce contexte. Nous avons pour le moment été arrêtés par une incertitude quant à la valeur de $\bar{s}$ chez le *VSV*, celle-ci variant fortement entre les deux études qui l'ont mesurée.

On a vu dans ces exemples qu'un modèle relativement simple pouvait rendre compte de trajectoires de fitness observées empiriquement. Toutefois, les situations empiriques modélisées étaient assez simples: environnement fixe et homogène dans l'espace, lignées initialement isogéniques (ou de faible effectif efficace pour l'expérience sur *D. melanogaster*). On peut donc se demander comment intégrer des phénomènes plus complexes dans les modèles d'adaptation. Nous avons vu (article 2) comment modéliser l'effet d'une mutation dans plusieurs environnements, j'ai initié un travail théorique pour prolonger cette approche (coûts de la spécialisation écologique) mais il n'est pas assez avancé pour être présenté ici. On peut aussi se demander comment varie l'effet d'une mutation dans plusieurs environnements génomiques (épistasie). Dans la dernière section de ce chapitre, je présente une extension simple de notre modèle permettant d'aborder cette question.

Distribution des interactions épistatiques entre mutations

Le modèle de taux d'adaptation présenté en Introduction (Eq. (1)) néglige l'existence d'interactions épistatiques entre plusieurs mutations. Au delà de ces seules prédictions, on sait que la force et la distribution des interactions épistatiques ont des implications importantes dans de nombreuses prédictions en Evolution (pour revue, voir Phillips et al. 2000). Parmi ces implications on peut citer notamment: l'évolution de la recombinaison et du sexe, la fixation de mutations délétères, l'importance du fardeau de mutation. L'épistasie entre deux allèles à deux loci mesure l'écart entre la valeur mesurée d'un trait dans un génotype portant ces deux allèles et celle attendue sous un modèle donné d'interaction entre les deux loci. Comme on peut concevoir plusieurs modèle de référence pour les interactions entre loci (par ex. additives ou multiplicatives) plusieurs définitions de l'épistasie existent. Je considèrerai l'épistasie multiplicative sur la fitness ε, car sa distribution a été directement étudiée empiriquement, et parce que c'est cette forme d'épistasie qui a l'impact le plus direct sur l'évolution de la recombinaison (voir chapitre suivant et (Lively and Peeters 2000)). Elle est définie comme l'écart à un effet multiplicatif des mutations sur la fitness, ce modèle multiplicatif étant celui que nous avons supposé pour calculer les taux d'adaptation (voir Eq. (1)). On sait que les interactions épistatiques peuvent être assez communes (pour une revue voir Phillips et al. 2000) et des modèles basés sur la théorie du contrôle métabolique prédisent également qu'elles sont variables entre paires de loci, et entre les paires d'allèles considérées à chaque locus. Les rares études ayant directement mesuré la distribution des interactions épistatiques entre des paires de mutations ont confirmé cette prédiction (de Visser and Hoekstra 1998; Elena and Lenski 1997; Sanjuan et al. 2004a), détectant environ autant de valeurs positives

que négatives, avec une assez forte variance. Les effets épistatiques semblent donc être très fréquents et difficilement négligeables en général. On peut donc se demander dans quelle mesure l'épistasie entre deux mutations prises au hasard dans le génome peut être modélisée sans avoir recours à la description spécifique des relations métaboliques entre différents enzymes et aux hypothèses de la théorie des contrôles métaboliques (Szathmary 1993). On peut aussi se demander comment et dans quel cas on doit prendre en compte ou négliger ces interactions, autrement dit quels facteurs affectent la distribution des épistasies $p(\varepsilon)$. Dans cette section, je montre comment une approche simple basée sur le paysages adaptatifs permet d'aborder ces questions.

I.11 DISTRIBUTION DE L'EPISTASIE ENTRE L'ENSEMBLE DES MUTATIONS

Je considère ici les même hypothèses que celles décrites en Chapitre I, pour décrire l'effet d'une mutation, mais je considère cette fois deux mutations aléatoires d'effets phénotypiques $\mathbf{dz}_1$ et $\mathbf{dz}_2$, tirés dans une multivariée gaussienne. Comme précisé en Chapitre I, le coefficient de sélection d'un effet phénotypique $\mathbf{dz}$ est définit par son effet sur la log-fitness $s(\mathbf{z}_o,\mathbf{dz}) = \mathrm{Log}(W(\mathbf{z}_o + \mathbf{dz})/W(\mathbf{z}_o))$, où $\mathbf{z}_o$ est le phénotype ancestral. Supposons dans un premier temps que les effets phénotypiques des mutations s'additionnent de sorte que le double mutant a un phénotype $\mathbf{z}_o + \mathbf{dz}_1 + \mathbf{dz}_2$. Son coefficient de sélection en échelle log est donc $s(\mathbf{z}_o , \mathbf{dz}_1+\mathbf{dz}_2)$. L'épistasie multiplicative entre les mutations $\mathbf{dz}_1$ et $\mathbf{dz}_2$, ε_{12}, est l'écart à un effet multiplicatif de chacune des mutations sur la fitness. Elle se traduit donc par un écart à l'additif en échelle log: $\varepsilon_{12} = s(\mathbf{z}_o , \mathbf{dz}_1+\mathbf{dz}_2) - s(\mathbf{z}_o , \mathbf{dz}_1) - s(\mathbf{z}_o , \mathbf{dz}_2)$. C'est la distribution de cette variable lorsque les mutations $\mathbf{dz}_1$ et $\mathbf{dz}_2$ sont tirées au hasard parmi un ensemble de mutants que nous allons étudier.

La fonction de fitness $W(\mathbf{z})$ est déterminée par la matrice d'interactions sélectives $\mathbf{S}$. L'effet de la mutation sur le phénotype est lui déterminé par la matrice de variance covariance mutationnelle $\mathbf{M}$. On peut utiliser la transformation du système vers un système diagonal (voir Chapitre I et Annexe 1), pour exprimer e_{12}. Ce système transformé définit un nouvel ensemble d'axes phénotypiques, tel que $\mathbf{z}_o$ et $\mathbf{dz}_i$ deviennent $\mathbf{x}_o$ et $\mathbf{dx}_i$ respectivement, $\mathbf{M}$ devient l'identité $\mathbf{I}$, et $\mathbf{S}$ devient une matrice diagonale $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$. A partir de la définition de ε_{12} et de l'expression (Chapitre I) du coefficient de sélection $s(\mathbf{dz}) = s(\mathbf{dx})$ pour une mutation unique d'effet $\mathbf{dz}$ ($\mathbf{dx}$ dans le système transformé), on obtient simplement l'expression de l'épistasie ε_{12}

$$\varepsilon_{12} = -\mathbf{dz}_1^{\mathrm{T}}\, \mathbf{S}\, \mathbf{dz}_2 = -\mathbf{dx}_1^{\mathrm{T}}\, \mathbf{\Lambda}\, \mathbf{dx}_2 . \qquad (13)$$

Notons d'abord que cette expression est indépendante du phénotype initial $\mathbf{x}_o$ (ou $\mathbf{z}_o$) donc de la distance à l'optimum (s_o). Ainsi, on s'attend à ce que la distribution des interactions épistatiques entre des paires de mutations prises au hasard ne soit pas dépendante du génotype sur lequel les mutations apparaissent.

Ensuite, notons que la distribution de e_{12} est une forme bilinéaire de vecteurs gaussiens ($\mathbf{dx}_1$ et $\mathbf{dx}_2$). Comme ces deux vecteurs sont tirés dans une multivariée gaussienne d'espérance $\mathbf{0}$ et de variance-covariance $\mathbf{I}$ (identité), la moyenne et la variance de la distribution de ε_{12} sont simplement données par

$$\begin{aligned} E(\varepsilon_{12}) &= 0 \\ V(\varepsilon_{12}) &= Tr(\mathbf{\Lambda}^2) = n\,\overline{\lambda}^2\,(1 + \mathrm{CV}(\lambda)^2) = 2\,V(s)\,* \end{aligned} \qquad (14)$$

Où $Tr(.)$ est la trace de la matrice, et $V(s)^*$ est la variance de la distribution de l'effet des mutations $f(s)$, prises à l'optimum (voir Eq. (3) avec $s_o = 0$). On peut donc tirer les conclusions suivantes quant à la distribution de ε entre deux mutations prises au hasard, lorsque les traits phénotypiques sous sélection interagissent de façon additive :

(i) la distribution de ε est de moyenne nulle et symétrique (par symétrie de $\mathbf{dx}_1$ et $\mathbf{dx}_2$).

(ii) la variance de ε est le double de la variance de l'effet s des mutations (isolées) à l'optimum (lorsque toutes les mutations sont délétères).

(iii) ε est indépendante du génotype ancestral.

I.12 PREDICTIONS POUR LA DISTRIBUTION EMPIRIQUE DE ε CHEZ LE VIRUS *VSV*

Ces résultats très simples peuvent être comparés aux distributions empiriques de l'épistasie rapportées dans la littérature. Sanjuan et al. (Sanjuan et al. 2004a) ont étudié la distribution de ε entre un ensemble de mutations ponctuelles dans le virus *VSV*. L'effet sur la fitness de chaque mutation isolée et des paires de mutations sont donnés dans le support online de l'article. A partir de ces données on obtient la distribution empirique de s, $f(s)$. J'ai utilisé la mesure exacte correspondant au modèle ($s = \log(W/W_o)$), mais les valeurs sont très proches de celles obtenues avec la définition classique du coefficient de sélection $s = (W - W_o)/W_o$. Cette distribution comprend certaines mutations avantageuses, et correspond donc à un génotype non parfaitement adapté aux conditions du laboratoire ($s_o>0$). Cela peut s'expliquer par le fait que le génotype de référence (ancestral) utilisé pour construire le lot de mutants résulte de l'association de portions de génome de deux souches standards, une « chimère » (Sanjuan et al. 2004b). Je n'ai pour le moment pas réussi à ajuster une gamma déplacée sur la distribution empirique $f(s)$, par maximum de vraisemblance. J'ai donc calculé (comme illustration) la gamma déplacée qui correspond à la même moyenne et variance de s que celles obtenues empiriquement et dont la valeur maximale s_o correspond à l'effet bénéfique maximal observé empiriquement (ici $s_o \approx 0.095$). On obtient une distribution assez proche de l'empirique, illustrée dans l'Encadré 11 figure a. J'ai considéré ici l'ensemble des effets de mutations uniques (non-léthales) s ayant servi de base pour les doubles mutants (124 au total, Table 1 de (Sanjuan et al. 2004a). Certaines de ces mutations sont répétées dans le jeu de données (utilisées pour plusieurs doubles mutants), mais j'ai inclus l'ensemble des valeurs de s pour partir du même lot de mutations dans l'analyse de $f(s)$ et de la distribution des épistasies.

A partir de la mesure de $V(s)$, s_o et $\bar{s}$ dans cette expérience, on peut calculer la variance de s attendue si le génotype ancestral était parfaitement adapté ($V(s)^*$ pour $s_o = 0$), à partir de l'Eq. (3) : $V(s)^* = V(s)/(1+2\, q\, s_o/\bar{s}\,)$, où q est la variable aléatoire d'espérance 1 définie en Chapitre 1. On prédit alors que la distribution des épistasies ε soit de moyenne 0 et de variance $V(\varepsilon) = 2V(s)^* = 0.107$ [0.0079 , 0.0166] (les valeurs entre crochets correspondent à l'enveloppe dépendant de la valeur de q pour $q \in [0.5,1.5]$). Je n'ai pas encore pu proposer une expression de la densité de probabilité $p(\varepsilon)$ de ε, mais je pense cela possible en utilisant la même approche que celle utilisée pour $f(s)$, *c.a.d.* en utilisant la densité de la distribution prédite pour n_e traits de même effet λ_e (voir Chapitre I). $p(\varepsilon)$ peut toutefois être assez bien approximée pour une gaussienne $N(0, \sqrt{2V(s)^*}\,)$. Cette distribution prédite est représentée en figure b de l'Encadré 11, elle est très proche de celle obtenue en ajustant une gaussienne aux données empiriques ($E(\varepsilon) = 0.0036$, $V(\varepsilon) = 0.0085$).

I.13 ECART AU MODELE ADDITIF POUR LES TRAITS D'ADAPTATION

Ces résultats suggèrent que l'on peut prédire approximativement la distribution des épistasies entre mutations aléatoires à partir de la distribution de leurs effets isolés $f(s)$. Notons que j'ai pour cela fait deux hypothèses restrictives. D'abord, j'ai estimé s_o comme la valeur maximale obtenue empiriquement, ce qui est une sous-estimation, et pourrait expliquer que l'on surestime en partie la variance empirique de ε. Ensuite, j'ai supposé l'additivité des effets phénotypiques des mutations (le double mutant a le phénotype $\mathbf{dz}_1 + \mathbf{dz}_2$). Il est possible de prendre en compte un écart au modèle additif en introduisant le vecteur $\mathbf{a}_{12}$ contenant l'ensemble des épistasies additives sur le phénotype entre $\mathbf{dz}_1$ et $\mathbf{dz}_2$, et tel que le double mutant est de phénotype $\mathbf{dz}_1 + \mathbf{dz}_2 + \mathbf{a}_{12}$. Dans ce cas, l'épistasie entre les mutations $\mathbf{dz}_1$ et $\mathbf{dz}_2$ peut s'écrire

$$\varepsilon_{12} = -\mathbf{dx}_1^T \boldsymbol{\Lambda}\, \mathbf{dx}_2 - \mathbf{a}_{12}^T \boldsymbol{\Lambda} \left(\mathbf{dx}_1 - \mathbf{dx}_2\right) - \mathbf{a}_{12}^T \boldsymbol{\Lambda}\, \mathbf{x}_o - \frac{\mathbf{a}_{12}^T \boldsymbol{\Lambda}\, \mathbf{a}_{12}}{2}. \qquad (15)$$

Cette épistasie dépend cette fois du phénotype ancestral ($\mathbf{x}_o$). L'épistasie additive sur le phénotype $\mathbf{a}_{12}$ est une variable aléatoire qui varie a priori selon la paire de mutations considérées. On voit qu'à l'optimum ($\mathbf{x}_o = \mathbf{0}$), cette épistasie sur le phénotype introduit un biais vers les ε négatifs d'ordre a^2 (le dernier terme de la somme). En revanche, si le phénotype ancestral n'est pas à l'optimum ($\mathbf{x}_o \neq \mathbf{0}$), la moyenne de ε peut être positive comme négative selon l'espérance de $\mathbf{a}_{12}^T \boldsymbol{\Lambda}\, \mathbf{x}_o - 1/2\, \mathbf{a}_{12}^T \boldsymbol{\Lambda}\, \mathbf{a}_{12}$ sur l'ensemble des paires de mutations. Notons aussi qu'on attend une variance de ε plus grande que celle prédite sous un modèle additif pour les phénotypes ($2V(s)^*$), puisque chacun des termes supplémentaires de la somme introduit une part de variance. Par conséquent, le fait que la prédiction simple $2V(s)^*$ semble bien correspondre à la variance empirique de ε, pour le *VSV* suggère qu'un modèle additif peut être une bonne approximation dans ce cas. A l'opposé, un excès de variance de ε par rapport à la prédiction pourrait suggérer qu'une part des mutations n'interagissent pas additivement sur le phénotype.

I.14 EPISTASIE ENTRE LES MUTATIONS AVANTAGEUSES

L'approche de Sanjuan et al. (2004a) sur le *VSV* a aussi permis de mettre en évidence une différence entre la distribution de ε parmi l'ensemble des mutations et parmi les seules mutations avantageuses. Il apparaît que la distribution de ε est biaisée vers les valeurs négatives entre les paires de mutations avantageuses. Un tel biais peut être expliqué à partir du modèle présenté ci dessus (et de tout modèle basé sur un paysage adaptatif de type MF). Je me bornerai à un argument qualitatif (géométrique) que j'illustrerai par des simulations, puisque je n'ai pour le moment pas pu proposer un traitement analytique. Je limite aussi le raisonnement au cas où les phénotypes interagissent de façon additive sur les traits d'adaptation ($\mathbf{a}_{12} = \mathbf{0}$).

Définissons la racine carrée de la matrice $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}^{1/2} = diag\left(\sqrt{\lambda_i}\right)$ et, pour chaque vecteur $\mathbf{dx}$, le vecteur $\mathbf{dy} = \boldsymbol{\Lambda}^{1/2}\, \mathbf{dx}$, dont la norme est dénotée $|\mathbf{dy}| = dy$. Comme pour la transformation de $\mathbf{dz}$ vers $\mathbf{dx}$, cette transformation constitue un simple changement de base de $\mathbf{dx}$ vers $\mathbf{dy}$, dans le paysage phénotypique. D'après l'Eq. (13), l'épistasie est le produit scalaire de $\mathbf{dy}_1$ et $\mathbf{dy}_2$: $\varepsilon_{12} = -\, \mathbf{dy}_1{}^T\, \mathbf{dy}_2 = -\, dy_1\, dy_2 \cos(\theta_{12})$, où θ_{12} est l'angle entre les vecteurs $\mathbf{dy}_1$ et $\mathbf{dy}_2$. De même, on peut définir le vecteur $\mathbf{y}_o = -\, \boldsymbol{\Lambda}^{1/2}\, \mathbf{x}_o$ (de norme y_o), et θ_1 et θ_2 les angles entre le vecteur $\mathbf{y}_o$ et les vecteurs $\mathbf{dy}_1$ et $\mathbf{dy}_2$ respectivement. Dans la base de l'espace phénotypique correspondant à $\mathbf{dy}$, θ_1 et θ_2 sont les angles que fait l'effet phénotypique des

mutations ($\mathbf{dy}_1$, $\mathbf{dy}_2$) avec le gradient de la fonction de fitness ($\mathbf{y}_o$), pris au niveau du phénotype ancestral. Ce gradient correspond à la direction de plus grande augmentation de fitness, partant du phénotype ancestral. Plus ces angles sont réduits, plus les mutations ont de chances d'être avantageuses (à l'extrême par exemple, $\theta_1 = 0$ et la mutation $\mathbf{dy}_1$ pointe vers l'optimum). On peut alors exprimer l'épistasie et les coefficients de sélection de chaque mutation en fonction de ces angles et des normes (positives) des vecteurs $\mathbf{y}_o$, $\mathbf{dy}_1$, $\mathbf{dy}_2$:

$$\begin{cases} \varepsilon_{12} = -dy_1\, dy_2 \cos(\theta_{12}) \\ s_1 = s(\mathbf{dy}_1) = dy_1 \left(y_o \cos(\theta_1) - \dfrac{dy_1}{2} \right) \\ s_2 = s(\mathbf{dy}_2) = dy_2 \left(y_o \cos(\theta_2) - \dfrac{dy_2}{2} \right) \end{cases} . \qquad (16)$$

Supposons pour simplifier que $dy_1 = dy_2 = dy$. On voit que la condition pour que les mutations $\mathbf{dy}_1$ et $\mathbf{dy}_2$ soient avantageuses est $\cos(\theta_i) > dy/(2y_o) > 0$, pour les deux angles θ_1 et θ_2. Ce critère se comprend assez facilement: une mutation d'une taille donnée (dy) ne peut être avantageuse si elle pointe dans une direction trop éloignée du gradient de fitness, cet angle seuil décroît d'autant que la mutation est de taille réduite par rapport à la distance à l'optimum ($dy/(2y_o)$ réduit). Ce principe est illustré dans l'Encadré 12.a. qui représente deux mutations avantageuses $\mathbf{dy}_1$ et $\mathbf{dy}_2$ formant les angles θ_1 et θ_2 avec la direction de l'optimum ($\mathbf{y}_o$). Le fait de considérer le sous-échantillon des mutations avantageuses crée un biais en faveur des mutations pointant vers l'optimum, et donc formant un angle θ réduit entre elles. Or cet angle entre $\mathbf{dy}_1$ et $\mathbf{dy}_2$ détermine le signe de ε, de faibles valeurs de θ correspondant à $\cos(\theta) > 0$ donc à $\varepsilon = -dy^2 \cos(\theta) < 0$. Plus précisément, en sommant les angles par rapport à $\mathbf{y}_o$ on a $\theta = \theta_1 - \theta_2$, donc $\cos(\theta) = \cos(\theta_1)\cos(\theta_2) + \sin(\theta_1)\sin(\theta_2)$. On sait que les cosinus sont nécessairement positifs parmi les mutations avantageuses. Dans la moitié des cas, $\sin(\theta_1)\sin(\theta_2) > 0$ (lorsque $\mathbf{dy}_1$ et $\mathbf{dy}_2$ sont du même côté de l'axe $\mathbf{y}_o$), de sorte que tous ces cas correspondent à $\cos(\theta) > 0$ et donc $\varepsilon < 0$. Dans l'autre moitié des cas, $\sin(\theta_1)\sin(\theta_2) < 0$, mais dans une partie de ces cas, $\cos(\theta_1)\cos(\theta_2) > -\sin(\theta_1)\sin(\theta_2) > 0$ de sorte que $\cos(\theta) > 0$ et donc $\varepsilon < 0$, également. On voit que la distribution de ε entre mutations avantageuses est nécessairement biaisée vers les épistasies négatives (antagonistes).

Ce raisonnement qualitatif peut s'étendre à l'ensemble des mutations de tailles dy variables. Il permet d'interpréter l'observation de Sanjuan et al. d'un biais vers les interactions antagonistes parmi les mutations avantageuses. On peut se demander comment la distance à l'optimum influence ce biais. Le raisonnement géométrique ci-dessus suggèrerait que plus on l'ancêtre est loin de l'optimum, moins le biais est important puisque la contrainte sur la direction de des effets phénotypiques est moins forte. Des simulations (du même type que celles décrite pour modéliser $f(s)$ en Chapitre I, voir Encadré 12.b.) confirment cette tendance: plus le génotype ancestral est adapté (s_o grand) moins le biais est important (*i.e.* plus il y a d'épistasies positives).

I.15 CONCLUSIONS

Les études empiriques suggèrent que la distribution des épistasies (multiplicatives) entre paires de mutations tend à présenter une variance importante et qu'aucune tendance nette en faveurs des épistasies positives ou négatives ne se dégage. Ces deux observations sont en accord avec un modèle basé sur les paysages adaptatifs. Plus précisément, le modèle

permet de prédire la distribution de ε d'après celle de *s*. La confrontations de nos prédictions aux données sur le *VSV* suggère que le modèle fait des prédictions approximativement correctes. Par ailleurs, le modèle prédit que la distribution de ε est indépendante du génotype ancestral, une prédiction a priori facilement testable. Des test supplémentaires sont nécessaires pour valider cette approche, particulièrement parce qu'elle néglige l'existence d'épistasies entres les traits phénotypiques sous sélection (effets additifs sur les phénotypes). De telles relations épistatiques devraient a priori entraîner (i) une augmentation de la variance de ε relativement à l'attendu sous un modèle phénotypique additif simple, (ii) introduire une dépendance de la distribution au niveau d'adaptation du génotype ancestral, et (iii) rendent possible l'existence d'un biais dans la distribution de ε.

Par ailleurs, on prédit également que la distribution de ε parmi les paires de mutations avantageuses est biaisé vers les valeurs négatives (comme observé sur le VSV par Sanjuan et al. (Sanjuan et al. 2004a)), et ce, d'autant que le génotype ancestral est bien adapté. Le test de cette dernière prédiction est possible, mais a priori compliqué par le fait que lorsque le génotype est bien adapté, le nombre de mutations avantageuses est plus réduit, et leurs effets plus faibles donc moins détectables.

Ces résultats ont plusieurs implications évolutives. D'abord, puisque la variance de ε est du même ordre de grandeur que celle de *s*, on ne peut a priori pas négliger ε devant *s*, une hypothèse souvent utilisée (par exemple, dans notre modèle de prédiction des taux d'adaptation). Toutefois, $V(\varepsilon)$ ne varie pas avec le génotype ancestral alors que $V(s)$ augmente a priori avec la maladaptation de celui-ci (Chapitre I et II). Ainsi, il est possible que l'on puisse négliger ε devant *s* lorsque le phénotype ancestral est loin de l'optimum. Cela pourrait expliquer que l'on ait pu obtenir des prédictions approximativement correctes pour les taux d'adaptation à un nouvel environnement (avec s_o grand) en négligeant ε. L'idéal serait de pouvoir incorporer une densité de probabilité pour ε dans les modèles évolutifs, et d'étudier l'impact de la variance de ε entre loci. Toutefois, selon le problème considéré, ce n'est pas la même distribution qui est intéressante. On sait que la variance de ε parmi les mutations délétères entraîne a priori une augmentation du fardeau de mutations (Phillips et al. 2000). La distribution proposée ici correspond à la distribution de ε parmi les mutations délétères lorsque $s_o = 0$. Elle peut donc être utilisée pour étudier (voire prédire ?) les fardeaux évolutifs dus aux mutations délétères récurrentes (fardeau de mutation et de fixation) dans le contexte d'équilibres mutation-sélection (donc près d'un optimum). En revanche, pour modéliser l'influence de l'épistasie sur les taux d'adaptation, ce sont surtout les épistasies entre mutations avantageuses qui sont importantes, or la distribution celles-ci n'est pas facilement prédictible. Il serait tout intéressant d'étudier l'influence du biais vers les ε négatives entre mutations avantageuses, biais suggéré par notre modèle et surtout par la distribution empirique chez le *VSV* (Sanjuan et al. 2004a).

Finalement, on sait que le signe mais aussi la variabilité de ε ont un impact déterminant sur l'évolution de la reproduction sexuée et la recombinaison (Lively and Peeters 2000). Ce point et plus généralement les implications de nos résultats pour l'évolution de la recombinaison sont développés dans la chapitre suivant.

CHAPITRE III: LIAISON GENETIQUE, DERIVE ET EVOLUTION DU SEXE

Articles:

Article 4: Martin, G., Otto, S. P. & Lenormand, T. Selection for recombination in structured populations. *Genetics* 172, 593 - 609 (2006).

D'une manière générale, on a vu dans le Chapitre II qu'il existe une forte interaction entre adaptation, mutations, et système de reproduction. Cette interaction est réciproque. Le type de reproduction (sexué ou asexué) détermine la vitesse de l'adaptation, parce qu'il détermine la probabilité de fixation des mutations avantageuses. Chez les asexués, cette probabilité est réduite par l'interférence avec d'autres mutations avantageuses (interférence clonale) et délétères (sélection de fonds). Nous avons vu dans la section III du Chapitre II que cet effet (notamment l'interférence clonale) pouvait entraîner une forte diminution du taux d'adaptation (parfois de deux ordres de grandeur) entre la drosophile et la bactérie asexuée *E. coli*. En retour, le système génétique peut lui-même être soumis à sélection via son effet indirect sur l'adaptation. L'étude de cette évolution indirecte est un pan important de la biologie évolutive, de nombreux modèles ont été proposés pour expliquer, par des mécanismes de sélection naturelle, la grande ubiquité du sexe. Dans ce chapitre, je m'intéresserai à un de ces mécanismes, basé sur l'interférence entre mutations avantageuses lorsque la recombinaison est réduite (similaire à l'interférence clonale). Je présente d'abord un bref aperçu des modèles d'évolution du sexe. Ensuite, je détaille une approche différente de celle utilisée jusqu'ici (modèle génétique et non phénotypique comme le MF) permettant d'étudier la sélection pour le sexe et la recombinaison dans une population subdivisée soumise à sélection directionnelle.

Résumé (rapide) des théories de l'évolution du sexe

La littérature théorique sur l'évolution du sexe est très fournie, j'essaie de donner ici un aperçu rapide des théories prédisant un avantage associé à la reproduction sexuée, en fonction de son effet sur l'adaptation (*c.a.d.* un avantage indirect). Pour une revue de l'ensemble de ces théories, on peut consulter (Barton and Charlesworth 1998; Kondrashov 1993; Otto and Lenormand 2002).

I.1 PRINCIPES GENERAUX

Par la méiose, le sexe permet la recombinaison et la ségrégation. La plupart des modèles se sont focalisés sur l'impact de la recombinaison. Cet impact dépend du déséquilibre de liaison (DL) qui mesure la tendance à l'association entre allèles à deux loci. Pour définir le DL, bornons-nous à des génotypes haploïdes (les raisonnements sont similaires pour des diploïdes avec des allèles codominants). Considérons deux loci biallélliques j et k, et dénotons p_j et p_k, la fréquences des allèles avantageux à chacun de ces loci respectivement, et p_{jk} la fréquence du génotype de fitness maximale (portant les deux allèles avantageux). Alors le DL est défini (par convention) comme $D = p_{jk} - p_j\, p_k$. Lorsque $D>0$ il y a un excès de génotypes de valeur extrêmes en fitness (celui portant les deux allèles avantageux et celui n'en portant aucun). Cet excès est défini relativement à la fréquence attendue de chaque

génotype si les loci étaient indépendants ($p_{jk} = p_j\, p_k$, $D = 0$). Inversement, lorsque $D<0$ il y a un excès de génotypes de fitness intermédiaires. L'effet génétique de la recombinaison se résume à réduire le DL. Si r est le taux de recombinaison entre j et k, la recombinaison, à chaque génération, réduit le DL de D à $(1 - r)D$. Ainsi, en l'absence de DL, la recombinaison n'a pas d'impact sur la réponse à la sélection car elle ne modifie pas les fréquences génotypiques (Maynard Smith 1971).

La majorité des résultats expérimentaux suggèrent que la recombinaison augmente la réponse à la sélection lors d'un changement environnemental (Rice 2002). Elles sont basées sur (i) l'observation d'une vitesse d'adaptation plus élevée dans des lignées recombinantes de microbes (Colegrave 2002; Goddard et al. 2005; Kaltz and Bell 2002) ou en manipulant la recombinaison dans des lignées de drosophiles (voir la review de Rice 2002). D'autres études ont démontré une augmentation de la recombinaison dans des populations soumises à forte sélection directionnelle (pour revue voir Otto and Lenormand 2002). Toutefois, différentes mécanismes peuvent favoriser la recombinaison, c'est à dire augmenter la fréquences de génotypes recombinants par rapport aux moins recombinants. On peut les résumer à deux cas.

Lorsque $D<0$, la recombinaison recrée des génotypes extrêmes en fitness, et augmente donc la variance en fitness, et la réponse à la sélection. Ce cas correspond à l'idée classique que le sexe « favorise la réponse à la sélection à long terme ». De fait, l'avantage existe à chaque génération, mais il est « à long terme », au sens où la recombinaison permet dans ce cas de produire le génotype qui sera amené à se fixer à terme (celui de fitness maximale). Un tel avantage favorise l'augmentation de la recombinaison *localement*. La notion de « local » est la suivante: considérons que certains loci (autre que j et k) influencent le taux de recombinaison entre j et k (ils codent pour la valeur de r). Ces loci peuvent être soumis à sélection pour augmenter ou diminuer r lorsque la recombinaison présente un avantage ou un désavantage. On appelle de tels loci des modifieurs de recombinaison. La recombinaison est favorisée *localement* si seuls des loci modifieurs liés aux loci j et k peuvent répondre à la sélection en faveur de la recombinaison. Cette notion est importante pour comprendre à partir de quelle base génétique le sexe et la recombinaison peuvent évoluer. Dans un contexte où $D<0$, la recombinaison présente un avantage sélectif, mais elle ne peut évoluer que si une mutation survient (ou un allèle est présent) qui permet plus de recombinaison, ET qui est liée aux loci j et k.

Alternativement, la recombinaison peut être avantagée si la moyenne (et non la variance) de la fitness est supérieure au sein des génotypes issus de recombinaison. C'est le cas lorsque l'épistasie additive est de signe opposé à D. De même que l'épistasie multiplicative (ε_{jk} entre j et k, voir chapitre II) est définie relativement au modèle multiplicatif à deux loci, l'épistasie additive a_{jk} est définie relativement à un modèle additif. Si s_j et s_k sont les coefficients de sélection aux loci j et k, $a_{jk} = s_j\, s_k + \varepsilon_{jk}$. Ces deux épistasies sont donc reliées et souvent de même signe. Or, Feldman (1972) a montré que dans une population infinie soumise à sélection directionnelle, l'épistasie multiplicative ε_{jk} génère un déséquilibre de liaison du signe de ε_{jk}. Ainsi, il est rare que D soit du signe contraire de a_{jk}, puisqu'il est du signe de ε_{jk}. Il est donc rare (dans une situation simple) que la recombinaison entraîne une augmentation de la moyenne de fitness des recombinants. Toutefois, dans de telles situations, l'avantage associé à la recombinaison n'est plus local: toute mutation augmentant r peut faire évoluer la recombinaison, quelle que soit sa position dans le génome.

En résumé, la recombinaison peut être avantagée parce qu'elle augmente la variance en fitness (effet sur la variance) ou la moyenne (effet sur la moyenne) des génotypes issus de recombinaison. Les différents modèles qui prédisent un avantage associé à la recombinaison correspondent donc à des situations où (i) $D<0$ (effet sur la variance) ou à des situations où (ii) D et a_{jk} sont de signes opposés (effet sur la moyenne). Le cas (i) entraîne une sélection locale pour la recombinaison (seuls des modifieurs liés répondent à la sélection), tandis que dans le cas (ii), tout locus modifieur peut répondre à la sélection.

I.2 THEORIES DETERMINISTES

Plusieurs théories prédisent un avantage associé à la recombinaison dans des populations de taille infinie (théories « déterministes » du sexe Kondrashov 1993). La recombinaison peut être favorable parce qu'elle permet d'éliminer efficacement les mutations délétères à plusieurs loci, ou parce qu'elle permet aux mutations avantageuses à plusieurs loci d'envahir la population. Dans un modèle déterministe et sous une sélection homogène dans le temps et l'espace, ces deux avantages supposent que l'épistasie multiplicative ε_{jk} est faible et négative (Barton 1995a). Ce résultat vient de l'interaction entre effet sur la variance et effet sur la moyenne. L'épistasie négative génère un DL négatif, de sorte que la recombinaison augmente la variance en fitness (effet sur la variance). Toutefois, lorsque ε_{jk} prend de fortes valeurs négatives, a_{jk} est également négative. Alors, l'épistasie additive est du même signe que D et l'effet sur la moyenne défavorise la recombinaison: la moyenne de la fitness des génotypes issus de recombinaison est inférieure à celles des non-recombinants. Ainsi, des valeurs positives ou trop fortement négatives de ε correspondent à un désavantage des recombinants. En outre, si la force et le signe des interactions épistatiques varient entre plusieurs paires de loci le long du génome, l'avantage associé à la recombinaison se trouve réduit, même si, en moyenne sur tous les loci, ε est faible et négative (Otto and Feldman 1997). L'avantage conféré à la recombinaison a un effet local puisqu'il provient d'un effet sur la variance. Il ne peut donc faire évoluer que des modifieurs liés.

La recombinaison peut aussi être favorisée en population infinies, sous certaines forme de sélection variable dans l'espace ou le temps. Plus précisément, si le signe de l'épistasie fluctue rapidement dans le temps (faisant des cycles de quelques générations) la recombinaison peut être favorisée (Barton 1995a). Ce mécanisme s'apparente à un effet sur la moyenne. En réponse aux fluctuations du signe de ε, le signe du DL fluctue également, mais avec un décalage temporaire. Ainsi, avec la bonne cyclicité, les épistasies additives (qui fluctuent en même temps que ε) peuvent se retrouver de signe opposé à D. Ainsi, la recombinaison peut être favorisée. Ce modèle suppose toutefois une cyclicité très rapide de l'épistasie, ce qui le rend a priori assez peu général (mais voir Lively and Peeters 2000; Peters and Lively 1999).

Par ailleurs, lorsque la sélection est hétérogène dans l'espace (entre plusieurs dèmes), les fréquences alléliques varient entre dèmes, à chaque locus. Dans ce cas, la migration est une source de DL importante qui vient s'ajouter à l'épistasie. Ainsi, on peut obtenir des situations où D est de signe opposé à a_{jk} (avantage lié à un effet sur la moyenne), et des situations où $D<0$ (avantage lié à un effet sur la variance) (Lenormand and Otto 2000). De telles situations sont prédites, par exemple, dans un contexte d'adaptation locale en environnement hétérogène.

Tous ces modèles supposent un mode de reproduction panmictique. La prise en compte de la consanguinité dans les modèles d'évolution du sexe peut amener à changer nombre des conclusions. Dans des populations partiellement ou complètement auto-fécondantes, la sélection pour la recombinaison peut être très forte, elle dépend cette fois du signe des épistasies dominance par dominance (entre loci et chromosomes) (Roze and Lenormand 2005). Un tel effet agit également lorsque les populations sont structurées (la structure créant une part de consanguinité même si la reproduction est panmictique au sein de chaque dème). Toutefois, on ne sait quasiment rien de la distribution de ces épistasies. Enfin, certains modèles montrent que le fait d'autoriser la ségrégation peut constituer également un avantage important associé au sexe (Otto 2003).

I.3 THEORIES STOCHASTIQUES

Alternativement, un ensemble de théories prédisent un avantage à la recombinaison dans des populations finies même sous sélection homogène dans l'espace ou le temps (« théories

stochastiques », Kondrashov 1993). Lorsque des mutations ségrègent simultanément à deux loci, un DL négatif tend à être généré par l'interaction entre dérive et sélection (même en l'absence d'épistasie, $\varepsilon = 0$). C'est l'effet Hill-Robertson (HRE) qui entraîne une réduction de la probabilité de fixations de mutations avantageuses à plusieurs loci lorsque la recombinaison est réduite (Barton 1995b; Hill and Robertson 1966). Ce processus sera détaillé plus loin. Il est similaire à l'interférence clonale chez des asexués (Gerrish and Lenski 1998), dans le cas limite où $r = 0$. Nous avons vu en chapitre II que cette interférence clonale réduisait le taux d'adaptation chez les asexués en l'absence d'épistasie (celle-ci n'était pas prise en compte dans les prédictions). La généralité du HRE (il biaise D vers les valeurs négatives quelle que soit l'épistasie) en fait une limite potentiellement importante à la réponse à la sélection d'un ensemble de loci liés dans une population de faible effectif. Dans ce contexte, la recombinaison peut être favorisée puisqu'elle réduit le HRE (Felsenstein 1974). Cet avantage associé à la recombinaison est basé sur un effet sur la variance (le HRE est une source de DL négatif), et ne fait donc évoluer la recombinaison que localement (à des modifieurs liés aux loci j et k).

Dans ce chapitre, je résume d'abord les implications de nos résultats précédents (Chapitre II) pour certaines des théories présentées ci-dessus. Ensuite, je présente un modèle permettant d'appréhender l'interaction entre dérive et la sélection (HRE) sur la réponse à la sélection à deux loci liés dans une population structurée. Je montre enfin comment un tel processus peut sélectionner pour le sexe et la recombinaison.

Implications de nos résultats précédents pour l'évolution du sexe

I.4 DISTRIBUTION DE L'EPISTASIE ET THEORIES DETERMINISTES DU SEXE

En populations infinie sous sélection directionnelle, la recombinaison n'est favorisée que si ε est faible, négative, et peu variable entre loci. Comme nous l'avons vu, les quelques distributions de ε observées empiriquement (Bonhoeffer et al. 2004; de Visser and Hoekstra 1998; Elena and Lenski 1997; Sanjuan et al. 2004a) tendent à être assez symétriques avec une moyenne proche de zéro et surtout une forte variance. Au vu de nos résultats théoriques, cette tendance devrait être assez générale, or une telle distribution de ε défavorise a priori la recombinaison.

Pour une paire de mutations donnée, l'avantage associé à la recombinaison par son effet sur les mutations avantageuses est quantitativement plus grand que celui lié à l'élimination des mutations délétères. Il est donc notable que l'on attend un biais vers les épistasies négatives entre les mutations avantageuses (voir Chapitre II), et que celui-ci est observé empiriquement (Sanjuan et al. 2004a). Ce biais vers les $\varepsilon<0$ entre mutations avantageuses pourrait favoriser la recombinaison. Toutefois, il faudrait pour cela que cet effet compense le désavantage associé à la recombinaison à cause des interactions variables entre les mutations délétères. En outre, nos simulations comme les données sur le VSV suggèrent que ε entre les mutations avantageuses peut atteindre de grandes valeurs négatives ce qui défavorise la recombinaison également. Déterminer l'effet net de toutes ces interactions nécessiterait un modèle explicite intégrant l'ensemble des interactions possibles, mais il me semble que le domaine de paramètre dans lequel la recombinaison est avantagée serait pour le moins réduit. En effet, l'avantage lié aux mutations avantageuses est plus important lorsque celles-ci sont fréquentes, donc pour de grandes valeurs de s_o. Or nous avons vu que dans ce cas la distribution de ε parmi les mutations avantageuses est plus symétrique (moins biaisée

vers ε<0), ce qui devrait là-aussi défavoriser la recombinaison. Ainsi, au vu du nombre limité d'études empiriques et du modèle théorique présenté en chapitre II, il semble peu probable que la recombinaison soit favorisée via la sélection épistatique en population infinie (Sanjuan et al. 2004a). D'autres modèles, basés sur la théorie du flux métabolique, amènent à des conclusions similaires (Phillips et al. 2000).

I.5 DERIVE ET THEORIES « STOCHASTIQUES » DU SEXE

Nous avons vu qu'en populations finies, le HRE pouvait constituer une forte limite à la réponse à la sélection à des loci liés. Pourtant, la prédiction du taux d'adaptation chez la drosophile (Chapitre II) a été faite en ignorant la liaison entre loci sous sélection, de sorte que le HRE était négligé malgré la taille efficace réduite des populations (N_e = 300). Cette prédiction semble pourtant cohérente avec les taux d'adaptation empiriques. Toutefois, on sait que le HRE est peu efficace en populations de taille très réduite (par exemple, N_e =300) parce que la probabilité que plusieurs mutations avantageuses ségrègent simultanément est réduite dans ce cas (Otto and Barton 2001). Cela pourrait expliquer la cohérence des taux d'adaptation observés dans ces populations avec celui prédit par notre modèle qui néglige le HRE. En revanche, nos résultats suggèrent un effet important de l'interférence clonale sur le taux d'adaptation chez *E. coli*. Cela suggère donc que l'interférence entre mutations avantageuses pourrait constituer une limite importante à la réponse à la sélection à des loci liés (ici à l'extrême puisque *E. coli* est asexuée).

En résumé, le HRE entraîne donc une limite à l'adaptation entre des loci liés, dans des populations de taille intermédiaire (assez grandes pour que des mutations avantageuses ségrègent simultanément, mais suffisamment réduites pour que la dérive ait une influence Otto and Barton 2001). Cet effet pourrait constituer une source assez générale de sélection en faveur de la recombinaison puisqu'elle agit indépendamment de l'épistasie et en l'absence de sélections variable dans l'espace ou le temps. Notons que dans une population finie, l'épistasie et le HRE devraient agir simultanément. Si, comme on l'a vu au chapitre II, ε entre les mutations avantageuses est biaisée vers les valeurs négatives, l'épistasie pourrait contribuer, en même temps que le HRE, à générer des DL négatifs. Dans ce contexte pourtant, la recombinaison pourrait être défavorisée si l'épistasie est trop négative en moyenne. En revanche, si l'épistasie moyenne est relativement faible (positive ou négative), le DL négatif généré par le HRE pourrait être suffisant pour que la recombinaison soit favorisée malgré la variance de ε.

Toutefois, les théories stochastiques ne s'appliquent pas aux populations de grandes tailles, or de nombreuses populations naturelles présentent des effectifs efficaces importants. Néanmoins, on peut également supposer que de nombreuses populations naturelles sont structurées (en dèmes de taille plus réduite, ou par un isolement par la distance). Les sections suivantes présentent une approche permettant de modéliser l'impact du HRE dans une population subdivisée. Le but est d'estimer si un tel processus peut entraîner, pour un niveau de recombinaison donné, une limite à l'adaptation quantitativement importante. Cette approche permet donc d'aller au delà de deux modèles extrêmes présentés précédemment (une reproduction purement asexuée ou l'absence totale de liaison) et d'appréhender comment la structuration des populations et la liaison génétique influencent conjointement la réponse à la sélection et, de là, l'évolution du sexe et de la recombinaison.

Liaison génétique et réponse à la sélection en populations subdivisées

Article 4 : Evolution of recombination in subdivided populations

I.6 DESEQUILIBRE DE LIAISON GENERE PAR L'EFFET HILL-ROBERTSON

L'interprétation intuitive de l'effet Hill-Robertson est difficile, et varie selon les articles. Historiquement, il fut décrit par ses « inventeurs » comme suit : le succès reproducteur d'un allèle à un locus donné est partiellement influencé par la fitness des allèles aux loci auxquels il se trouve lié. Cette variance de succès reproducteur, indépendante de l'effet sélectif de l'allèle considéré, est assimilable à une réduction de l'effectif efficace local (dans la portion de génome liée au du locus considéré) qui entraîne une réduction des probabilités de fixation des allèles avantageux. Toutefois, Barton (Barton) a montré que cette interprétation n'est plus valide lorsque la recombinaison est très faible car les différences de succès reproducteur entre allèles à un même locus sont autocorrélées dans le temps (ce qui n'est pas le cas sous l'effet de dérive décrit par une réduction locale de N_e). Une interprétation alternative est proposée par Barton et Otto (2004) en terme de production de DL négatif. La dérive, par simple échantillonnage, crée de la variance de DL (sans le biaiser en moyenne). Lorsque par hasard le DL est négatif entre les loci sous sélection, la réponse à la sélection s'en trouve réduite. Le DL reste alors longtemps présent, jusqu'à ce que les allèles avantageux se soient fixés. Inversement, lorsqu'un DL positif est initialement créé, celui-ci disparaît vite car la réponse à la sélection accrue entraîne la fixation rapide des allèles avantageux. Sur l'ensemble des évènements possibles de balayages sélectifs (ou dans plusieurs populations isolées), le DL moyen se trouve donc biaisé vers les valeurs négatives. Cette interprétation permet de comprendre comment l'interaction de la sélection et de la dérive produit un DL négatif en moyenne, proportionnellement à la variance par dérive donc à $1/N_e$.

La modélisation analytique du HRE est problématique car les équations de diffusion à deux loci liés n'ont pas de solution analytique (Hill and Robertson 1966). Un modèle basé sur les processus de branchements (« branching process ») a permis de prédire le déséquilibre de liaison généré par le HRE lors d'une phase de sélection à deux loci liés, en considérant des mutations initialement rares (Barton 1995b). Sa transcription dans un modèle structuré pour obtenir des résultats explicites et intuitifs ne m'est pas apparue évidente. Un autre modèle permet d'obtenir des expressions simples pour le DL généré durant la fixation d'allèles avantageux (Barton and Otto 2004). Comme les autres modèles, il s'appuie sur une approximation en population de grande taille (mais finie). On étudie les déviations stochastiques par rapport à une trajectoire déterministe (des fréquences alléliques et de D) prédite en population infinie. En suivant l'espérance et la variance de ces déviations, des expressions peuvent être obtenues pour le déséquilibre moyen E(D) attendu à chaque moment du balayage sélectif des allèles avantageux, ainsi que pour la réduction de la vitesse de ce balayage relativement au cas non lié. Si les fréquences des allèles avantageux (d'effet sélectif s_j et s_k) aux deux loci (j et k) sont p_j et p_k à un moment donné du balayage, le DL entre ces loci dans une population de taille N vaut en espérance

$$E(D) = s_j s_k \frac{p_j(1-p_j)p_k(1-p_k)}{N r^3} \qquad (17)$$

Où r est le taux de recombinaison (Barton and Otto 2004). Cette approximation suppose que r est suffisamment grand ($r>>s$) pour que l'approximation de « quasi-linkage equilibrium »

(QLE) s'applique (Barton 1995a). L'équation (17) montre de façon assez simple que le déséquilibre de liaison généré est (i) proportionnel au coefficient de sélection de chacun des allèles avantageux (ii) inversement proportionnel à la taille de la population (N) et au taux de recombinaison (r). Le HRE constitue donc une limite forte à l'adaptation dans des populations de taille limitée soumises à de fortes pressions de sélection (par ex. après un changement environnemental abrupt). La réduction de la réponse à la sélection due à la liaison se traduit ici par une réduction de la vitesse de fixation des allèles avantageux et non par une réduction de leur probabilité de fixation (qui est supposée égale à 1 s'ils ne sont pas trop rares initialement). Elle est proportionnelle à ce déséquilibre.

La principale limite de cette approche est qu'elle suppose que les fréquences alléliques ne sont pas trop basses initialement (≥ 1%), ce qui tend à limiter son application au cas d'une réponse à la sélection partant d'une variance génétique préexistante.

I.7 INFLUENCE DE LA STRUCTURATION ET DE LA LIAISON SUR E(D) ET LA REPONSE A LA SELECTION

Nous avons étendu ce modèle à une population subdivisée en n îles de taille N échangeant une proportion m de migrants à chaque génération de façon équiprobable (modèle en îles fini). Le modèle est simplifié en faisant l'hypothèse que les fréquences alléliques initiales varient peu entre dèmes, ce qui revient à une hypothèse de faible structuration. Dans ce cas la trajectoire déterministe des fréquences alléliques et de D est la même dans tous les dèmes et le système à n îles peut être réduit à deux dèmes : un dème focal et un « dème moyen » représentant le nuage de migrants (ses fréquences génotypiques sont la moyenne des fréquences dans la métapopulation). Le déséquilibre de liaison moyen par dème (et les autres variables stochastiques) est celui qui serait obtenu dans le dème focal s'il échangeait $m_e = m\,n/(n\text{-}1)$ migrants uniquement avec ce « dème moyen ». Les récursions montrent que le système se comporte comme un dème isolé d'une taille « équivalente » intermédiaire (selon la valeur de m_e) entre la taille d'un dème(N) et la taille totale($n\,N$) pour l'ensemble des variables.

Ce résultat semble assez intuitif, mais l'effet de la structure diffère toutefois ici de son effet prédit en l'absence de liaison. D'abord, dans un modèle en île, les probabilités de fixation de mutations codominantes en l'absence de liaison ne sont pas affectées par la structure, à taille efficace égale (Maruyama 1974). Ici, nos simulations ont montré qu'elles étaient réduites lorsque l'on considère une forte structure et de faibles fréquences initiales (*i.e.* hors des hypothèses du modèle). Ce résultat n'est pas détaillé dans l'article 4, qui traite surtout du DL et de l'évolution du sexe. Ensuite, la taille équivalente ne dépend pas que de la structure, mais aussi du taux de recombinaison entre les deux loci sous sélection. Ainsi, l'effet de la structure sur la limitation de la réponse à la sélection multilocus n'est pas homogène sur tout le génome, même à coefficient de sélection égal. Cet effet apparaît dans le modèle analytique: pour une liaison faible (QLE, $r \geq 0.05$) le déséquilibre de liaison moyen par dème $E(\overline{D})$ est égal à celui obtenu dans un dème isolé de taille N_{qle} (*i.e.* en remplaçant N par N_{qle} dans l'équation (17)) :

$$N_{qle} = \frac{nN}{1+(n-1)\alpha} \xrightarrow{n\to\infty} \frac{N}{\alpha} \qquad (18)$$

Où α dépend à la fois de m et de r. En outre, ce résultat reste exact pour tout taux de recombinaison lorsque l'on considère un modèle en île infini. Ce dernier point suppose une structuration faible, mais l'approximation donne le bon ordre de grandeur même pour m = 0.001.

La conséquence principale de ce résultat est que, quelle que soit la taille totale de la population (nN), le HRE influe sur le DL et sur la réponse à la sélection comme dans une

population finie de taille N_{qle}. Ce N_{qle} peut être proche de la taille du dème si la population est assez fortement structurée.

Par ailleurs, nos simulations dans le cas de populations structurées en dèmes de petite taille (N<1000) révèlent que dans ce cas, la migration entre dèmes (même réduite) restaure le polymorphisme de sorte que le DL entre loci sous sélection se maintient. Si les populations sont isolées, ce DL est éliminé puisque l'un des deux voire les deux allèles avantageux sont rapidement perdus par dérive en très petites populations (Otto and Barton 2001). Ainsi, dans ce contexte, la liaison génétique constitue une limite à l'adaptation qui n'est présente ni en l'absence de migration, ni en population panmictique. Il est donc probable que des populations de faibles effectifs locaux mais partiellement connectées soient également sensibles au HRE, et présentent des réponse à la sélection réduites sur des gènes liés.

I.8 EVOLUTION DU SEXE ET DE LA RECOMBINAISON EN POPULATION SUBDIVISEE

Evolution de la recombinaison: Une conséquence logique du fait que la liaison limite la réponse à la sélection est que la recombinaison peut évoluer pour diminuer cette liaison. Toutefois, la réponse de la recombinaison à cette sélection indirecte dépend de la base génétique du « trait » recombinaison. Plus précisément, on peut modéliser l'évolution de la recombinaison en incluant au modèle présenté précédemment un troisième locus i, codant pour le taux de recombinaison entre les loci j et k (locus modifieur, présenté en partie I de ce chapitre). Dénotons r_{ij} le taux de recombinaison de ce locus avec le locus j. Supposons que ce locus est biallélique avec un allèle A codant pour un taux de recombinaison $r + dr$ entre les loci j et k, et l'autre allèle a codant pour un taux r (dr est l'effet du modifieur). Un tel locus peut rendre compte de la base génétique réelle de la variation du taux de recombinaison, ou être vu comme un marqueur de la sélection sur la recombinaison. On peut calculer la variation de la fréquence p_i, dans la métapopulation, de l'allèle A à chaque génération (prise en espérance sur plusieurs répliquas de métapopulations). Ce changement de fréquence moyen E(dp_i) détermine la force de la sélection pour la recombinaison. Même dans les conditions de « quasi-linkage equilibrium » (r, r_{ij} >> s_j , s_k), l'expression de E(dp_i) est compliquée. Toutefois, elle est bien approximée par le changement moyen attendu dans une population panmictique de taille N_{qle} définie en Eq. (18). Au QLE, on obtient donc

$$E(dp_i) \propto \frac{dr}{r^3 r_{ij}^2} \frac{s_j^2 s_k^2}{N_{qle}} pq_{ijk}, \qquad (19)$$

Où pq_{ijk} est le produit des variances génétiques ($p(1 - p)$) aux trois loci (Eq. 7a de Barton and Otto 2004).

Il y a donc une équivalence approximative entre l'effet de la structure sur le DL et sur la sélection en faveur de la recombinaison. Comme pour le DL moyen (E(D), Eq. (17)), la sélection pour la recombinaison croît avec la force de la sélection directionnelle (d'autant plus que celle-ci est égale aux deux loci (s_j^2 s_k^2)), et elle décroît avec le taux de recombinaison entre les loci sous sélection ($1/r$). Par ailleurs, cette sélection favorise des modifieurs liés aux loci j et k (la sélection est proportionnelle à $1/r_{ij}$).

On peut tirer les même conclusions que pour D : la sélection pour la recombinaison sous l'effet joint de la dérive et de la sélection (du HRE) dépend non pas de la taille totale de la population, mais de son niveau de structure. La sélection directionnelle dans une population grande mais structurée peut favoriser la recombinaison à un niveau équivalent à celui attendu dans des populations de tailles intermédiaires, si m est réduit. Ces conditions constituent justement le domaine de paramètre où la recombinaison est le plus favorisée dans les théories stochastiques.

Evolution du sexe face au coût de deux: Enfin, nous avons aussi étudié par simulation l'évolution de modifieurs de sexe. Le but étant de déterminer dans quelle mesure l'avantage conféré par la recombinaison pouvait dépasser le coût de deux associé à la reproduction sexuée. Nous avons considéré une population initialement asexuée, et étudié dans quelles circonstances un génotype produisant une proportion réduite de sa descendance par reproduction sexuée pouvait envahir la population, le nombre de descendant issus de reproduction sexuée étant décompté d'un facteur ½ par rapport aux nombre de descendants issus de reproduction asexuée. Ce « coût » du sexe rend compte du fait que l'apparentement d'un descendant sexué avec sa mère n'est que la moitié de celui qu'elle aurait eu avec un descendant issu de reproduction asexuée. Dans ces simulations, un génotype se reproduisant de façon partiellement sexuée peut envahir, malgré ce coût de deux, mais seulement s'il se reproduit de façon sexuée marginalement (c'est à dire rarement ou pour une faible fraction de ses descendants).

I.9 CONCLUSIONS

Théories stochastiques du sexe: La conclusion principale de cette étude est que les théories stochastiques du sexe ne se limitent pas aux populations de taille intermédiaire. En effet, en l'absence de structuration la recombinaison n'est pas favorisée dans les populations qui sont trop grandes (où l'effet de la dérive n'est pas sensible) ou trop petites (où plusieurs allèles avantageux ne ségrégent pas simultanément), tandis qu'avec en population structurée, la recombinaison peut être significativement avantagée tant que la taille des dèmes n'est pas trop grande, ce qui est beaucoup moins restrictif. Les populations de petite taille mais connectées par migration évoluent en effet comme une population plus grande (mais qui reste finie). Inversement, les populations de grande taille totale mais subdivisées en dèmes de taille assez réduite peuvent évoluer quasiment comme des populations isolées d'une taille comparable à celle des dèmes. Ces conclusions dépendent évidemment du niveau de structuration des populations naturelles. La plupart des études sur ce sujet mettent cependant en évidence que l'immense majorité des populations naturelles sont structurées, de manière presque automatique quand les habitats sont fragmentés ou discontinus, mais aussi lorsque les habitats sont plus continus (isolement par la distance).

Une limite importante de nos résultats (et plus généralement des théories stochastiques) est que le HRE ne favorise vraiment la recombinaison que lorsqu'elle est faible (modifieurs liés). Comme nous l'avons vu en section I du chapitre, cette limite est inhérente à tout modèle basé sur un effet sur la variance (donc la majorité d'entre eux). Ainsi, il est difficile de prédire l'évolution de forts taux de recombinaison par ces modèles ou de forts taux de reproduction sexuée (par exemple dans nos simulations d'un modifieur de sexe). Pourtant, de nombreuses espèces se reproduisent exclusivement de façon sexuée (la nôtre y compris…à ma connaissance). Cependant, tous ces modèles (y compris le nôtre) sont souvent basés sur une dynamique à deux loci (Iles et al. 2003). Pourtant, comme on l'a vu en Introduction et en chapitre I, il peut arriver que, lors d'un changement environnemental par exemple, un grand nombre de mutations avantageuses ségrègent simultanément. Nous avons constaté par ailleurs que chez *E. coli*, l'interférence entre mutations avantageuses semble avoir un effet quantitatif très important sur le taux d'adaptation (Chap. II, section III). Il semble donc que pour comprendre, quantitativement, l'importance du sexe et de la recombinaison, il faille prendre en compte de nombreux loci. Illes et al. (2003) ont en effet montré que, lorsque de nombreux loci ségrègent simultanément, le HRE peut être quantitativement important, même dans de grandes populations. Par ailleurs, la structuration génère de la consanguinité au sein de chaque dème (même chez des espèces allogames), ce qui pourrait aussi entraîner d'autres avantages associés au sexe, notamment si les épistasies

dominance par dominance sont négatives (Roze and Lenormand 2005), ou via un avantage associé à la ségrégation (Otto 2003).

HRE et adaptation : Du point de vue des théories de l'adaptation, nos résultats suggèrent que le HRE pourrait être une force importante agissant sur la réponse à la sélection *in natura*. Or, les approches sur des populations de laboratoire en petits effectifs, par exemple avec la drosophile, ne détectent pas cet effet car les populations sont de taille trop réduite. Alternativement, l'utilisation de populations microbiennes (donc potentiellement de grande taille) pourrait permettre de vérifier l'impact de la structure en imposant un régime démographique de population structurée au laboratoire. Le modèle biologique qui me semblerait le plus approprié est la levure. En effet, on peut imposer un système de reproduction sexué ou asexué dans cette espèce, et il a déjà été utilisé pour mesurer l'impact du sexe sur l'adaptation (Colegrave 2002; Goddard et al. 2005). On pourrait alors comparer les taux d'adaptation entre sexués ou asexués selon différents niveaux de structuration. Pour correspondre au modèle, il semblerait plus logique de se baser sur des populations initialement variables (et non des lignées isogéniques). En effet notre approche étudie l'adaptation à partir d'allèles avantageux relativement fréquents initialement (par ex. p_j,p_k>0.1%) donc s'appliquerait plus à un processus adaptatif à partir de variance existante.

CONCLUSIONS ET PERSPECTIVES

J'ai discuté les résultats de mon travail dans chaque chapitre. J'essaie donc ici de dégager et résumer les apports principaux, à mon avis, de cette thèse. Ensuite, j'envisagerai certaines limites et perspectives de mon travail théorique. Je finirai par ouvrir sur les perspectives de recherche qui me semblent possibles sur mon modèle biologique *Artemia*, pour l'étude de l'adaptation *in natura*.

Bilan des résultats obtenus

I.1 MUTATION ET ADAPTATION

Le but principal était de proposer un modèle pour $f(s)$ et de le valider empiriquement. Le modèle que nous avons proposé (Chapitre I) permet (i) de faire des prédictions testables (puisqu'il repose sur des grandeurs mesurables), (ii) de rendre compte de la variation de $f(s)$ entre génotypes et entre environnements et (iii) de relier $f(s)$ à des paramètres biologiques explicites (importance des corrélations phénotypiques, niveau d'adaptation du génotype). Par ailleurs, il permet de prendre en compte une part importante de la complexité de l'interaction entre mutation, phénotype et fitness et de la résumer à un ensemble limité de paramètres. Nous avons pu tester certaines des prédictions du modèle (Chapitre II), à partir des données empiriques existantes. On peut tirer plusieurs conclusions de ces analyses, quant à la variation de $f(s)$ entre espèces et entre environnements:

(i) L'effet moyen des mutations délétères augmente avec le nombre de traits sous sélection (avec le nombre de gènes du génome), *i.e.* des espèces peu complexes aux organismes « supérieurs ».

(ii) $f(s)$ pour les mutations délétères est approximativement une gamma dont le paramètre de forme (β_o) varie peu entre espèces phylogénétiquement très éloignées.

(iii) La variance de $f(s)$ augmente avec la maladaptation du génotype ancestral à son environnement.

(iv) L'effet moyen et le nombre des mutations délétères exprimées varient peu entre environnements.

Toutes ces observations sont en accord avec le modèle que nous avons proposé. Indépendamment de toute interprétation à partir du modèle (ou d'un autre), les revues des données empiriques proposées ici (Chapitre II), semblent dégager des tendances assez générales quant à la façon dont $f(s)$ varie entre espèces ou entre environnements. Elles reposent sur des analyses comparatives entre des espèces très différentes (bien qu'en nombre limité), ce qui donne donc une certaine généralité à ces conclusions.

Néanmoins, ces validations sont basées essentiellement sur les distributions empiriques de l'effet des mutations délétères (parce que ce sont les mieux documentées). Or, il est important de tester aussi la validité des prédictions pour la distribution des effets avantageux, puisque ceux-ci sont la source de l'adaptation. Nous avons pu proposer un test de ces prédictions, en étudiant les trajectoires de fitness moyenne pendant une expérience d'adaptation, à long

terme, dans un environnement fixé et contrôlé (Chapitre II, section III). Notre modèle permet en effet de prédire (dans les espèces pour lesquelles les paramètres mutationnels sont connus) le taux d'adaptation initial à un nouvel environnement en connaissant la fitness finale (lorsque l'optimum pour le nouvel environnement est atteint), ou inversement, de prédire la fitness finale à partir du taux d'adaptation initial. Les comparaisons de nos prédictions aux trajectoires de fitness empiriques suggèrent que le modèle permet de faire des prédictions assez correctes quantitativement. Cela semble donc valider les prédictions du modèle pour les mutations avantageuses, mais aussi les approximations faites dans des travaux théoriques antérieurs pour calculer les probabilités de fixation de ces mutations. Toutefois, notre test ne s'est basé que sur trois expériences indépendantes, et dans seulement deux espèces (*E. coli* et *D. melanogaster*), ce qui est clairement insuffisant. Néanmoins, le modèle reste facilement testable à partir d'autres expériences et dans d'autres espèces modèles.

Tous ces tests sont indirects, notre modèle constitue donc une approche « minimale » pour rendre compte de ces données. Je n'ai pu faire un test direct des prédictions du modèle. Comme expliqué en Chapitre II (section II), celui-ci pourrait être réalisé par une comparaison de $f(s)$, pour un même génotype ancestral, le long d'un gradient continu pour une variable environnementale (par ex. salinité, température). Cette variable devrait définir une niche écologique pour le génotype ancestral, avec une valeur optimale bien caractérisée. Le crustacée *Artemia* constitue un bon modèle biologique pour une telle approche puisqu'il présente des niches claires pour la salinité et la température (Browne, 2000 #713, pour revue voir Lenz and Browne 2000), deux variables facilement contrôlables au laboratoire (au moins la salinité). Un apport empirique de ma thèse a donc été de développer divers protocoles d'élevage et de mesures de traits et d'accumulation de mutations induites pour ce modèle biologique (Annexe 2). Je n'ai pas pu produire de lignées mutantes fixées d'*Artemia*. Cependant, les protocoles de mutagenèse proposés pour *Artemia* me semblent pouvoir être appliqués efficacement, en modifiant légèrement le système d'élevage pour obtenir une bonne reproduction au laboratoire.

Au final, j'espère que ce travail puisse constituer un premier pas vers la modélisation quantitative de l'adaptation dans des systèmes bien maîtrisés au laboratoire. La validité de mes approches sera peut-être infirmée par la confrontation à un plus grand nombre de données expérimentales, mais elle constitue au moins un ensemble de prédictions testables. Il me semble aussi qu'il est nécessaire de proposer un test plus direct du modèle, du type de celui proposé sur *Artemia*. Il serait également possible de réaliser un tel test sur d'autres modèles biologiques, comme *E. coli*, la drosophile, le VSV, ou la levure. En effet dans chacune de ces espèces, des lignées mutantes portant une unique mutation on été produites. Cela présente donc l'avantage qu'on peut directement observer $f(s)$ (par opposition aux expériences MA, qui mesurent l'effet cumulé de plusieurs mutations dont le nombre varie entre lignées, voir Introduction). En revanche il serait nécessaire de caractériser une variable environnementale continue et bien contrôlée, pour laquelle ces espèces présentent une niche claire (la température par exemple ?).

I.2 MUTATION, DERIVE ET SYSTEME DE REPRODUCTION

Au delà de la mutation, je me suis également intéressé dans cette thèse à deux facteurs affectant l'adaptation: la dérive et surtout le système de reproduction, ainsi qu'à la façon dont ce dernier peut évoluer sous sélection naturelle.

Les prédictions de trajectoires de fitness présentées plus haut (et Chapitre II, section III) permettent de quantifier l'impact relatif de ces deux facteurs (dérive et système de reproduction) sur les taux d'adaptation. En comparant les taux observés aux prédictions prenant ou non en compte ces facteurs, on peut ainsi déterminer quantitativement leur effet sur les taux d'adaptation. Par ce moyen (et si la validité du modèle est confirmée !), on peut

donc faire un premier pas vers une compréhension de l'impact *quantitatif* de la dérive ou de la reproduction sexuée sur l'adaptation, à partir de données empiriques. On peut ainsi appréhender, notamment, la force de la sélection pour la reproduction sexuée. Nos premiers résultats sur *E. coli* suggèrent que l'interférence clonale (due à la reproduction asexuée) constitue dans cette espèce un forte limite, qui réduit de plusieurs ordres de grandeurs les taux d'adaptation en grande population (Chapitre II, section III). Cet effet semble nettement plus important que celui induit par la sélection de fonds (due aux mutations délétères).

Cependant, d'autres facteurs peuvent influencer l'évolution de la reproduction sexuée, en particulier l'épistasie ε. Bien qu'elle ait un impact important pour les théories de l'évolution du sexe, on connaît assez peu la distribution de ε entre paires de mutations aléatoires. J'ai donc étendu mon modèle de $f(s)$ pour prédire cette distribution (Chapitre II, section IV). La distribution prédite de ε présente une assez forte variance et une moyenne nulle sous un modèle minimal (*i.e.* si les traits phénotypiques interagissent de façon additive). Une telle distribution défavorise *a priori* la recombinaison et la reproduction sexuée. Un autre point important est que la distribution prédite n'est pas irréaliste. En effet, sous les hypothèses du modèle on prédit une relation simple entre la distribution de ε mesurée par paire de mutations et celle de s mesurée pour des mutations uniques, et cette relation semble confirmée par des données empiriques sur le virus *VSV*.

Deux théories classiques de l'évolution du sexe et de la recombinaison peuvent être discutées à partir de ces résultats (Chapitre III). Une première prédit un avantage au sexe, en populations infinies (« théories déterministes »), si l'épistasie ε est en général négative, faible, et peu variable. Cette théorie semble difficile a concilier avec les distributions empiriques de ε (Bonhoeffer et al. 2004 ; Sanjuan et al. 2004a) ou par nos prédictions (Chapitre II, section IV). Une seconde prédit que le sexe est avantageux parce qu'en populations finies, les mutations avantageuses tendent à être associées négativement indépendamment de ε (effet Hill-Robertson qui englobe l'interférence clonale pour les asexués). Le sexe permet alors, via la recombinaison, de réunir ces mutations avantageuses dans un même individu, augmentant alors la réponse à la sélection (« théories stochastiques »). Nos résultats précédents, basés sur les travaux de Gerrish & Lenski (1998), et plusieurs expériences sur les micro-organismes (Goddard, 2005 #1278, pour revue voir, De Visser and Rozen 2005; Kaltz and Bell 2002; Miralles et al. 1999) suggèrent que l'interférence clonale peut en effet constituer une limite forte à l'adaptation chez les asexués.

Toutefois, l'effet Hill-Robertson n'est important a priori que dans des populations de tailles efficaces intermédiaires. J'ai donc étudié la possibilité que cet effet constitue un avantage plus général associé au sexe et à la recombinaison, dans des populations subdivisées, quelle que soit leur taille totale. Pour cela, j'ai utilisé un modèle explicite d'évolution de la recombinaison (modèle avec modifieur) entre deux loci sous sélection directionnelle, dans une population structurée (modèle en île). Le modèle suggère que l'impact du HRE sur l'adaptation (et l'avantage correspondant associé au sexe) dépend plus de la structuration de la population que de sa taille efficace totale ou locale (par dème). Ainsi, le HRE peut être important même dans des populations infinies pourvu qu'elles soient suffisamment structurées, et même dans de très petites populations si elles sont interconnectées par migration. J'ai discuté en Chapitre III, la généralité de ces résultats et de leurs limites. Une limite évidente est que le modèle ne permet pas de faire évoluer de forts taux de recombinaison ou de sexe (qu'on observe pourtant dans de nombreuses espèces). Cette limite est inhérente à d'autres modèles classiques d'évolution du sexe, et il est possible qu'elle ne soit pas si importante lorsqu'on prend en compte des situations plus réalistes (par ex. de nombreux loci sous sélection au lieu de seulement deux, l'effet de la consanguinité chez les diploïdes).

Ces travaux sont une illustration (de plus ?) que le système de reproduction joue de façon importante sur l'adaptation. Par ailleurs, ils suggèrent que le sexe peut être favorisé dans des

conditions assez générales. Un résultat important, il me semble, de ce travail, est qu'il suggère un tri entre les différents mécanismes qui déterminent cet avantage associé au sexe. On s'attend à ce que celui-ci soit particulièrement important lorsque la sélection et la dérive (en populations finies comme structurées) interagissent sur de nombreuses mutations avantageuses (la sélection de fonds ou l'épistasie étant peut-être moins importantes dans cette situation).

Dans la section suivante, je détaille certaines limites de mes résultats sur la distribution de l'effet des mutations et comment on pourrait envisager d'élargir l'approche.

Critiques et perspectives theoriques

La plus grande partie de mes résultats analytiques se focalise sur $f(s)$ dans un unique environnement, fixe dans le temps. Ils ne permettent donc pas d'appréhender l'adaptation dans un milieu hétérogène dans l'espace ou le temps.

Comme nous l'avons vu en Article 2 et Chapitre II section III, notre modèle (et le MF en général) constitue une bonne trame pour traiter des interactions mutationnelles G×E pour la fitness, celles-ci pouvant concerner les mutations avantageuses comme délétères. Toutefois, deux types de G×E ont un impact évolutif particulièrement important : les mutations à effet antagonistes (avantageuses dans un environnement, désavantageuses dans un autre) ou conditionnellement délétères (neutres dans un environnement et délétères dans un autre). En effet, ces deux types d'interactions peuvent jouer un rôle important dans l'évolution de la spécialisation écologique (Fry 1996; Kawecki et al. 1997), bien que l'influence de la dérive puisse aussi avoir un impact (Whitlock 1996). Il serait donc intéressant de caractériser la proportion et la distribution des effets de ces deux types de mutations à partir du modèle. J'ai commencé à aborder ce problème par une question relativement simple : quel est l'effet moyen, dans un environnement e_1, d'une mutation d'effet s dans un environnement e_2 ? Au-delà de l'effet moyen, il serait intéressant de prédire la *distribution* des effets des mutations dans e_1, sachant leur effet dans e_2. Ces questions ont trait au coût de l'adaptation: le fait, souvent observé empiriquement que des mutations avantageuses dans un nouvel environnement ont un effet pléiotrope délétère dans l'environnement d'origine. Ce problème est donc directement relié à la pléiotropie antagoniste. Son traitement analytique est en cours.

Une autre limite de mon approche sur l'effet de l'environnement, est que l'on néglige la plasticité phénotypique. Suivant cette hypothèse, aucun généraliste écologique ne peut évoluer, seul peut évoluer un déplacement de la niche vers un nouvel optimum (chaque génotype code pour un phénotype unique, proche d'un unique optimum, donc il est spécialiste). Comme on l'a vu (Article 2, Chapitre II section II), l'effet net sur $f(s)$, de traits ou gènes à expression environnement-dépendante semble réduite. Nos résultats suggèrent donc que les traits plastiques ont un impact réduit dans *chaque* environnement, pris *séparément*. Toutefois, ils ont potentiellement un impact important sur la corrélation de fitness *entre* environnements. Ce sont ces traits plastiques seuls qui peuvent faire évoluer la largeur de la niche écologique (donc le fait d'être plus ou moins généraliste). En effet, un génotype généraliste correspondrait dans notre modèle à un génotype capable de produire, *selon l'environnement*, le phénotype optimal correspondant à cet environnement. On peut là aussi, intégrer la plasticité dans un modèle de paysage adaptatif comme celui que j'ai proposé. On peut par exemple définir, pour chaque trait et chaque génotype, deux grandeurs au lieu d'une seule (z_i) : une valeur du trait dans un environnement de référence plus une pente de réponse plastique du trait à une variable environnementale donnée, les deux grandeurs pouvant changer par mutation, produisant des génotypes plus ou moins plastiques pour ce trait. Là aussi, je n'ai pas avancé ce travail suffisamment pour le moment, mais on voit qu'il est possible de modéliser l'évolution d'un généraliste dans un modèle de paysages adaptatif.

Enfin, plusieurs autres types de sélection ne sont pas pris en compte par les approches de type paysages adaptatifs. D'abord, comme nous l'avons vu en Introduction, le polymorphisme au sein d'une population est souvent négligé: il est supposé que les populations se résument à un point dans l'espace des phénotypes. Ainsi, l'adaptation se produit par la fixation successive de mutations avantageuses, dont l'effet est fort relativement au polymorphisme à chaque génération. Cette simplification est utile mais elle ne permet pas de modéliser la sélection fréquence-dépendante. De même elle ne permet pas de modéliser l'adaptation à partir de variation pré-existante. La génétique quantitative permet une modélisation de ce type de situation, toutefois, comme nous l'avons vu en Introduction, il peut être plus difficile de l'appliquer à l'évolution de la fitness (par opposition à un trait morphologique par ex.), car la distribution des fitness est rarement gaussienne (voir l'Annexe 1, sur la possibilité de prendre en compte des distributions non-gaussiennes des traits d'adaptation). Enfin, les modèles présentés ici considèrent un environnement fixe dans le temps (*i.e.* un optimum fixe). On sait que la sélection peut varier (par exemple lors de processus coévolutifs), et il serait intéressant de pouvoir modéliser, voire prédire, des taux d'adaptation dans ce contexte. Il existe des approches pour prendre en compte un optimum « mobile » dans des modèles de paysages adaptatifs (voir par ex. Jones et al. 2004), qui pourraient permettre de développer de tels modèles plus généraux.

En résumé, je pense que l'approche à partir de paysages adaptatifs présente de nombreux potentiels au-delà de ce qui est présenté dans cette thèse. On a vu qu'elle permet de faire des prédictions testables sur diverses variables centrales en biologie évolutive (on a vu ici la distribution de l'effet des mutations et de l'épistasie). Il serait donc intéressant de tenter d'élargir son domaine d'applications à des situations plus complexe que les modèles d'adaptations minimaux présentés dans cette thèse. Par ailleurs, je pense que l'utilisation des matrices aléatoires comme un modèle nul de covariances phénotypiques (**P**-matrix) ou génétiques (**G**-matrix) pourrait se révéler utile en génétique quantitative, notamment pour caractériser la variation et l'évolution de ces matrices (pour une revue sur cette question, voir Steppan et al. 2002).

De même qu'il serait intéressant de modéliser des situations d'adaptation plus complexes que celles que j'ai traitées dans cette thèse, il serait utile de pouvoir étudier empiriquement de telles situations, *c.a.d.* de pouvoir étudier l'adaptation dans des milieux plus complexes, *in natura*. Je finis donc par présenter comment de telles études pourraient être entreprises avec *Artemia*.

Etude de l'adaptation *in natura* avec *Artemia*

Comme nous l'avons vu en Introduction, la plupart des études de l'adaptation à long terme ont été faites sur des populations de laboratoire, souvent des micro-organismes, et dans des environnement contrôlés. Bien que ce n'en était pas le but premier, mon travail sur *Artemia* m'a amené à considérer que ce modèle biologique présente de nombreux atouts pour l'étude de l'adaptation à long terme *in natura*. Je conclurai donc en détaillant les particularités qui font d'*Artemia* un bon modèle pour ce genre d'étude.

Accès aux générations passées : *Artemia* présente la particularité de produire des œufs de diapause (cystes) qui peuvent conserver un pouvoir d'éclosion pendant de nombreuses années (plus de 20 ans !). L'Artemia Reference Center (ARC) maintient une collection de cystes récoltés dans divers sites partout dans le monde et à diverses époques. Par ailleurs, nous avons mis au point une méthode pour avoir accès aux générations passées par des échantillonnages sur le terrain. Les cystes flottent à la surface de l'eau et sont poussés par le vent de sorte qu'ils s'accumulent sur les plages des salines. Si le milieu n'est pas perturbé, ils peuvent alors

sédimenter au cours des années. Ainsi, en faisant un carottage du sédiment et en extrayant les cystes contenus à différentes profondeurs, on a accès à une série temporelles des cystes pondus dans la saline. Etant donné que l'on peut extraire de l'ADN de ces cystes, on a donc a priori accès aux génotypes passés (il existe des protocoles publiés d'analyse par AFLP et RFLP Bossier et al. 2004; Sun et al. 1999). Par ailleurs, il est possible d'éclore des cystes anciens et ainsi de mesurer directement des traits phénotypiques sur des individus vivants, pondus il y a plus de 20 ans (soit de l'ordre de 200 générations). Le taux d'éclosion de cystes anciens est faible et dépend des conditions de conservation, mais on dispose en général de grandes quantités de cystes et on peut donc obtenir plus de 100 individus « vieux » de 10 à 20 ans. Ceci est aussi valable pour des cystes contenus dans des sédiments, auquel cas on a potentiellement accès à une séries temporelle sur plusieurs centaines de générations. J'ai pu obtenir des éclosions pour des cystes probablement vieux de plus de 10 ans, et F. Amat a obtenu des éclosions pour des cystes plus anciens encore. On peut donc étudier, par exemple, l'évolution de la niche écologique *in natura*, sur plusieurs centaines de générations, ce qui est rarement possible avec des organismes supérieurs. On peut par ailleurs dater les sédiments (éléments radioactifs) pour obtenir une calibration (approximative toutefois) en nombre d'années puis en nombre de générations.

Milieux présents et passés bien connus : Pour nombre des milieux de vie d'*Artemia*, les paramètres physiques du milieu sont régulièrement mesurés (salinité, T° etc.), que ce soient des milieux anthropisés (marais salants) ou naturels (par ex. Great Salt Lake, Utah). Ainsi, on peut avoir accès aux caractéristiques du milieu abiotique passé et présent, et à leurs variations spatiales et temporelles, notamment pour les deux variables majeures de la niche d'*Artemia* : la salinité et la température.

Bases génétiques de l'adaptation connues : Plusieurs adaptations clé d'*Artemia* ont été bien caractérisées au niveau moléculaire (pour revue voir Marco et al. 2000). C'est le cas, par exemple des protéines heat-shock impliquées dans la tolérance aux forts changements de températures survenant dans les salines (Crack and MacRae 1999; Liang and MacRae 1999). C'est le cas aussi de la pompe Na-K ATPase, impliquée dans l'osmorégulation, et dont la séquence est également connue (voir aussi une étude récente du polymorphisme de cette séquence chez *A. franciscana* et *A. parthenogenetica* Saez et al. 2000). On peut donc envisager, à partir de cystes anciens d'étudier l'évolution moléculaire de séquences déterminant partiellement la niche pour la salinité ou la température. Par ailleurs la séquence mitochondriale d'*Artemia franciscana* a été publiée (Garesse et al. 1997).

Suivi démographique: En plus de ces particularités, le fait qu'*Artemia* soit une ressource importante en aquaculture fait que certaines populations (par ex. Great Salt Lake, Utah ou les salins d'Aigues-Mortes en France) font l'objet de suivis précis depuis de nombreuses années, on peut notamment avoir accès au rendement des pêches de cystes ou d'individus adultes ce qui donne une estimation des variations démographiques des populations.

Des expériences in natura : Enfin, un autre intérêt de l'utilisation d'*Artemia* en aquaculture est que de nombreuses introductions ont été réalisées par l'Homme dans des milieux divers. Ces changement brusques d'environnement sont autant d'expériences grandeur nature, avec un changement du milieu dont on connaît l'époque lorsque c'est l'Homme qui l'a provoqué. Je citerai deux exemples qui me semblent intéressants (parmi d'autres).

Aigues-Mortes: colonisation d'une population asexuée par une population sexuée. La côte méditerranéenne est le siège, depuis le début des années 70, d'une invasion par l'espèce sexuée d'Amérique du Nord (*A. franciscana* pour revue voir Amat et al. 2005; Green et al. 2005). Aux Salins d'Aigues-Mortes, la population endémique est asexuée diploïde et, malgré le succès invasif d'*A. franciscana* (souche San Francisco Bay), elle est toujours présente dans

les Salins. Pour étudier l'évolution de ces deux espèces en compétition, nous avons fait des carottages des sédiments de plusieurs bassins des Salins d'Aigues-Mortes (dans des zones sédimentaires non perturbées). Nous avons extrait les cystes des différentes couches (époques) et F. Amat a obtenu des éclosions de cystes issus de couches à -20 cm de profondeur (correspondant en première approximation à 20 ans d'âge). A partir de ces cystes, il est possible d'étudier l'évolution de la niche pour la salinité et la température dans les deux espèces. Par ailleurs, il est possible d'estimer la proportion de chaque espèce dans un lot de cystes d'une époque donnée. En effet, ces deux espèces sont caractérisées par des diamètres des cystes nettement différents (Vanhaecke and Sorgeloos 1980). Ainsi, la distribution des diamètres des cystes dans les différents échantillons donne une estimation indirecte de la proportion de chaque espèce, et de sa variation avec le temps, donc de la dynamique temporelle de l'invasion. Le travail réalisé par Luis Miguel CHEVIN en stage au laboratoire a montré que la proportion d'*A. franciscana* a augmenté fortement dans les premières années après l'introduction puis s'est stabilisée (avec des fluctuations cycliques). Il semblerait donc que l'état soit maintenant stationnaire. Les questions principales auxquelles on peut répondre expérimentalement sont donc: y a-t-il eu séparation des niches écologiques permettant ainsi une coexistence des deux espèces ? et comment a varié la vitesse d'évolution chez les sexués et les asexués ?

Lavalduc : Le saline de Lavalduc (près d'Istres, France) est un étang fermé qui a subi un augmentation brutale de salinité (due à un changement d'utilisation par les Salins du Midi), passant d'un niveau assez stable autour $120g.l^{-1}$ à quasi-saturation en près de 2 ans (1977-1979). C'est donc une situation idéale pour étudier un changement brusque de l'environnement. La population d'*Artemia* (parthénogénétiques polyploïdes), qui était très importante jusqu'à la fin des années 70 à fortement décliné après cette augmentation de salinité. Toutefois, une population se maintient, avec une reproduction très irrégulière, dépendant a priori des pluies (D. Facca, *com .pers.* + observations de terrain). Ainsi, à chaque pluie d'orage, les cystes présents éclosent dans la zone superficielle d'eau peu salée. Ces individus se développent jusqu'à maturité et laissent une nouvelle génération de cystes, puis meurent au fur et à mesure que la couche d'eau superficielle se sature à nouveau en sel par mélange avec les couches sous-jacentes. Un tel fonctionnement a permis le maintien de la population dans des conditions extrêmes pendant plus de 35 ans. D. Facca m'a fourni une grande quantité de cystes de 1978 (une des dernières récoltes dans cette saline) correspondant environ à l'époque du changement de salinité. Par ailleurs, nous avons pu échantillonner une faible quantité de cystes actuels. Il serait intéressant de refaire un échantillonnage puis de comparer les populations de 1978 aux actuelles, pour quantifier l'évolution de la niche pour la salinité, de la tendance à la reproduction ovipare (cystes), des trait d'histoire de vie. On pourrait également étudier l'adaptation au niveau moléculaire en séquençant la pompe Na/K ATPase dans des individus des deux époques.

On pourrait aussi citer d'autres « expériences » *in natura* comme diverses introductions depuis des milieux tempérés (Great salt lake, San Francisco Bay) vers des salines situées dans des zones tropicales aux températures beaucoup plus stables comme le Vietnam (Clegg et al. 2000), ou le Brésil (Camara 2001). Ces introductions constituent une occasion unique d'étudier l'évolution de la spécialisation écologique pour la température.

Bien que l'ensemble de ces points ne soient que des perspectives, j'espère avoir montré qu'*Artemia* présente un ensemble de particularités (par sa biologie, son écologie, mais aussi par sa situation de ressource aquacole) qui en font un modèle biologique unique (et sous – exploité) pour l'étude de l'adaptation *in natura*. Les protocoles d'élevage et de mesures de traits d'histoire de vie au laboratoire (voir Annexe 2) pourraient donc être complétées par des analyses de terrain et permettre l'étude de l'adaptation et de l'évolution des traits d'histoire de vie chez *Artemia*, en fonction de l'environnement et du système de reproduction.

BIBLIOGRAPHIE

Abatzopoulos, T. J., J. A. Beardmore, J. S. Clegg, and P. Sorgeloos. 2002. Artemia: basic and applied biology. Springer verlag

Abatzopoulos, T. J., B. Zhang, and P. Sorgeloos. 1998. *Artemia tibetiana*: preliminary characterization of a new *Artemia* species found in Tibet (People's Republic of China). International Study on *Artemia*. LIX. International Journal of Salt Lake Research 7:41-44.

Abramoff, M. D., P. J. Magelhaes, and S. J. Ram. 2004. Image Processing with ImageJ. Biophotonics International 11:36-42.

Abreu-Grosbois, F. A. 1987. A review of the genetics of *Artemia*. Pp. 61-99 *in* P. Sorgeloos, D. A. Bengston, W. Decleir and E. Jaspers, eds. Artemia research and its applications. Universa Press, Wetteren, Belgium.

Amat, F., F. Hontoria, O. Ruiz, A. J. Green, F. Hortas, and J. Figuerola. 2005. The American brine shrimp as an exotic invasive species in the western Mediterranean. Biological Invasions 7:37-47.

Arnold, S. J., M. E. Pfrender, and A. G. Jones. 2001. The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica 112:9-32.

Bai, Z. D. 1999. Methodologies in spectral analysis of large dimensional random matrices, a review. Statistica Sinica 9:611-662.

Barata, C., F. Hontoria, F. Amat, and R. Browne. 1996. Demographic parameters of sexual and parthenogenetic Artemia: Temperature and strain effects. Journal of Experimental Marine Biology and Ecology 196:329-340.

Barigozzi, C. 1974. Artemia: a survey of its significance in genetic problems. Pp. 221-252 *in* T. Dobzhansky, ed. Evolutionary biology.

Barton, N., and S. P. Otto. 2004. Evolution of recombination due to random drift. in prep

Barton, N., and L. Partridge. 2000. Limits to natural selection. Bioessays 22:1075-1084.

Barton, N. H. 1995a. A general model for the evolution of recombination. Genetical Research 65:123-144.

Barton, N. H. 1995b. Linkage and the limits to natural selection. Genetics 140:821-841.

Barton, N. H., and B. Charlesworth. 1998. Why sex and recombination? Science 281:1986-1990.

Barton, N. H., and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nature Reviews Genetics 3:11-21.

Barton, N. H., and M. Turelli. 1991. Natural and Sexual Selection on Many Loci. Genetics 127:229-255.

Barton, N. H., and M. Turelli. 2004. Effects of genetic drift on variance components under a general model of epistasis. Evolution 58:2111-2132.

Barton, N. H., and M. C. Whitlock. 1997. The evolution of metapopulations. Pp. 183-210 *in* H. I.A. and G. M.E., eds. Metapopulation Biology. Academic Press.

Bataillon, T. 2000. Estimation of spontaneous genome-wide mutation rate parameters: whither beneficial mutations? Heredity 84:497-501.

Bataillon, T. 2003. Shaking the 'deleterious mutations' dogma? Trends in Ecology & Evolution 18:315-317.

Bateman, A. J. 1959. The Viability of near-Normal Irradiated Chromosomes. International Journal of Radiation Biology and Related Studies in Physics Chemistry and Medicine 1:170-180.

Beaumont, M. A. 2001. Conservation genetics. Pp. 779-812 *in* D. J. Balding, M. Bishop and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons.

Bjedov, I., O. Tenaillon, B. Gerard, V. Souza, E. Denamur, M. Radman, F. Taddei, and I. Matic. 2003. Stress-induced mutagenesis in bacteria. Science 300:1404-1409.

Bonhoeffer, S., C. Chappey, N. T. Parkin, J. M. Whitcomb, and C. J. Petropoulos. 2004. Evidence for positive epistasis in HIV-1. Science 306:1547-1550.

Bossier, P., X. M. Wang, F. Catania, S. Dooms, G. Van Stappen, E. Naessens, and P. Sorgeloos. 2004. An RFLP database for authentication of commercial cyst samples of the brine shrimp Artemia spp. (International Study on Artemia LXX). Aquaculture 231:93-112.

Browne, R. A. 1980. Reproductive Pattern and Mode in the Brine Shrimp. Ecology 61:466-470.

Browne, R. A. 1992. Population-Genetics and Ecology of Artemia - Insights Into Parthogenetic Reproduction. Trends in Ecology & Evolution 7:232-237.

Browne, R. A., and C. W. Hoopes. 1990. Genotype diversity and selection in the Brine Shrimp (*Artemia)*. Evolution 44:1035-1051.

Browne, R. A., and S. E. Sallee. 1984. Partitioning genetic and environmental components of reproduction and lifespan in Artemia. Ecology 65:949-960.

Browne, R. A., P. Sorgeloos, and C. N. Trotman. 2000. Artemia Biology. CRC Press

Camara, M. R. 2001. Dispersal of Artemia franciscana Kellogg (Crustacea; Anostraca) populations in the coastal saltworks of Rio Grande do Norte, northeastern Brazil. Hydrobiologia 466:145-148.

Champion, C. J. 2003. Empirical Bayesian estimation of normal variances and covariances. Journal of Multivariate Analysis 87:60-79.

Charlesworth, B., and D. Charlesworth. 1998. Some evolutionary consequences of deleterious mutations. Genetica 103:3-19.
Charlesworth, D., M. T. Morgan, and B. Charlesworth. 1993. Mutation accumulation in finite outbreeding and inbreeding populations. Genetical Research, Cambridge 61:39-56.
Clarke, B., and W. Arthur. 2000. What constitutes a 'large' mutational change in phenotype? Evolution & Development 2:238-240.
Clegg, J. S., S. A. Jackson, N. V. Hoa, and P. Sorgeloos. 2000. Thermal resistance, developmental rate and heat shock proteins in Artemia franciscana, from San Francisco Bay and southern Vietnam. Journal of Experimental Marine Biology and Ecology 252:85-96.
Colegrave, N. 2002. Sex releases the speed limit on evolution. Nature 420:664-666.
Coutteau, P., L. Brendonck, P. Lavens, and P. Sorgeloos. 1992. The use of manipulated baker's yeast as an algal substitute for the laboratory culture of Anostraca. Hydrobiologia 234:25-32.
Coutteau, P., P. Lavens, and P. Sorgeloos. 1990. Baker's yeast as a potential substitute for live algae in aquaculture diets: *Artemia* as a case study. Journal of the world aquaculture society 21:1-9.
Crack, J. A., and T. H. MacRae. 1999. Molecular analysis of p26, a small heat shock/alpha-crystallin protein from artemia. Faseb Journal 13:A1399-A1399.
Darwin, C. 1859. On the origin of species by means of natural selection. John Murray, London.
De Stasio, E. A., and S. Dorman. 2001. Optimization of ENU mutagenesis of *Caenorhabditis elegans*. Mutation Research-Genetic Toxicology and Environmental Mutagenesis 495:81-88.
de Visser, J. 2002. The fate of microbial mutators. Microbiology-Sgm 148:1247-1252.
de Visser, J., and R. F. Hoekstra. 1998. Synergistic epistasis between loci affecting fitness: evidence in plants and fungi. Genetical Research 71:39-49.
De Visser, J., and D. E. Rozen. 2005. Limits to adaptation in asexual populations. Journal of Evolutionary Biology 18:779-788.
de Visser, J., C. W. Zeyl, P. J. Gerrish, J. L. Blanchard, and R. E. Lenski. 1999. Diminishing returns from mutation supply rate in asexual populations. Science 283:404-406.
Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. Genetics 148:1667-1686.
Elena, S. F., L. Ekunwe, N. Hajela, S. A. Oden, and R. E. Lenski. 1998. Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica 103:349-358.
Elena, S. F., and R. E. Lenski. 1997. Test of synergistic interactions among deleterious mutations in bacteria. Nature 390:395-398.
Elena, S. F., and R. E. Lenski. 2003. Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. Nature Reviews Genetics 4:457-469.
Feldman, M. W. 1972. Selection for linkage modification: I. Random mating populations. Theoretical Population Biology 3:324-346.
Felsenstein, J. 1974. The evolutionary advantage of recombination. Genetics 78:737-756.
Fisher, R. A. 1930. The genetical theory of natural selection. Oxford University Press, Oxford.
Forrester, P. J., N. C. Snaith, and J. J. M. Verbaarschot. 2003. Developments in random matrix theory. Journal of Physics a-Mathematical and General 36:R1-R10.
Frankham, R. 2005. Stress and adaptation in conservation genetics. Journal of Evolutionary Biology 18:750-755.
Fry, J. D. 1996. The evolution of host specialization: Are trade-offs overrated? American Naturalist 148:S84-S107.
Fry, J. D., and S. L. Heinsohn. 2002. Environment dependence of mutational parameters for viability in *Drosophila melanogaster*. Genetics 161:1155-1167.
Garcia-Dorado, A., C. Lopez-Fanjul, and A. Caballero. 1999. Properties of spontaneous mutations affecting quantitative traits. Genetical Research 74:341-350.
Garcia-Dorado, A., and J. M. Marin. 1998. Minimum distance estimation of mutational parameters for quantitative traits. Biometrics 54:1097-1114.
Garesse, R., J. A. Carrodeguas, J. Santiago, M. L. Perez, R. Marco, and C. G. Vallejo. 1997. Artemia mitochondrial genome: Molecular biology and evolutive considerations. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology 117:357-366.
Gerrish, P. 2001. The rhythm of microbial adaptation. Nature 413:299-302.
Gerrish, P. J., and R. E. Lenski. 1998. The fate of competing beneficial mutations in an asexual population. Genetica 103:127-144.
Gillespie, J. H. 1984. Molecular Evolution over the Mutational Landscape. Evolution 38:1116-1129.
Gilligan, D. M., and R. Frankham. 2003. Dynamics of genetic adaptation to captivity. Conservation Genetics 4:189-197.
Gingerich, P. D. 2001. Rates of evolution on the time scale of the evolutionary process. Genetica 112:127-144.
Goddard, M. R., H. Charles, J. Godfray, and A. Burt. 2005. Sex increases the efficacy of natural selection in experimental yeast populations. Nature 434:636-640.
Goho, S., and G. Bell. 2000. Mild environmental stress elicits mutations affecting fitness in Chlamydomonas. Proceedings of the Royal Society of London Series B-Biological Sciences 267:123-129.

Green, A. J., M. I. Sanchez, F. Amat, J. Figuerola, F. Hontoria, O. Ruiz, and F. Hortas. 2005. Dispersal of invasive and native brine shrimps Artemia (Anostraca) via waterbirds. Limnology and Oceanography 50:737-742.
Gupta, A. K., and W. J. Huang. 2002. Quadratic forms in skew normal variates. Journal of Mathematical Analysis and Applications 273:558-564.
Halligan, D. L., A. D. Peters, and P. D. Keightley. 2003. Estimating numbers of EMS-induced mutations affecting life history traits in *Caenorhabditis elegans* in crosses between inbred sublines. Genetical Research 82:191-205.
Hendry, A. P., and M. T. Kinnison. 2001. An introduction to microevolution: rate, pattern, process. Genetica 112:1-8.
Hereford, J., T. F. Hansen, and D. Houle. 2004. Comparing strengths of directional selection: How strong is strong? Evolution 58:2133-2143.
Hill, W. G., and A. Robertson. 1966. The effect of linkage on the limits to artificial selection. Genet. Res. 8:269-294.
Ihaka, I., and G. Robert. 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5:299-314.
Iles, M. M., K. Walters, and C. Cannings. 2003. Recombination can evolve in large finite populations given selection on sufficient loci. Genetics 165:2249-2258.
Imhof, M., and C. Schlotterer. 2001. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. Proceedings of the National Academy of Sciences of the United States of America 98:1113-1117.
Jaschke, S. 2002. The Cornish Fisher expansion in the context of the delta-gamma-normal approximations. Journal of Risk 4:33-52.
Jaschke, S., C. Kluppelberg, and A. Lindner. 2004. Asymptotic behavior of tails and quantiles of quadratic forms of Gaussian vectors. Journal of Multivariate Analysis 88:252-273.
Johnson, T., and N. H. Barton. 2002. The effect of deleterious alleles on adaptation in asexual populations. Genetics 162:395-411.
Jones, A. G., S. J. Arnold, and R. Burger. 2004. Evolution and stability of the G-matrix on a landscape with a moving optimum. Evolution 58:1639-1654.
Joseph, S. B., and D. W. Hall. 2004. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: More beneficial than expected. Genetics 168:1817-1825.
Kaltz, O., and G. Bell. 2002. The ecology and genetics of fitness in Chlamydomonas. XII. Repeated sexual episodes increase rates of adaptation to novel environments. Evolution 56:1743-1753.
Kassen, R., and T. Bataillon. 2005. The distribution of fitness effects among beneficial mutations prior to selection in experimental populations of bacteria. In review
Kauffman, S., and S. Levin. 1987. Towards a General-Theory of Adaptive Walks on Rugged Landscapes. Journal of Theoretical Biology 128:11-45.
Kawecki, T. J., N. H. Barton, and J. D. Fry. 1997. Mutational collapse of fitness in marginal habitats and the evolution of ecological specialisation. Journal of Evolutionary Biology 10:407-429.
Keightley, P. D. 1994. The Distribution of Mutation Effects On Viability in *Drosophila melanogaster*. Genetics 138:1315-1322.
Keightley, P. D. 2004. Comparing analysis methods for mutation-accumulation data. Genetics 167:551-553.
Keightley, P. D., E. K. Davies, A. D. Peters, and R. G. Shaw. 2000. Properties of ethylmethane sulfonate-induced mutations affecting life-history traits in *Caenorhabditis elegans* and inferences about bivariate distributions of mutation effects. Genetics 156:143-154.
Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science 290:331-333.
Keightley, P. D., and M. Lynch. 2003. Toward a realistic model of mutations affecting fitness. Evolution 57:683-685.
Keightley, P. D., and O. Ohnishi. 1998. EMS-induced polygenic mutation rates for nine quantitative characters in *Drosophila melanogaster*. Genetics 148:753-766.
Kibota, T. T., and M. Lynch. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E-coli*. Nature 381:694-696.
Kidwell, M. G., and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. Proceedings of the National Academy of Sciences of the United States of America 94:7704-7711.
Kimura, M. 1979. Model of Effectively Neutral Mutations in Which Selective Constraint Is Incorporated. Proceedings of the National Academy of Sciences of the United States of America 76:3440-3444.
Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.
Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. American Naturalist 157:245-261.
Kinnison, M. T., and A. P. Hendry. 2001. The pace of modern life II: from rates of contemporary microevolution to pattern and process. Genetica 112:145-164.

Kishony, R., and S. Leibler. 2003. Environmental stresses can alleviate the average deleterious effects of mutations. Journal of Biology 2:14.
Kivisaar, M. 2003. Stationary phase mutagenesis: mechanisms that accelerate adaptation of microbial populations under environmental stress. Environmental Microbiology 5:814-827.
Kondrashov, A. S. 1993. Classification of hypotheses on the advantage of amphimixis. Journal of Heredity 84:372-387.
Korona, R. 2004. Experimental studies of deleterious mutation in *Saccharomyces cerevisiae*. Research in Microbiology 155:301-310.
Kuonen, D. 1999. Saddlepoint approximations for distributions of quadratic forms in normal variables. Biometrika 86:929-935.
Lande, R. 1979. Quantitative Genetic-Analysis of Multivariate Evolution, Applied to Brain - Body Size Allometry. Evolution 33:402-416.
Lande, R. 1980. The Genetic Covariance between Characters Maintained by Pleiotropic Mutations. Genetics 94:203-215.
Latter, B. D. H., and J. C. Mulley. 1995. Genetic Adaptation to Captivity and Inbreeding Depression in Small Laboratory Populations of *Drosophila melanogaster*. Genetics 139:255-266.
Lenormand, T., D. Bourguet, T. Guillemaud, and M. Raymond. 1999. Tracking the evolution of insecticide resistance in the mosquito Culex pipiens. Nature 400:861-4.
Lenormand, T., and S. P. Otto. 2000. The evolution of recombination in a heterogeneous environment. Genetics 156:423-38.
Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-Term Experimental Evolution in Escherichia-Coli .1. Adaptation and Divergence During 2,000 Generations. American Naturalist 138:1315-1341.
Lenski, R. E., and M. Travisano. 1994. Dynamics of Adaptation and Diversification - a 10,000-Generation Experiment with Bacterial-Populations. Proceedings of the National Academy of Sciences of the United States of America 91:6808-6814.
Lenz, P. H., and R. A. Browne. 2000. Ecology of *Artemia*. Pp. 237-253 *in* R. A. Browne, P. Sorgeloos and C. N. Trotman, eds. Artemia Biology. CRC Press.
Liang, P., and T. H. MacRae. 1999. The synthesis of a small heat shock/alpha-crystallin protein in Artemia and its relationship to stress tolerance during development. Developmental Biology 207:445-456.
Lively, C. M., and A. D. Peeters. 2000. Epistasis and the maintenance of sex. Pp. 99-112 *in* J. B. Wolf, E. D. Brodie and M. J. Wade, eds. Epistasis and the evolutionary process. Oxford University press.
Loewe, L., V. Textor, and S. Scherer. 2003. High deleterious genomic mutation rate in stationary phase of *Escherichia coli*. Science 302:1558-1560.
Lyman, R. F., F. Lawrence, S. V. Nuzhdin, and T. F. C. Mackay. 1996. Effects of single P-element insertions on bristle number and viability in *Drosophila melanogaster*. Genetics 143:277-292.
Lynch, M., J. Blanchard, D. Houle, T. Kibota, S. Schultz, L. Vassilieva, and J. Willis. 1999. Perspective: Spontaneous deleterious mutation. Evolution 53:645-663.
Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. Science 302:1401-1404.
Marco, R., R. Garesse, J. Cruces, and J. Renart. 2000. *Artemia* molecular genetics. Pp. 1-21 *in* R. A. Browne, P. Sorgeloos and C. N. Trotman, eds. Artemia Biology. CRC Press.
Maruyama, T. 1974. A simple proof that certain quantities are independent of the geographical structure of population. Theoretical Population Biology 5:148-154.
Massey, R. C., and A. Buckling. 2002. Environmental regulation of mutation rates at specific sites. Trends in Microbiology 10:580-584.
Mathai, A. M., and S. B. Provost. 1992. Quadratic forms in random variables. Marcel Dekker, New York.
Matos, M., T. Avelar, and M. R. Rose. 2002. Variation in the rate of convergent evolution: adaptation to a laboratory environment in Drosophila subobscura. Journal of Evolutionary Biology 15:673-682.
Maynard Smith, J. 1971. What use is sex? Journal of theoretical Biology 30:319-335.
Metalli, P., and E. Ballardin. 1970. Radiobiology of artemia: radiation effects and ploidy. Current topics in radiation research quarterly 7:181-240.
Miralles, R., P. J. Gerrish, A. Moya, and S. F. Elena. 1999. Clonal interference and the evolution of RNA viruses. Science 285:1745-1747.
Miralles, R., A. Moya, and S. F. Elena. 2000. Diminishing returns of population size in the rate of RNA virus adaptation. Journal of Virology 74:3566-3571.
Mukai, T. 1964. Genetic Structure of Natural Populations of *Drosophila Melanogaster* .1. Spontaneous Mutation Rate of Polygenes Controlling Viability. Genetics 50:1-&.
Mukai, T., S. I. Chigusa, L. E. Mettler, and J. F. Crow. 1972. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. Genetics 72:335-355.
Mullins, M. C., M. Hammerschmidt, P. Haffter, and C. Nussleinvolhard. 1994. Large-Scale Mutagenesis in the Zebrafish - in Search of Genes-Controlling Development in a Vertebrate. Current Biology 4:189-202.

Novella, I. S., E. A. Duarte, S. F. Elena, A. Moya, E. Domingo, and J. J. Holland. 1995. Exponential Increases of Rna Virus Fitness During Large Population Transmissions. Proceedings of the National Academy of Sciences of the United States of America 92:5841-5844.
Orr, H. A. 1998. The population genetics of adaptation: The distribution
of factors fixed during adaptive evolution. Evolution 52:935-949.
Orr, H. A. 1999. The evolutionary genetics of adaptation: a simulation
study. Genetical Research 74:207-214.
Orr, H. A. 2000a. Adaptation and the cost of complexity. Evolution 54:13-20.
Orr, H. A. 2000b. The rate of adaptation in asexuals. Genetics 155:961-968.
Orr, H. A. 2001. The "sizes" of mutations fixed in phenotypic evolution: a response to Clarke and Arthur. Evolution & Development 3:121-123.
Orr, H. A. 2002. The population genetics of adaptation: The adaptation of DNA sequences. Evolution 56:1317-1330.
Orr, H. A. 2003. The distribution of fitness effects among beneficial mutations. Genetics 163:1519-1526.
Orr, H. A. 2005a. The genetic theory of adaptation: A brief history. Nature Reviews Genetics 6:119-127.
Orr, H. A. 2005b. Theories of adaptation: what they do and don't say. Genetica 123:3-13.
Otto, S., and N. Barton. 2001. Selection for recombination in small populations. Evolution 55:1921-1931.
Otto, S. P. 2003. The advantages of segregation and the evolution of sex. Genetics 164:1099-1118.
Otto, S. P., and M. W. Feldman. 1997. Deleterious mutations, variable epistatic interactions and the evolution of recombination. Theoretical Population Biology 51:134-147.
Otto, S. P., and T. Lenormand. 2002. Resolving the paradox of sex and recombination. Nature Genetics 3:252-261.
Otto, S. P., and M. C. Whitlock. 1997. The probability of fixation in populations of changing size. Genetics 146:723-733.
Otto, S. P., and P. Yong. 2002. The evolution of gene duplicates. Pp. 451-483. Homology Effects.
Parsons, P. A. 1991. Evolutionary Rates - Stress and Species Boundaries. Annual Review of Ecology and Systematics 22:1-18.
Peck, J. R. 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137:597-606.
Peters, A. D., and C. M. Lively. 1999. The red queen and fluctuating epistasis: a population genetic analysis of antagonistic coevolution. American Naturalist 154:393-405.
Phillips, P., S. P. Otto, and M. C. Whitlock. 2000. Beyond the average. Pp. 20-38 *in* J. B. Wolf, E. D. Brodie and M. J. Wade, eds. Epistasis and the evolutionary process. Oxford University press.
Remold, S. K., and R. E. Lenski. 2001. Contribution of individual random mutations to genotype-by-environment interactions in *Echerichia coli*. Proceedings of the National Academy of Science 98:11388-11393.
Reznick, D. N., and C. K. Ghalambor. 2001. The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. Genetica 112:183-198.
Rice, W. R. 2002. Experimental tests of the adaptive significance of sexual recombination. Nature Reviews Genetics 3:241-251.
Rokyta, D. R., P. Joyce, S. B. Caudle, and H. A. Wichman. 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nature Genetics 37:441-444.
Rousset, F. 2001. Inferences from spatial population genetics. Pp. 239-270 *in* D. J. Balding, M. Bishop and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons.
Roze, D., and T. Lenormand. 2005. Self-fertilization and the evolution of recombination. Genetics 170:841-857.
Rozen, D. E., J. de Visser, and P. J. Gerrish. 2002. Fitness effects of fixed beneficial mutations in microbial populations. Current Biology 12:1040-1045.
Saez, A. G., R. Escalante, and L. Sastre. 2000. High DNA sequence variability at the alpha 1 Na/K-ATPase locus of Artemia franciscana (brine shrimp): Polymorphism in a gene for salt-resistance in a salt-resistant organism. Molecular Biology and Evolution 17:235-250.
Sanjuan, R., A. Moya, and S. F. Elena. 2004a. The contribution of epistasis to the architecture of fitness in an RNA virus. Proceedings of the National Academy of Sciences of the United States of America 101:15376-15379.
Sanjuan, R., A. Moya, and S. F. Elena. 2004b. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences of the United States of America 101:8396-8401.
Schneider, D., and R. E. Lenski. 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Research in Microbiology 155:319-327.
Schultz, S. T., M. Lynch, and J. H. Willis. 1999. Spontaneous deleterious mutation in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America 96:11393-11398.
Shaver, A. C., P. G. Dombrowski, J. Y. Sweeney, T. Treis, R. M. Zappala, and P. D. Sniegowski. 2002. Fitness evolution and the rise of mutator alleles in experimental *Escherichia coli* populations. Genetics 162:557-566.

Shaw, F. H., C. J. Geyer, and R. G. Shaw. 2002. A comprehensive model of mutations affecting fitness and inferences for *Arabidopsis thaliana*. Evolution 56:453-463.
Shaw, R. G., D. L. Byers, and E. Darmo. 2000. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. Genetics 155:369-378.
Silverstein, J. W. 1999. Methodologies in spectral analysis of large dimensional random matrices, a review - Comment: Complements and new developments. Statistica Sinica 9:667-671.
Sniegowski, P. 2004. Evolution: Bacterial mutation in stationary phase. Current Biology 14:R245-R246.
Squire, R. D. 1970. Effects of Acute Gamma Irradiation On Brine Shrimp, Artemia .2. Female Reproductive Performance. Biological Bulletin 139:375-&.
Squire, R. D. 1973. Effects of Acute Gamma-Irradiation On Brine Shrimp, Artemia .3. Male F1 Reproductive Performance Following Paternal Irradiation of Mature Sperm. Biological Bulletin 144:192-199.
Squire, R. D., and D. S. Grosch. 1970. Effects of Acute Gamma Irradiation On Brine Shrimp, Artemia .1. Life Spans and Male Reproductive Performance. Biological Bulletin 139:363-&.
Steppan, S. J., P. C. Phillips, and D. Houle. 2002. Comparative quantitative genetics: evolution of the G matrix. Trends in Ecology and Evolution 17:320-327.
Sun, Y., W.-Q. Song, Y.-C. Zhong, R.-S. Zhang, T. J. Abatzopoulos, and R.-y. Chen. 1999. Diversity and genetic differentiation in *Artemia* species and populations detected by AFLP markers. International Journal of Salt Lake Research 8:341-350.
Szathmary, E. 1993. Do Deleterious Mutations Act Synergistically - Metabolic Control-Theory Provides a Partial Answer. Genetics 133:127-132.
Tachida, H. 1991. A Study on a Nearly Neutral Mutation Model in Finite Populations. Genetics 128:183-192.
Turelli, M. 1985. Effects of Pleiotropy on Predictions Concerning Mutation-Selection Balance for Polygenic Traits. Genetics 111:165-195.
Turelli, M. 1988. Phenotypic Evolution, Constant Covariances, and the Maintenance of Additive Variance. Evolution 42:1342-1347.
Turelli, M., and N. H. Barton. 1990. Dynamics of Polygenic Characters under Selection. Theoretical Population Biology 38:1-57.
Vanhaecke, P., and P. Sorgeloos. 1980. International study on artemia IV. The biometrics of *Artemia* strains from different geographical origins. Pp. 394-405 *in* G. Persoone, P. Sorgeloos, O. Roels and E. Jaspers, eds. The Brine Shrimp Artemia. Universa Press, Wetteren, Belgium.
Vassilieva, L. L., A. M. Hook, and M. Lynch. 2000. The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. Evolution 54:1234-1246.
Via, S., and R. Lande. 1985. Genotype-Environment Interaction and the Evolution of Phenotypic Plasticity. Evolution 39:505-522.
Welch, J. J., and D. Waxman. 2003. Modularity and the cost of complexity. Evolution 57:1723-1734.
Welch, J. J., and D. Waxman. 2005. The nk model and population genetics. Journal of Theoretical Biology 234:329-340.
Whitlock, M. C. 1996. The red queen beats the jack-of-all-trades: The limitations on the evolution of phenotypic plasticity and niche breadth. American Naturalist 148:S65-S77.
Yang, H. P., A. Y. Tanikawa, W. A. Van Voorhies, J. C. Silva, and A. S. Kondrashov. 2001. Whole-genome effects of ethyl methanesulfonate-induced mutation on nine quantitative traits in outbred *Drosophila melanogaster*. Genetics 157:1257-1265.
Zeyl, C. 2004. Capturing the adaptive mutation in yeast. Research in Microbiology 155:217-223.

RÉCAPITULATIF DES ABRÉVIATIONS

$f(s)$: distribution de l'effet des mutations sur la fitness relative

$f_w(dw)$: distribution de l'effet des mutations sur la fitness absolue

MF : Modèle de Fisher

PM : modèles de paysages mutationnels

EVT : théorie des valeurs extrêmes

HC : modèles « en châteaux de cartes »

ε : épistasie multiplicative (ε_{jk} : entre les loci j et k)

a_{jk} : épistasie additive (entre les loci j et k)

HRE : effet Hill-Robertson

THEORETICAL APPENDIX

Properties of Quadratic forms

Let $\mathbf{dz} = \{dz_i\}_{i\in[1,n]}$ be a random vector drawn into a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{M}$, let $\boldsymbol{\Delta}$ be a constant vector and let θ be a constant scalar. Then the scalar random variable defined by

$$Q = \theta + \boldsymbol{\Delta}^{\mathrm{T}}\,\mathbf{dz} + \frac{1}{2}\mathbf{dz}^{\mathrm{T}}\,\mathbf{S}\,\mathbf{dz} \qquad (20)$$

Is a general quadratic form in random gaussian vectors. The distribution of Q has been much studied (Mathai and Provost 1992) in particular in econometrics, where authors have developed various approximations for its probability density, tail behaviour etc. (Jaschke et al. 2004; Kuonen 1999). The general expression of the cumulants of any order r (κ_r) of the distribution of Q can be expressed in a simple formulation (see e.g. Jaschke 2002):

$$\begin{aligned}\kappa_1 &= E(Q) = \theta + \frac{1}{2}Tr\,(\mathbf{SM}) \\ \kappa_r &= \frac{1}{2}(r-1)!\,Tr\left((\mathbf{SM})^r\right) + \frac{1}{2}r!\,\boldsymbol{\Delta}^{\mathrm{T}}\,\mathbf{M}(\mathbf{SM})^{r-2}\boldsymbol{\Delta}\end{aligned} \qquad (21)$$

As the cumulants equal the central moments for r = 1, 2 and 3, these directly give simple expressions for the first moments of Q. In our context, the distribution of s is simply obtained by taking negative values ($\mathbf{S}$ becomes $-\mathbf{S}$), using $\theta = 0$ and $\boldsymbol{\Delta}^{\mathrm{T}} = -\,\mathbf{z}_o^{\mathrm{T}}\mathbf{S}$. This can also be expressed in the transformed (diagonalized) system, $\mathbf{S}$ becomes the diagonal matrix $\boldsymbol{\Lambda} = \mathbf{diag}(\lambda_i)$, $\mathbf{M}$ becomes the identity, and $\mathbf{z}_o$ becomes $\mathbf{x}_o$. This gives the simplified expressions given in the appendix of Article 3.

Various approximations have been proposed for this distribution, most of them fit the first three moments or more moments (Mathai and Provost 1992). However none of these approximation could be easily interpreted biologically in the context of our model (*i.e.* defining n_e , for example) so we used the displaced gamma approximation detailed in Article 1.

More general expressions can be derived for the case where the distribution of $\mathbf{dz}$ has a non zero mean (to account for biased mutations effects on the phenotype). Quadratic forms in non-gaussian vectors have also been studied. As we argued in Chapter I and article 1, given proper transformations, the distribution of $\mathbf{dz}$ can often be "made gaussian". However, it may not be possible when the distribution of $\mathbf{dz}$ is skewed (not symmetric). Therefore, of particular interest are quadratic forms of vectors with a skewed distribution, e.g. the skew normal distribution for which a complete treatment can be found in (Gupta and Huang 2002). Interestingly, When $\boldsymbol{\Delta} = \mathbf{0}$ (in our case corresponding to $\mathbf{z}_o = 0$, only deleterious mutations), the moments of a quadratic form in multivariate skew-normal vectors are exactly the same as those of Q given in Eq. (28).

The distribution of the product of independent gaussian vectors $e = \mathbf{dz}_1^{\mathrm{T}} \mathbf{S}\, \mathbf{dz}_2$ drawn in the same multivariate distribution with mean **0** and variance **M** is a bilinear form in Gaussian vectors, which has mean and variance

$$\begin{aligned} E(e) &= 0 \\ V(e) &= 2\, Tr\,(\mathbf{SM}) \end{aligned} \quad . \qquad (22)$$

(Mathai and Provost 1992). This result is used in chapter II section IV for the computation of the distribution of epistatic interactions.

Random Matrix Theory

Random Matrix Theory (RMT) was initially developed in Quantum physics to model interactions in large nuclei. The idea behind this approach is well illustrated by this quote of F. J. Dyson (1962) in a recent review of the use of RMT in physics (Forrester et al. 2003):

"What is here required is a new kind of statistical mechanics, in which we renounce exact knowledge not of the state of the system but of the system itself. We picture a complex nucleus as a "black box" in which a large number of particles are interacting according to unknown laws. The problem then is to define in a mathematically precise way an ensemble of systems in which all possible laws of interaction are equally possible."

As we can see, this philosophy may also be applied to the study of complex phenotypes where many traits are connected by unknown mutational or selective interactions. The state of the system would refer to a given mutation while the system itself is the matrix of mutational covariances, that is unknown and may change with the environment or the species. I do not have the theoretical background to push the analogy any further, but I will present here the basic results that I used. A technical review of results in RMT is given in Bai (1999). In the following, I will refer to various equations from this article, however, Bai uses n as another quantity than the dimension of covariance matrices. For clarity of this text, I will keep the notations I have used in the rest of the document: n will refer to the dimension of covariance matrices (the number of traits). However, to make the correspondence with the notations of (Bai 1999), one should invert p and n in this text.

Let $\mathbf{X} = \{x_{jk} \mid j \in [1, n],\ k \in [1, p]\}$ be an $n \times p$ matrix which elements are drawn independently into an arbitrary distribution with mean 0 and variance σ^2. The matrix $\mathbf{S} = 1/p\ \mathbf{X}\,\mathbf{X}^{\mathrm{T}}$ is an $n \times n$ random symmetric matrix that can be used to model a random covariance matrix. Denote $y_p = n/p$ the ratio index of **S** (y is the ratio of the dimensions of **X**). Then if $y_p \to y$ (a finite value) when $p \to \infty$ the distribution of the eigenvalues of **S** converges to a known distribution. In other words, as the dimensions p and n get large, the eigenvalue distribution of any large **S** converges to a unique distribution known as the Marçenko-Pastur distribution. If λ_i are the eigenvalues of **S**, the probability density of λ_i is asymptotically (large p and n)

$$p(\lambda) = \begin{cases} \dfrac{1}{2\pi\, \lambda\, y\, \sigma^2} \sqrt{(b-\lambda)(a-\lambda)} & \text{if } a \le \lambda \le b \\ 0\ , & \text{otherwise} \end{cases} , \qquad (23)$$

Where $a = \sigma^2\left(1-\sqrt{y}\right)^2$ and $b = \sigma^2\left(1+\sqrt{y}\right)^2$, from Eq. 2.12 in (Bai 1999). This distribution has a point mass 1-1/y at $y = 0$ which means that only $n(1/y) = p$ eigenvalues of **S** are non-zero. This distribution has only two parameters, σ and y. This means that the distribution of the eigenvalues of various (random) large covariance matrices converges to the same two parameter distribution given above.

For a matrix **S** of dimension n, the expected value of the average of λ^h across eigenvalues can then be computed for any h , we denote this average $\overline{\lambda^h}$. As p gets large E(λ^h) converges to a unique value. For large n this time, the mean $\overline{\lambda^h}$ over n dimensions converges to its expectation:

$$\overline{\lambda^h} \xrightarrow{n\to\infty} E\left(\overline{\lambda^h}\right) = E\left(\frac{1}{n}Tr(\mathbf{S}^h)\right) = \alpha_h,$$
$$\text{where} \quad \alpha_h = \sigma^{2h}\sum_{i=1}^{h-1}\frac{y_p}{i+1}\binom{h}{i}\binom{h-1}{i} + O\left(p^{-1}\right) \qquad (24)$$

From Eq. 2.15 in (Bai 1999). The variance around this expectation is $O(p^{-2})$ so that the convergence to the asymptotic value is rather quick with increasing p (e.g. with $p > 10$ and $n > 50$, the approximation is good). We can use this result in our context to compute the moments of the eigenvalues of e.g. the selection covariance matrix (**S**), assuming it is drawn randomly as described above. The most important quantities in our model are the mean and coefficient of variation of eigenvalues, which are simply given by $\overline{\lambda} = \sigma^2$ and $CV(\lambda)^2 = y_p$. At this point, we can see why a large CV(λ) results in a reduced dimensionality of the phenotypic space. For a given number of traits n, a large CV(λ) corresponds to a small p, which means few eigenvalues are non-zero, hence the reduction in effective dimensionality ($n_e << n$) for large CV(λ).

When the entries of **X** are drawn into a Gaussian $N(0, \sigma)$, **S** is called a Wishart matrix. In this case, the ratio index y_p is related to the distribution of the correlations ρ_{ij} within matrix **S** (see the appendix of article 1 and Champion 2003). Therefore, we can link the parameter y_p that determines the moments of the distribution of the λ_i in Eqs. (23) and (24) with the level of phenotypic correlations. The standard deviation of this distribution is $\rho = \sqrt{E(\rho_{ij}^2)} = 1/p$, so that the ratio index of **S** is $y_p = n/p = n\,\rho^2$. We can see here that when correlations are strong (large ρ) p will be less than n so that matrix **S** is then positive *semi-definite*. Again, as these strong correlations correspond to smaller p, they result in a reduction of the dimensionality (only p non-zero eigenvalues).

However, in our model, the distribution of s depends on the distribution of the eigenvalues of a *product* of two random covariance matrices **S** and **M**, not of a single matrix. These moments can also be derived if **S** is assumed to be Wishart. The results of Bai (1999 theorem 2.10) refer to the product of **S** (which can be semi-definite) with a positive *definite* matrix (our **M**), but Silverstein (1999) shows in the comments of this review that the results are valid for positive *semi-definite* **M**. The general expression of the moments of order h of the eigenvalues $\lambda_{\mathbf{SM}}$ of matrix **S M** is:

$$E\left(\overline{\lambda_{\mathbf{SM}}{}^h}\right) = \sigma^{2h}\sum_{s=1}^{h} y_p^{h-s}\sum_{\Psi(s)}\frac{h!}{s!}\prod_{m=1}^{s}\frac{\alpha_m^{i_m}}{i_m!} + O(1), \qquad (25)$$

Where y_p is the ratio index of **S** defined above, from Eq. 2.24 in (Bai 1999) (Note that s here is an index). The inner sum is taken over $\psi(s)$ gathering all the sets of index (non-negative integers) $\{i_1,..,i_s\}$ that are solutions of the system of two equations $\{i_1+..+ i_s = h + 1 - s$ and $i_1 + 2\ i_2 + 3\ i_3....+ s\ i_s = h\}$. A set $\{i_1,..,i_s\}$can contain repeated values e.g. $i_1 = i_2$. The α_m are the m^{th} moments of the eigenvalue distribution of **M**.

The general expression above is a complicated sum but it simplifies greatly for the first moments (h = 1, 2, 3). **M** and **S** are constructed in same way: $\mathbf{M} = 1/q\ \mathbf{X}_\mathrm{M}\ \mathbf{X}_\mathrm{M}^\mathrm{T}$ where the entries of **X** have zero mean and variance σ_M. These entries need not be Gaussian deviates, but the entries of **S** have to be Gaussian (so that **S** be Wishart). The ratio index of **M** is denoted $y_q = n/q$. Then, the α_m in Eq. (25) are directly given by Eq (24) replacing y_p by y_q and σ by σ_M. The first moments of the distribution of λ_SM are

$$\begin{aligned} E\left(\overline{\lambda_{\mathbf{SM}}}\right) &= \sigma^2\ \sigma_{\mathbf{M}}^2 + O(1) \\ E\left(\overline{\lambda_{\mathbf{SM}}^2}\right) &= \sigma^4\ \sigma_{\mathbf{M}}^4\ (1 + y_p + y_q) + O(1) \\ E\left(\overline{\lambda_{\mathbf{SM}}^3}\right) &= \sigma^6\ \sigma_{\mathbf{M}}^6\ (1 + y_p^2 + y_q^2 + 3(y_p + y_q + y_p\ y_q)) + O(1) \end{aligned} \qquad (26)$$

If **M** is also a Wishart matrix and if we denote ρ_M the standard deviation of the correlations in **M**, then $y_q = n\ \rho_\mathrm{M}^2$. From Eq. (25) it is then easy to show that $\mathrm{CV}(\lambda)^2 = n\ (\rho^2+\rho_\mathrm{M}^2)$ so that $n_e = n/(1+\mathrm{CV}(\lambda)^2) = n/(1+ n\ (\rho^2+\rho_\mathrm{M}^2))$.

Therefore, RMT allows us to make a relation between the distribution of phenotypic correlations (its variance) and the effective complexity in a simple way. Furthermore, the value of n_e remains roughly constant when **M** and **S** vary randomly (e.g. with environments or across species with the same n value and same variance of correlations). We also see the limit of this approach. Strong correlations (large ρ), correspond to small values of p, in which case these asymptotic results are not accurate. However, simulations showed that Eq. (25) is still a good approximations for p as small as 10 (*i.e.* ρ as large as 0.3).

Criterion for pertaining to the Gumbel type III extreme value distribution

The displaced gamma used in our approach is the distribution of s_o - γ, where γ follows a gamma with scale α and shape β. The cumulative density function (cdf) of this distribution is

$$F(s) = 1 - F_\Gamma(s_\mathrm{o} - s) = \frac{\Gamma\left(\beta, \frac{s_o - s}{\alpha}\right)}{\Gamma(\beta)} \qquad (27)$$

where, $\Gamma(.,.)$ is the incomplete gamma function defined as $\Gamma(\beta, x) = \int_x^\infty e^{-t} t^{\beta-1} dt$. By taking the derivative of this cdf with respect to s, one finds the probability density function (pdf) of the distribution of s, $f(s)$, given in Eq (5).

The criterion for applying Orr's (2002; 2003) results derived from Extreme Value Theory is related to the distribution of absolute fitness effects (dw), not relative fitness effects s. $dw = s$

w, where *w* is the current fitness of the genotype from which the mutants are derived (wild-type or ancestor). The CDF $F_w(.)$ of *dw* is simply given by $F_w(dw) = F(dw/w)$ where $F(.)$ is given above (Eq. (27)), corresponding to a pdf $f_w(dw)$. The *necessary* and sufficient condition for the distribution of *dw* to be in the domain of attraction of the Gumbel extreme value distribution (eq. A.2 of (Orr 2003)) can then be evaluated for this distribution. If dw_f is the rightmost endpoint of the distribution ($dw_f = s_o\, w$ in our case), then this condition is that their exist a strictly positive function g(*t*) such that

$$\lim_{t \to dw_f} \frac{1 - F_w(t + dw\, g(t))}{1 - F_w(t)} = e^{-dw} \qquad (28)$$

The domain of values in which extreme value theory applies for $f_w(dw)$ is all the values of *dw* where this condition is fulfilled. In the case of the displaced gamma cdf, the left hand term is given by

$$\frac{1 - F_w(t + x\, g(t))}{1 - F_w(t)} = \frac{\Gamma(\beta) - \Gamma\left(\beta, 1/\alpha \left(s_o - \dfrac{t + dw\, g(t)}{w}\right)\right)}{\Gamma(\beta) - \Gamma\left(\beta, 1/\alpha \left(s_o - \dfrac{t}{w}\right)\right)}. \qquad (29)$$

The denominator of this expression tends towards 0 when $t \to dw_f = s_o w$. Therefore, in order to have it equal to e^{-dw}, we must necessarily have the numerator tending towards 0 too, which means $s_o - 1/w(t + dw\, g(t)) \to 0$, when $t \to dw_f$. We can thus use the series expansion of the incomplete gamma function ($\Gamma(\beta, z) \approx \Gamma(\beta) - z^\beta/\beta$ for small *z*, for both the numerator and denominator when *t* is close to dw_f. Solving Eq. (28) with this asymptotic expression yields an expression for g(*t*) in the neighbourhood of $t = dw_f$

$$g(t) \underset{t \to dw_f}{\approx} (s_o\, w - t)\, \frac{e^{-dw/\beta}}{dw} (e^{dw/\beta} - 1). \qquad (30)$$

Whenever $t < dw_f = s_o w$, g(*t*) exists and is strictly positive if $dw > 0$. Therefore the displaced gamma corresponds to a distribution of mutant absolute fitness effects that is in the domain of attraction of the Gumbel III distribution for positive values of *s*.

EMPIRICAL APPENDIX: METHODS AND RESULTS WITH THE BRINE SHRIMP *ARTEMIA*

Brief presentation of the brine shrimp Artemia

Ecology: *Artemia* (Crustacea, Branchiopoda, Anostraca) or "brine shrimp" is a primitive crustacean characterized by adaptation to hyper saline, highly variable aquatic environments. *Artemia* populations are found in about 500 natural salt lakes and man-made salterns scattered throughout the tropical, subtropical and temperate climatic zones, along coastlines as well as inland. Its physiological adaptations allow them to inhabit these "extreme environments" characterized by high concentrations of NaCl and other ions (e.g. carbonates), with wide fluctuations of both salinity and temperature (6-35°C), and occasional periods of desiccation. These adaptations are mainly:

(i) A very efficient osmoregulatory system (with among the widest range of salinity tolerance in all multi-cellular organisms), based on very efficient Na-K ATPase pumps. *Artemia* dies off at salinities close to NaCl saturation (*i.e.* >300 $g.l^{-1}$) but can survive in any range below this limit.

(ii) The capacity to synthesize very efficient respiratory pigments to cope with the low O_2 levels at high salinities.

(iii) The ability to tolerate strong changes in temperature by efficient heat shock proteins.

(iv) The ability to produce dormant cysts under stressful environmental conditions (e.g. low oxygen levels, high salinity, food shortage, low or very high temperatures etc.).

Artemia is a non-selective filter feeder of organic detritus, microscopic algae as well as bacteria. The *Artemia* biotopes typically show a very simple trophical structure and low species diversity; the absence of aquatic predators and food competitors allows brine shrimp to develop into monocultures (of billions of individuals!). Although brine shrimps can survive and reproduce in low salinity water (even less than natural seawater), they are usually not able to overcome predation and competition with other filter feeders. *Artemia* therefore, is only found at salinities where its predators (mostly fish) cannot survive (~70 $g.l^{-1}$). Concentrated seawaters with NaCl as major salt make up most, if not all, of the coastal *Artemia* habitats but also some of the inland habitats, such as the Great Salt Lake in Utah, USA. Other *Artemia* biotopes are located inland and have an ionic composition that differs greatly from that of natural seawater: sulphate waters, carbonate waters and potassium-rich waters (particularly in North America). These different conditions often correspond to different *Artemia* species or subspecies. One species (*Artemia tibetiana*) is also found in low temperature mountain salt lakes of Tibet (Abatzopoulos et al. 1998).

As *Artemia* is incapable of active dispersion, wind and waterfowl (especially flamingos) are the most important natural dispersion vectors; the floating cysts (diapause eggs) adhere to feet and feathers of birds, and when ingested they remain intact for at least a couple of days in the digestive tract of birds.

Taxonomy and reproductive mode: The genus *Artemia* is a complex of sibling species and superspecies, defined by the criterion of reproductive isolation. Generally, different names are assigned to reproductively isolated populations or clusters of populations. The bisexual species are
A. salina, Northern Mediterranean area, also called *A. tunisiana* for the southern Mediterranean populations
A. urmiana, Iran, Lake Urmia, the probable origin of the genus.
A. sinica, Central and Eastern Asia;
A. persimilis, Argentina
A. franciscana superspecies: Americas, Carribean and Pacific islands, including populations reproductively isolated in nature like *A.(franciscana) franciscana* and *A.(franciscana) monica* adapted to carbonate lakes (Mono Lake, California). There is a strong sexual dimorphism between males and females, males having two large frontal appendages that allow them to grab the female's ovisac. This pre-copulatory behaviour seems an important aspect of sexual reproduction, and may have led to strong sexual selection as it is not rare to observe two males grabbing the same female for several minutes.

There are also parthenogenetic strains found only in the Old World (Europe, Africa, Asia, Australia). They are all denoted *A. parthenogenetica* but correspond to either polyploids that reproduce by apomixis (clonal reproduction) or diploids that reproduce by automixis. This last mode of reproduction is thought to be an equivalent to selfing in diploid asexual *Artemia*, based on their low level of heterozygosity (for review, see Abreu-Grosbois 1987). In fact, the genetic system of these diploid parthenogens is not completely clear (Barigozzi 1974). I illustrate in Box 1 the different types of reproduction observed in different *Artemia* species (bisexuals, diploid asexuals, and polyploid asexuals). The problem is that there are two possible types of automixis (implying the absence of expulsion of one polar body to restore diploidy), both of which have been observed in cytologic analyses of diploid asexuals (Metalli and Ballardin 1970). Non-expulsion of the first polar body leads to clonality (e.g. in a strain from Italy), while non-expulsion of the second polar body should lead to complete homozygosity in the absence of recombination, or at least to strong homozygosity (e.g. in a strain from Sète, France). It is possible that diploid asexuals use a mixture of these two types of reproduction, leading to an overall low heterozygosity. In what follows I have considered diploid asexuals to be equivalent to a selfing species, based on a consensus from the literature (Abreu-Grosbois 1987).

In all these species, a single female reaches sexual maturity in about 8 to 30 days according to the strain and conditions. It can then produce several broods of about 100 individuals each (with a total reproductive output of up to 1000 individual in a few month Browne 1980). Each brood of a given female can consist of either ovoviviparous nauplii or oviparous dormant cysts, with a proportion of each over her lifetime that varies with the strain and environmental conditions. The cysts can be kept for months or years in desiccated state, into saturated brine or under vacuum. They may then be hatched within 24 hrs by putting them into aerated tubes with strong light at about 37°C.

Applications: *Artemia* is also the main food source of aquaculture. As it is possible to store large quantities of live eggs (cysts) and to hatch them within 24h, nauplii are used as a source of live food for fish larvae and adult brine shrimps are also used for the growth of later stages in fish development. Last but not least, it is an important character of children literature and entertainment (see the Sea Monkeys websites and (Pif Gadget N°60, 1970; Pif Gadget N° 1, 2nd edition 2004)).

Overall, the biology, physiology, development and growth conditions of *Artemia* are well characterized. In particular, the Artemia Reference Center (in Ghent, Belgium) gathers many information and publications on *Artemia*, and also keeps a large collection of cysts from many distinct periods and areas worldwide. Samples are available to researchers and can be

sent by mail. There have also been many studies on the ecology of *Artemia*, with a focus on the competition between sexual and asexual species (for review see Browne 1992; Lenz and Browne 2000). The ecological niche of *A. franciscana* and *A. parthenogenetica* diploids from the Mediterranean area are well characterized (Barata et al. 1996; Browne and Hoopes 1990; Browne and Sallee 1984). It is mostly determined by the salinity and temperature of the waters, and by the amount of food available (which also depends on these abiotic parameters).

Many information on *Artemia* can be found on the Artemia Reference Center (ARC) website (http://www.aquaculture.ugent.be/index.htm). In particular, a concise but thorough presentation of *Artemia* biology and ecology, and of its use in aquaculture can be found on the the FAO manual available online at http://www.fao.org/DOCREP/003/W3732E/W3732E00.HTM, from which the first part of this brief presentation is inspired. Finally, several books present general features of *Artemia* biology from the molecular to the species levels (Abatzopoulos et al. 2002; Browne et al. 2000).

In the following, I present the techniques I developed in order to do a mutation accumulation experiment on a diploid parthenogenetic strain of *Artemia*.

Protocols developped for breeding and measurements of life history traits

The main goal of my experimental work on *Artemia* was to produce mutant lines by a mutation accumulation experiment, and to assay their fitnesses along a continuous gradient of salinities. To do so, I produced a single isogenic line from a parthenogenetic diploid female sampled in the salterns of Salins de LaPalme (Southern France), denoted LAPX. I used a parthenogenetic line because of the ease with which it allowed to quickly produce a large number of isogenic individuals. I chose a diploid strain based on the assumption that they reproduced by selfing, so that I could expect to observe mutations in homozygous state. A mutation accumulation experiment also requires being able to breed individually large numbers of lines in the most permissive conditions, and to be able to measure fitness traits precisely. As our lab had never worked on *Artemia* previously, I developed a set of techniques for this purpose.

Breeding system: The principle of the experimental setting is illustrated in Box 2. Females were raised individually, in test tubes with a filter bottom, in batches of 70 tubes. This ensures a common environment within blocks and saves much time for the replacement of used water, allowing a single person to raise up to 2800 individuals together.

Food was given as individual doses of the unicellular alga *Dunaliela salina*, the most widely used food source in experimental studies of *Artemia* (Coutteau et al. 1990). However, as our goal was to study the effects of mutations in various salinities, we wanted to avoid giving live food to the *Artemia*. Indeed, if live algae tend to multiply according to the salinity imposed in different treatment, the (controlled) salinity parameter may also induce variation in an (uncontrolled) biotic factor (algae concentration), introducing a confounding factor. We therefore used dried (lyophilised) algae, mixed with ~ 20% of live cells previously frozen (*i.e.* assumed to be dead or at least not physiologically active). The admixture of fully compartmented cells was needed to allow reproductive maturity. Indeed, *Artemia* fed only dried *D. salina* show very good survival but remain in immature stage (they do not develop ovisacs). The feeding schedule during the course of development was taken from (Coutteau et al. 1992), computing the equivalence between live and dried algae based on biomass. This breeding system allows a good survival of individual females until the onset of reproduction

(80%). However, we will see at the end of this section that it may raise some difficulties regarding reproduction.

Experimental asexual line LAPX: The LAPX line was produced by four generations of multiplication from a single female. Multiplication took place in 3L aerated bottles with salt concentration 90g.l^{-1}, populations of approximately 100-200 adults were fed with live algae under constant light. The nauplii produced in each bottle were isolated in new 3L bottles with the same conditions except lower food supply (to avoid overfeeding in the early stages). The cysts produced were gathered and kept under vacuum at 4°C, for each generation. By the fourth generation, the populations were starved and aeration was stopped for ~1 week to stimulate cyst production. This led to the production of 20 000 cysts kept under vacuum at 4°C. This line produces a large proportion of cysts in the conditions of the laboratory (>60-70 %), at least for the first to third broods that I used.

Trait measurements: To allow precise estimation of the effects of mutation accumulation, repeatable measurement protocols are needed. The traits measured were fitness traits (survival curves with time, number of cysts per brood), and a morphological trait (size at day 24). I developed a technique to count automatically the number of cysts in a brood by image analysis of photos of the cysts produced in a tube (using the free software ImageJ Abramoff et al. 2004). This method is illustrated in Box 3 a. It allows to count the number of cysts and to measure individual parameters of the cysts (diameter, area etc.) although I did not assess the precision of these last measures. In any case, the biometric parameters of the cysts showed little variation among lines or across cysts within lines. I did not check the validity of the method for counting nauplii because almost no nauplius was produced in my experiment (this will be discussed in the next section). I think that a similar method may be developed to count nauplii but it may be less precise as the software is particularly well suited for identifying circular objects (cysts). The size of females at day 24 was measured as body length from the frontal naupliar eye (or its inferred position) to the extreme part of the furca, as illustrated in Box 3 b.

mutagenesis experiments

Protocols: As the generation time that we obtained in the laboratory was larger than 25 days, (longer in the beginning of my phD), a standard mutation accumulation (at least 15-20 generations) may have taken too long to be undertaken during my phD. Therefore, I chose to produce mutants by random mutagenesis. There is no protocol for random mutagenesis in *Artemia* apart from studies of the effects of gamma rays (Squire 1970; Squire 1973; Squire and Grosch 1970). Such irradiation mutagenesis is known to provoke almost only severely deleterious mutations which may not reflect the array of spontaneous mutation effects. Typically, studies of random mutagenesis now rely on alkylating agents that are known to produce mainly point mutations. This method has been used successfully in *C. elegans* (Keightley et al. 2000) and *D. melanogaster* (Keightley and Ohnishi 1998) to accelerate mutation accumulation experiments and obtained larger effects. These studies were based on the mutagen Ethyl methane-sulfonate (EMS). Other studies have used Ethyl nitro-sourea (ENU) to produce random mutations (e.g. De Stasio and Dorman 2001) although this mutagen was never used in the context of the analysis of mutation fitness effects. During my PhD, I did five trials of mutagenesis using the isogenic line LAPX, the first four were based on EMS and the last one, which I present here briefly, used both EMS and ENU, with two doses for each mutagen.

The critical point in any mutagenesis protocol is to manage to mutate the germ line, as only mutations occurring in the germ line are passed to the F_1 (as opposed to somatic mutations). In my first attempts of mutagenesis, I used EMS on instar I nauplii (as F_o). At this stage, numerous cell (including germ line cells) are undergoing mitosis. However, this proved inefficient and has the drawback that the F_o has to be grown to reproductive maturity to collect F1 offspring, despite the injuries caused by somatic mutations. This may induce selection for the nauplii having incorporated only small doses of mutagen which F_1 offspring may be mutation-free. I thus turned to a protocol on adult females using results from a thorough review of radiobiology experiments in *Artemia* (Metalli and Ballardin 1970 and advices from Godelieve Criel). Females were put in mutagen at stage D of oogenesis (illustrated in Box 4 a.). At this easily recognizable stage, meiosis for the first brood is in metaphase I while the germ line is undergoing mitotic cell divisions that will produce the second brood. This pattern has been described with cytologic observations of this stage (Metalli and Ballardin 1970). Therefore we mutagenised adult females in stage D, and used the second brood occurring after mutagenesis as the F_1 mutagenised lines.

The protocols I developed for the mutagenesis were inspired by a classic protocol on zebrafish (*Danio rerio*), as an example of mutagenesis in a seawater environment (Mullins et al. 1994) and with doses taken from a protocol for *C. elegans* (De Stasio and Dorman 2001). They are presented in Box 4 b. for both ENU and EMS. These two mutagens are highly hazardous chemicals and proper decontamination must be handled with much care.

Results: I do not give the results of the ENU which were unexpected. Indeed, the control showed higher mortality over the 24 first days than the two mutagenised lines. However, a deleterious effect of the mutagen was observed in the 1st brood of the F_o lines (the one in metaphase I at the time of mutagenesis). I did not find any satisfactory explanation for this result.

The results of the mutagenesis on the F_1 survival and reproductive output are presented in Box 5. The mutagenised lines showed larger mortality in the first period of development (until day 27) but the difference did not persist on the long term (from day 35 to 45). Similarly, the distribution of reproductive early reproductive output in mutagenised F_1 is biased to the left relative to that of the control, but no such trend is observed in the late output, which distribution is very similar among mutagenised and control lines.

The distribution of sizes (body length at day 24, Box 6) seemed to be little affected by the mutagenesis. The distributions are very close between treatments, apart from three individuals from the 30 mM EMS treatment, that showed abnormal development. The sizes do not correlate with total reproductive output (late + early) in both the control and 15 mM treatments, but a significantly positive correlation (ρ = 0.30) was detected in the 30 mM treatment.

The summary of statistics of the reproductive output and size distribution of mutagenised and control lines is given in Box 7. I will discuss below the difficulties raised by this experiment. Overall the variance on all traits does increase among the mutagenised lines relative to the control (dV >0) and the mean reproductive outputs are decreased (dM<0). I give the corresponding Bateman-Mukai estimates of $\bar{s}$ and U that *would* correspond to this dataset (s_{BM} and U_{BM}). They do not seem unrealistic, for the early reproductive output (but not for the late output). The estimated average effect of mutations on fitness s_{BM} (early output) are of the order of those known for drosophila (several percents). The estimates of U_{BM} (early output) are about 100 times the ones observed for spontaneous mutations, which is in agreement with data from EMS mutagenesis on *D. melanogaster* and *C. elegans* (Keightley et al. 2000; Keightley and Ohnishi 1998).

To study these data, I used a linear mixed model ("lme" package in R Ihaka and Robert 1996) with block and F_o mother as fixed effects and a random effect of mutagenesis nested within F_o mother (the same dose for a given mother). The effect of mutagenesis is significant

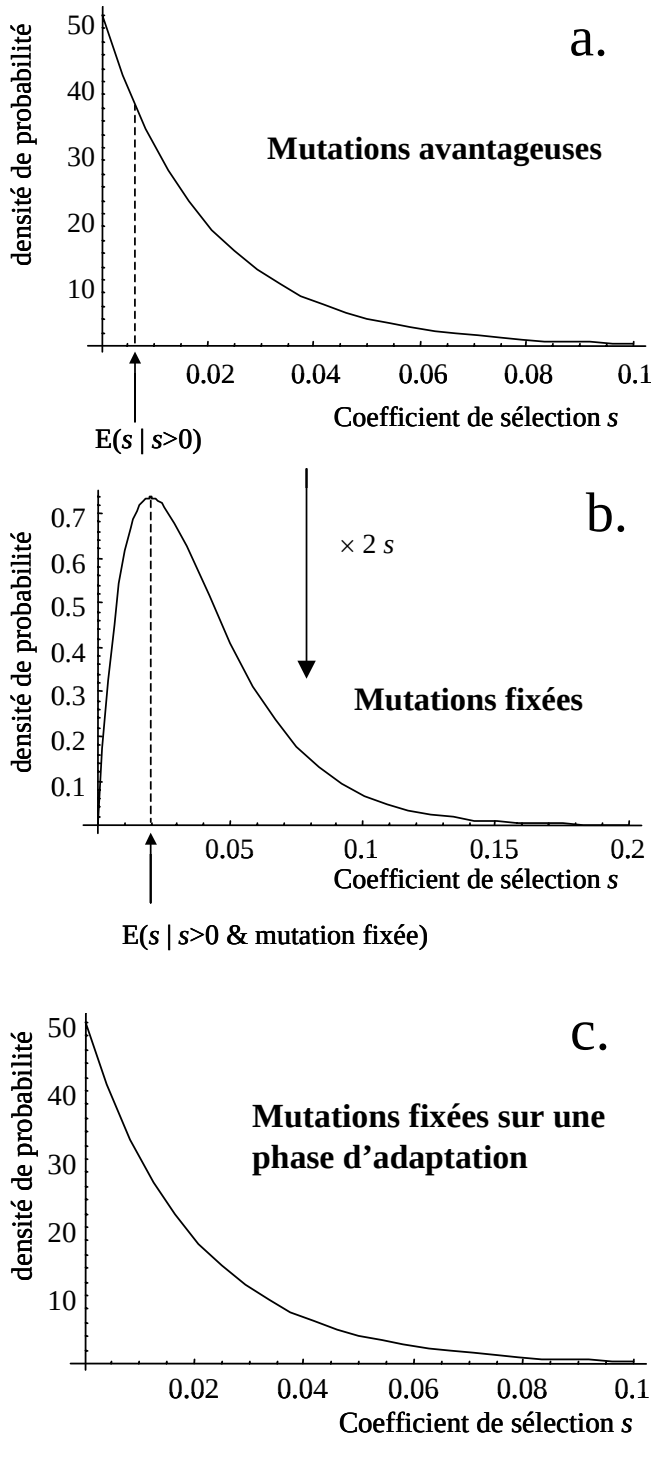
on early reproductive output (p = 0.034) when pooling the two doses (15mM and 30mM, which do not statistically differ in their effect) in a single random effect. Mother and block effects are not significant. The effect of mutagenesis is not significant for late output or body length (while block effects are significant). I did not do a complete analysis of these data because of the difficulties explained below. I wish to undertake a more rigorous analysis with an explicit maximum likelihood model.

Difficulties: The obvious limit of these results is that they correspond to F_1 lines data, so we only assessed a single individual per line. Therefore, we cannot estimate line effects, and (ii) we cannot rule out maternal effects: if the mutagenesis has lowered the condition of the F_o females, the F_1 offspring may be less fit because of the bad condition of their mother and not because of the genetic effects of mutations. In this regard, the fact that the deleterious effects of mutations appear mostly on early reproduction (Box 5 a. and Box 7), but not later may suggest that such maternal effects are responsible for at least part of the effects detected. An obvious solution to this problem is to assess F_2 line fitnesses. However, although many lines produced offspring in rather large quantities (~100 per brood, see Box 5 a.), they were all cysts, and I did not manage to hatch them so far. This lack of hatchability is even observed for cysts from the control lines, suggesting that it is not due to a severely deleterious effect of mutations on F_2 lines. I imagine one possible explanation for this problem: the food used may have been suboptimal. Indeed, almost all lines reproduced oviparously (cysts) which may be evidence that they were stressed (oviparity is favoured by stressful conditions), even if the LAPX line tends to produce a large proportion of cysts. This may also have been caused by the mutagenesis (as such a tendency for oviparity has been observed after gamma irradiation Metalli and Ballardin 1970), but the controls also showed increased oviparity. However, the survival of the F_1 lines was rather good with ~ 75 - 85% survival at the onset of reproduction (see Box 5 a.). Severely stressful conditions should have induced a larger mortality in the first stages. However, as we have seen, reproduction is highly dependent on the quality not the quantity of food supply (whereas survival is more dependent on the quantity, from what I observed). Indeed, a food regime based only on dried algae results in a good survival but in the absence of sexual maturation. It is possible that the frozen algae were a suboptimal complement, and should be replaced by live algae, which we know give good results for reproduction in the conditions of our breeding system.

Conclusion: our experiment detected a limited effect of mutagenesis on F_1 lines of the LAPX strain, acting only on early reproductive output, with no significant effect on late output or on body length. This suggests two possibilities: either the mutagenesis was inefficient, or the control trait values were too variable to detect a strong effect of mutagenesis. It is possible that the feeding protocol induced some stress as we have seen above, so that it may have generated too much variance in the control traits (the fitness of the control is highly variable, see Box 5). However, it is also possible that the mutagenesis was partly inefficient, for two possible reasons. Mutations may have appeared but not be detected if the LAPX line reproduces by clonality (so that most mutations were in heterozygous state at least in F_1). This is possible as we have seen that the reproductive system of diploid asexuals is not clear, although assumed to be equivalent to selfing. The penetration of the chemical mutagen in the reproductive tissue is a priori efficient as I could observe penetration of toluidine blue in these tissues, and this stain is a much larger molecule than EMS. However, it is also possible that the LAPX line be strongly resistant to mutagenesis. Indeed, I observed very few lethal mutations, which is unexpected in a mutagenesis experiment considering the doses that I used. It is possible that asexual diploids (that are quite ancient asexuals) tend to be robust to mutation. One way to overcome these possible difficulties would be to do a mutagenesis with sexual strains.

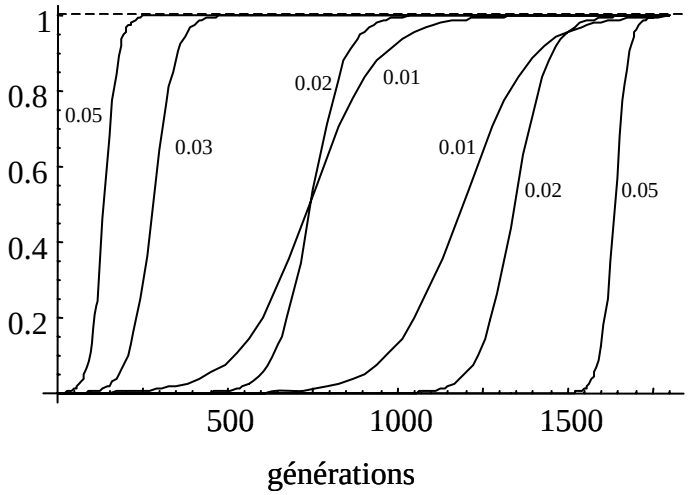
Encadré 1: Effets des mutations responsables de l'adaptation

Je montre ici un exemple de distribution des effets sélectifs des mutations avantageuses (ici une exponentielle (**a.**). Celle-ci est tronquée sur la gauche lorsqu'on considère la distribution de l'effet des mutations fixées lors d'un évènement de substitution (correction pour la probabilité de fixation : 2*s*) (**b.**). Quelle que soit la distribution des effets avantageux (en **a.**), la distribution des effets fixés pendant une phase d'adaptation (plusieurs substitutions successives) est toujours approximativement exponentielle (**c.**).



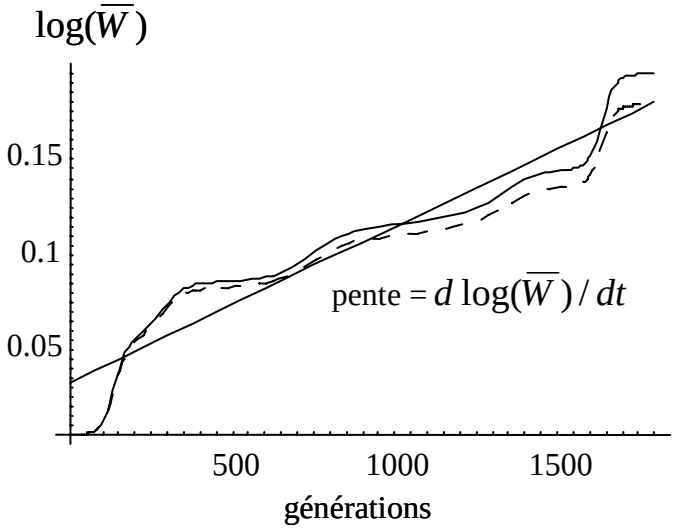
Encadré 2 : Approximation continue pour le taux d'adaptation

Je représente ci-dessous l'augmentation en fréquence de mutations avantageuses apparues successivement au cours d'une phase d'adaptation (ici, dans une population sexuée de grande taille). Pour chaque mutation d'avantage s_i (donné sur la courbe) la fréquence $p_i(t)$ est donnée en utilisant la trajectoire déterministe (logistique).



Balayages sélectifs successifs

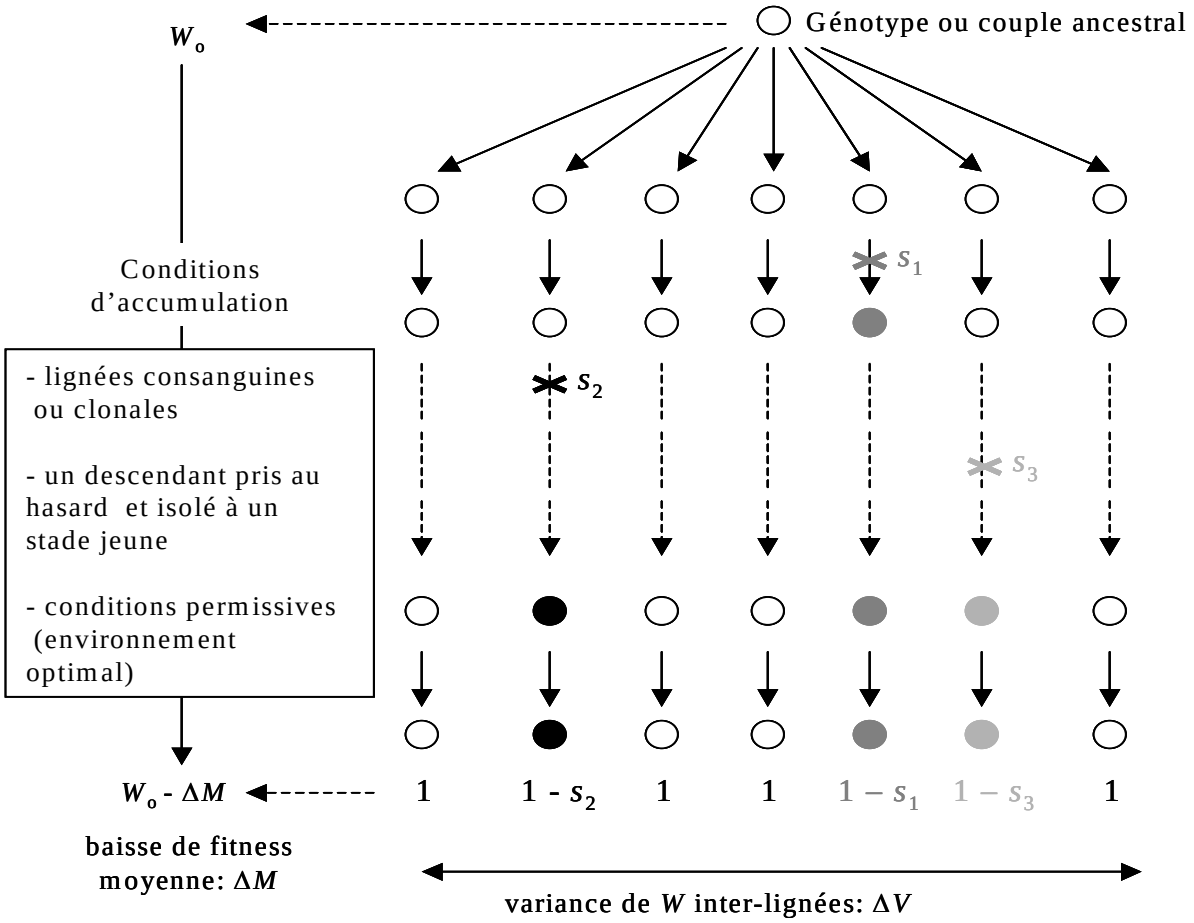
Ce processus détermine la dynamique temporelle de la fitness moyenne (ci-dessous). Son logarithme ($\log(\overline{W})$, courbe pleine). Si les mutations avantageuses ont des effets faibles (e.g. $s < 0.1$) $\log(\overline{W})$ est bien approximé par $\sum s_i \, p_i(t)$ (courbe en pointillés). Le changement de fitness moyenne au cours du temps est ensuite approximé par une fonction linéaire du temps, donc par la dérivée de $\log(\overline{W}(t))$:



Taux net d'augmentation de la fitness moyenne

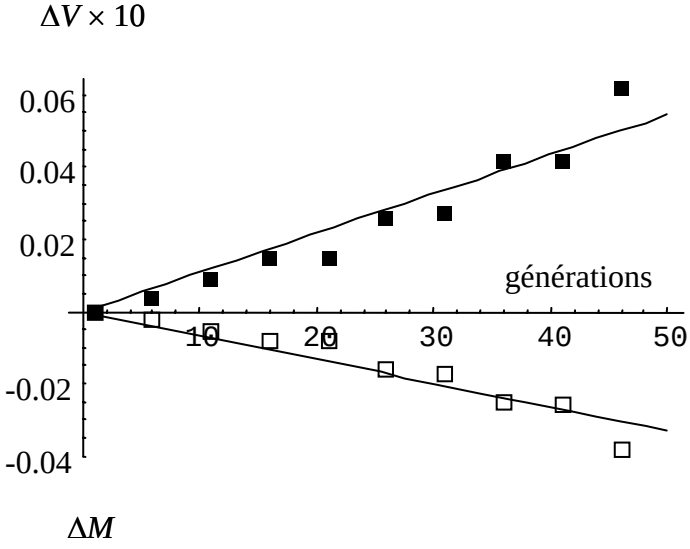
Encadré 3 : accumulation de mutations

Les niveaux de gris correspondent à différentes mutations.



s: mutation d'effet délétère s

Calcul de l'incrément de variance en fitness *dV* et de la perte de fitness moyenne *dM* dues aux mutations, par génération

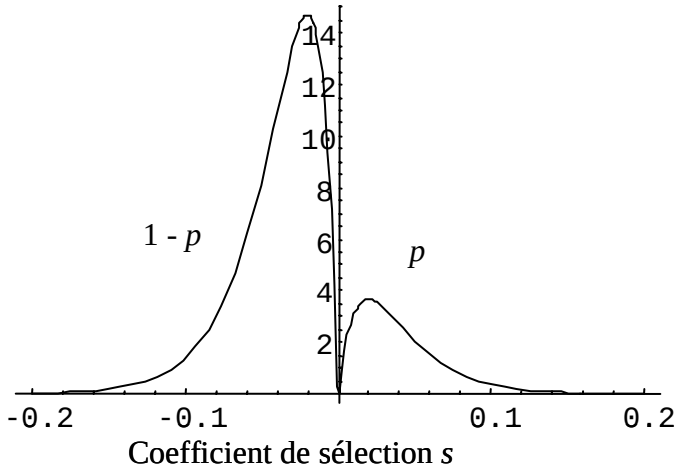


$$dV = \frac{\Delta V}{\Delta t} = U\,\bar{s}^2(1 + CV(s)^2)$$

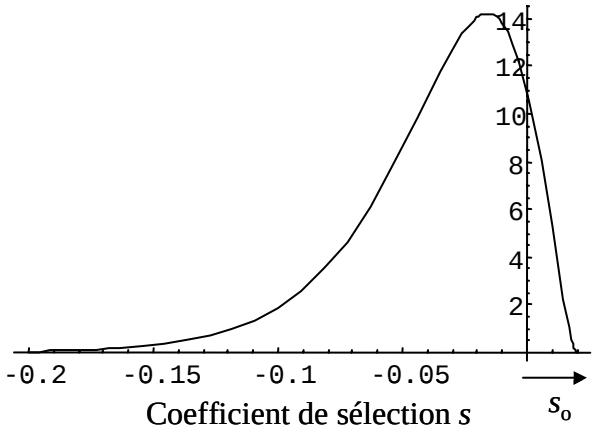
$$dM = \frac{\Delta M}{\Delta t} = U\,\bar{s}$$

Encadré 4 : Distributions de s utilisées a priori pour l'analyse d'expériences d'accumulation de mutations

Gamma réfléchie: on suppose que les mutations délétères et avantageuses sont distribuées suivant des gamma (négatives et positives) de mêmes paramètres, avec une proportion *p* de mutations avantageuses (1 - *p* de délétères) (Keightley, 1994).



Gamma déplacée: on suppose que les mutations sont distribuées comme la somme d'une constante (s_o est le paramètre de déplacement) et d'une gamma négative. Cette distribution ne contient pas de discontinuité en zéro mais suppose que les mutations avantageuses sont bornées ($s < s_o$) (Shaw et al. 2002).

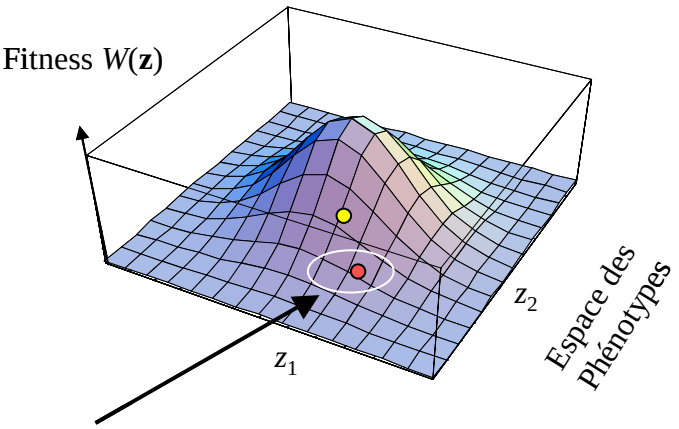


Encadré 5 : Modèle de Fisher et distribution de l'effet des mutations

a. Modèle de Fisher classique (hypothèses de symétrie entre tous les traits)

Ce schéma représente un paysage adaptatif à deux dimensions. Les deux traits $\mathbf{z} = (z_1, z_2)$ définissent un espace phénotypique. La valeur du phénotype **z** détermine la fitness (fonction $W(\mathbf{z})$) de façon isotrope: les deux traits sont soumis à la même force de sélection stabilisante. Les phénotypes mutants sont distribués de façon également isotrope autour du phénotype ancestral $\mathbf{z}_o$.

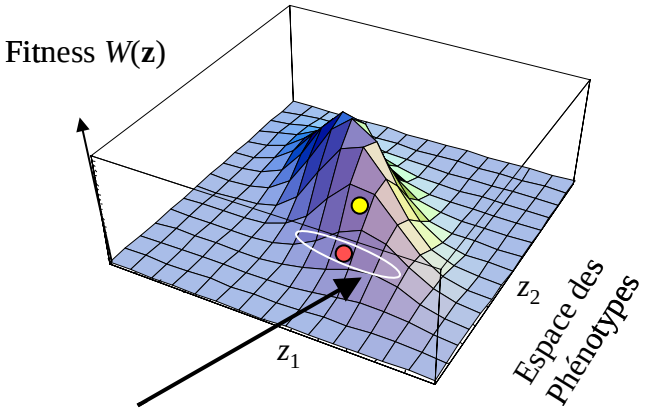
- Optimum phénotypique $\mathbf{z} = \mathbf{0}$
- Phénotype ancestral $\mathbf{z}_o$



Distribution de phénotypes mutants

b. Modèle de Fisher généralisé pour prendre en compte les covariances phénotypiques

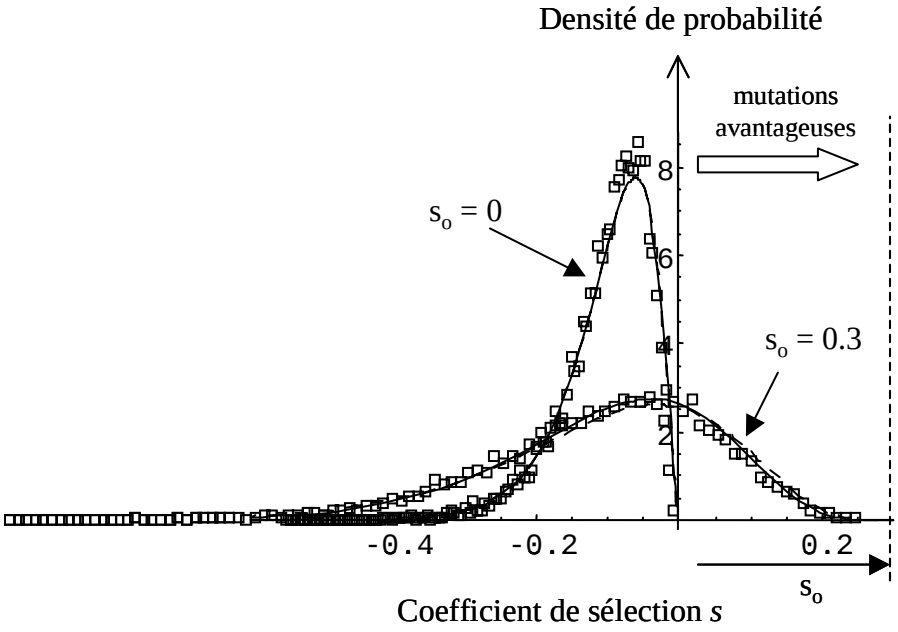
ce schéma représente l'extension du modèle pour prendre en compte l'existence de covariances mutationnelles et sélectives. La force de la sélection stabilisante sur chaque trait diffère entre les deux traits, et ils covarient dans leur effet sélectif. De même, la distribution des phénotypes mutants n'est pas isotrope: ici, la variance mutationnelle est plus grande sur le trait z_1, et il y a une covariance mutationnelle entre z_1 et z_2.



Distribution de phénotypes mutants

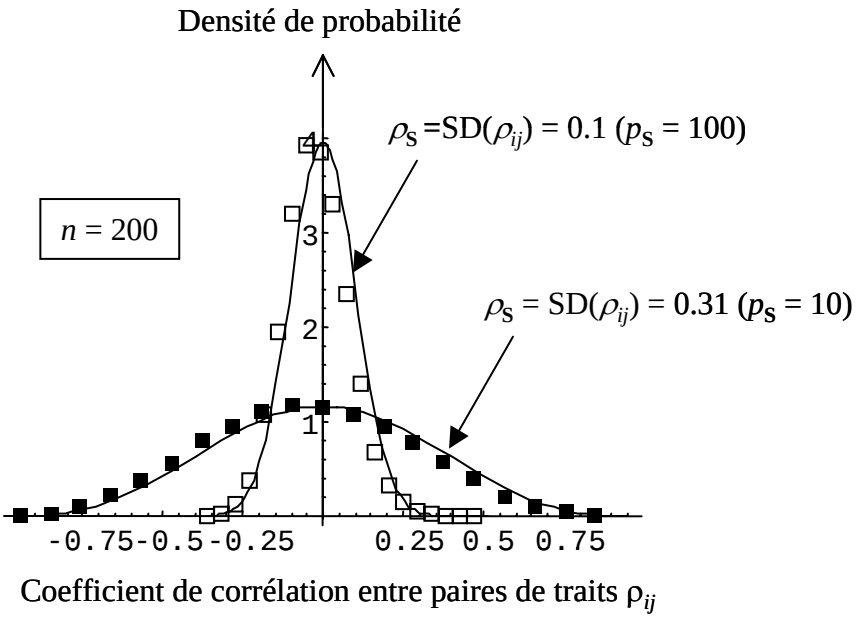
Encadré 6: Distribution prédite par notre modèle

Cette figure représente la distribution de s obtenue par simulations (distribution exacte, carrés) et l'approximation par la gamma déplacée donnée en Eq. (5) (lignes, q = 1). Les matrices **M** et **S** ont été tirées aléatoirement (voir la partie II.3) pour n = 200 traits phénotypiques. On voit que l'approximation est correcte, et que malgré une grande valeur de n, la distribution est très asymétrique lorsque $s_o = 0$. Lorsque le génotype ancestral est maladapté ($s_o = 0.3$) elle est plus proche d'une gaussienne (le paramètre de forme β est plus grand). Enfin il apparaît que la distribution des effets avantageux est bornée par s_o.



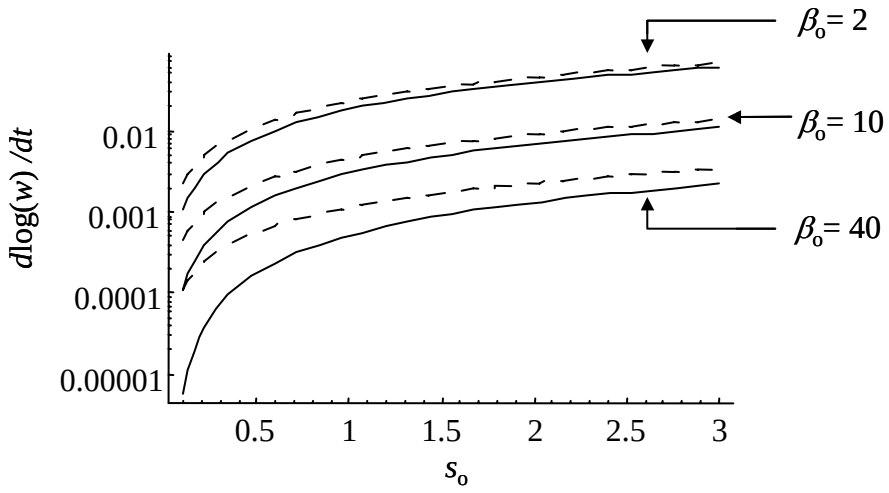
Encadré 7: Distribution des corrélations phénotypiques dans une matrice de Wishart

On montre ici la distribution des coefficients de corrélation ρ_{ij} (entre paires de traits $\{i,j\}$) au sein d'une matrice **S** tirée dans une distribution de Wishart de dimension n = 200 traits, de paramètre d'échelle σ_S = 1 (les corrélations ne dépendent pas de ce paramètre d'échelle) et avec deux valeurs du paramètre p_S ($\mathbf{S} = \mathbf{X_S}\,\mathbf{X_S}^T$ où $\mathbf{X_S}$ est une matrice $n \times p_S$). La force des corrélations phénotypiques (indépendamment de leur signe) est estimée par la déviation standard de la distribution des corrélations dans la matrice $\rho_S = SD(\rho_{ij})$. ($\rho_S = \sqrt{1/p_S}$), qui est proche de la valeur absolue moyenne de ρ_{ij} $E(|\rho_{ij}|)$. Les carrés donnent la distribution obtenue pour une matrice simulée et les lignes donnent l'attendu théorique. On voit que le paramètre p_S détermine bien de fortes différences dans la distribution des corrélations ρ_{ij}.



Encadré 8: Approximation pour le taux d'adaptation loin de l'optimum

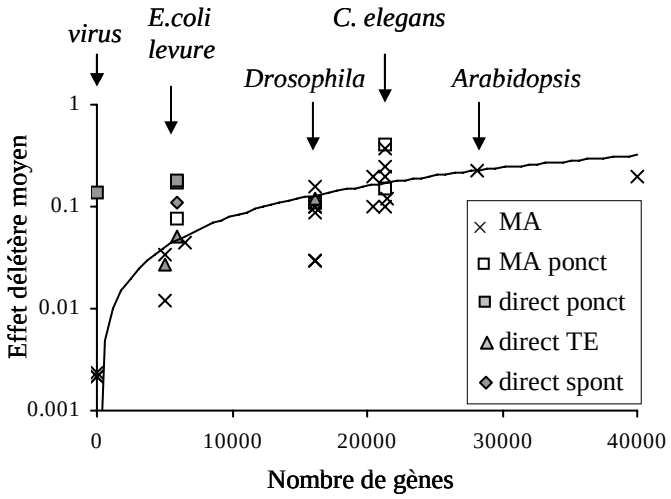
Le taux d'adaptation (augmentation de la log fitness moyenne par génération) est représenté (en échelle log) en fonction de la distance à l'optimum. Les lignes pleines donnent la valeur calculée à partir de la densité de probabilité de la gamma déplacée. Les pointillés montrent l'approximation donnée en Eq. (9). Les courbes sont données pour trois valeur de β_o. On voit que l'approximation est meilleure lorsque s_o est grand. Elle est également meilleure pour de faibles valeurs de β_o. Les valeurs des autres paramètres sont $N_e = 300$, $U = 0.015$ et $\bar{s} = 0.01$.



Encadré 9: Distribution de l'effet des mutations délétères entre espèces.

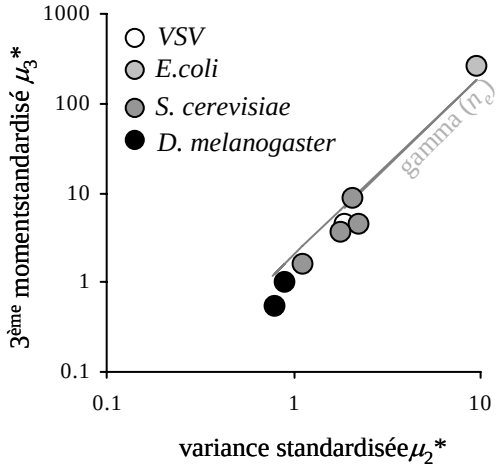
a. Relation entre nombre de gènes et effet délétère moyen des mutations

La figure ci-dessous représente la relation entre l'effet moyen délétère des mutations ($\bar{s}$ estimé, en échelle log) pour différentes espèces (indiquées sur la figure), en fonction de leur nombre de gènes. Les différents codes correspondent à différentes méthodes de production des mutants (voir article) et à des mesures de $\bar{s}$ directes (direct) ou indirectes (MA, estimation par la méthode de Bateman-Mukai en général). La ligne représente la relation linéaire observée (ici en échelle log).



b. Relation entre deuxième et troisième moments de *f*(*s*) (test de la distribution gamma)

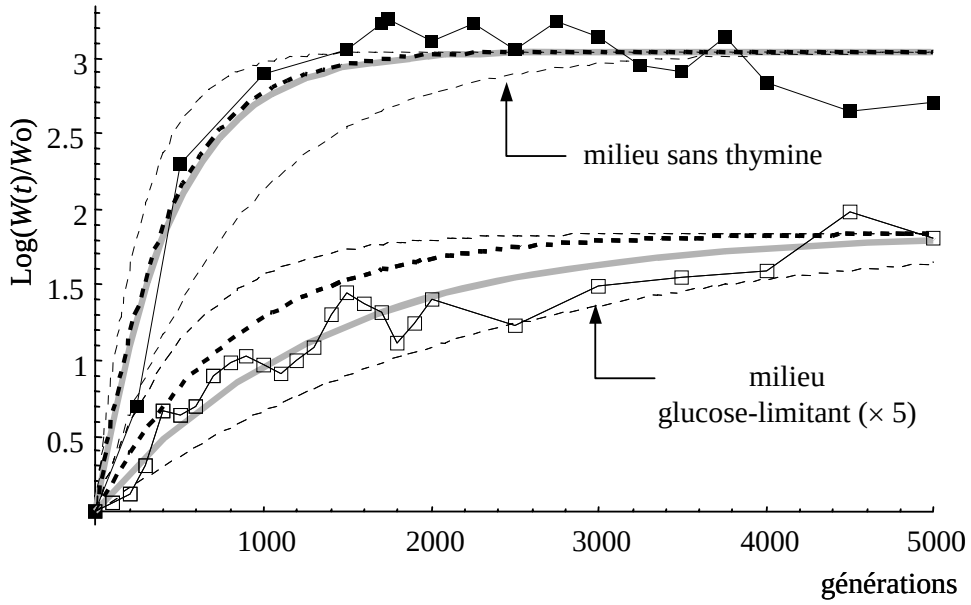
Pour différentes espèces, relation entre la variance et le 3ème moment de *f*(*s*), standardisés par l'effet moyen $\bar{s}$: μ_2* et μ_3* respectivement (voir article 1). La ligne indique la relation attendue sous l'hypothèse que *f*(*s*) est une gamma.



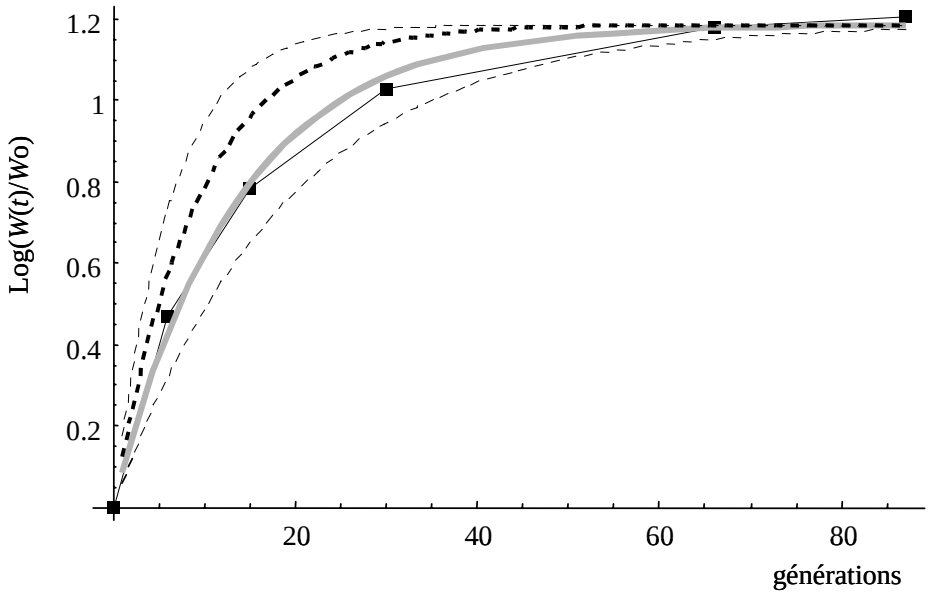
Encadré 10: Trajectoires de fitness observées et prédites chez *E. coli* et *D. melanogaster*

On représente le changement de la fitness moyenne ($W(t)$) avec le temps (en nombre de générations dans le milieu de sélection) relativement à la fitness initiale (W_o). Les carrés correspondent aux fitness mesurées empiriquement. Les lignes pleines grises correspondent à la trajectoire ajustée aux données, en utilisant le modèle décrit par l'Eq. (12). Les lignes pointillées représentent les trajectoires prédites d'après la valeur du plateau de fitness final (donc avec la valeur prédite de a). Les pointillés fins correspondent à « l'enveloppe » associée à cette prédiction pour des valeurs de q variant entre 0.5 et 1.5.

a. adaptation d'*E . coli* à des changements de source de carbone (Lenski et Travisano 1994) ou d'acides aminés (De Crécy Lagard et al. 2001)

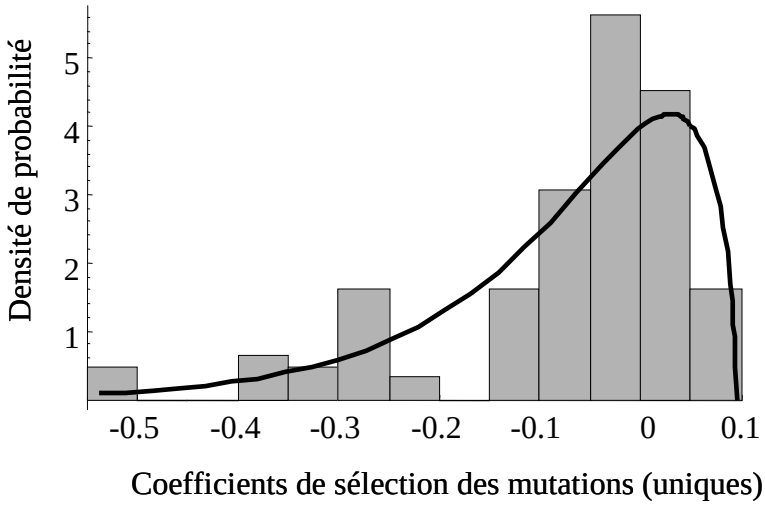


b. Adaptation de D. melanogaster à la captivité (Gilligan et Frankham 2003)

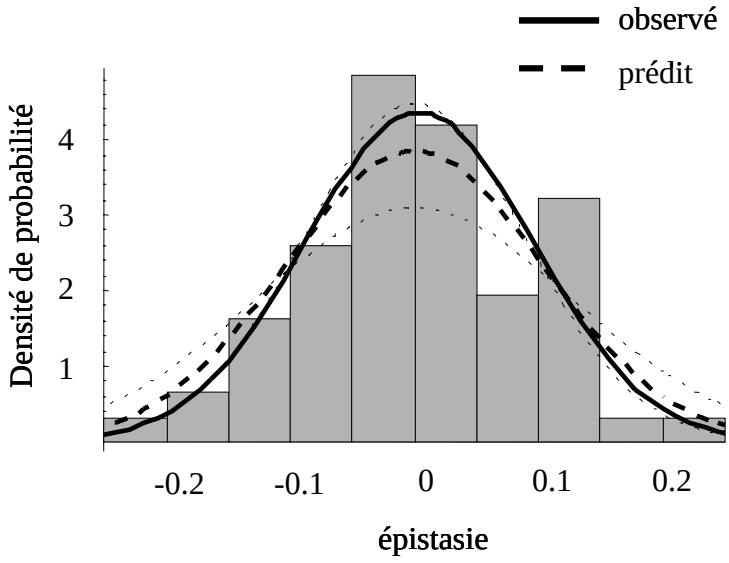


Encadré 11: Distribution des interactions épistatiques chez le virus VSV et prédictions à partir de *f*(*s*)

a. Distribution de l'effet des mutations individuelles, *f*(*s*) chez le virus *VSV*
Pour 91 mutations ponctuelle (d'après [Sanjuan, 2004 #1331]). La ligne pleine donne la densité de probabilité d'une gamma déplacée ajustée à partir des moments et en supposant que s_o est l'effet avantageux maximal observé.

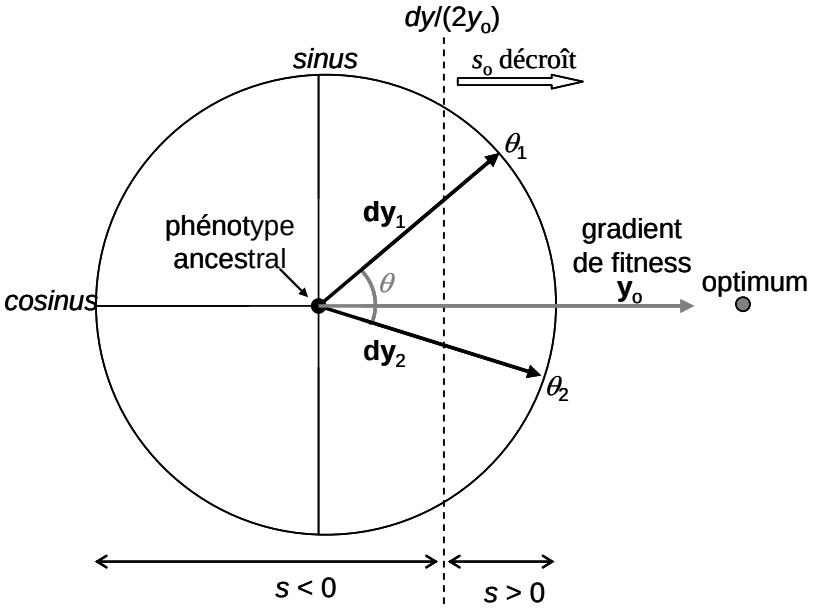


b. Distribution des interactions épistatiques *p*(*ε*) chez les virus *VSV*
Distribution des interactions épistatiques entre paires de mutations choisies au hasard (d'après [Sanjuan, 2004 #1331]). Les lignes correspondent à : la distribution gaussienne ajustée aux données à partir de la moyenne et la variance de $p(\varepsilon)$ (ligne pleine), la distribution prédite d'après la variance de $f(s)$, sa moyenne et s_o en supposant $q = 1$ (pointillés épais), et l'enveloppe correspondant à $q \in [0.5, 1.5]$ (pointillés fins).



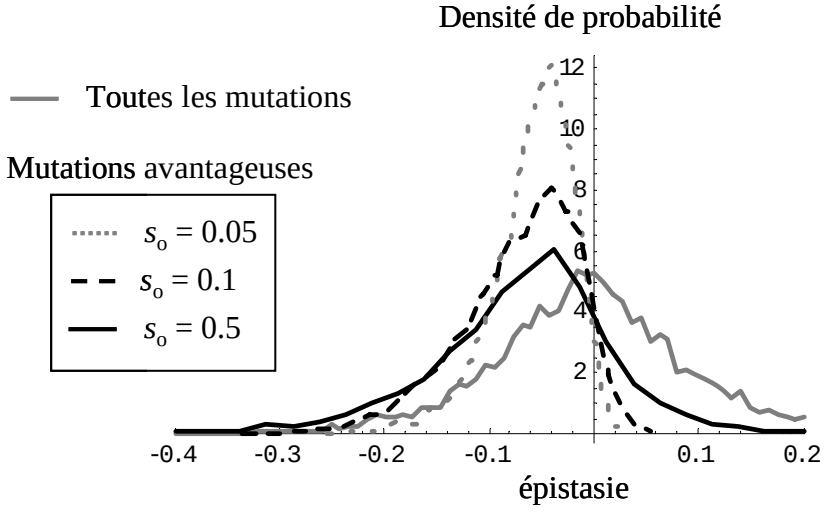
Encadré 12: Distribution des interactions épistatiques parmi les mutations avantageuses

a. Principe du raisonnement géométrique : $\mathbf{dy}_1$ et $\mathbf{dy}_2$ sont deux mutations avantageuses représentées sur un cercle trigonométrique. Les mutations avantageuses sont contraintes à avoir un angle avec la direction de l'optimum tel que $\cos(\theta_i) > dy/(2y_o)$ (la limite indiquée par la ligne pointillée). L'angle θ entre les deux mutations est donc réduit, ce qui biaise les valeurs de $\cos(\theta)$ vers des valeurs positives, donc les épistasies $\varepsilon = -dy^2 \cos(\theta)$ vers les valeurs négatives.



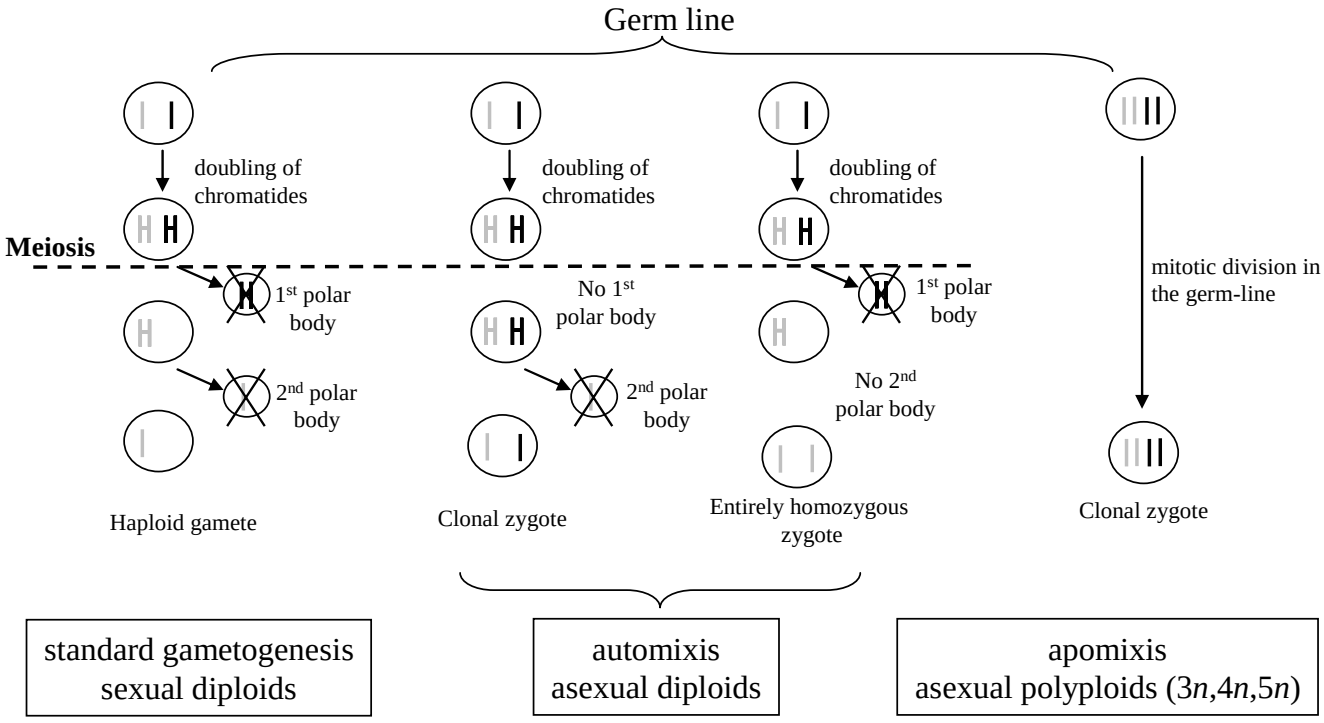
cos(θ) > 0 entre les mutations avantageuses
donc $e = -dy^2 \cos(\theta) < 0$

b. Simulations : Distribution des interactions épistatiques parmi un ensemble de mutations aléatoires (pointillés) et parmi les seules mutations avantageuses (lignes pleines), pour différentes valeurs de la distance à l'optimum (s_o indiqué sur la figure). On constate un biais vers les épistasies négatives (antagonistes) parmi les mutations avantageuses. Ce biais augmente quand la distance à l'optimum (s_o) décroît. Paramètres des simulations comme en Encadré 8.



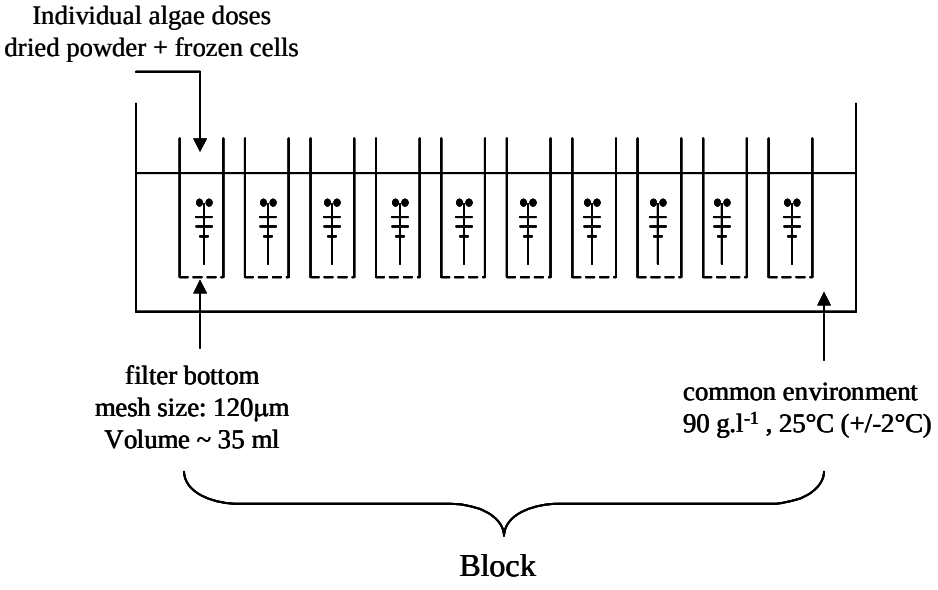
Box 1: alternative reproductive systems in *Artemia*

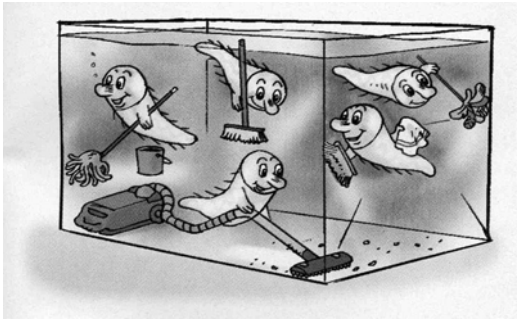
The reproductive mode differs in the *Artemia* genus between bisexual, asexual diploid and asexual polyploid species. I represent here the possible alternative modes of production of a gamete or a zygote by sexual and asexual females. The bisexual female reproduces by meiosis followed by chromosome reduction leading to a haploid oocyte. The asexual diploid female reproduces through automixis, which means one of the polar bodies is not expelled thus maintaining diploidy. According to whether the 1st or 2nd polar body is kept, the genetic system differs greatly; both modes of automixis have been observed in parthenogenetic females, but the outcome is thought to be approximately equivalent to selfing. The polyploid asexuals reproduce through apomixis which means that oocytes are produced by mitotic division of the germ line (see Metalli and Ballardin 1970).



Box 2: Individual breeding system

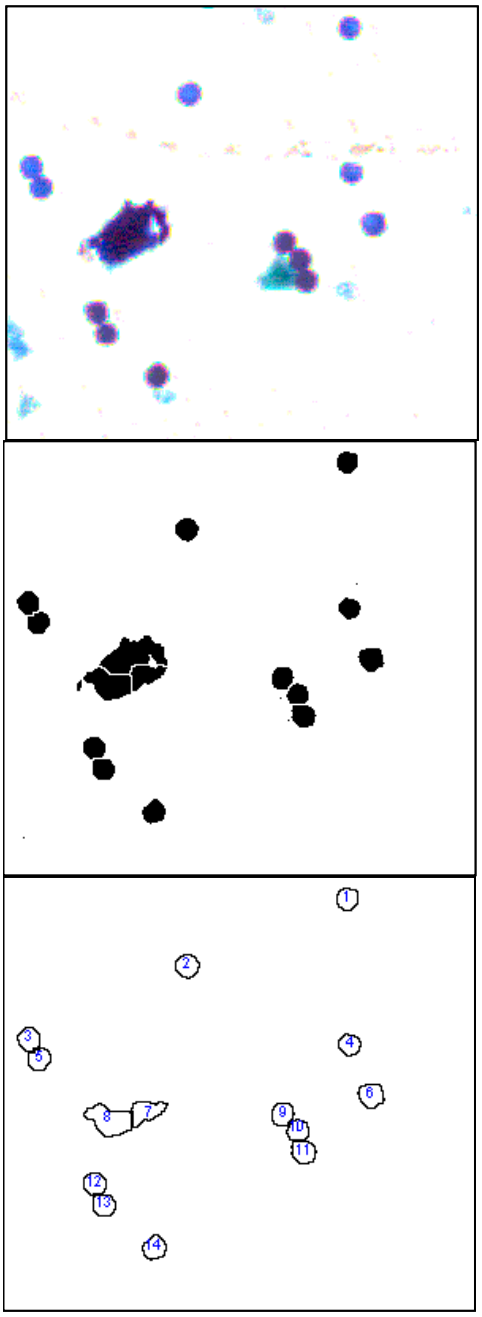
Females were raised individually into test tubes with a filter bottom allowing water to flow within each tube of a batch containing 70 individuals. This system allows to replace water of all 70 individuals at once, and to maintain common conditions within a block. Food was given every two to three days as individual doses of a mixture of dried algae (*Dunalliela salina*) complemented with frozen (inert) cells of the same species. Water was replaced every 3 to 4 days in all blocks with fresh medium from a single batch source (seawater + unrefined salt to obtain a salinity of 90 $g.l^{-1}$).

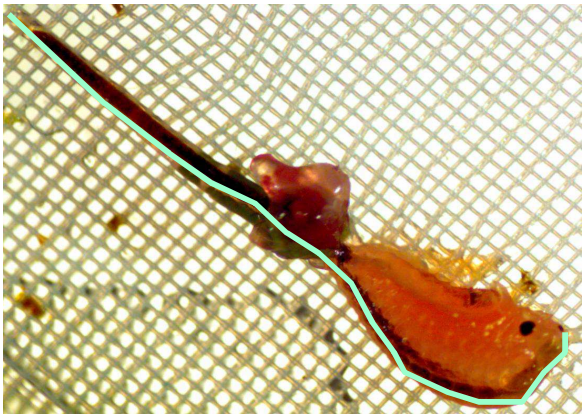




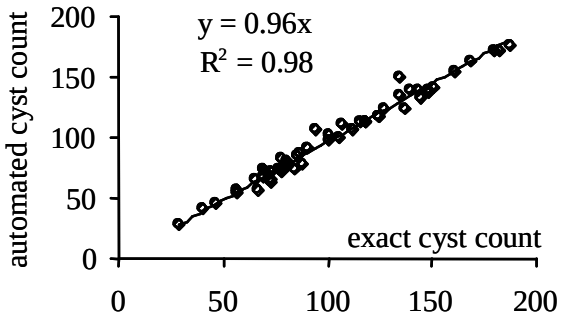
(source: Pif Gadget, 2nde édition N°1 Juillet 2004)

Box 3 : Measuring traits in *Artemia*





b. Example of a measure of body size under ImageJ:
The measure was made by following the digestive tractus. Starting from the extreme part of the cerebral ganglions and until the extreme part of the furca.

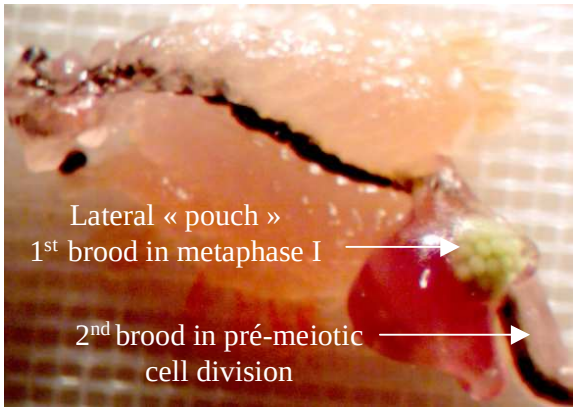


a. Automated cyst count using ImageJ:
By color-thresholding the original photo, most particules (e.g. algae etc.) can be removed. Different pluggins then allow to separate objects that are adjacent. The analysis then excludes all objects that are too large, and we can finally exclude the detected objects that are not circular enough. The methods detects 98% of the variance in cyst number (counted directly) as illustrated on the graph.

Box 4 : Mutagenesis protocols

a. chosen stage for the mutagenesis stage D

The picture below shows a mature female at stage D (« lateral pouch ») of oogenesis for her first brood. The lateral pouch contains matured oocytes in metaphase I of meiosis. At this easily recognizable stage (which lasts no longer than 6-7 hours), the germ line is undergoing pre-meiotic cell divisions that will produce the oocytes of the 2nd brood. We kept this *2nd brood as the one susceptible of carrying heritable mutations.*



b. Protocols for EMS and ENU mutagenesis of adult females in stage D

20 females in stage D (see below) were put into test tubes containing 200 ml of artificial seawater (10 $g.l^{-1}$ crystal Reef ® salt buffered to pH = 6.5 with dihydrogen phosphate), with small amount of food supply (lyophilised *D. salina* algae).

Under the hood:

ENU: 85 ml of 10mM acetic acid were injected into a 1 g ENU isopac and agitated until dissolution of the powder (~ 2 hrs) to obtain a 100 mM ENU solution. For the 1 mM (resp 0.6 mM) treatment, 2 ml (resp. 1.2 ml) of the 100mM ENU solution were injected into the test tubes containing the F_o females. For the control treatment, 2 ml of the 100 mM acetic acid solvant was added to the tube.

EMS: for the 30 mM (resp. 15 mM) treatments, 600 µL (resp. 300 µL) of EMS from a 1 ml isopack were added to the 200 ml test tubes containing F_o females.

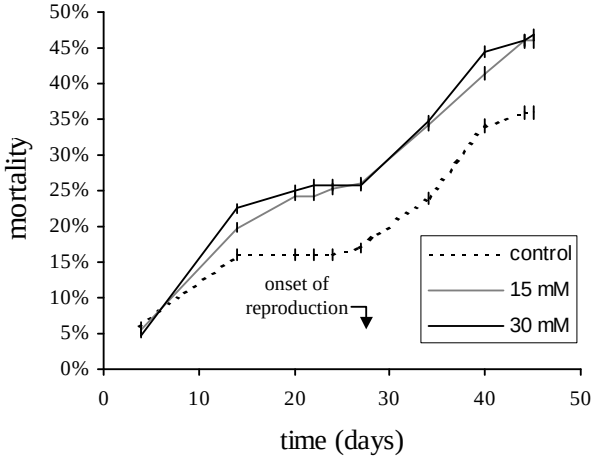
For all treatments, two tubes were used per dose, and the tubes were gently agitated for 4 hrs under the hood.

Decontamination: The mutagenized females were transferred to a tube with a 400 µm filter bottom and decontaminated into three successive bath of decontaminating solution (1M thiosulfate + 0.1M NaOH at pH = 10) of ~30s each, agitating to favour the dilution of the mutagen (the thiosulfate baths must be rather short because of a harmful osmotic pressure). The females were then passed through three seawater baths of 20 mn each. After decontamination, females were left overnight with oxygenation and live algae to recover from stress. All used objects and individuals dying in the next two days were left into decontaminating solution and left for 48h under the hood before discarding.

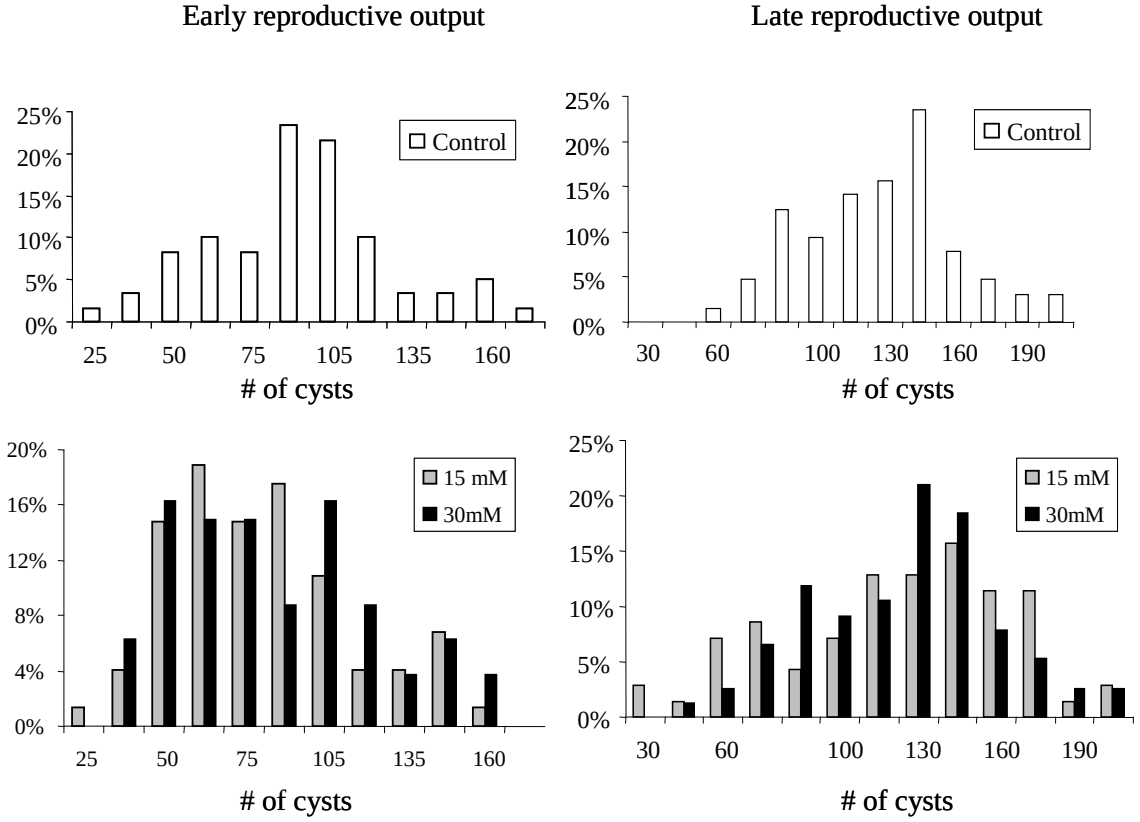
The first brood of the F_o females were left aside and the 2nd brood was used as the F_1 lines for the assays.

Box 5: effects of EMS mutagenesis on survival and reproductive output

a. *The effect of EMS mutagenesis on the mortality of F_1 females (2^{nd} brood of mutagenised F_o females) is shown for the pre-reproductive period. Bars indicate 5% confidence intervals. This graph corresponds to pooled data from five blocks of 70 individuals (*n *is the total number of individuals per treatment).*

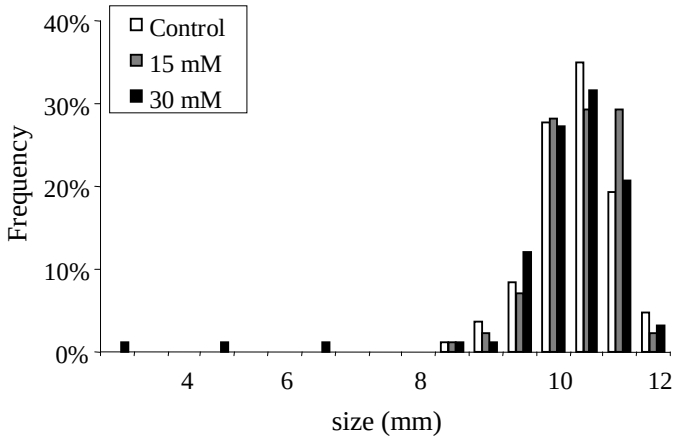


b. The graph below shows the distribution of reproductive output for control and mutagenized F_1 lines, measured for early lifetime (total number of cysts laid until day 35) and later (from day 35 to 45). The number of cysts was measured by the semi-automated method explained in Box 1: alternative reproductive systems in *Artemia* and text.

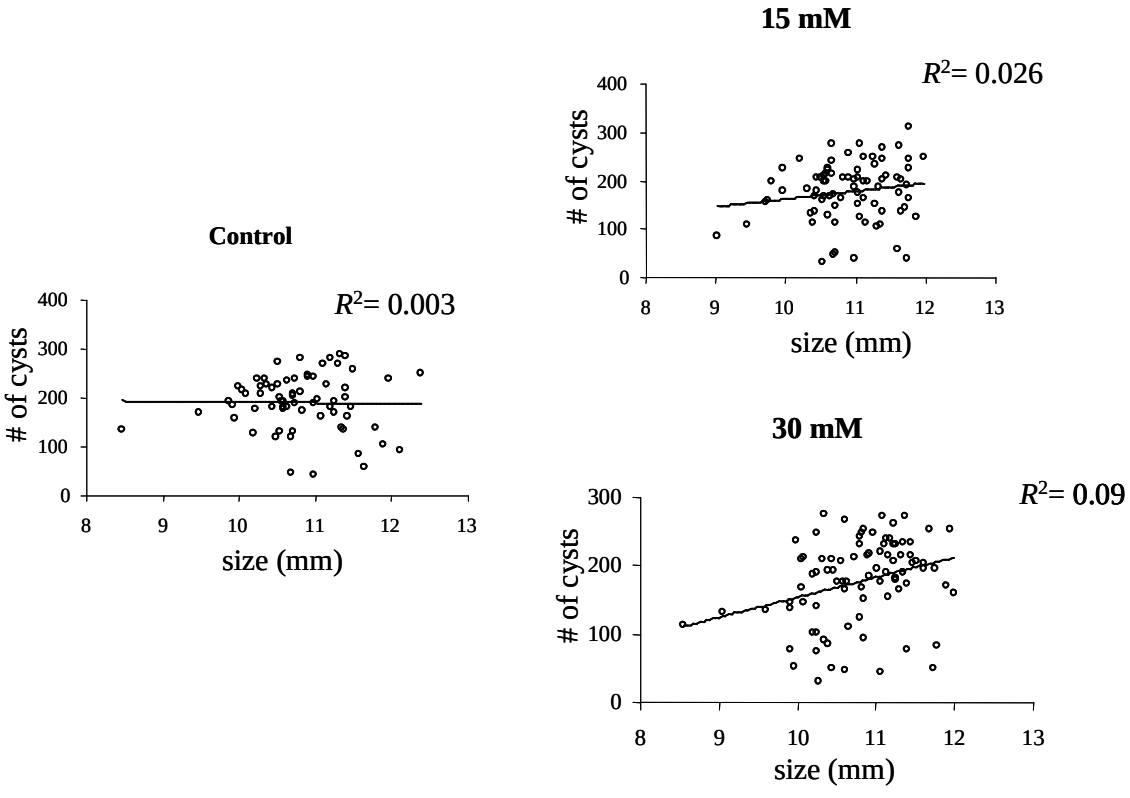


Box 6 : Correlation between size en reproductive output among mutagenised lines

The distribution of sizes at maturity in the control and mutagenised F_1 lines is given below (with the corresponding EMS dose). Sizes were measured from photos (method explained in Box 1: alternative reproductive systems in *Artemia* and text), at day 24 for all lines (corresponding to the approximate onset of reproduction).



The graphs below show the correlation between size at day 24 and total reproductive output. The correlation is not significant for the control (p = 0.88) and for the 15mM EMS treatment (p = 0.15), but is significant for the 30 mM EMS treatment (Pearson's ρ = 0.30, p = 0.005). This correlation is not significant for either early or late reproductive output in any treatment, so the observed correlation for total output in the 30mM treatment is most probably due to the fact that the smaller females tended to have no late output.



Box 7: Summary of data for the EMS mutagenised lines

These tables give the value of the reproductive output (number of cysts laid) during early lifetime (before day 35) and later (from day 36 to 45). The "total" column is simply the sum of these outputs. The "size" column gives the size of each individual measured at day 24 (i.e. approximately at the onset of reproduction). Almost no line produced nauplii (at least none was detected) so we do not report any viviparous reproductive output. Absent data correspond to cases where no brood was produced in the period considered (early or late) as space is left in the case.

.I.1.1 Summary of the data for the lines from the control treatment

summary statistics				
Variable	Early output	Late output	Total output	Size
min	24	55	41	8.47
max	173	200	290	12.38
mean	88	120	191	10.82
variance	990	1038	3253	0.42

These statistics are the minimum, maximum, mean value end variance among lines for the trait considered, for .

.I.1.2 Summary of the data for the lines from the 15 mM EMS treatment

summary statistics				
Variable	Early output	Late output	Total output	Size
min	24	29	32	9.03
max	163	202	311	11.97
mean	79	117	177	10.91
variance	1007	1701	3639	0.361
dM	-0.108	-0.024	-0.070	
dV	0.002	0.046	0.011	
s_{BM}	-0.021	-1.957	-0.153	
U_{BM}	5.27	0.01	0.46	

In addition to the statistics given for the control, this table gives the mutational parameters as estimated from the data, for relative fitness (measured either by early, late or total reproductive output). *dM*: the change in mean relative to the control, *dV*: the increment of variance (standardised by the control mean), $s_{BM} = dV/dM$, the Bateman - Mukai estimate of the average deleterious effect of mutations (assumed to be in homozygous state) and $U_{BM} = dM^2/dV$, the Bateman - Mukai estimate of the mutation rate induced by one generation of EMS mutagenesis.

.I.1.3 Summary of the data for the lines from the 30 mM EMS treatment

summary statistics				
Variable	Early output	Late output	Total output	Size
min	26	34	31	8.56
max	155	191	274	12.00
mean	80	116	178	10.80
variance	1052	1200	3877	0.408
dM	-0.087	-0.031	-0.069	
dV	0.008	0.011	0.017	
s_{BM}	-0.093	-0.360	-0.250	
U_{BM}	0.93	0.09	0.27	

Evolution, 60(5), 2006, pp. 893–907

A GENERAL MULTIVARIATE EXTENSION OF FISHER'S GEOMETRICAL MODEL AND THE DISTRIBUTION OF MUTATION FITNESS EFFECTS ACROSS SPECIES

GUILLAUME MARTIN[1,2] AND THOMAS LENORMAND[1]
[1]*Centre d'Ecologie Fonctionnelle et Evolutive, Centre National de la Recherche Scientifique, UMR 5175, 1919 Route de Mende, 34 293 Montpellier, France*

Abstract.—The evolution of complex organisms is a puzzle for evolutionary theory because beneficial mutations should be less frequent in complex organisms, an effect termed ''cost of complexity.'' However, little is known about how the distribution of mutation fitness effects ($f(s)$) varies across genomes. The main theoretical framework to address this issue is Fisher's geometric model and related phenotypic landscape models. However, it suffers from several restrictive assumptions. In this paper, we intend to show how several of these limitations may be overcome. We then propose a model of $f(s)$ that extends Fisher's model to account for arbitrary mutational and selective interactions among n traits. We show that these interactions result in $f(s)$ that would be predicted by a much smaller number of independent traits. We test our predictions by comparing empirical $f(s)$ across species of various gene numbers as a surrogate to complexity. This survey reveals, as predicted, that mutations tend to be more deleterious, less variable, and less skewed in higher organisms. However, only limited difference in the shape of $f(s)$ is observed from *Escherichia coli* to nematodes or fruit flies, a pattern consistent with a model of random phenotypic interactions across many traits. Overall, these results suggest that there may be a cost to phenotypic complexity although much weaker than previously suggested by earlier theoretical works. More generally, the model seems to qualitatively capture and possibly explain the variation of $f(s)$ from lower to higher organisms, which opens a large array of potential applications in evolutionary genetics.

Key words.—Complexity, Fisher model, mutation fitness effect distribution, mutational and selective covariances, random matrices, survey.

Received July 22, 2005. Accepted February 18, 2006.

Mutations are the raw material for evolution. As a consequence, predicting and estimating the distribution of the effects of mutations on fitness (hereafter $f(s)$) is a central issue for many aspects of evolutionary theory (Lynch et al. 1999; reviewed in Charlesworth and Charlesworth 1998; Bataillon 2000). Considerable effort has been devoted, mostly empirically, to address this issue. Yet, we are still largely ignorant of how this distribution varies among organisms and how to explain this variation (Lynch et al. 1999; Bataillon 2003; Keightley and Lynch 2003; Shaw et al. 2003), although significant differences across species can be observed, even at small phylogenetic scales (Baer et al. 2005). From a theoretical standpoint, adaptive landscape modeling, such as Fisher's model (1930), is the most commonly used framework to predict $f(s)$. Fisher modeled an organism as a vector of n trait values (i.e., a position in the phenotypic space), whose fitness is determined by the distance of this phenotype to a given optimum. Mutation randomly displaces phenotypes in this n-dimensional space, which allows one to compute $f(s)$. Fisher's model is appealing because it predicts $f(s)$ based on selective and mutational assumptions on the underlying phenotypic traits (as in, e.g., Welch and Waxman 2003). Other approaches have been suggested to predict some features of $f(s)$. In particular, extreme value theory can be used to characterize the right-tail behavior of $f(s)$ (advantageous mutations) with a minimal set of assumptions regarding $f(s)$ itself (Orr 2003, 2005a). However, by avoiding describing $f(s)$, this approach also has the weakness of disconnecting properties of advantageous and deleterious mutations, which may jointly affect some evolutionary process (e.g., the rate of adaptation in asexuals; Gerrish and Lenski 1998). In contrast, Fisher's model predicts the full distribution with both deleterious and advantageous effects under a given set of assumptions. Its prediction may thus be compared with most empirical data on $f(s)$, which mainly describe the distribution of deleterious mutation effects (Lynch et al. 1999).

Fisher's model makes simplifying assumptions, whose realism may be questioned (Clarke and Arthur 2000; Orr 2000, 2001). As a consequence, this model is often considered to have a heuristic but not a quantitative value (Orr 2005a). For example, its predictions regarding $f(s)$ have never been confronted to empirical distributions although qualitative predictions on the adaptive process have received empirical support (Burch and Chao 1999; Imhof and Schlotterer 2001; Rozen et al. 2002; Rokyta et al. 2005). Similarly, its predictions regarding $f(s)$ are rarely used within theoretical models (but see Poon and Otto 2000) because its lack of realism may then compromise the models' conclusions. Therefore, many theoretical approaches are limited to those that do not depend too much on $f(s)$ (Orr 2005b; Otto 2004) which is a desirable property but potentially restrains the scope of theoretical investigation: when variation in $f(s)$ affects theoretical predictions, the problem is still ultimately to find what $f(s)$ really is. Paradoxically, the problem of $f(s)$ has been much more discussed for the statistical analysis of mutation accumulation experiments (Keightley 1994; Keightley and Lynch 2003; Shaw et al. 2003), without much guidance from a ''realistic'' theoretical expectation. More generally, although several papers sought to explain the variation in rates of mutation across taxa (Drake et al. 1998; Lynch et al. 1999; Keightley and Eyre-Walker 2000), variation in its fitness effect and its potential causes has received much less attention (but see Lynch et al. 1999; Bataillon 2000). The main goal of this paper is

[2] Present address: Department of Ecology and Evolution, Université de Lausanne, CH1015 Lausanne, Switzerland; E-mail: guillaume.martin@unil.ch.

to describe this variation with a survey of the empirical literature, and to analyze it in the light of a generalized version of Fisher's model.

Building and testing a model for $f(s)$ necessarily involves simplifying assumptions because an exhaustive description of the effect of all possible mutations on phenotype and their fitness consequences is obviously intractable. Fisher's model makes several specific simplifying assumptions that have been criticized. In this paper, we propose a model relaxing some of these assumptions, but before describing our approach, we discuss the main criticisms of Fisher's model that have been put forward. This brief presentation draws on previous discussions of this issue (Orr 1998, 2005a; Poon and Otto 2000; Welch and Waxman 2003).

Fisher's Model Realism

Optimum and fitness

The first key assumption of the model is that there is a single phenotypic optimum, so that it can only be used to model stabilizing selection (around a single optimum) or directional selection (away from the optimum). Quantitative genetic studies (reviewed in Kingsolver et al. 2001) and experimental evolution experiments (reviewed in Elena and Lenski 2003) provide evidence for both types of selection, but also for disruptive selection with alternative possible optima. However, using Fisher's model to predict $f(s)$ only requires that most of the range of possible mutations lies on the slope of a single local fitness peak. It may indeed be quite rare that a population stays long in a fitness valley. In any case, predicting local $f(s)$ is different from predicting long term evolutionary change, and Fisher's model may be less appropriate for the latter than for the former. Similarly, although the pattern of adaptation cannot be directly predicted with a moving optimum (Orr 2005a), $f(s)$ can still be predicted at any given time. An additional restriction is that a given fitness function must be chosen to map fitness with the phenotypic distance from the optimum. When close to the optimum, a quadratic or Gaussian fitness function is a straightforward local approximation for many arbitrary fitness functions (Lande 1980) and is therefore widely used. Overall, these assumptions of Fisher's model on the way phenotype determines fitness are not so unrealistic when considering a population close to a local optimum, but may be less accurate under strong environmental change.

Mutation and traits

In Fisher's model, fitness is determined by the phenotype, which is made of a set of n phenotypic traits describing adaptations (hereafter ''adaptation traits''). A more specific assumption is that the effect of mutation on these traits is continuous and symmetric. This assumption is also used in quantitative genetics and is useful for the mathematical analysis. It has been criticized (Clarke and Arthur 2000), but is consistent with some empirical evidence (Garcia-Dorado et al. 1999; Orr 2001); for example, based on the effect of single P-element inserts on bristle numbers (Lyman et al. 1996). There is obviously a qualitative difference between such adaptation traits for which we can reasonably assume symmetrical mutation effects and fitness traits (e.g., survival, fecundity, functional efficiency traits, etc.) which are known to decrease on average by mutations, since most mutations are deleterious (Lynch et al. 1999). For example, molecular properties of the active site of an enzyme (e.g., volume, charge, etc.) can be reasonably expected to change symmetrically by mutation. There are indeed no clear reasons why mutations should bias toward higher or lower charge or volume; and if such a bias is present, it is likely to be small compared to the mutational variance. In contrast, when considering a trait that measures a distance from a given value (e.g., departure from the charge or volume corresponding to maximal affinity), mutation effects cannot be symmetric since a distance is never negative: both an increase and a decrease in charge from the optimal value will decrease affinity. The same difference applies between adaptation traits (the axes in Fisher's model) and fitness traits: the former have an optimal value which is defined by the latter. The aim of the model is precisely to predict the distribution of the effect on fitness traits of mutations affecting adaptation traits.

The distribution of mutation phenotypic effects is also often assumed to be Gaussian (we will use this assumption too), again mainly for mathematical convenience. Empirically, this assumption is approximately valid for some traits, but clearly fails for others (Garcia-Dorado et al. 1999). We suggest here that the measure of each trait is somehow arbitrary, so that, with appropriate scaling, it is possible to reduce kurtosis as needed (in the same way as transforming a response variable in a linear model, to conform to the hypothesis of normal errors). Such a scaling would affect the fitness function accordingly, but this new fitness function could still be approximated by a quadratic function close to the optimum. In all that follows, it will be important to keep in mind the distinction between the effect of mutations on phenotype (i.e., on adaptation traits) and on fitness (i.e., $f(s)$): the former will be assumed Gaussian, whereas the latter will be predicted by the model.

A perhaps stronger assumption of Fisher's model is that a single mutation can potentially affect all the phenotypic traits of the organism (universal pleiotropy). There is empirical evidence showing that pleiotropy and compensatory mutations are widespread (reviewed in Poon and Otto 2000). However, studies of development and genetic regulatory networks strongly suggest that the genotype-phenotype map may be organized into modules. As a consequence, it is often argued that compensation can only occur within modules or that a change within a module leaves adaptation in other modules undisturbed (Wagner and Altenberg 1996). Although this view is certainly correct for phenotypic compensation, it may not hold when considering fitness compensation. For instance, if two traits in different modules are correlated in their effect on fitness, then the fitness effect of a mutation in a given module may be compensated by mutation in another module. Similarly, with selective correlation among traits in different modules, a mutation in a given module can cause maladaptation in another module. The modularity of mutation effects on phenotypes does not ensure the modularity of their fitness effects. In any case, Fisher's model only requires a description of the net phenotypic effect of all mutations averaged over modules and can thus accommodate

modularity or partial pleiotropy (as in Welch and Waxman 2003).

Finally, the distribution of the phenotypic effects of mutations in Fisher's model is always assumed to be independent of the genotype in which they arise (the genetic background). This assumption is not a problem when considering the distribution of mutations effects on a single genotype. However, considering that mutational effects vary with the background may alter long-term predictions in which the background changes. There is qualitative support for the idea that a given allele has similar phenotypic effects when introduced in different genetic backgrounds as demonstrated by the success of genetic engineering and improvement of domesticated species. It is probably quantitatively inexact, but remains a good working assumption, in the absence of alternative models.

Symmetry assumptions

Another strongly restrictive assumption in Fisher's model is that all traits are equivalent and independent of each other with respect to both mutation and selection (Welch and Waxman 2003). This spherical symmetry is indeed an oversimplification that can hardly apply to real organisms. Given appropriate scaling, it is possible to account for some correlations between traits, but such scaling can only be performed for either selection, or mutation effects, not for both (Orr 1998, 2005a; Poon and Otto 2000). There is empirical evidence showing that phenotypic correlations between traits are widespread for both selection (Kingsolver et al. 2001) and mutation (Garcia-Dorado et al. 1999; Keightley et al. 2000), so that relaxing the symmetry assumption would significantly improve the model's realism and the range of its potential applications. Recently, Waxman and Welch (2005) have proposed the first model to account for selective interactions between traits, but still neglecting mutational correlations. One of the aims of this paper is to relax the symmetry assumptions for both mutation and selection (using a different approach from that of Waxman and Welch 2005).

Scaling and empirical predictions

The last problem when trying to use Fisher's model is to find an appropriate scaling in terms of measurable quantities. For instance, we ignore the distribution of the size of mutation in the phenotypic space, the phenotypic distance to an optimum, or the number n of adaptation traits. As a consequence, even if it has a strong heuristic value, the whole model may appear arbitrary and of little use when it comes down to real and testable predictions (Orr 2005a). Fisher's model can make predictions that are scale independent and testable; for instance, regarding the distribution of factors fixed during a bout of adaptation. These predictions have received empirical support (Imhof and Schlotterer 2001; Rokyta et al. 2005). However, the agreement is qualitative and does not unambiguously support Fisher's model since several models make the same predictions (Orr 2005a).

Another prediction of the model with perhaps important evolutionary implication is that the number of traits, n, influences the fitness effect distribution of mutations in such a way that beneficial mutations are less likely and less favorable in more complex organisms (Orr 1998; Barton and Partridge 2000). An important consequence of this is that more complex organisms should adapt at slower rates, an effect dubbed ''cost of complexity'' (Orr 2000). This prediction has not been tested empirically. Similarly, it has not been tested whether the distribution of mutation fitness effects varies with complexity as predicted by this model because testing this prediction would require scaling mutation fitness effects to measurable quantities.

Possible Improvements and Tests

The first aim of this paper is to propose a model predicting the fitness effect distribution of mutations without assuming equivalence and independence between traits. This model attempts to predict in a simple analytic form how the moments of $f(s)$ should vary with phenotypic complexity (i.e., n, the number of adaptation traits under selection) and with the level of covariation between traits. We then make approximations that allow one to use the empirical distribution of mutation effects measured in a given environment (e.g., in the laboratory) to predict the new distribution in another environment. The second aim of this paper is to survey the available empirical data on $f(s)$ across taxa to test our model's predictions with the appropriate scaling. More precisely, we use gene number as a surrogate estimate of n and test for correlations between gene number and empirical moments of s across species. The third aim of this paper is to use our model and survey to quantify the cost of complexity by comparing predicted rates of adaptation across species, based on their empirical distribution of deleterious mutations. Overall, our survey and model indicate that complexity—as measured by gene number—and phenotypic correlations are critical factors shaping the fitness effect of mutations across taxa.

Model and Predictions

As with other models based on Fisher's geometric approach, we consider that fitness is determined by n adaptation traits. As explained in the introduction, we use a Gaussian distribution of mutation phenotypic effects on these traits. Furthermore, we assume that these traits are under Gaussian stabilizing selection around a fixed phenotypic optimum, which will work best when close to the optimum. These Gaussian assumptions allow the mathematical treatment of the model for any arbitrary selective and mutational covariance matrices for adaptation traits. This model of multivariate stabilizing selection and mutation is similar to that introduced by Zhang and Hill (2003) for the study of mutation-selection balance on a quantitative trait, but extended to account for beneficial mutations (i.e., phenotypes are not necessarily at their optimum). The presentation of the model takes several steps: we first derive the exact distribution of s under our assumptions, then we give a general exact expression for the moments of $f(s)$ at the optimum (i.e., when there are only deleterious mutations). Then we formulate testable predictions on the effect of n on these moments. Next, we derive an approximation for the probability density function of s, $f(s)$ at any distance to a new optimum defined by a new environment. We show how the parameters of this distribution can be estimated empirically and we use this approximation to compute the rate of adaptation in this new envi-

ronment. Next, because all our theoretical results depend on the strength of mutational and selective correlations between traits, we introduce a null model of phenotypic interactions to evaluate their influence on $f(s)$, based on random matrix theory. Next, we present simulations to validate our approximations. Finally, we confront our predictions with empirical data from the literature.

Description of the Model

Phenotypes are modeled as a set of n continuous phenotypic traits represented by a column vector $\mathbf{z}$. The fitness $W(\mathbf{z})$ of phenotype $\mathbf{z}$ is a multivariate Gaussian function of the distance between $\mathbf{z}$ and a phenotypic optimum that is set to zero for all traits, without loss of generality: $W(\mathbf{z}) = \text{Exp}(-\frac{1}{2}\mathbf{z}^t\mathbf{S}\mathbf{z})$, where t denotes transposition. $\mathbf{S}$ is the $n \times n$ matrix of the selective effects of all traits. Diagonal elements in $\mathbf{S}$ measure the selection intensity on each trait while nondiagonal elements measure selective interactions between trait pairs. To describe the most general situation of stabilizing selection, we assume that $\mathbf{S}$ is positive semidefinite (not strictly definite as often assumed), which ensures that no phenotype has a higher fitness than phenotype $\mathbf{z} = 0$. Note that we will only consider traits under direct stabilizing selection (i.e., with a strictly positive diagonal term), although some linear combinations of these traits may be neutral because of selective interactions.

Consider now the effect of mutation on a single genotype, referred to as the initial genotype (with phenotype $\mathbf{z}_o$). We assume that the distribution of mutant phenotypes around $\mathbf{z}_o$ is multivariate Gaussian with mean zero and covariance matrix $\mathbf{M}$. As explained in introduction, this Gaussian assumption is valid as long as there is a transformation from the ''real'' traits to a set of traits that are distributed as a multivariate Gaussian. As for selection, the model allows both for differences in mutational variances across traits and for mutational correlations between traits (nondiagonal elements in $\mathbf{M}$). We assume that $\mathbf{M}$ is positive semidefinite (the most general structure for a covariance matrix). The selection coefficient s of a mutant phenotype $\mathbf{z}_o + \mathbf{dz}$ is defined relative to the initial phenotype $\mathbf{z}_o$ as $W(\mathbf{z}_o + \mathbf{dz})/W(\mathbf{z}_o) - 1$. We assume that s is small enough that $s \approx \log(1 + s)$ and we define $s_o \equiv -\log(W(\mathbf{z}_o)/W(\mathbf{0}))$, the selective disadvantage of the initial phenotype $\mathbf{z}_o$ relative to the optimal phenotype. Under these assumptions, the joint effects of all selective and mutational covariances (matrices $\mathbf{M}$ and $\mathbf{S}$) reduce to the n eigenvalues of the product $\mathbf{S}\cdot\mathbf{M}$. The exact distribution of s is a quadratic form in Gaussian vectors (Mathai and Provost 1992). The distribution is entirely determined by the distance to the optimum s_o, the direction to the optimum in the phenotypic space, and the n eigenvalues $\{\lambda_i\}_{i\in[1,n]}$ of $\mathbf{S}\cdot\mathbf{M}$, as shown in Appendix 1, available online only at http://dx.doi.org/10.1554/05-412.1.s1. Each λ_i corresponds to a phenotypic direction (a linear combination of biological traits z_i) on which mutation and selection act independently with a net effect λ_i on fitness. Thus, a large λ_i corresponds to a combination of traits that displays a large mutational variance and is under strong selection.

Exact Moments of f(s) at the Optimum

A distribution can be fully characterized by its moments. Since the central moments of quadratic forms of Gaussian vectors have analytic expression for any order (Mathai and Provost 1992; online Appendix 1), $f(s)$ is fully specified in our model. However, we focus on the first three central moments of s that are the most available empirically. When most mutations are deleterious, it can be assumed that the initial genotype is close to the optimum ($s_o \ll 1$). Then, defining the raw moments of the λ_i across traits i as $\overline{\lambda^r} \equiv 1/n \sum_{i=1}^{n} \lambda_i^r$, the three first central moments ($E(s)$, $V(s)$, and $\mu_3(s)$) of $f(s)$ are given by:

$$E(s) = -\frac{n}{2}\bar{\lambda}$$

$$V(s) = \frac{n}{2}\overline{\lambda^2}$$

$$\mu_3(s) = -n\overline{\lambda^3}. \qquad (1)$$

A more general expression for any distance to the optimum (i.e., with beneficial mutations) can be found in equation (A3) of Appendix 1 (available online). At the optimum, the moments of s depend only on the number of phenotypic traits n and on the distribution of the eigenvalues of $\mathbf{SM}$ (the λ_i).

Predicting the Effect of n on the Moments of f(s)

Equation (1) yields simple predictions on the three first moments of empirical distributions of mutation effects when there are few beneficial mutations ($s_o \ll 1$).

First, $E(s) = -n\bar{\lambda}/2$, so that the average deleterious effect of mutations should be larger in organisms with a presumably larger number of traits (e.g., in *Drosophila* vs. *Escherichia coli*). Larger $E(s)$ in *Drosophila* than in *E. coli* could be due either to a larger n (larger number of traits under selection in fruit flies than in bacteria) or a larger $\bar{\lambda}$ (same number of traits in both species, but a larger effect of mutation on each trait, in *Drosophila*) but the latter seems much less parsimonious. Note that, in addition, the expression for $E(s)$ in equation (1) is still valid when $s_o \neq 0$, that is, if beneficial mutations also occur (see eq. A3, Appendix 1 online). This outcome results directly from the Gaussian approximation (quadratic in log scale) of the fitness function: at the optimum, all mutations are weakly deleterious, whereas, when away from the optimum, deleterious mutations are more severe but compensated by some beneficial mutations. The net outcome depends only on the local curvature of the fitness function around the initial phenotype. This curvature is constant at any distance from the optimum with our quadratic fitness function, so that $E(s)$ is independent of s_o.

Second, rearranging equation (1) shows that both the coefficient of variation and the skewness of s should decrease with n. More precisely, if we note $\mu_2^* = V(s)/E(s)^2$ and $\mu_3^* = \mu_3(s)/E(s)^3$, the second and third moments of s scaled to the mean effect $E(s)$, we obtain from equation (1) two independent quantities that should increase linearly with n:

$$\frac{1}{\mu_2^*} = n\frac{\bar{\lambda}^2}{2\overline{\lambda^2}} \quad \text{and} \quad \frac{1}{\sqrt{\mu_3^*}} = n\sqrt{\frac{\bar{\lambda}^3}{8\overline{\lambda^3}}}. \qquad (2)$$

These predictions are based on scale invariant measures (scaled to $E(s)$), which should make them more robust for comparisons across species. Again, finding that $1/\mu_2^*$ and $1/\sqrt{\mu_3^*}$ are larger, for example in *Drosophila* than in *E. coli*, is more likely due to a larger number of traits in *Drosophila* than to variation in the distribution of the λ_i between these species.

We propose below (see A Model of Random Phenotypic Correlations) a null model for **S** and **M** to better understand the possible dependence between n and the distribution of the λ_i. Under this model, $E(s)$ increases linearly with n while $1/\mu_2^*$ and $1/\sqrt{\mu_3^*}$ increase but plateau with large n. If this model holds, we should indeed find larger $E(s)$ in more complex organisms (first prediction) but may not detect such an increase for $1/\mu_2^*$ and $1/\sqrt{\mu_3^*}$ (second prediction).

Distribution of Deleterious Effects and ''Effective Complexity'' n_e

Approximating the probability density of s is required for computing the rate of adaptation and may be useful in maximum likelihood analyses of empirical distributions of s. Equation (A2) in Appendix 1 (online), gives the exact distribution in the general case but its probability density is not known. However, an approximation can be obtained from the moments of s given above, using the moment matching method, which requires choosing an a priori distribution for the density. We chose the negative gamma distribution because it is the exact distribution of s corresponding to the simplest situation: when all $\lambda_i = \lambda$ are equal and $s_o = 0$, $f(s)$ is a negative gamma distribution with scale λ and shape $n/2$. Staying at the optimum ($s_o = 0$) but with the λ_i varying across traits, with coefficient of variation $CV(\lambda)$, $f(s)$ can be approximated by a negative gamma with scale $\lambda_e = \bar{\lambda}(1 + CV(\lambda)^2)$ and shape $n_e/2$ where n_e is the effective number of traits. n_e is defined as the number of traits that would generate the same mean and variance of $f(s)$ (i.e., the same parameters for the approximate gamma distribution), if all traits where independent and of equal effect (λ_e) as in the original Fisher model. From equation (1), it is given by

$$n_e = \frac{n}{1 + CV(\lambda)^2}. \tag{3}$$

This n_e is lower than n and decreases relative to n when the heterogeneity among traits (measured by $CV(\lambda)$) increases. This effect is strongest when only a few linear combinations of traits display both large mutational variance and are under strong selection (i.e., correspond to major λ_i). A small n_e means that the distribution of s is highly skewed, while a high n_e corresponds to more symmetrical distributions, closer to the Gaussian.

Based on a distribution of deleterious mutation effects (i.e., at the optimum $s_o = 0$), both n_e and λ_e can be directly estimated from the mean and variance of s as

$$n_e = 2\frac{E(s)^2}{V(s)} \quad \text{and} \tag{4a}$$

$$\lambda_e = \frac{V(s)}{-E(s)}. \tag{4b}$$

This will be used later to estimate the effective number of traits n_e across species from empirical distributions of deleterious mutation effects.

Predicting f(s) Away from the Optimum

When the initial genotype is at the optimum, empirical distributions of deleterious mutation effects can be used to estimate n_e and λ_e. It is thus tempting to ask whether we can then use this information to predict $f(s)$ in any new environment, in which the initial genotype is not at the optimum but at a distance s_o from the optimum. When the initial genotype is away from the optimum ($s_o > 0$), we use a similar approximation as above, and $f(s)$ becomes a ''displaced gamma'' (Shaw et al. 2002)—the sum of a negative gamma and the constant s_o: $s = s_o - \gamma$, where γ is approximately gamma distributed with scale α and shape β (see online Appendix 1). We will denote $f_\Gamma(s)$ this approximation for the probability density $f(s)$ of s, because it rests on the approximation that the random part γ of the distribution of s is a gamma deviate:

$$f_\Gamma(s) = \frac{e^{-(s_o - s)/\alpha}(s_o - s)^{\beta-1}\alpha^{-\beta}}{\Gamma(\beta)}. \tag{5}$$

Such a distribution can account for both advantageous and deleterious mutations in a continuous manner and with only three parameters and can be implemented in maximum likelihood analysis of empirical $f(s)$ (Shaw et al. 2002). As for the case $s_o = 0$, α and β must be set to match the mean and variance of s away from the optimum $E(s)$' and $V(s)$'. The resulting approximation is accurate but depends on both the distance to the optimum s_o and the particular direction $\mathbf{z}_o$ of the initial genotype (see eq. A3 in online Appendix 1). The distance to the optimum s_o could be estimated in principle (e.g., using long-term experimental evolution), whereas the direction ($\mathbf{z}_o$) may not be measurable. As a consequence, this approximation may be of little interest. Fortunately, $f(s)$ does not vary too much with the direction of $\mathbf{z}_o$, so that we can find a less accurate approximation for $f(s)$ that depends only on the distance to the optimum s_o and on the moments of $f(s)$ at the optimum (eq. A4 in online Appendix 1). The resulting α and β of the displaced gamma approximation in equation (5) can then be predicted based on estimable quantities (λ_e, n_e, and s_o) yielding

$$\beta = \frac{n_e}{2}\frac{(1 + \varepsilon)^2}{1 + 2\varepsilon} \quad \text{and} \tag{6a}$$

$$\alpha = \lambda_e\frac{1 + 2\varepsilon}{1 + \varepsilon}, \tag{6b}$$

where n_e and λ_e can be estimated from a distribution of deleterious effects (equation 4) , and $\varepsilon = s_o/|E(s)| = 2s_o/(n_e\lambda_e)$ is the distance to the optimum relative to the average fitness effect of a mutation $E(s)$. With this approximation, it is possible to predict $f(s)$ in a new environment (with a given $s_o > 0$) from the mean and variance of mutation effects measured close to the optimum (i.e., on deleterious mutations). Importantly, this means that we can predict the approximate $f(s)$ for any species in which deleterious mutation effects have been measured and in an environment for which s_o is known.

Rates of Adaptation and Cost of Complexity

The approximation $f_\Gamma(s)$ above (eq. 5) can now be used to compute the rate of adaptation to a new environment (corresponding to a given s_o), defined as the per generation increase in mean fitness, $d\bar{W}/dt$. In a population of very large size N, only beneficial mutations ($s > 0$) reach fixation (with probability $2s$, we ignore possible complications due to linkage) so that using the displaced gamma approximation yields:

$$\frac{d\bar{W}}{dt} \approx NU \int_{s>0} 2s^2 f_\Gamma(s)\, ds = N\mu E(s)^2 F(n_e, \varepsilon), \qquad (7)$$

where U is the per generation per genome mutation rate. $E(s)$, $\varepsilon = s_o/|E(s)|$, and n_e have been defined above. The function $F(\cdot)$ has a complicated expression but can be computed simply from equations (5) and (7) Interpretation of equation (7) is consistent with previous studies based on the Fisher-Orr model (Orr 2000; Welch and Waxman 2003). First, F is an increasing function of ε, meaning that the rate of adaptation increases with the maladaptation of the initial genotype. Second, F decreases with increasing complexity n: this is Orr's (2000) cost of complexity. The effect of n on the rate of adaptation may also be influenced by any covariation of other parameters (N, U, s_o, $E(s)$) with n. In particular, $|E(s)|$ may increase with n, as suggested by equation (1) (and confirmed by our survey below). This has antagonistic effects on the rate of adaptation by increasing $E(s)^2$ but decreasing ε. Most importantly, our model shows that it is n_e, not n, that determines the rate of adaptation. Therefore, including heterogeneity between traits greatly reduces the cost of complexity by reducing the effective number of traits n_e. Overall, variation in $f(s)$ affects the rate of adaptation by changing $E(s)^2 F(n_e, s_o/E(s))$, in which both $E(s)$ and n_e can be estimated from empirical distributions of deleterious mutation effects (see eq. 4).

A Model of Random Phenotypic Correlations

We considered so far arbitrary mutational and selective matrices **S** and **M**. To understand the relationship between the strength of phenotypic correlations and $f(s)$, we propose here a model of random interactions between many traits, using results from random matrix theory, a mathematical tool widely used in physics and finance to model complex interactions (Forrester et al. 2003). The available evidence suggests that both positive and negative phenotypic correlations are widespread (Lynch and Walsh 1998), so we considered a case where correlations of both signs are equally probable. Mutational and selective covariance matrices **S** and **M** can, for example, be assumed to be drawn randomly into independent Wishart distributions (a classic model for random covariance matrices, see Appendix 2, available online only at http://dx.doi.org/10.1554/05-412.1.s2). These matrices are built by drawing the elements of a first matrix into the standard Gaussian distribution, and multiplying it by its transpose to obtain a symmetric positive semidefinite matrix. Standard Wishart matrices contain a random set of both positive and negative correlations with a zero average. If the number of traits is sufficiently large (e.g., $n > 15$), the distributions of phenotypic correlations in **S** and **M** converge to a simple distribution with known probability density (see eq. A5 in online Appendix 2). Let ρ_S and ρ_M be the standard deviations of these asymptotic distributions around zero. ρ_S and ρ_M measure the strength of selective and mutational correlations averaged over all traits: a large ρ_S (respectively, ρ_M) means that there are many large correlations (of any sign) within matrix **S** (respectively, **M**). Similarly with large n, the distribution of the eigenvalues of $\mathbf{S}\cdot\mathbf{M}$, for any random draw of **S** and **M**, converges to an asymptotic distribution with known moments such that $CV(\lambda)^2 = n(\rho_S^2 + \rho_M^2)$ (online Appendix 2). We can then directly obtain an expression of n_e (from eq. 3) in terms of phenotypic correlations:

$$n_e = \frac{n}{1 + n(\rho_S^2 + \rho_M^2)} \xrightarrow{n\to\infty} \frac{1}{\rho_S^2 + \rho_M^2}. \qquad (8)$$

Equation (8) shows that with random selective and/or mutational correlations between traits, trait heterogeneity $CV(\lambda)$ increases with the number of traits n. This effect drastically reduces n_e and hence the cost of complexity: as n increases indefinitely, n_e reaches a plateau $1/\rho_S^2 + \rho_M^2$ that depends only on the strength of phenotypic correlations and can be very small. This behaviour is consistent with the observation that empirical distributions of s are typically more asymmetric than the Gaussian. This suggests that n_e is typically small although n is expected to be very large in most species.

Simulations and Scaling Issues

Simulations were run using the software R (Ihaka and Robert 1996) to jointly check the displaced gamma approximation for $f(s)$ in equation (5), and the approximation for n_e based on random matrix theory in equation (8). Mutational and selective covariance matrices (**S** and **M**) were randomly drawn as Wishart matrices with fixed correlation strength ρ_S and ρ_M, and scaled to obtain a given value of $E(s) = -\frac{1}{2}\mathrm{Tr}(\mathbf{S}\cdot\mathbf{M}) = -n\bar{\lambda}/2$, where Tr(.) denotes matrix trace. The phenotypic distance to the optimum $\mathbf{z}_o$ was then drawn as a vector of n independent Gaussian deviates $n(0,0.1)$, and scaled to obtain a given value of $s_o = \frac{1}{2}\mathbf{z}_o^t\mathbf{S}\mathbf{z}_o$. Each mutant phenotype was then drawn from a multivariate Gaussian with mean zero and covariance **M** and the corresponding s was computed following the exact distribution in equation (A1) of Appendix 1 (available online).

Figure 1 shows how the approximate distribution matches simulations of the exact distribution of s when mutational and selective covariances (**M** and **S**) are drawn randomly. The approximation fits the exact distribution very well when only deleterious effects are considered ($s_o = 0$, Fig. 1a for $n = 40$ traits). The shape and scale of the gamma approximation on Figure 1 were computed using the asymptotic result from random matrix theory: $CV(\lambda)^2 = n(\rho_S^2 + \rho_M^2)$. This approximation is almost as accurate as if using the exact $CV(\lambda)$ computed from the eigenvalues of simulated matrices. The left top panels in Figure 1a also show the distribution of phenotypic correlations within the simulated Wishart matrices together with the predicted asymptotic distribution from equation (A5). The prediction remains accurate even with a limited number of traits (e.g., $n = 15$, not shown). Finally, Figure 1 shows that stronger correlations can considerably reduce n_e, resulting in more skewed $f(s)$ (a gamma

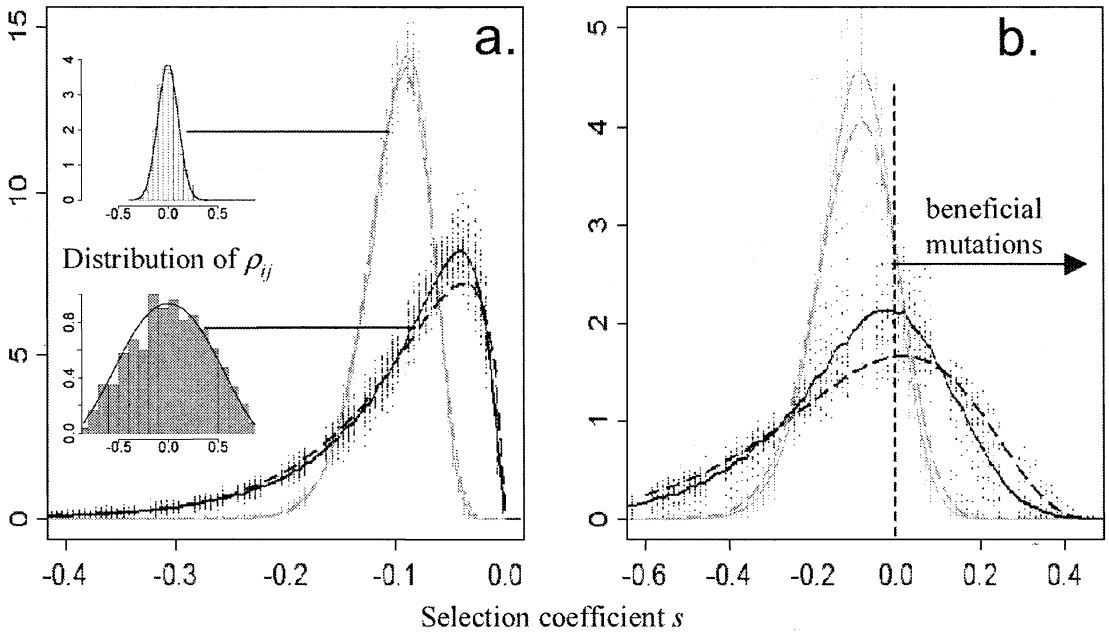


FIG. 1. Distribution of s with random phenotypic covariances and displaced-gamma approximation. Distribution of mutation fitness effect with $n = 40$ traits and the initial genotype perfectly adapted (a; $s_o = 0$) or away from the optimum (b; $s_o = 0.5$; i.e., the fitness of the initial genotype is 50% that of the best genotype). Dots show the density of s from 50 simulations (each with 20,000 mutants and independent draws of **S** and **M** as 40×40 Wishart matrices, see A Model of Random Phenotypic Correlations). Plain lines show the mean over these 50 simulated densities. In each panel, $f(s)$ is illustrated with strong (black) or weak (gray) phenotypic correlations, whose distributions are illustrated with histograms. Strong and weak phenotypic correlations are $\rho_S = \rho_M = 0.38$ (corresponding to $n_e = 3.2$) or 0.1 (corresponding to $n_e = 22$), respectively. Dashed lines show the displaced gamma approximation in equation (5) with n_e values computed as in equation (8), $n_e = n/(1 + n(\rho_S^2 + \rho_M^2))$.

with smaller shape), at the optimum ($s_o = 0$, Fig. 1a) or away from it ($s_o = 0.5$, Fig. 1b).

When $s_o > 0$ (e.g., $s_o = 0.5$ in Fig. 1b.) the displaced gamma approximation for $f(s)$ is reasonably accurate, although it tends to overestimate the proportion of advantageous mutations. The discrepancy is almost entirely due to our approximation for the variance of s at a given fitness distance to the optimum s_o (eq. A4 in online Appendix 1). Using the exact expression for this variance (eq. A3 in online Appendix 1) results in a very accurate displaced gamma approximation. However, this more accurate approximation involves some nonestimable quantities and was thus discarded here.

TESTING THE MODEL WITH EMPIRICAL F(S)

Survey of the Moments of Empirical f(s) across Species

Principle of the approach

To test our predictions on the effect of n on the moments of s (eq. 2), we compared the moments of empirical distributions of mutational fitness effects across species ranging from viruses to higher plants using the number of protein genes per haploid genome as a surrogate measure of complexity. We chose this measure mainly because it is the only one available for all the species considered in our survey, and to avoid any ranking of species according to an a priori qualitative ''complexity.'' However, we do not equate genes and traits in this approach. By definition, a single mutation affects only one gene sequence but may have many pleiotropic effects on other gene products, so that a simple proportional relationship between gene number and phenotypic complexity is likely to be rather unrealistic. Instead, we assume that, at the phylogenetic scale of our comparisons, gene number is positively correlated with the number of adaptation traits, because gene number is expected to determine the number of gene products and their interactions, and because the evolution of new functions is currently sought to occur by the addition of new genes (Vassilieva et al. 2000). This quantitative measure also provides an explicit way to describe the (dis-)similarity between species more precisely and objectively than some a priori attributed rank of complexity between species. We therefore test our model only by assuming some ordered relationship: the larger the gene number, the larger the number of traits. Such a relationship may not hold at a finer scale of comparison (e.g., the plant *Arabidopsis thaliana* has more genes than the fruit fly *D. melanogaster*, but which has the largest number of traits under selection?). However, it does not seem too unrealistic when comparing viruses, microbes and higher organisms.

Estimates of $E(s)$ were drawn from the literature on mutation accumulation (MA), discarding cases with a large proportion of beneficial mutations, suggesting that the initial genotype was far from the optimum, in which case our Gaussian fitness function may not apply. Our survey updates the review by Bataillon (2000) including recent studies on *Drosophila, Caenorhabditis,* vesicular stomatitis virus (VSV), *Saccharomyces cerevisiae* and the fungus *Cryptococcus neoformans,* to a total of 33 $E(s)$ estimates in eight taxa. Esti-

TABLE 1. Gene count and average effect of mutations from mutation accumulation (MA) experiments. VSV, vesicular stomatitis virus; spont, accumulation of spontaneous mutation; mmr, MA on mismatch repair deficient strains; ems, MA using EMS mutagenesis. II or III refer to mutations on the second or third chromosomes in *Drosophila*. The fitness traits measured were: relative viability in competition (viab), lifetime reproductive success (LRS), intrinsic growth rate (r) or relative growth rate in competition (RGR). GC is haploid gene count. $E(s)$ is the average haploid or homozygous effects of mutations.

Species	Method	Trait	GC	$E(s)$	Reference
VSV	spont	RGR	5	−0.0022	Elena and Moya 1999
VSV	spont	RGR	5	−0.0024	Elena and Moya 1999
VSV	spont	RGR	5	−0.0022	Elena and Moya 1999
Escherichia coli	spont	*r*	4970	−0.034	Loewe et al. 2003
E. coli	spont	*r*	4970	−0.012	Bataillon 2000
Saccharomyces cerevisiae[1]	mmr	*r*	5855	−0.075	Zeyl and DeVisser 2001
Cryptococcus neoformans[2]	spont	*r*	6475	−0.045	Xu 2004
Drosophila melanogaster	spont	viab	16,130	−0.16	Bataillon 2000
D. melanogaster	spont (II)	viab	16,130	−0.11	Bataillon 2000
D. melanogaster[3]	spont (II)	viab	16,130	−0.03	Bataillon 2000
D. melanogaster[3]	spont (II)	viab	16,130	−0.03	Bataillon 2000
D. melanogaster	spont (II)	RGR	16,130	−0.1	Avila and Garcia-Dorado 2002
D. melanogaster	spont (II)	viab	16,130	−0.08	Chavarrias et al. 2001
D. melanogaster[4]	spont (III)	viab	16,130	−0.1	Charlesworth et al. 2004
D. melanogaster[5]	ems	viab	16,130	−0.11	Keightley and Ohnishi 1998
Caenorhabditis elegans	spont	*r*	21,357	−0.1	Bataillon 2000
C. elegans	spont	*r*	21,357	−0.2	Bataillon 2000
C. elegans	mmr	*r*	21,357	−0.413	Estes et al. 2004
C. elegans	ems	*r*	21,357	−0.15	Keightley et al. 2000
C. elegans	spont	*r*	21,357	−0.364	Baer et al. 2005
C. elegans	spont	*r*	21,357	−0.25	Baer et al. 2005
C. briggsae	spont	*r*	<21,357	−0.1	Baer et al. 2005
C. briggsae	spont	*r*	<21,357	−0.198	Baer et al. 2005
Arabidopsis thaliana	spont	LRS	28,159	−0.23	Bataillon 2000
Triticum durum	spont	LRS	40,000	−0.2	Bataillon 2000

[1] M grande lines in Zeyl and DeVisser (2001), that is, keeping only ''petite'' mutations in mmr strain. The normal ''F'' strain produced only a single ''grande'' mutant and is not reported here. Only heterozygous effect $E(hs)$ are reported in Zeyl and DeVisser (2001). Homozygous effects in the table are corrected using $h = 0.2$ based on dominance estimates for point mutations in *S. cerevisiae* (Korona 2004).
[2] Average effect in optimal environment (YEPD/37°) from all MA lines in table 2 of Xu (2004).
[3] Mukai-Ohnishi studies.
[4] Average $E(s)$ for all nonlethal mutations in table 4 of Charlesworth et al. (2004).
[5] Maximum likelihood estimate in table 4 of Keightley and Ohnishi (1998).

mation of higher moments of s is very difficult using MA data (Lynch et al. 1999; Keightley 2004), so only few estimates of $V(s)$ and $\mu_3(s)$ are available. To obtain them, we surveyed empirical studies that directly measured the distribution of fitness among lines carrying a single mutation, so that the observed distribution of mutant fitnesses directly gives estimates of the first moments of s. We found nine estimates from five taxa.

Gene number

For most species, we used the number of open reading frames (available on the Kyoto Encyclopedia of Genes and Genomes [KEGG] website http://www.genome.jp). The value used for wheat (*Triticum durum*) was a recent estimate from the rice genome (40,000; Bennetzen et al. 2004) based on strong similarities between cereal genomes (Ware and Stein 2003). The number of protein genes in *C. briggsae* was considered equal to that of *C. elegans* for the statistical analysis (Stein et al. 2003). The number of protein genes in VSV is five (Sanjuan et al. 2004a).

Empirical moments of s

Estimates of $E(s)$ in Table 1 were obtained by surveying mutation accumulation experiments, using either Bateman-Mukai (BM) estimates or maximum likelihood estimates when the latter were significantly better (i.e., in Keightley and Ohnishi 1998; Vassilieva et al. 2000). We discarded two recent studies on yeast (Joseph and Hall 2004) and *Arabidopsis* (Shaw et al. 2002) in which a very large proportion of mutations were beneficial, suggesting that the initial genotype was far from the optimum. $E(s)$ values in Table 2 are direct measures from single mutation effects, except in two cases. For *C. elegans* we used the corrected BM estimate given in Vassilieva et al. (2000). For transposable element (TE) single inserts on chromosomes II and III of *Drosophila* (Lyman et al. 1996), $E(s)$ is biased by a direct TE effect and not given in this study. Therefore, we used the per insert viability effect of third chromosome TE insertions for chromosome III given by Mackay et al. (1992). For chromosome II, we used the average $E(s)$ of all second chromosome viability effects in Table 1. This average estimate was not, of course, included in the statistical analyses of $E(s)$. We did not include the Mukai et al. (1972) and Ohnishi (1977) results for the computation of this average $E(s)$, because their validity has been questioned (Garcia-Dorado et al. 1999), but we did include them in the statistical analysis of $E(s)$ estimates.

We surveyed estimates of higher moments of s from single mutations and from one study reporting a precise estimate (i.e., with limited confidence interval) of CV(s) obtained by maximum-likelihood analysis of MA data in *C. elegans* (Vas-

TABLE 2. Empirical distributions of s and empirical estimates of first moments of $f(s)$. In all these studies, all moments of s (haploid or homozygous effects) are estimated directly using single mutation lines except in *Caenorhabditis elegans* (see Empirical Moments of s). Abbreviations as in Table 1 except subst, single point substitutions; TE, single transposable element insertion. μ_2^*, scaled variance (i.e., squared coefficient of variation) $CV(s)^2 = V(s)/E(s)^2$; μ_3^*, scaled third moment $\mu_3(s)/E(s)^3$.

Species	Method	Trait	GC	$E(s)^2$	$CV(s)^2$	μ_3^*	$\hat{n}_e = 2/\mu_2^*$	Reference
VSV[1]	subst	RGR	5	−0.139	1.87*	4.58	1.07	Sanjuan et al. 2004a
Escherichia coli	TE	*r*	4970	−0.0275	9.55	259	0.21	Elena et al. 1998
Saccharomyces cereviae[1]	spont	*r*	5855	−0.109	2.07	8.59	0.96	Wloch et al. 2001
S. cereviae[1]	TE	*r*	5855	−0.05	2.25	4.59	0.89	Thatcher et al. 1998
S. cereviae[1]	ems	*r*	5855	−0.171	1.80	3.69	1.11	Wloch et al. 2001
S. cereviae[1]	mmr	*r*	5855	−0.183	1.12	1.64	1.78	Wloch et al. 2001
Drosophila melanogaster	TE II	viab	16,130	−0.098	0.79	0.54	2.23	Lyman et al. 1996
D. melanogaster	TE III	viab	16,130	−0.122	0.90	1.03	2.52	Lyman et al. 1996
Caenorhabditis elegans[2]	spont	*r*	21,537	−0.12	0.77	?	2.58	Vassilieva et al. 2000

[1] Moments of s directly computed from fitness effect values in supporting information of Sanjuan et al. (2004a), table 1 of Thatcher et al. (1998), and provided by Wloch et al. (2001).
[2] Maximum likelihood estimate of μ_2^* and correspondingly corrected BM estimate of $E(s)$ given in Vassilieva et al. (2000).

silieva et al. 2000). Studies on *E. coli* (Elena et al. 1998), *S. cerevisiae* (Thatcher et al. 1998) and *D. melanogaster* (Lyman et al. 1996) used TE insertions to generate single mutations. Note that results on *Drosophila* are based on the viability effects of either second or third chromosome TE inserts (as indicated in Table 2), instead of the whole genome. In another study on *S. cerevisiae*, Wloch et al. (2001) used tetrad analysis to isolate single spontaneous or induced mutation events and measure their fitness effect in the haploid stage. In the VSV (Sanjuan et al. 2004a), single nucleotide substitutions are produced by site-directed mutagenesis. No or very few beneficial mutations were detected in these studies, except in VSV with 4% of advantageous mutations (Sanjuan et al. 2004a). The initial genotype was therefore considered well adapted to the laboratory environment ($s_o \ll E(s)$) so that equation (2) applies.

Statistical analyses

We tested our predictions on $E(s)$ using a linear model accounting for gene number and four other potentially confounding variables that could covary with gene number and produce false positive correlations. First, the measure of s might be biased between microbes and higher organisms because a more integrative measure of fitness is used in the former (e.g., growth rate over many generations) instead of fitness components (e.g., viability) over one generation in the latter. The level of integration of the fitness measure was included as an ordered variable with values: 4, growth rate in competition; 3, intrinsic growth rate; 2, lifetime reproductive success; and 1, viability. Second, MA experiments in microbes may underestimate $E(s)$ because selection within sublines is more likely when several generations occur between population bottlenecks (i.e., severely deleterious mutations may not be detected in MA on microbes; Kibota and Lynch 1996). Therefore, we included a factor discriminating microbes versus nonmicrobes to avoid detecting a correlation between $E(s)$ and gene number that would in fact reflect a bias in estimates between lower and higher organisms. Third, the type of mutation was included as a factor: spontaneous versus TE versus point mutations, the latter referring to experiments based on single nucleotide substitutions, or mutagenesis by EMS and mismatch repair deficiency, which are known to cause mainly point mutations (Wloch et al. 2001). Note that spontaneous mutations, as accumulated in standard MA experiments, are a mix of different types of mutations including TE and point mutations. Fourth, the type of estimate and method of mutation accumulation, that is, MA (Table 1) versus direct (Table 2) estimates of $E(s)$, was included as a factor. This factor discriminates between measures based on an unknown number of mutational events (MA) versus a single mutation per line (direct). The full model including pairwise interactions was simplified backward to isolate the significant factors. Regarding our predictions on μ_2^* and μ_3^*, we only tested for a correlation to gene number due to the limited number of estimates available. Finally, we did not correct for phylogenetic independence given the phylogenetic scale of our comparisons. Similarly, we did not pool different estimates for the same species, which would artificially mask the quite large within-species variation of the estimates. However, pooling estimates per species does not qualitatively alter our conclusions below.

The Effect of n on Empirical Moments of Deleterious Mutation Effects

As predicted from equation (1) $E(s)$ (33 estimates from Tables 1 and 2) increases with gene number, our surrogate measure of n ($R^2 = 0.33$, $F_{1,31} = 16.1$, $P = 0.0004$, Fig. 2), and the trend remains significant when discarding viruses ($R^2 = 0.27$, $F_{1,27} = 10$, $P = 0.0035$) or among only eukaryotes ($R^2 = 0.18$, $F_{1,24} = 5.2$, $P = 0.031$). It has been suggested that, for *D. melanogaster*, the results of Mukai et al. (1972) and Ohnishi (1977) may be less reliable than more recent ones (Garcia-Dorado et al. 1999). When removing these estimates, the correlation is stronger ($R^2 = 0.38$, $F_{1,29} = 18$, $P = 0.0002$). Finally, in the VSV, direct and BM estimates are very different, and the most reliable measure is probably the direct estimate (S. F. Elena, pers. comm.). Including only this measure for the VSV in the analysis does not remove the global trend of an increase in $E(s)$ with gene number ($R^2 = 0.24$, $F_{1,28} = 8.8$, $P = 0.006$).

The effect of gene number also remains significant ($F_{1,30} = 24.7$, $P < 0.0001$) when including potentially confounding factors (see Statistical Analyses). Among these, only the method to obtain mutants has a significant effect: point mu-

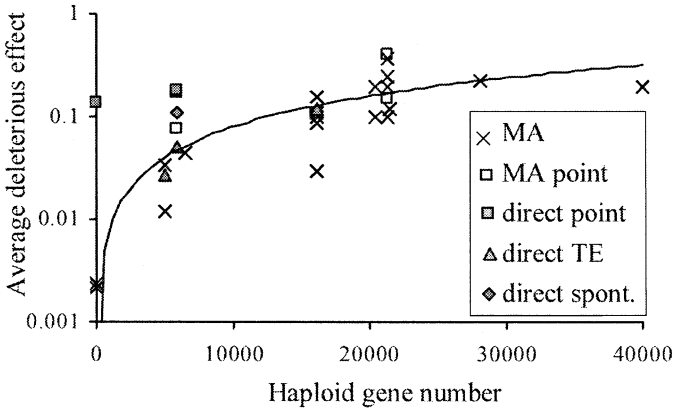


FIG. 2. Variation of the average deleterious effect of mutations with gene number. Estimates of $E(s)$ (y-axis, log-scale) are either indirect estimates from mutation accumulation (MA) experiments (MA, Table 1) or direct estimates from mutagenesis experiments (direct, Table 2). The type of mutation is indicated on the graph as: spont (spontaneous), TE (transposons), point (point mutations; i.e., ems, mmr, and subst in Tables 1 and 2). The line indicates the regression with intercept set to zero on all 33 estimates. $E(s)$ increases by about two orders of magnitude from viruses to higher organisms.

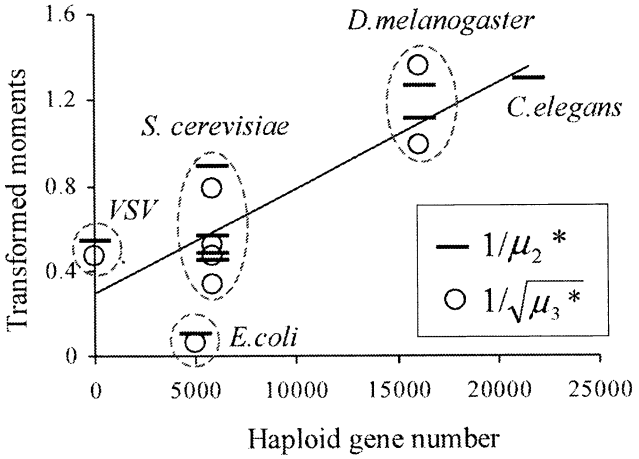


FIG. 3. Variation of second and third moments of $f(s)$ with gene number and of the inverse of the scaled second moment, $1/\mu_2^*$ (dashes), and the square root of the scaled third moment $1/\sqrt{\mu_2^*}$ (circles) with gene count. These measures are expected to correlate with n in equation (2). μ_2^*, μ_3^*, and gene count are from Table 2. The line shows the regression of $1/\mu_2^*$ on gene number. The second and third moments of $f(s)$ ($\mu_2^*$and μ_3^*) tend to decrease with gene number.

tations tend to have more deleterious effects than spontaneous mutations and transposable element insertions ($F_{1,30} = 9.5$, $P = 0.004$). However, point mutations were often obtained by methods that artificially increase the mutation rate (using EMS mutagenesis or mismatch repair deficient strains) so that the effect detected might be due to increased number of mutation per line (through, e.g., negative epistatic interactions), rather than to the molecular nature of point mutations. Our model also predicts that $E(s)$ should not depend on the level of adaptation of the initial genotype (whenever the Gaussian fitness approximation is still valid; i.e., not too far from the optimum). This is consistent with the estimates of $E(s)$ in the MA experiment on VSV (Table 1), which are very similar for three initial genotypes that differed in fitness (0.8, 1, 2.5) (Elena and Moya 1999). The second prediction (eq. (2) was that both $1/\mu_2^*$ and $1/\sqrt{\mu_3^*}$ should increase linearly with n. Although limited, data from Table 2 indicate that both these quantities positively correlate with gene number: Pearson's $\rho = 0.84$, $n = 9$, $P = 0.0043$ and $\rho = 0.80$, $n = 8$, $P = 0.018$, respectively (see Fig. 3). The VSV is clearly an outlier in the dataset, as it has much larger $1/\mu_2^*$ and $1/\sqrt{\mu_3^*}$ values than would be expected from its gene number. This might be due to the fact that $E(s)$, which scales both μ_2^* and μ_3^* may have been overestimated in the VSV direct measure, as could be suggested by the fact that the $E(s)$ estimate is much larger in the direct measure than in the MA experiments (see Fig. 2).

Shape of Empirical Distributions and Estimates of n_e

We now turn to testing whether our approximation for $f(s)$ in terms of a gamma is consistent with the available empirical data. If $f(s)$ is gamma distributed, we expect a quadratic relationship between the scaled second and third moments such that $\mu_3^* = 2\mu_2^{*2}$. The values of μ_2^* and μ_3^* (Table 2) exhibit such a relationship across species ($\mu_3^* \approx 3.07\ \mu_2^{*2}$; 95% bootstrap slope CI 1.01–3.09; $R^2 = 0.99$; $P < 0.0001$, Fig. 4). This strong relationship indicates that empirical $f(s)$ belong to a gamma-like distribution family. Under such a gamma approximation, the shape of the empirical $f(s)$ for deleterious effects (i.e., when $s_o = 0$), is simply $\frac{1}{2}n_e$ (see above), where n_e can be estimated as $n_e = 2E(s)^2/V(s) = 2/\mu_2^*$ (eq. 4). The resulting n_e estimates (reported in Table 2) increase with increasing gene number (same P-values as $1/\mu_2^*$, above), as predicted among species of increasing complexity. Overall, with the exception of the VSV, increasing gene number (complexity) results qualitatively in gamma $f(s)$ of increasing

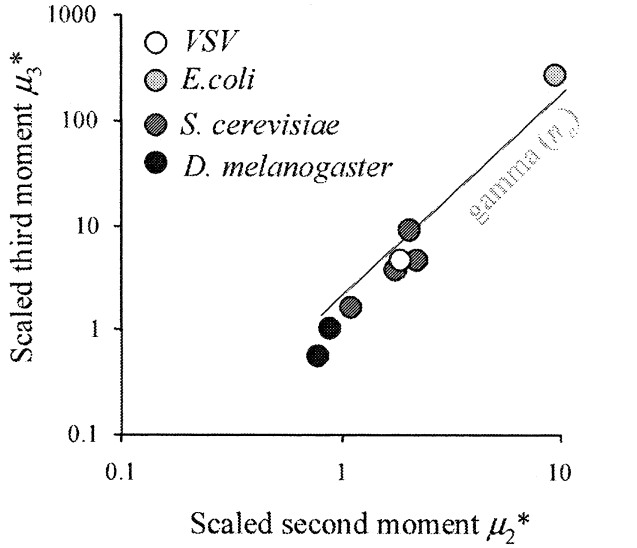


FIG. 4. Variation of the second and third scaled moments of $f(s)$ (μ_2^* and μ_3^* respectively, both in log-scale, from Table 2) across species. The line indicates the predicted relationship if s follows a gamma distribution with shape $n_e/2$ and $s_o = 0$, where n_e is $2/\mu_2^*$. Second and third moments show a strong quadratic relationship consistent with a gamma-like distribution.

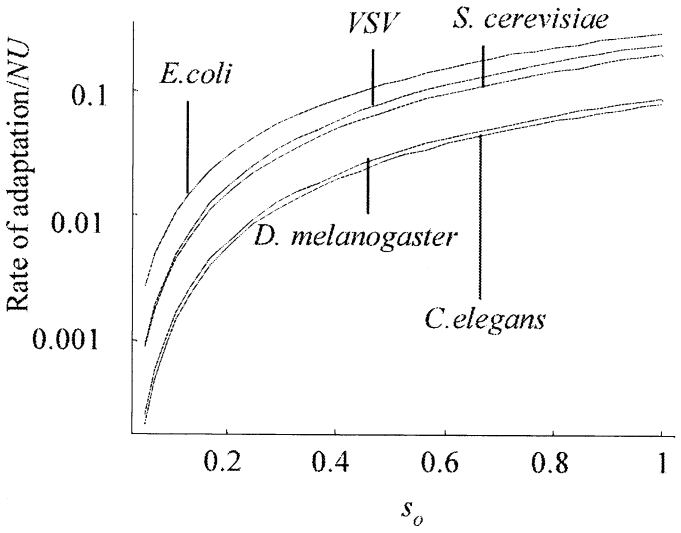


FIG. 5. Predicted rates of adaptation across species. Rates of adaptation (scaled to *NU*) for different species for different distance of the initial genotype relative to the optimum (s_o, x-axis). They are computed from equation (7) using $E(s)$ and n_e estimates in Table 2. For yeast and *Drosophila*, the mean over 4 (2, respectively) estimates is indicated. Predicted rates of adaptation decrease by an order of magnitude from bacteria to fruit flies or nematodes.

shape parameter ($\frac{1}{2}n_e$, see Fig. 4). In the following section we use this finding to quantify the cost of complexity by estimating how the effects of phenotypic complexity that we observe in our survey may influence the rate of adaptation across species.

Rates of Adaptation and Cost of Complexity across Species

Equation (7) can be used to evaluate approximately how variation in $f(s)$ across species may influence their rates of adaptation for a given fitness distance to the optimum s_o and a given flow of mutations $N\mu$. Rates of adaptation (scaled to $N\mu$), based on $E(s)$ and n_e estimates in Table 2 for the different species, are shown in Figure 5 for s_o varying from 0.05 to 1. First, Figure 5 shows that the increase in n_e values from *E. coli* to higher organisms indeed translates into a reduction in rates of adaptation, confirming the existence of a cost of complexity. Second, Figure 5 also shows that only a modest difference in the rate of adaptation is predicted among species (as expected due to the small variation in n_e), with an increase of only an order of magnitude from *E. coli* to *C. elegans*. In the next section we discuss how phenotypic correlations may explain such a reduced cost of complexity.

Why Such a Low Cost? The Effect of Phenotypic Correlations

A main conclusion of our survey is that the effective number of dimensions n_e increases with complexity as predicted by our model, but varies only by an order of magnitude from bacteria to fruit flies, and remains very small (0.2–2.5). This finding is surprising because our model predicts that $n_e = n/[1 + \mathrm{CV}(\lambda)^2]$ (eq. 3) should be proportional to the actual number of traits under selection, which, intuitively, should vary by more than an order of magnitude across the species considered. One possible explanation may be that trait heterogeneity $\mathrm{CV}(\lambda)$ increases with the number of traits n, leading to a diminishing return of n_e on n. We showed that such a phenomenon occurs in the simplest case where mutational and selective covariance matrices, **S** and **M**, are drawn randomly (see eq. 8). For example, n_e values as small as 2.5 (as observed for *C. elegans* or *D. melanogaster*) can be obtained for any large number of traits with correlation strength $\rho_S^2 = \rho_M^2 \approx 0.2$. Note, however, that the very small n_e value estimated for *E. coli* ($n_e = 0.21$) cannot be explained by this model, even assuming very strong correlations (in eq. 8, n_e cannot be less than 0.5 for large n). Nevertheless, the model predicts that large variation in the total number of traits n may translate into limited differences in n_e values, a pattern fully consistent with our observations. It may be argued the other way around that all species have a very large number of traits but that higher organisms tend to have weaker phenotypic correlations, resulting in a higher n_e (see eq. 8, limit $n \to \infty$). This hypothesis seems less parsimonious but remains to be tested.

DISCUSSION

Predicting how the distribution of mutation fitness effects should vary across species is a key issue in evolutionary biology but has received little attention so far. In particular, no study has been proposed to estimate and explain variation in $f(s)$ across species. In this paper, we proposed a model of mutation fitness effects, that extends Fisher's (1930) geometric approach to take into account (co)variation between traits in the effect of both mutation and selection (into a multivariate Gaussian framework). The shape of the predicted distribution depends on the number of phenotypic traits under selection (n). We predicted how n should influence the three first moments of $f(s)$, and we tested these predictions using gene number as a surrogate measure of phenotypic complexity n. A comparison of data collected from several species confirmed all three predicted trends: as n increases, the average s increases (Fig. 2), while the second and third moments of $f(s)$ decrease (Fig. 3). These results show that phenotypic complexity has a strong influence on the distribution of s. Our survey is based on a limited amount of data. However, we find that different experiments for the same species are consistent with each other except for VSV, where the direct measure of $E(s)$ differs markedly with BM estimates. More studies are required to obtain consistent estimates of mutation fitness effects in viruses, but it is probable that the direct estimate is more reliable (S. F. Elena, pers. comm.), in which case VSV indeed represents a striking exception in the observed correlation. Alternatively, the quantitative genetic framework used here might also not apply to very small viral genomes. Next, we show that these empirical $f(s)$ are gamma-like (Fig. 4), as expected under our model (Fig. 1), and that their shape measures an effective number of traits, n_e. Although rather indirect and qualitative, the results of this review validate three clear predictions of the model, each of which could have been invalidated by the data. Therefore, this review provides a first test of fitness landscape models, which calls for further confrontation with empirical data.

In our model, this effective number of traits is reduced

relative to n when trait heterogeneity increases. We use this finding to measure how variation in $f(s)$ across species translates into variation in rates of adaptation; that is, to predict the resulting cost of complexity. We predict only modest decrease of this rate from lower to higher organisms in the range of species for which empirical $f(s)$ are available (Fig. 5). This moderate decrease confirms the existence of a cost of complexity but minimizes its quantitative importance. Finally, we use random matrix theory to determine how $f(s)$ should vary with the number of traits if mutational and selective covariance matrices are drawn randomly. We find that under this simple model, trait heterogeneity scales with the number of traits such that even organisms with a very large number of traits exhibit a skewed and gamma-like $f(s)$ and pay a very low cost to complexity.

Comparison with Previous Theoretical Results

In a recent paper, Waxman and Welch (2005) proposed a model of mutation that accounts for selective interactions between traits (i.e., an arbitrary matrix **S** in our framework). We briefly compare the two approaches. Waxman and Welch's approach allows for the effect of biased mutation effects on phenotypes, which is neglected in ours. Mutational bias may be included in our analysis, but the results rely on some extra parameters which may be difficult to measure empirically. Our prediction of $f(s)$ based on deleterious mutations and a measure of s_o also neglects the correlation between the z_i and λ_i across traits i (see online Appendix 1), which are not neglected in Waxman and Welch (2005). This was done to propose results in terms of measurable quantities, but analytic arguments and simulations (G. Martin and T. Lenormand, unpubl. data) suggest that this correlation may have quantitatively limited impact on our results.

In addition to selective correlations, our model also accounts for mutational correlations, which are ignored in Waxman and Welch (2005). However, we show in Appendix 1 (available online) that our more general model can be reduced to a simpler model with spherically symmetrical mutation and heterogeneous selection across traits (with selection effects equal to the λ_i, the eigenvalues of $\mathbf{S}\cdot\mathbf{M}$), as is assumed in Waxman and Welch (2005). We therefore believe that their results should hold in our context.

The link between the two models and results can be made simply by considering the simplest situation common to both models: initial genotype at the optimum, unbiased, and uncorrelated effects of mutation on phenotypes. The most remarkable fact arising from this comparison is that both approaches yield very similar expressions for the effective number of dimensions, n_e. When the initial phenotype is at the optimum ($\mathbf{z}_o = 0$), both predict the same reduction of n_e relative to n, when traits are heterogeneous: $n_e = n/(1 + \mathrm{CV}(\lambda)^2)$. This can be seen by comparing equation 27 of Waxman and Welch (2005) and equation (3) of the present paper: $\mathrm{CV}(\lambda)^2 = f_{z,\sigma}$ in Waxman and Welch's notations, when $\mathbf{z} = 0$. When not at the optimum, our definition of n_e differs from that of Waxman and Welch (2005): ours is based on $f(s)$ alone, whereas that of Waxman and Welch (2005) is based on the rate of adaptation, which depends on both $f(s)$ and the distance to the optimum. Overall, these similarities in the predicted effect of phenotypic correlations on n_e give strong support to the idea that they have a very large impact on $f(s)$ and on rates of adaptation in general.

The Structure of Phenotypic Interactions

We find that random and independent mutational and selective covariance matrices (**M** and **S**) generate distributions of mutation fitness effects that are consistent with the available empirical $f(s)$. However, this agreement does not rule out that mutational and selective covariance matrices be in fact nonrandom and/or interdependent. We used random matrices to exhibit a simple situation in which the observed pattern is predicted. However, we believe that random matrix theory, which has proven fruitful in the analysis of complex systems in physics and finance (Forrester et al. 2003), is a promising avenue to analyze models of phenotypic interactions with many traits, or selective interactions with many genes (e.g., distribution of epistatic interactions Bonhoeffer et al. 2004; Sanjuan et al. 2004b).

Why and when should we expect mutational and selective covariance matrices to be nonrandom and mutually dependent? We can discuss two extreme situations that reflect the range of possibilities. In the first situation, $\mathbf{M} \propto \mathbf{S}^{-1}$, (where $\mathbf{S}^{-1}$ is the inverse of **S**) such that the traits under the strongest selective pressure exhibit the lowest mutational variance. Consequently, $\mathbf{S}\cdot\mathbf{M} \propto \mathbf{I}$, and all λ_i are equal, so that $n_e = n$, even if traits are very heterogeneous within both **M** and **S**. This situation would be theoretically expected under strong canalization (Rice 1998). In the opposite situation, $\mathbf{M} \propto \mathbf{S}$, such that traits under strong selection exhibit the highest mutational variance. Consequently, $\mathbf{S}\cdot\mathbf{M} \propto \mathbf{M}^2$, so that $\mathrm{CV}(\lambda_i)$ is maximized and $n_e \ll n$. This situation, on the contrary, would be expected under decanalization (Rice 1998). The moderate increase in n_e among species of presumably large variation in phenotypic complexity (or at least gene number) could thus be explained if more complex organisms tend to show less canalization. Obviously, more work is needed to determine how selection can shape mutational and selective covariance matrices.

Small n_e Values

Even assuming strong correlations, the very small n_e values obtained for *E. coli* ($n_e < 0.5$) cannot be explained by our model. In addition, in *E. coli*, unlike in the VSV, the estimates of $E(s)$ from two independent MA studies and from a direct measure are consistent (Fig. 1). Therefore, an overestimation of $\mathrm{CV}(s)$ (hence an underestimation of n_e) due to imprecision in $E(s)$ estimates in this species seems unlikely at first glance. However, the agreement between BM and direct estimates of $E(s)$ in *E. coli* is in fact surprising, because the former are biased upward relative to the latter proportionately to $1 + \mathrm{CV}(s)^2$ (Lynch et al. 1999). Given the large values of $\mathrm{CV}(s)$ reported for this species (Table 2), we would expect direct measures of $E(s)$ to be smaller than the corresponding BM estimates (whereas this effect should be limited in yeast or *Drosophila*, where $\mathrm{CV}(s)$ is much smaller). It is possible that a direct deleterious effect of transposition biased upward the estimates of $E(s)$ in the direct measure based on single TE insertions—such an effect has been reported in a later study

using the same lines (Remold and Lenski 2001). This could then lead to the unexpected agreement between BM and direct estimates of $E(s)$ in *E. coli*. In any case, correcting for such overestimation of $E(s)$ could only lead to an even smaller n_e estimate than the one reported here. More generally, if $CV(s)$ decreases with complexity, as suggested by our survey, the BM estimates of $E(s)$ should particularly overestimate $E(s)$ in lower organisms. As a consequence, the increase of $E(s)$ with gene number should be more radical than the one we report here, based mostly on BM estimates.

A possible explanation to the very small n_e estimate in *E. coli* could be that we considered universal pleiotropy. Modularity may affect the distribution of s (Wagner and Altenberg 1996; Welch and Waxman 2003). However, as noted in the introduction, with both mutational and selective covariances, such ''modularity'' requires having matching blocks in both **S** and **M**, such that mutations in a given module only affect fitness in the same module. Such ''matching blocks'' modularity could explain low values of n_e among species (i.e., not only in *E. coli*), and lead to a small value in this species. For example, with m exactly equivalent modules, the three first moments of s are obtained by simply replacing n by n/m in equation (1) , which can help to explain the very low n_e values. However, the limited cost of complexity that we predict across species in our survey is due to the limited variation in n_e across species rather than to n_e values themselves. To explain such limited variation with modularity, a very specific relationship between the number of modules and the total number of traits would be required. A more general model including modules of variable sizes and trait heterogeneity within modules would be necessary to compare their relative impact on $f(s)$. Such a model could be developed by considering a set of covariance matrices describing the mutational phenotypic effects of each module. Weighting each matrix by the probability of a mutation in the corresponding module would yield the net effect of all modules and could be summarized with a single matrix **M**.

Evolution of Complex Organisms

The relationship between gene number and phenotypic complexity might be rather weak, beyond coarse phylogenetic divisions (for discussion see Otto and Yong 2002). However, at the phylogenetic scale of our study, gene number provides, at least, an intuitive measure of complexity by ranking viruses, unicellular and multicellular organisms. In any case, beyond our interpretation in terms of phenotypic complexity, our survey shows that gene number is a good predictor of differences in $f(s)$: increasing gene number results in larger average deleterious effects and in approximately gamma $f(s)$ with increasing shape parameter, VSV being an exception (Fig. 4).

Because the term ''complexity'' may have various meanings according to authors, it is important to recall that the definition we refer to here is a number of adaptation traits under selection, as in the Fisher-Orr approach. However, even in this context, the notion of complexity remains somewhat vague and only defined in reference to a measurable quantity; for instance, in reference to rates of adaptation. It may be more straightforward to define it in reference to $f(s)$ because other factors may influence the rate of adaptation either favoring (Orr 2000; Welch and Waxman 2003) or disfavoring adaptation in higher organisms (e.g., longer generation time and smaller population sizes or mutation rates; Lynch and Conery 2003). Complexity in reference to $f(s)$ is our effective number of traits n_e, which may be radically different from complexity as perceived from organismal organization. This idea of an ''effective dimensionality'' (Orr 1998, 2000; Barton and Keightley 2002) has already been put forward; we intended here to provide a formal analysis of the influence of explicit biological assumptions on this quantity. We found that n_e and consequently, predicted rates of adaptation, differ little among species, and we showed that increasing the number of traits can have almost no effect on the rate of adaptation if they are not independent. If mutational and selective covariances are drawn randomly, the outcome is even more extreme: when the number of traits is large, n_e (as both $f(s)$ and rates of adaptation) is determined primarily by phenotypic correlations and tends to a finite limit as the number of trait increases. We therefore expect that $f(s)$ in more ''complex'' organisms should be similar to $f(s)$ in fruit flies or nematodes in our survey. In any case ''complexity,'' as defined by a number of traits under selection (adaptation traits, defined in the introduction), may not pose such an evolutionary paradox as previously suggested.

Conclusions

Our model and analysis is an attempt to predict distributions of mutation fitness effects based on explicit biological hypotheses and to validate it with empirical data. We also intended to show that the Fisher-Orr geometric approach may not be as unrealistic as it is sometimes suggested, provided phenotypic correlations are accounted for. However, important limitations remain (apart from the issue on phenotypic modularity discussed above). First, frequency-dependent or disruptive selection cannot be taken into account. While this may not be a problem to predict $f(s)$ in the context of laboratory studies where each line's fitness is assayed individually, it may limit the generality of predictions on the adaptation of natural populations. Second, predictions far from the optimum may be less robust, as we outlined in the introduction. However, such predictions can still be made in this situation (eq. A3 in online Appendix 1 shows the predicted effect of maladaptation on moments of $f(s)$ in our model). These predictions could be tested, by considering, for example, the effect of stress on the distribution of mutation fitness effects. Finally, we note that it remains difficult to critically test Fisher's model based on empirical data because of the lack of alternative models. One possibility would be to compare rates of adaptation predicted using extreme value theory (Orr 2002) or Fisher's model, but this is beyond the scope of this paper.

Our results suggest that mutation fitness effect distributions have a predictable shape and variation from lower to higher organisms, and that phenotypic landscape models may capture this variation. Furthermore, our approach suggests that distributions of deleterious mutation effects can be used to predict the distribution of beneficial ones for a given environmental change, which is open to further empirical tests.

This opens a wide array of perspectives, as these distributions may be important to a large diversity of questions in evolutionary genetics.

Acknowledgments

We thank P. Jarne, Y. Michalakis, S. F. Elena, D. Waxman, and M. Kirkpatrick. We also thank R. Korona for kindly providing the estimates of selection coefficients in the yeast from Wloch et al. (2001). This work was supported by an Action Concertée Incitative grant from the French Ministry of Research to TL.

Literature Cited

Avila, V., and A. Garcia-Dorado. 2002. The effects of spontaneous mutation on competitive fitness in *Drosophila melanogaster*. J. Evol. Biol. 15:561–566.

Baer, C. F., F. Shaw, C. Steding, M. Baurngartner, A. Hawkins, A. Houppert, N. Mason, M. Reed, K. Sinnonelic, W. Woodard, and M. Lynch. 2005. Comparative evolutionary genetics of spontaneous mutations affecting fitness in rhabditid nematodes. Proc. Natl. Acad. Sci. USA 102:5785–5790.

Barton, N., and L. Partridge. 2000. Limits to natural selection. BioEssays 22:1075–1084.

Barton, N. H., and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nat. Rev. Genet. 3:11–21.

Bataillon, T. 2000. Estimation of spontaneous genome-wide mutation rate parameters: whither beneficial mutations? Heredity 84:497–501.

———. 2003. Shaking the ''deleterious mutations'' dogma? Trends Ecol. Evol. 18:315–317.

Bennetzen, J. L., C. Coleman, R. Y. Liu, J. X. Ma, and W. Ramakrishna. 2004. Consistent over-estimation of gene number in complex plant genomes. Curr. Opin. Plant Biol. 7:732–736.

Bonhoeffer, S., C. Chappey, N. T. Parkin, J. M. Whitcomb, and C. J. Petropoulos. 2004. Evidence for positive epistasis in HIV-1. Science 306:1547–1550.

Burch, C. L., and L. Chao. 1999. Evolution by small steps and rugged landscapes in the RNA virus *phi*6. Genetics 151: 921–927.

Charlesworth, B., and D. Charlesworth. 1998. Some evolutionary consequences of deleterious mutations. Genetica 103:3–19.

Charlesworth, B., H. Borthwick, C. Bartolome, and P. Pignatelli. 2004. Estimates of the genomic mutation rate for detrimental alleles in *Drosophila melanogaster*. Genetics 167:815–826.

Chavarrias, D., C. Lopez-Fanjul, and A. Garcia-Dorado. 2001. The rate of mutation and the homozygous and heterozygous mutational effects for competitive viability: a long-term experiment with *Drosophila melanogaster*. Genetics 158:681–693.

Clarke, B., and W. Arthur. 2000. What constitutes a ''large'' mutational change in phenotype? Evol. Dev. 2:238–240.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. Genetics 148:1667–1686.

Elena, S. F., and R. E. Lenski. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat. Rev. Genet. 4:457–469.

Elena, S. F., and A. Moya. 1999. Rate of deleterious mutation and the distribution of its effects on fitness in vesicular stomatitis virus. J. Evol. Biol. 12:1078–1088.

Elena, S. F., L. Ekunwe, N. Hajela, S. A. Oden, and R. E. Lenski. 1998. Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica 103:349–358.

Estes, S., P. C. Phillips, D. R. Denver, W. K. Thomas, and M. Lynch. 2004. Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. Genetics 166:1269–1279.

Fisher, R. A. 1930. The genetical theory of natural selection. Oxford Univ. Press, Oxford, U.K.

Forrester, P. J., N. C. Snaith, and J. J. M. Verbaarschot. 2003. Developments in random matrix theory. J. Physics A 36: R1–R10.

Garcia-Dorado, A., C. Lopez-Fanjul, and A. Caballero. 1999. Properties of spontaneous mutations affecting quantitative traits. Genet. Res. 74:341–350.

Gerrish, P. J., and R. E. Lenski. 1998. The fate of competing beneficial mutations in an asexual population. Genetica 103: 127–144.

Ihaka, I., and G. Robert. 1996. R: a language for data analysis and graphics. J. Comput. Graphic. Stat. 5:299–314.

Imhof, M., and C. Schlotterer. 2001. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. Proc. Natl. Acad. Sci. USA 98:1113–1117.

Joseph, S. B., and D. W. Hall. 2004. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: more beneficial than expected. Genetics 168:1817–1825.

Keightley, P. D. 1994. The distribution of mutation effects on viability in *Drosophila melanogaster*. Genetics 138:1315–1322.

———. 2004. Comparing analysis methods for mutation-accumulation data. Genetics 167:551–553.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science 290:331–333.

Keightley, P. D., and M. Lynch. 2003. Toward a realistic model of mutations affecting fitness. Evolution 57:683–685.

Keightley, P. D., and O. Ohnishi. 1998. EMS-induced polygenic mutation rates for nine quantitative characters in *Drosophila melanogaster*. Genetics 148:753–766.

Keightley, P. D., E. K. Davies, A. D. Peters, and R. G. Shaw. 2000. Properties of ethylmethane sulfonate-induced mutations affecting life-history traits in *Caenorhabditis elegans* and inferences about bivariate distributions of mutation effects. Genetics 156: 143–154.

Kibota, T. T., and M. Lynch. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E-coli*. Nature 381: 694–696.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245–261.

Korona, R. 2004. Experimental studies of deleterious mutation in *Saccharomyces cerevisiae*. Res. Microbiol. 155:301–310.

Lande, R. 1980. The genetic covariance between characters maintained by pleiotropic mutations. Genetics 94:203–215.

Loewe, L., V. Textor, and S. Scherer. 2003. High deleterious genomic mutation rate in stationary phase of *Escherichia coli*. Science 302:1558–1560.

Lyman, R. F., F. Lawrence, S. V. Nuzhdin, and T. F. C. Mackay. 1996. Effects of single P-element insertions on bristle number and viability in *Drosophila melanogaster*. Genetics 143: 277–292.

Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. Science 302:1401–1404.

Lynch, M., and B. Walsh. 1998. Correlations between characters. Pp. 629–655. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA.

Lynch, M., J. Blanchard, D. Houle, T. Kibota, S. Schultz, L. Vassilieva, and J. Willis. 1999. Perspective: Spontaneous deleterious mutation. Evolution 53:645–663.

Mackay, T. F. C., R. F. Lyman, and M. S. Jackson. 1992. Effects of P-Element insertions on quantitative traits in *Drosophila melanogaster*. Genetics 130:315–332.

Mathai, A. M., and S. B. Provost. 1992. Quadratic forms in random variables. Marcel Dekker, New York.

Mukai, T., S. I. Chigusa, L. E. Mettler, and J. F. Crow. 1972. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. Genetics 72:335–355.

Ohnishi, O. 1977. Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in *Drosophila melanogaster*. II. Homozygous effects of polygenic mutations. Genetics 87: 529–545.

Orr, H. A. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52: 935–949.

———. 2000. Adaptation and the cost of complexity. Evolution 54:13–20.
———. 2001. The ''sizes'' of mutations fixed in phenotypic evolution: a response to Clarke and Arthur. Evol. Dev. 3:121–123.
———. 2002. The population genetics of adaptation: the adaptation of DNA sequences. Evolution 56:1317–1330.
———. 2003. The distribution of fitness effects among beneficial mutations. Genetics 163:1519–1526.
———. 2005a. Theories of adaptation: what they do and don't say. Genetica 123:3–13.
———. 2005b. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6:119–127.
Otto, S. P. 2004. Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proc. R. Soc. Lond. B 271:705–714.
Otto, S. P., and P. Yong. 2002. The evolution of gene duplicates. Pp. 451–483. Homology effects. Academic Press, San Diego, CA.
Poon, A., and S. P. Otto. 2000. Compensating for our load of mutations: freezing the meltdown of small populations. Evolution 54:1467–1479.
Remold, S. K., and R. E. Lenski. 2001. Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. Proc. Natl. Acad. Sci. 98:11388–11393.
Rice, S. H. 1998. The evolution of canalization and the breaking of von Baer's laws: modeling the evolution of development with epistasis. Evolution 52:647–656.
Rokyta, D. R., P. Joyce, S. B. Caudle, and H. A. Wichman. 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat. Genet. 37: 441–444.
Rozen, D. E., J. de Visser, and P. J. Gerrish. 2002. Fitness effects of fixed beneficial mutations in microbial populations. Curr. Biol. 12:1040–1045.
Sanjuan, R., A. Moya, and S. F. Elena. 2004a. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. USA 101:8396–8401.
———. 2004b. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc. Natl. Acad. Sci. USA 101: 15376–15379.
Shaw, F. H., C. J. Geyer, and R. G. Shaw. 2002. A comprehensive model of mutations affecting fitness and inferences for *Arabidopsis thaliana*. Evolution 56:453–463.
Shaw, R. G., F. H. Shaw, and C. Geyer. 2003. What fraction of mutations reduces fitness? A reply to Keightley and Lynch. Evolution 57:686–689.
Stein, L. D., Z. R. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. S. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, and many others. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. Plos Biol. 1: 166.
Thatcher, J. W., J. M. Shaw, and W. J. Dickinson. 1998. Marginal fitness contributions of nonessential genes in yeast. Proc. Natl. Acad. Sci. USA 95:253–257.
Vassilieva, L. L., A. M. Hook, and M. Lynch. 2000. The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. Evolution 54:1234–1246.
Wagner, G. P., and L. Altenberg. 1996. Perspective: Complex adaptations and the evolution of evolvability. Evolution 50: 967–976.
Ware, D., and L. Stein. 2003. Comparison of genes among cereals. Curr. Opin. Plant Biol. 6:121–127.
Waxman, D., and J. J. Welch. 2005. Fisher's microscope and Haldane's ellipse. Am. Nat. 166:447–457.
Welch, J. J., and D. Waxman. 2003. Modularity and the cost of complexity. Evolution 57:1723–1734.
Wloch, D. M., K. Szafraniec, R. H. Borts, and R. Korona. 2001. Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. Genetics 159: 441–452.
Xu, J. P. 2004. Genotype-environment interactions of spontaneous mutations for vegetative fitness in the human pathogenic fungus *Cryptococcus neoformans*. Genetics 168:1177–1188.
Zeyl, C., and J. DeVisser. 2001. Estimates of the rate and distribution of fitness effects of spontaneous mutation in *Saccharomyces cerevisiae*. Genetics 157:53–61.
Zhang, X. S., and W. G. Hill. 2003. Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. Evolution 57:1761–1775.

Corresponding Editor: P. Phillips

Appendix 1: Exact distribution of *s*

In this appendix, we first show how our general model of mutation can be analyzed using a transformation of the original phenotypic space, which is always possible under our assumptions. Next, we describe the exact and approximate distribution of *s*. Finally, we give an approximate expression that allows one to predict the moments of $f(s)$ in a new environment as a function of the moments of $f(s)$ measured in an optimal environment.

Definition of an equivalent system

A given mutant has phenotype $\mathbf{z}_o + \mathbf{dz}$ where $\mathbf{z}_o$ is the initial phenotype and $\mathbf{dz}$ is drawn from a multivariate Gaussian distribution with mean $\mathbf{0}$, and covariance matrix $\mathbf{M}$. The selection coefficient *s* of a mutant, assuming small *s*, is given by

$$s \approx \log(1+s) = \log\left(\frac{W(\mathbf{z}_o + \mathbf{dz})}{W(\mathbf{z}_o)}\right) = -\frac{1}{2}(\mathbf{z}_o + \mathbf{dz})^t \mathbf{S}\ (\mathbf{z}_o + \mathbf{dz}) - \frac{1}{2}\mathbf{z}_o^t \mathbf{S}\, \mathbf{z}_o . \qquad \text{(A1)}$$

The right hand term in Eq. (A1) is distributed as a quadratic form in random gaussian vectors much studied in e.g. statistics and finance (Mathai and Provost 1992). There is always a linear transformation of the phenotypic vector $\mathbf{z}$ into vector $\mathbf{x}$ such that, in this new definition of traits, (i) all traits are independent and have equal variance by mutation ($\mathbf{M}$ becomes the identity matrix $\mathbf{I}$ so that mutation becomes spherically symetrical) and (ii) all traits are independent for selection (the arbitrary matrix $\mathbf{S}$ becomes a diagonal matrix $\mathbf{\Lambda}$). This linear transformation is characterised by the constant matrix $\mathbf{Q}$ such that $\mathbf{z} = \mathbf{Q}\,\mathbf{x}$ where $\mathbf{Q}$ is obtained by solving the generalized eigenvalue problem (Mathai and Provost 1992). More precisely, $\mathbf{Q}$ is the matrix that verifies the two conditions (i) $\mathbf{Q}.\mathbf{Q}^t = \mathbf{M}$ and (ii) $\mathbf{Q}^t.\mathbf{S}.\mathbf{Q} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix. Matrix $\mathbf{Q}$ always exists when both $\mathbf{M}$ and $\mathbf{S}$ are positive semi-definite and can be computed as follows: as $\mathbf{M}$ is positive semi-definite there is a matrix $\mathbf{B}$ (not necessarily square) such that $\mathbf{M} = \mathbf{B}.\mathbf{B}^t$. Then, if $\mathbf{C}$ is the eigenmatrix of the product $\mathbf{B}^t.\mathbf{S}.\mathbf{B}$, it contains all the (orthogonal) eigenvectors of $\mathbf{B}^t.\mathbf{S}.\mathbf{B}$ such that $\mathbf{C}^t.(\mathbf{B}^t.\mathbf{S}.\mathbf{B}).\mathbf{C} = \mathbf{\Lambda}$ is diagonal and $\mathbf{C}^{-1} = \mathbf{C}^t$ ($\mathbf{C}$ is orthogonal). Then, it is easy to show that the matrix $\mathbf{Q} = \mathbf{B}.\mathbf{C}$ verifies the conditions (i) and (ii) defined above. By definition, the matrix $\mathbf{\Lambda}$ has diagonal elements equal to the eigenvalues of $\mathbf{B}^t.\mathbf{S}.\mathbf{B}$ which are also the non-zero eigenvalues λ_i of $\mathbf{S}.\mathbf{M}$ (since $\mathbf{S}.\mathbf{M} = \mathbf{S}.\mathbf{B}.\mathbf{B}^t$). Therefore, the diagonal elements of matrix $\mathbf{\Lambda}$ are the λ_i, and they define the selective effect of each trait in the new system ($\mathbf{S}$ becomes $\mathbf{\Lambda}$).

Note that if $\mathbf{M}$ is only positive semi-definite (instead of positive definite), $\mathbf{Q}$ is not square. In general, the transformed phenotypic vector is given by $\mathbf{x} = \mathbf{Q}'\,\mathbf{z}$ where $\mathbf{Q'}$ is the left inverse of $\mathbf{Q}$ (such that $\mathbf{Q'}.\mathbf{Q} = \mathbf{I}$), and the dimension of the transformed system (of $\mathbf{x}$) reduces from *n* to the number of non-zero eigenvalues of $\mathbf{S}.\mathbf{M}$. This transformation is similar to that proposed by Zhang and Hill (Zhang and Hill 2003) for definite $\mathbf{S}$ and $\mathbf{M}$, but we show above that it applies to a wider range of covariance matrices (semi-definite), which allows us to account for stronger correlations (see Appendix 2, available online only at http://dx.doi.org/10.1554/05-412.2.s2).

Distribution of s

In the transformed space corresponding to vector **x**, any selection coefficient *s* can be expressed as a simple sum of *n* independent random components. First, denote $\mathbf{x}_o = \{x_i\} = \mathbf{Q}'\mathbf{z}_o$, the vector characterising the initial phenotype in the transformed space, and $s_o \equiv \sum_i \lambda_i x_i^2 / 2$ = -log($W(\mathbf{z}_o)/W(\mathbf{0})$) = -log($W(\mathbf{x}_o)/W(\mathbf{0})$) the selective disadvantage of the initial phenotype relative to the optimal phenotype. The equivalent of Eq. (A1) in the transformed space is:

$$s \approx \sum_i \frac{\lambda_i}{2} x_i^2 - \sum_i \frac{\lambda_i}{2} \left(dx_i + x_i\right)^2 \equiv s_o - \gamma, \tag{A2}$$

Where each dx_i are independent standard normal deviates ($dx_i \rightarrow N(0,1)$), so that the random term in Eq. (A2), $\gamma \equiv \sum_i \lambda_i \left(dx_i + x_i\right)^2 / 2$, is a linear combination of independent non-central chi-squares with one degree of freedom and non-centrality parameter x_i^2. Because all $\lambda_i \geq 0$ (since $\mathbf{S}\cdot\mathbf{M}$ is positive semi-definite), both γ and s_o are positive. The distribution of *s* given in Eq. (A2) encompasses both deleterious and advantageous mutations. Deleterious effects are unbounded (the distance to the optimum can take any value) whereas the best advantageous effect is $s = s_o$ corresponding to a mutation compensating all phenotypic deviations of the initial genotype (i.e. $\mathbf{dx} = -\mathbf{x}_o$ or equivalently $\mathbf{dz} = -\mathbf{z}_o$). Eq. (A2) shows that, close to the optimum (small x_i) and with all λ_i equal to λ, $s \approx -\gamma = -½\,\lambda \sum_{i=1}^{n} dx_i^2$ follows a negative Gamma distribution with scale λ and shape *n*/2 (by definition, because the sum $\sum_{i=1}^{n} dx_i^2$ follows a central chi-squared distribution with *n* degrees of freedom, which is a gamma deviate). Therefore, we chose the negative gamma as our *a priori* distribution to approximate *f*(*s*) in the more general case $s_o \neq 0$. This displaced gamma approximation corresponds to approximating $\sum_{i=1}^{n} (dx_i + x_i)^2$ by a gamma deviate. We develop this approximation in the next paragraph.

Effect of s_o on the moments of f(s)

The general expression of the cumulants of the distribution in Eq. (A1) is given by Mathai and Provost (Mathai and Provost 1992), and the three first central moments of *s* equal the three first cumulants. If we denote by a bar any average across the *n* traits (e.g. $\overline{\lambda^2} = \frac{1}{n}\sum_i \lambda_i^2$), the mean *E*(*s*)', variance *V*(*s*)' and third moment $\mu_3(s)$' of *s* can be written

$$\begin{aligned} E(s)' &= -\frac{n}{2}\overline{\lambda} \\ V(s)' &= \frac{n}{2}\left(\overline{\lambda^2} + 2\,\overline{\lambda^2 x^2}\right) \\ \mu_3(s)' &= -n\left(\overline{\lambda^3} + 3\,\overline{\lambda^3 x^2}\right) \end{aligned} \quad . \tag{A3}$$

The ' indicates that the initial genotype is not at the optimum. At the optimum (s_o = 0, all x_i^2 = 0), $E(s)'$, $V(s)'$ and $\mu_3(s)'$ in Eq. (A3) reduce to $E(s)$, $V(s)$ and $\mu_3(s)$ given in Eq.(1). Note that the average effect of mutations, $E(s)'$, is independent of the distance to the optimum (of x_i^2).

To obtain an approximate probability density for s, we approximate the distribution of the random term, γ in Eq. (A2), by the gamma distribution that matches the two first moments of γ (i.e. $E(\gamma) = s_o - E(s)'$ and $V(\gamma) = V(s)'$ using Eq. (A3)). The accuracy of this approximation for $f(s)$ can be evaluated by the ratio of the third moment $\mu_3(s)'$ of the exact and approximate distributions. This ratio depends on both the level of maladaptation $\overline{x^2}$ and the distribution of λ_i values across traits i (not shown). For example, when the λ_i are gamma distributed with any shape (i.e. for any CV(λ)), this ratio lies within [1/2, 4/3]. The moments given in Eq. (A3) depend on the x_i^2 and on their covariance with the λ_i. Therefore, for a given distance to the optimum (measured in fitness effect $s_o = n\overline{\lambda x^2}/2$), the value of $V(s)'$ and $\mu_3(s)'$ may vary depending on the direction of $\mathbf{x}_o$. As this direction cannot be measured, we propose an approximation for these exact moments, depending only on s_o and the moments at the optimum $E(s)$, $V(s)$ and $\mu_3(s)$ given in Eq. (1). Neglecting the correlation between the x_i^2 and the λ_i across traits, such that $\overline{\lambda x^2} \approx \bar{\lambda}\,\overline{x^2}$, $\overline{\lambda^2 x^2} \approx \overline{\lambda^2}\,\overline{x^2}$ and $\overline{\lambda^3 x^2} \approx \overline{\lambda^3}\,\overline{x^2}$, the moments of s given in Eq. (A3) can be approximated by

$$\begin{aligned} E(s)' &= E(s) \\ V(s)' &\approx V(s)\left(1+2\varepsilon\right) \\ \mu_3(s)' &\approx \mu_3(s)\left(1+3\varepsilon\right) \end{aligned} \quad , \qquad \text{(A4)}$$

where we define $\varepsilon \equiv s_o / |E(s)|$. ε approximately equals $\varepsilon \approx \overline{x^2}$ when we neglect the correlation between the x_i^2 and the λ_i across traits, such that $\overline{\lambda x^2} \approx \bar{\lambda}\,\overline{x^2}$. The displaced gamma approximation is less accurate when using the approximate $V(s)'$ given above, than when using the exact $V(s)'$ given in Eq. (A3). However, simulations (Fig. 1 and results not shown) suggest that the approximation stays reasonably accurate. Note that the approximation in Eq. (A4) assumes that in a new environment, the distance to the optimum on each trait is independent (overall across traits) of their net fitness effect λ_i. It may not be true when considering the distance to the optimum of a genotype that has adapted to the environment for some time (e.g. close to the mutation-selection equilibrium). Indeed, in this case, we expect the largest x_i^2 (distance to the optimum on trait i) to correspond to the most weakly selected traits (smallest λ_i), which corresponds to a negative covariance between x_i^2 and λ_i across traits. However, in this case, s_o should be quite small (after generations of stabilizing selection for the same optimum) so that this effect is unlikely to have a strong influence on $f(s)$.

REFERENCES:

Mathai, A. M., and S. B. Provost. 1992. Quadratic forms in random variables. Marcel Dekker, New York.

Zhang, X. S., and W. G. Hill. 2003. Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. Evolution 57:1761-1775.

Appendix 2: Random covariance matrices

Random Matrix Theory is a powerful tool to study the eigenvalue distribution of certain classes of large covariance matrices (see Bai 1999for a review). For example, let **V** be an $n \times m$ matrix with elements v_{ij} drawn independently from the same arbitrary distribution with mean 0 and variance σ^2. The matrix $\mathbf{W} = 1/m\ \mathbf{V}\mathbf{V}^t$ (t denotes transpose) is positive semi-definite (definite if $n \leq m$) and can be used as a model of random covariance matrix. When the distribution of v_{ij} is a gaussian, **W** is called a Wishart matrix of dimension n with m degrees of freedom and scale parameter σ. The parameter m determines the distribution of the correlation coefficients ρ_{ij} of **W** (which includes both positive and negative ρ_{ij}, see histograms on Fig. 3). When **W** is a Wishart matrix, this distribution has probability density ((see e.g. Champion 2003)

$$p(\rho_{ij}) = (1-\rho_{ij}^{\ 2})^{(m-3)/2} \frac{\Gamma(m/2)}{\sqrt{\pi}\,\Gamma((m-1)/2)}, \qquad \text{(A1)}$$

so that the correlation strength can be simply expressed as $\rho = \sqrt{E(\rho_{ij}^2)} = 1/\sqrt{m}$. When m and n are large enough (here for a large number of traits and not too strong correlations), the distribution of the eigenvalues λ_i of **W** tends to a simple distribution (independent of the exact distribution of the v_{ij}) known as the Marçenko-Pastur law with ratio index n/m and scale σ^2 (Bai 1999). Note that when **W** is a Wishart matrix, the ratio index is $n\,\rho^2$ where ρ is the correlation strength. Moreover, the eigenvalue distribution of the product of two such random (positive semi-definite) covariance matrices also tends to a known distribution (Bai 1999; Silverstein 1999). In our context, these results can be used to obtain the moments of the eigenvalues λ_i of $\mathbf{S}\cdot\mathbf{M}$ when **S** and **M** are Wishart matrices with correlation strength ρ_S^2 and ρ_M^2, and scale parameter $\sigma_S^{\ 2}$ and $\sigma_M^{\ 2}$, respectively. Using Eq. 2.24 in (Bai 1999), it can be shown that (i) $\bar{\lambda} \approx \mathrm{E}(\lambda) = \sigma_S^{\ 2}\sigma_M^{\ 2}$ (where $\bar{\lambda}$ converges to $\mathrm{E}(\lambda)$ as n gets large) and (ii) $\mathrm{CV}(\lambda)^2 = n\,(\rho_S^2 + \rho_M^2)$, which yields the expression of n_e in Eq. (8). Therefore, under this model of random correlations for mutation and selection, (i) $\bar{\lambda}$ is independent of n and (ii) heterogeneity between traits increases with n for a given strength of correlations.

REFERENCES:

Bai, Z. D. 1999. Methodologies in spectral analysis of large dimensional random matrices, a review. Statistica Sinica 9:611-662.

Champion, C. J. 2003. Empirical Bayesian estimation of normal variances and covariances. Journal of Multivariate Analysis 87:60-79.

Silverstein, J. W. 1999. Methodologies in spectral analysis of large dimensional random matrices, a review - Comment: Complements and new developments. Statistica Sinica 9:667-671.

The fitness effect of mutations across environments: a survey in the light of fitness landscape models

Guillaume Martin* and Thomas Lenormand

Centre d'Ecologie Fonctionnelle et Evolutive (CEFE-CNRS UMR 5175)
1919 Rte de Mende, 34 293 Montpellier, France.
E-mail addresses: guillaume.martin@unil.ch, thomas.lenormand@cefe.cnrs.fr

*Corresponding author
Present address : Dpt d'Ecologie et Evolution, Université de Lausanne, CH 1015 Lausanne, Suisse

ABSTRACT

The fitness effects of mutations on a given genotype are rarely constant across environments to which this genotype is more or less adapted, i.e. between more or less stressful conditions. This can have important implications, especially on the evolution of ecological specialization. Stress is thought to increase the variance of mutations' fitness effects, their average or the number of expressed mutations. Although empirical evidence is available for these three mechanisms, their relative magnitude is poorly known. In this paper, we propose a simple approach to discriminate between these mechanisms, using a survey of empirical measures of mutation effects in contrasted environments. This survey, across various species and environments, shows that stress mainly increases the variance of mutations' effects on fitness, with a much more limited impact on their average effect or on the number of expressed mutations. This pattern is consistent with a simple model in which fitness is a Gaussian function of phenotypes around an environmentally-determined optimum. These results suggest that a simple, mathematically tractable landscape model may not be quantitatively as unrealistic as previously suggested. They also suggest that mutation parameter estimates may be strongly biased when measured in stressful environments.

KEYWORDS: G*E interaction, mutation, fitness landscape, environmental stress, survey

INTRODUCTION

Spontaneous mutation is the ultimate source of variation, and influences the evolutionary fate of a wide range of phenomena (Charlesworth and Charlesworth 1998; Lynch et al. 1999). More specifically, the evolutionary role of mutation depends on the genomic mutation rate, U and the distribution of mutations' fitness effects, $f(s)$. In principle, U and $f(s)$ may vary among genotypes within species, among species and among environments, which may obscure both the interpretation of empirical measurements and theoretical predictions. However, we still have little insights into how U and $f(s)$ may vary in general. In a recent paper (Martin and Lenormand 2006), we showed that $f(s)$ may vary in a predictable way between more or less complex organisms. In this paper, we focus on the variation of mutation effects in different environments.

Since the work of Kondrashov and Houle (1994), several studies have documented differences in mutation effects across various environments, for quantitative traits that are more or less related to fitness (reviewed in Chang and Shaw 2003; Fry and Heinsohn 2002; Hermisson and Wagner 2004; Korona 2004; Lenormand 2002; Lynch et al. 1999). The implications of this environment-dependence have been debated in an empirical and a theoretical context. First, it might explain the discrepancies among estimates of mutational parameters within species, in particular in drosophila (Garcia-Dorado et al. 1999; Kondrashov and Houle 1994) and more generally it has been studied to determine whether laboratory measures could be extrapolated *in natura*. Second, the environment-dependant variation of mutation effects has been debated in the context of ecological specialization (Fry 1996; Kawecki et al. 1997). In both cases, environmental variation is often measured or qualified in terms of more or less stressful conditions, which is also the approach chosen in this article. The definition of a stressful environment is not straightforward and varies according to

authors and fields. Stressful environments can for example be defined as environments imposing some constraints on metabolism (e.g. desiccation, high temperature etc.) that can only be "coped with" at some energetic cost. However, species adapted to "extreme" conditions, may have a lower fitness in less "extreme" conditions (Parsons 1991). Following a widely used definition in studies of mutational effects (Kishony and Leibler 2003; Korona 1999; Szafraniec et al. 2001), we will rather consider that an environment is stressful for a given genotype if it reduces its fitness, relative to that achieved in a benign (reference) environment.

Mutational parameters are in most cases measured using mutation accumulation (MA) experiments in which several sublines are maintained over several generations under minimal selection. Per generation differences in the mean (ΔM) and variance (ΔV) of fitness among sublines during the accumulation can be estimated to infer mutational parameters. In general, we have

$$\begin{aligned} \Delta M &= U\,\bar{s} \\ \Delta V &= U\,\bar{s}^2\left(1+CV(s)^2\right) \end{aligned} \qquad (1)$$

where U is the genomic rate of non-neutral mutation, $\bar{s}$ is the average fitness effect of single mutations and CV(s) its coefficient of variation (Mukai et al. 1972). This result relies on only two assumptions: (i) the number of mutation events per genotype is Poisson distributed (with parameter U) and (ii) the effects of mutations on fitness are additive (epistasis is neglected). Equations in (1) are the basis of the Bateman-Mukai method of estimation of U and $\bar{s}$ when CV(s) is neglected (constant mutation effects), but we will use them in their general formulation above (i.e. without neglecting CV(s)). The same sublines may be assayed in different environments, yielding several estimates of ΔM and ΔV. From Eq. (1), these

estimates may vary because U, $\bar{s}$ and/or CV(s) differ between environments but there is little agreement on the prevalence of either scenario (Fry and Heinsohn 2002). In the following, we give some details on the biological implications of each of the three scenarios.

Intuitively, variation in the genomic mutation rate U between environments should reflect the fact that some mutations have a detectable fitness effect in some environments but are neutral in others. This can happen if the mutational target varies between environments (by mutational target we mean the fraction of expressed genes or more generally the fraction of genes affecting fitness). Note that the list of genes in this mutational target may vary between environments while leaving the overall size of the mutational target (and hence also U) almost constant. However, since we see no reasons for the mutational target size to be identical in all environments, environment-dependent expression should in principle be detected by variation of U between environments. We will label this hypothesis "conditional expression", hereafter CE.

Variation in $\bar{s}$ would reflect that selection intensity differs in more 'stressful' environments (as suggested e.g. in Kishony and Leibler 2003). For instance, a given mutation impairing maltose metabolization may have milder fitness consequences in a 'benign' environment where several sugars are available (it may even be neutral in the absence of maltose) than in a stressful one where maltose is the only source of carbon available. We will label this hypothesis "conditional average", hereafter CA.

Finally, variation in CV(s) between environments would reflect that mutational effects are more or less variable (among mutations) in 'stressful' environments. Consider for instance a population of bacteria that has adapted in a given environment for a long period of time (e.g. with glucose as the carbon source), so that the average phenotype is close to an optimum on glucose. Most mutations are then likely to be deleterious in glucose whereas a mixture of deleterious and beneficial mutations may be expected in a new environment (e.g. in maltose).

Mutation effects are therefore likely to be less variable in glucose than in maltose. We will label this hypothesis "conditional variance", hereafter CV. This hypothesis has been proposed by Fry and Heinsohn (2002) and also by Remold and Lenski (2001) to describe a situation similar to the example cited above.

In principle these different scenarios may be distinguished by estimating U, $\bar{s}$ and CV(s) in different environments. However, it is difficult to disentangle them from MA data, since environmental variation may affect both the number of non-neutral mutations, and their fitness effect (Lynch et al. 1999). For instance, Bateman-Mukai estimates of U and $\bar{s}$, derived from Eq. (1),

$$\begin{aligned} U_{BM} &\equiv \Delta M^2 / \Delta V = U/\left(1 + CV(s)^2\right) \\ \bar{s}_{BM} &\equiv \Delta V / \Delta M = \bar{s}\left(1 + CV(s)^2\right) \end{aligned} \qquad (2)$$

are biased by the variance of mutational effects (i.e. by CV(s)), which may itself vary across different environments. Therefore, it is impossible to directly assess whether U or $\bar{s}$ actually change across environments (Fry and Heinsohn 2002). Maximum-likelihood (Keightley 1994) or minimum distance (Garcia-Dorado and Marin 1998) methods can partially address this problem. However, with these methods, it is still difficult to state how much U and CV(s) vary relative to one another (Keightley 2004), and Bateman-Mukai estimates remain the most widely available in the literature, so that the empirical issue is still unresolved.

In this paper, we develop a simple approach to discriminate among CE, CA and CV hypotheses. We then test these hypotheses using a survey of data obtained with MA experiments where mutant fitness was assayed in different environments. We focus in particular on environments that are more or less stressful, i.e. in which fitness is reduced compared to a benign (reference) environment (see above).

After presenting, in a fist part, the results of our survey and conclusions about CE, CA and CV hypotheses, we explain, in a second part, when to expect these different scenarios and how the observed patterns can be interpreted. In particular, we interpret these patterns in terms of fitness landscape models that allow one to predict how mutation fitness effects may vary in different environments.

METHODS

Simple predictions based on Bateman-Mukai estimates

The three extreme hypotheses that either U, $\bar{s}$ or CV(s) differ between environments generate distinct and straightforward predictions that can be tested with measures of ΔM and ΔV in different environments. Let us consider two environments (1 and 2) in which ΔM (ΔM_1 and ΔM_2) and ΔV (ΔV_1 and ΔV_2) are measured for a given genotype. Define the measurable ratios $\rho_V \equiv \Delta V_1 / \Delta V_2$ and $\rho_S \equiv \bar{s}_{BM1} / \bar{s}_{BM2}$ (using the definition of s_{BM} in Eq. (2)), and the non-measurable ratios ρ_U, ρ_s and ρ_{CV} of U, $\bar{s}$ and 1+CV(s)^2 (respectively) in environment 1 *vs*. 2. Then from Eq. (1), $\log(\rho_V) = \log(\rho_U) + 2\log(\rho_s) + \log(\rho_{CV})$, and from Eq. (2), $\log(\rho_S) = \log(\rho_s) + \log(\rho_{CV})$. Therefore, we can predict distinct relationships between $\log(\rho_V)$ and $\log(\rho_S)$, for each of the extreme scenarios considered, according to which of ρ_U, ρ_s and ρ_{CV} is assumed to depart from one (expected in the absence of environment-dependent variation). If only U varies between environments (CE hypothesis), $\log(\rho_s)$ and $\log(\rho_{CV})$ should remain negligible relative to $\log(\rho_U)$, so that $\log(\rho_S) = 0$. Similarly, if only $\bar{s}$ varies between the two environments, (only $\log(\rho_s)$ is non-zero, CA hypothesis), then $\log(\rho_S) = \log(\rho_s) = ½ \log(\rho_V)$. Finally, if only CV($s$) varies between environments (only $\log(\rho_{CV})$ is non-zero, CV hypothesis), then $\log(\rho_S) = \log(\rho_{CV}) = \log(\rho_V)$. With several pairs of ρ_S and ρ_V estimates

(ratios from several pairs of environments and/or several studies), we can discriminate among the three hypotheses depending on the slope of the empirical relationship between $\log(\rho_S)$ and $\log(\rho_V)$, i.e. no relationship (CE) or a linear relationship with slope ½ (CA) or 1 (CV). These predictions are summarized in Table 1. Of course, this empirical relationship (if any) may differ from the three predicted trends (for instance if U, $\bar{s}$ and CV(s) all vary between environments or differently so in different species or experiments). Therefore, all three extreme hypotheses could easily be rejected, either by any non-linear relationship between $\log(\rho_S)$ and $\log(\rho_V)$, or by a linear relationship with a slope that differs from ½ or 1.

Stressful versus benign environments

In experiments measuring mutation fitness effects in two environments, one can often be considered more stressful than the other. The most stressful environment is the one in which the non-mutated initial genotype has the lowest absolute fitness. As above, U, $\bar{s}$ or CV(s) may differ between stressful and benign environments. However, these parameters may vary in a consistent direction with stress. For instance, it may be argued that U, $\bar{s}$ or CV(s) should increase in more stressful environments. To detect such a trend, we can take the same approach as above but systematically standardizing our ratios by values in the most benign environment, if such information is available. Denoting with or without a star the value in the benign or stressful environment, respectively, we can therefore compute $\rho_V \equiv \Delta V / \Delta V^*$, $\rho_M \equiv \Delta M / \Delta M^*$, and the corresponding $\rho_S \equiv \bar{s}_{BM} / \bar{s}_{BM}{}^*$. In this way, we can therefore determine whether U, $\bar{s}$ or CV(s) are systematically changed in stressful *vs*. benign environments and in which direction. For the sake of clarity, all ratios will be computed in this way (i.e. relative to the most benign environment) in the paper.

Survey of mutational G×E interactions for fitness

Mutation accumulation (MA) experiments is the most widely used method to generate a set of mutants from a single isogenic line. The fitness of these mutants can then be estimated, providing estimates of ΔM (the per generation average change in relative fitness due to mutation), and of ΔV (the per generation increment in relative fitness variance due to mutation). In addition, these moments can be measured for a given set of lines, in different environmental conditions, providing a measure of the change of ΔM and ΔV in different environments. We surveyed nine MA experiments (some using mutagenesis) reporting such variation, most of which are also discussed in (Fry and Heinsohn 2002). Our survey is summarized in Table 2. When it was not directly provided, we computed the mutational variance in *relative* fitness ΔV as the squared mutational coefficient of variation of the fitness trait measured, i.e. the increase in variance among mutant lines relative to control lines divided by the mean value of the control. Similarly, when not directly provided, we computed the average mutational change in relative fitness ΔM as the difference between the mean value of the fitness measure among mutants and the control value, divided by the control value. We considered that the least stressful environment was the one with the highest absolute fitness of the non-mutated (control) genotype (reported in Table 2). In some cases, fitness was measured in competition with a reference strain, in which case it was not possible to identify the least stressful environment for the control (as stress also affects the competitor). In these cases, the benign environment was defined according to the authors, usually as the "standard" laboratory environment to which the control line has adapted for generations, or in some studies, the low density environment. In most cases, the original papers provided unambiguously the required information. However, in some cases, we had to make some choices or to read some of the data on figures presented in the papers.

In the study of Fry et *al.* (1996) on *D. melanogaster*, fitness of the non-mutated genotype is not provided. We considered the strain used for the competitive assays as a

surrogate for this control genotype. The reproductive output/vial for this strain was read from Fig. 2A and 2B in the paper. This strain is not related to the MA lines so that it is not a proper control, but it was assumed free of mutation, based on its higher fitness (fig 2B), and on it having not undergone MA. It may however differ from the exact control (ancestor of MA lines) in its level of adaptation to some of environments, which could explain the strong difference observed between MA and "control" mean under low temperature (Fry and Heinsohn 2002; Fry et al. 1996). In any case, (i) this use of an improper control should only bias ΔM estimates (not ΔV) since this strain was isogenic or nearly so, and (ii) removing the estimates from this study does not affect any of our conclusions. In the study of Korona (1999) on *S. cerevisiae*, the among line variance of MA haploid lines ("M lines") was directly read on Fig. 1 in the paper, and the variance among control ("F") lines was set to 0 (from Fig. 1). For diploid strains, the among line variances of M/M and F/F strains were read on Fig. 3 in the paper. In the study of Xu (2004) on *C. neoformans*, two environments (37° and 25°) were used during the mutation accumulation itself. We pooled results from all MA lines (to a total of 16 lines). The absolute fitness of the controls which is not available in the paper was provided by the author. In (Fernandez and Lopez-Fanjul 1997), no control was available and we did not find an alternative control measure as in (Fry et al. 1996), so we only report effects of mutations on ΔV computed by neglecting the variance among control lines. This study is therefore not used in the relationship between $\log(\rho_S)$ and $\log(\rho_V)$ (Fig. 1), but only to assess the effect of stressful conditions on the sign of $\log(\rho_V)$ (Fig. 2), see Results. Conversely, in (Kishony and Leibler 2003), ΔV is not given, so that we only report ΔM. Finally, in two other studies of mutation effects across environments (Chang and Shaw 2003; Kavanaugh and Shaw 2005), there is no clear evidence of any fitness variance induced by mutation, in any environment, so we discarded these studies.

As explained above, to estimate the effect of the environment on ΔM and ΔV we used log-ratios estimates of ΔM (log (ρ_M)) and ΔV (log(ρ_V)) in a given stressful environment relative to the estimate in the benign environment (denoted ΔM^* and ΔV^* respectively). These measures are therefore standardised within each study, which allows to compare different experiments that may differ in the species used, the experimental design, the fitness measure, the number of MA generations (which may even be poorly known, e.g. in mutagenesis or microbe studies) or to specific features of the organism studied (e.g. ploidy or genome size that affect the mutational target size). Note also that the results in each experiment are based on a single set of lines having accumulated mutations in a common controlled environment so that there is no influence of environment-dependent molecular mutation rates (except potentially in Xu 2004).

RESULTS

On Fig. 1, we illustrate how log(ρ_S) varies with log(ρ_V) in the surveyed experiments. We find a clear linear relationship between log(ρ_S) and log(ρ_V). Since both log(ρ_S) and log(ρ_V) are measured with error, we use the reduced major axis to measure the slope of the relationship between the two variables (regression type II, Sokal and Rohlf 1995). Because there is an obvious outlier, we report the estimated slope with or without it. In addition, we report the slope assuming or not a zero intercept. Table 3 summarizes the estimates and their 95% bootstrap confidence limits. The fitted linear relationships between log(ρ_S) and log(ρ_V) give a good fit to the data, explaining 55 % or 88 % of the total variance (with or without the outlier, respectively), and the estimated slope ranges between 0.86 and 1.02 (depending on the model). This slope is not significantly different from 1 (expected under the CV hypothesis) except for the model with zero intercept excluding the outlier for which the slope 95% confidence interval is [0.80, 0.99]. In all cases, the estimated slope is significantly

different from ½ (CA hypothesis) and from 0 (CE hypothesis) (P-value < 0.0001). These results indicate that the observed pattern is very close to that expected under the CV hypothesis — the corresponding predicted relationship, $\log(\rho_S) = \log(\rho_V)$ explains 82% of the total variance, when excluding the outlier — with a slope perhaps slightly less than 1, however. This result strongly supports the idea that changing environments mainly changes the variance of mutation fitness effects (CV hypotheses) rather than their average effect (CA hypothesis) or their net expression level over the genome (CE hypothesis). Finally, note that this first conclusion does not depend on correctly assessing stressful *vs*. benign environments: the observed relationship $\log(\rho_S) \approx \log(\rho_V)$ is expected even when standardizing with a non-benign environment.

To scale the range of variation of the mean and variance of mutation effects, we considered the standard deviation $\sigma = \Delta V^{1/2}$, and its relative change under stressful conditions: $\log(\rho_\sigma) = ½ \log(\rho_V)$. Fig. 2 shows the distribution of $\log(\rho_M)$ and $\log(\rho_\sigma)$ in our survey. In the large majority of experiments (25 out of 30 estimates), stressful conditions result in an increase of the mutational variance (or σ) in fitness (i.e. CV hypothesis with a directional effect of stress): most $\log(\rho_\sigma)$ values are positive (two-tailed Wilcoxon signed-rank test, $p < 0.0001$). On the contrary, $\log(\rho_M)$ does not show the same pattern, increasing in only half of the cases (14 out of 28 estimates) and not showing any significant positive or negative sign (two-tailed Wilcoxon signed-rank test, $p = 0.73$). Therefore, contrary to their effect on ΔV, stressful conditions do not result in a consistent trend towards increased or decreased ΔM. Finally, note that the variation of $\log(\rho_M)$ is also smaller than that of $\log(\rho_\sigma)$ (see Fig. 2), although means and standard deviations are of the same scale. 96% of $\log(\rho_M)$ estimates fall in the range [-0.5,0.5] (i.e. all but the outlier mentioned above) whereas more than 23% (7/30) of $\log(\rho_\sigma)$ estimates fall outside this range.

INTERPRETATION IN TERMS OF FITNESS LANDSCAPES

Our survey reveals that stressful conditions tend to inflate the variance in mutational fitness effects (i.e. CV(s)) while leaving almost unaffected either U or $\bar{s}$. This pattern is consistent across species and experiments. Overall, this result suggests that the CV hypothesis is the prominent explanation for environmental variation of mutation fitness effects. The next step is of course to interpret this result: when do we expect such a pattern? Is it compatible with a mutation fitness effect model? In this section, we briefly present a fitness landscape model which provides a framework to interpret these empirical patterns.

A fitness landscape model of mutation fitness effects

A straightforward way to evaluate the effect of the environment on the distribution of mutation fitness effects $f(s)$, is to consider fitness landscape models, similar to Fisher's (1930) geometric model, whereby the fitness of a given phenotype falls off with the phenotypic distance to an optimum determined by the environment. Assuming a distribution of mutational effects on phenotypic traits, this approach provides a natural way to predict $f(s)$ at a given distance from the phenotypic optimum (Orr 2000; Welch and Waxman 2003). We can model a phenotype as a set of n phenotypic traits z_i represented by a column vector $\mathbf{z} = \{z_i\}_{i\in[1,n]}$, with fitness given by an arbitrary (twice differentiable) fitness function $W(\mathbf{z})$. Each MA line accumulates mutations causing a phenotypic displacement $\mathbf{dz} = \{dz_i\}_{i\in[1,n]}$ from an initial phenotype $\mathbf{z}_o$. In MA experiments, $\mathbf{z}_o$ can be thought of as the phenotype of the strain from which MA lines are derived. The fitness of this initial phenotype $\mathbf{z}_o$ is $W(\mathbf{z}_o)$ and a mutant line has phenotype $\mathbf{z}_o + \mathbf{dz}$ and fitness $W(\mathbf{z}_o + \mathbf{dz})$. W is the *absolute* fitness but our review focuses on the effect of mutation accumulation on *relative* fitness w, i.e. on the distribution of the fitness deviation of MA lines relative to the initial genotype. This deviation

for a line with phenotype $\mathbf{z}_o + \mathbf{dz}$ is $dw = (W(\mathbf{z}_o + \mathbf{dz}) - W(\mathbf{z}_o))/W(\mathbf{z}_o)$, which is the selection coefficient of the mutant line relative to its wild-type ancestor. Note that we do not denote it "s", which refers to the effect of single mutations, whereas *dw* may result from several mutations accumulated in a given line. If the deviations remain small, the effect of **dz** on *relative* fitness approximately equals its effect on *absolute* log-fitness ($\ln(W(\mathbf{z}))$ denoted $\ln W(\mathbf{z})$), $dw \approx \ln(1 + dw) = \ln W(\mathbf{z}_o + \mathbf{dz}) - \ln W(\mathbf{z}_o)$. Under the same assumption of small deviations, **dz** remains small around the initial phenotype $\mathbf{z}_o$, so that $\ln W(\mathbf{z}_o + \mathbf{dz}) - \ln W(\mathbf{z}_o)$ can be approximated by a second order multivariate Taylor Series around $\mathbf{z}_o$, yielding

$$dw \approx \ln(1+dw) = \sum_{i=1}^{n} \frac{\partial \ln W(\mathbf{z}_0)}{\partial z_i} dz_i \; + \; \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 \ln W(\mathbf{z}_0)}{\partial z_i \partial z_j} dz_i dz_j + o(dz^2) \qquad (3)$$

From this equation, we can compute the mean and variance of *dw* over mutant line effects **dz**, which are the quantities ΔM and ΔV, respectively, reported in our survey. From here, we assume that mutation effects, on phenotypic traits (**z**), are unbiased ($\mathrm{E}(dz_i) = 0$). This appears to be the most parsimonious assumption, as there is no clear trend expected or observed for the effect of mutations on e.g. morphological traits (for further discussion, see Martin and Lenormand 2006). Note that we make no assumption on the effect of mutation on fitness, the latter being derived from the model, not assumed. Then, keeping only terms up to the second order in *dz*, Eq. (3) yields

$$\begin{aligned} \Delta M &\approx \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 \ln W(\mathbf{z}_0)}{\partial z_i \partial z_j} \mathrm{E}\left(dz_i dz_j\right) + o\left(\mathrm{E}(dz^2)\right) \\ \Delta V &\approx \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \ln W(\mathbf{z}_0)}{\partial z_i} \frac{\partial \ln W(\mathbf{z}_0)}{\partial z_j} \mathrm{E}\left(dz_i dz_j\right) + o\left(\mathrm{E}(dz^2)\right) \end{aligned} \qquad (4)$$

This approximation shows that both the average and variance of the relative fitness of MA lines can be decomposed into two parts reflecting the genotype – phenotype relationship ($E(dz_i\ dz_j)$), and the phenotype – (log)fitness relationship (derivatives of $\ln W(\mathbf{z})$ at phenotype $\mathbf{z}_o$).

The terms $E(dz_i\ dz_j)$ are the variances and covariances of the effects of mutation accumulation on the underlying phenotypic traits (because $E(dz_i) = E(dz_j) = 0$). They quantify globally how and how much mutation accumulation affects the phenotype distribution among MA lines. They absorb the mutation rate and the way in which mutations have a phenotypic effect, in a given environment (patterns of expression and pleiotropy). Eq. (4) shows that the mean and variance of mutation effects are both proportional to $E(dz_i\ dz_j)$: both ΔM and ΔV scale with the amount of phenotypic change produced by mutation accumulation.

$\ln W(\mathbf{z})$ describes how phenotypic changes (dz_i) translate into relative fitness changes (dw) in a given environment. First, from Eq. (4), the average relative fitness effect of mutations depends on the local curvature of the log-fitness function $(\partial^2 \ln W(\mathbf{z}) / \partial z_i \partial z_j)$. Indeed, symmetrical variation in dz can only translate into a bias in dw if the phenotype-fitness relationship $\ln W(\mathbf{z})$ is non-linear. Second, the mutational variance in relative fitness (ΔV) is proportional to the product of first derivatives of $\ln W(\mathbf{z})$ taken at $\mathbf{z}_o$. This result is intuitively simple: variance in the underlying phenotypic traits (z_i) transforms into variance in fitness according to the local slope of the fitness function (to $\partial \ln W(\mathbf{z})/\partial z_i$) irrespective of its sign (hence the square).

For example, if the log-fitness is concave around $\mathbf{z}_o$ $(\partial^2 \ln W(\mathbf{z}) / \partial z_i \partial z_j < 0)$, (i) mutations are deleterious on average ($\Delta M < 0$, see (4)), and (ii) the variance ΔV increases with the distance to the optimum $|\mathbf{z}_o|$. Indeed, the absolute slope of $\ln W$ $\left(|\partial \ln W(\mathbf{z}) / \partial z_i|\right)$ increases as the initial phenotype $\mathbf{z}_o$ gets away from the optimum. This argument is illustrated

on Fig. 3 for the case of a quadratic log-fitness function. Note that concavity or convexity is defined locally in the range of phenotypes produced by mutations and in the environment where fitness is measured. The landscape may be rugged at a finer scale (i.e. with changing concavity), for instance at the level of DNA sequences. However we focus here on the fitness function measured at the scale of newly arising phenotypic variation. Note also that, from Eq. (4), there should be no variance in fitness at the optimum (where all $\partial \ln W/\partial z_i$ are zero). However, this conclusion only arises from our approximation in $o(dz^2)$. When all the first order derivatives are zero, the higher order terms ($O(dz^3)$, $O(dz^4)$, etc.) become leading order terms in ΔV, so that there still is some mutational variance in fitness at the optimum.

Overall and to summarize, ΔM is proportional to the curvature of $\ln W(\mathbf{z})$ at $\mathbf{z}_o$, whereas ΔV is proportional to the square of the slope of $\ln W(\mathbf{z})$ at $\mathbf{z}_o$, and both ΔM and ΔV are proportional to the mutational variance accumulated on the underlying phenotypic traits (hence to U). We now turn to interpret the effect of the environment in this landscape approximation.

Fitness effects in different environments in a fitness landscape model

Let us first consider that the main effect of the environment is to affect the genotype-phenotype relationship (the $E(dz_i\ dz_j)$). Such effect may be expected when some phenotypic traits are plastic, so that the same mutation has different phenotypic effects depending on the environment; then the (co)variances of mutation effects *on phenotypic traits* ($E(dz_i\ dz_j)$) may be environment-dependent. In this case, from Eq. (4) the variations of ΔM and ΔV across environments should be proportional, i.e. both proportional to the environmentally determined variation of $E(dz_i\ dz_j)$. Quite logically, this corresponds to the expected pattern under the CE hypothesis (see Table 1), which assumes that only U is affected by the

environment. Such variation of U may therefore reflect environmental variation of the proportion of expressed mutations or of their effect variance on the underlying traits (z_i).

However, different environments may also correspond to different optima for the phenotypic traits z_i. If we assume that the environment only alters the optimum and not the fitness function $\ln W(\mathbf{z})$ around this optimum, then different environments equivalently correspond to different positions of the initial phenotype $\mathbf{z}_o$. In particular, and following our definition, stressful and benign environments correspond to situations where the initial phenotype is at a small or large distance from the optimum, respectively. As the mean ΔM and variance ΔV of mutant relative fitnesses depend on the first and second derivatives of $\ln W$ at $\mathbf{z}_o$, the way in which ΔM and ΔV vary with the environment gives us information on the way the derivatives of $\ln W$ vary with $\mathbf{z}_o$. This in turn gives us information on the type of log-fitness function $\ln W(\mathbf{z})$. Therefore, different log-fitness functions correspond to distinct predictions regarding the relationship between ΔM and ΔV across environments (or equivalently between $\log(\rho_S)$ and $\log(\rho_V)$). Just as environmental effects on the genotype-phenotype relationship correspond to variation of U (CE hypothesis), variation of the distance to the optimum corresponds to variation of $\bar{s}$ and var(s) (CE and CV hypotheses respectively).

We illustrate the above argument on Fig. 4, showing the influence of the distance to the optimum on the mean ($\bar{s}$, Fig. 4.a) and variance (var(s), Fig. 4.b) of single mutation fitness effects, for various log-fitness functions. Specifically, we consider that $\ln W$ is a power function (of order k) of the distance to the optimum on each trait, i.e. a linear combination of $|z_i|^k$, which provides an easy way to consider different shapes for $\ln W$ by varying a single parameter (k), and reduces to the Gaussian case with $k = 2$. The distance to the optimum is defined as the log-fitness of a genotype lying at the optimum (W_{max}) relative to that of the initial genotype $W(\mathbf{z}_o)$: $s_o = \log(W_{max}/W(\mathbf{z}_o))$ (Martin and Lenormand 2006). We also allow

the direction of $\mathbf{z}_o$ (for a given s_o) to vary randomly across environments, as well as the coefficients of each of the $|z_i|^k$, with the constraint that their sum remains constant across traits z_i. Consequently, the variance among points (for a given s_o), mainly reflects that all directions in the phenotypic landscape are not equivalent (each point represents a particular direction $\mathbf{z}_o$, with different coefficients of each of the $|z_i|^k$). Fig 4.a. shows that varying the shape of the log-fitness function (i.e. varying the parameter k) strongly affects the change in $\bar{s}$ with the distance to the optimum (with s_o). In all these examples, as ln$W$ is concave, mutations are deleterious on average ($\bar{s}$ <0) and the variance of their effect increases with s_o, as predicted by Eq. (4). In the particular case k = 2, which corresponds to a Gaussian $W(\mathbf{z})$, $\bar{s}$ does not change with s_o because ln$W(\mathbf{z})$ is quadratic and therefore has a constant second derivative.

Interpreting the empirical patterns

Our survey revealed that mutations are always deleterious on average in all environments and that stressful conditions tend to inflate the variance in mutational fitness effects (i.e. CV(s) or equivalently var(s)) while leaving almost unaffected either U or $\bar{s}$. This corresponds to the CV hypothesis. Under our landscape model, this pattern would be expected if E($dz_i\ dz_j$) does not change in different environments and with a globally quadratic ln$W(\mathbf{z})$, that is a Gaussian function $W(\mathbf{z})$ with a constant width across environments and with an environment-dependent optimum. With this specific model, (i) U is constant in different environments because E($dz_i\ dz_j$) is constant, (ii) var(s) increases in more stressful environments because ln$W(\mathbf{z})$ is concave and (iii) $\bar{s}$ is negative and constant because ln$W(\mathbf{z})$ is concave with constant second derivatives (see Fig. 4, k = 2).

A continuum of *log*-fitness functions ln$W(\mathbf{z})$ would predict a continuum of relationships between log(ρ_S) and log(ρ_V). Fig. 5 illustrates this by showing the different relationships between log(ρ_S) and log(ρ_V) simulated for various types of log-fitness functions

(i.e. different k values). Fig. 5 shows that the relationships between $\log(\rho_S)$ and $\log(\rho_V)$ differs strongly whether k = 1, 1.5 or 2 but that it is more difficult to discriminate among higher k values (k = 2, 2.5 or 3). The relationship obtained from our survey $\log(\rho_S) = 0.87 \log(\rho_V)$, is close to the CV hypothesis $\log(\rho_S) = \log(\rho_V)$ (corresponding to k = 2). It is very different from the relationship expected for k values lower than 2. However, it may be consistent with k values slightly larger than 2 (e.g. maybe k = 2.5) since it is more difficult to distinguish among larger values of k, at least in very stressful conditions (i.e. away from the origin on Fig. 5). However, it is important to keep in mind that the number of mutations sampled in a typical MA experiment is not very large, which introduces an additional source of noise in the patterns that can be observed. Fig. 6 illustrates the expected relationship between $\log(\rho_S)$ and $\log(\rho_V)$ for a Gaussian fitness function (k = 2), when only 40 mutations are sampled around the initial phenotype (instead of 1000 in figures 4 and 5). It shows that there is considerable uncertainty in the precise relationship between $\log(\rho_S)$ and $\log(\rho_V)$ and that this uncertainty is much larger in very stressful conditions (i.e. away from the origin on the figure). Overall, given these various sources of uncertainty and given the uncertainty of the empirical estimates themselves, we believe that it is not justified to make a very precise statement about $W(\mathbf{z})$. However, the observation that the main effect of stress is to increase CV(s) with relatively little effect on U and $\bar{s}$ suggest that a landscape model with a constant $E(dz_i\, dz_j)$ and a Gaussian fitness function would be consistent with the data available.

DISCUSSION

The two main findings of this study are that stressful conditions tend to inflate var(s) while leaving almost unaffected either U or $\bar{s}$ and that this pattern is consistent with a simple Gaussian fitness landscape (quadratic log-fitness). In such landscape model, (i) the fitness

function is Gaussian (or nearly so) around an optimum that is determined by each environment, (ii) the parameters of this Gaussian fitness function around each optimum are little affected by the environment and (iii) the mutational variances and covariances on phenotypic traits do not vary much across environments. Points (i) and (ii) ensure that only var(s) and not $\bar{s}$ changes across environments and point (iii) ensures that U does not change with the environment. We now discuss the plausibility of this interpretation, and to what extent our findings are consistent with less restrictive assumptions.

Plausibility of the fitness landscape model proposed

It may seem surprising that a simplified fitness landscape model explains well empirical patterns across different species and environments. In the following section, we discuss the realism of some of the assumptions underlying this model, in particular, the existence of a single optimum and the constancy of parameters across environments. There are several lines of evidence indicating that simple fitness landscape models (or equivalently stabilizing selection models on many traits) may be more realistic than sometimes claimed. First the idea that an environmental change determines a new phenotypic optimum is supported by the long term dynamics of experimental adaptation to new environments. In both microbes (reviewed in Elena and Lenski 2003) and *Drosophila* (Gilligan and Frankham 2003), fitness typically plateaus in the long run, suggesting an approach to a new optimum. Second, the distribution of mutation fitness effects is gamma-like in many species, which is consistent with the predictions of an approximately Gaussian fitness landscape model (Martin and Lenormand 2006). Third, an increase in the proportion of beneficial mutations under stressful conditions as predicted by fitness landscape models has been documented in *E. coli* (Remold and Lenski 2001; Remold and Lenski 2004). Fourth, there is ample evidence for stabilizing selection on various traits (Hereford et al. 2004; Kingsolver et al. 2001) or

landscape-like relationships between enzymatic activities and fitness (Dean et al. 1986; Dean et al. 1988; Dykhuizen and Dean 1990; Dykhuizen et al. 1987; Dykhuizen and Hartl 1983).

Overall, the idea that adaptation may be modelled by some simple Gaussian or quadratic fitness function around an environment-dependant optimum seems not so unrealistic, although it would have been difficult to predict that a Gaussian fitness function assumption would *quantitatively* match the data so closely. However, to explain the empirical pattern, it is also necessary to assume that mutational and selective covariances on phenotypic traits remain constant across environments (points (ii) and (iii) above). This seems quite unrealistic, as we know that different genes are expressed in different environments and that a given trait may be strongly or weakly selected depending on the environment (i.e. the width of the Gaussian, not only its maximum, may change with the environment). First, it is possible that only a small portion of the genome has environment-dependent expression. Micro-array studies in micro-organisms suggest that only about 1% of the genome is activated or repressed in many specific stress responses (Wright 2004). However, this observation does not rule out that mutations phenotypic effects (if not expression) may strongly depend on the environment. Nevertheless, conditions (ii) and (iii) may be less restrictive than it seems. The only requirement is in fact that the *net* effect of the environment on *all* traits remains approximately constant, not that selective and mutational parameters on each of them be invariant. This is much less restrictive, particularly when considering a large number of traits. With the Gaussian fitness landscape model, $W(\mathbf{z})$ can be fully specified with a covariance matrix of selection intensity **S** (which is the multivariate measure of width of the fitness function). As a consequence the distribution of mutation fitness effects $f(s)$, in a given environment depends on this matrix **S**, on the mutational covariance matrix **M** describing the (co)variance of all traits by mutation (the $E(dz_i\ dz_j)$ in Eq. (4)) and on the optimal values for each trait. More specifically, $f(s)$ depends on the distribution of the eigenvalues of the matrix

S.**M**, on the number of traits and the distance to the optimum (Martin and Lenormand 2006). Conditions (ii) and (iii) are therefore overly restrictive. The empirical pattern that we observe would be consistent with different **S** and **M** matrices in different environments as long as the distribution of **S**.**M** eigenvalues and the number of traits stay approximately constant across environments. This condition would be met asymptotically (i.e. with many traits) if elements of **S** correspond to random draws in a distribution, which is arbitrary but identical in all environments, with the same requirement for **M** (Martin and Lenormand 2006). This argument which stems from random matrix theory (Bai 1999), indicates that random variation of **S** (width of the fitness function) and **M** (mutational variance) across environments (in addition to a change in the optimum) may be undistinguishable from exactly constant **S** and **M** as far as mutations' fitness effects are concerned. This fact was illustrated in Fig. 4 and 5, where we allowed for such random variation of mutational and selective parameters, and found robust relationships between $\bar{s}$, var(s) and the fitness distance to the optimum (s_o). In any case, although a Gaussian fitness landscape model does not fully describe the relationship between phenotype, fitness and the environment, it may nonetheless be a sufficiently robust simplification as it captures empirical patterns surveyed in this paper. As such, it may be a useful and reasonably accurate model for mutation fitness effects.

However, a Gaussian fitness landscape model does not account for all the data. For instance, in our survey (see Table 2, Fig. 1 and 5), a minority of experiments found that ΔV was lower in more stressful conditions (instead of higher in a Gaussian fitness landscape). In many cases, the difference is not large and may be simply accounted for by measurement error on the relative fitness of the mutant lines or to the low number of mutations sampled among the mutant lines. It is also possible that the stressful and benign environments were ill-attributed in some cases (i.e. when no stress measure was provided, see Table 2). However, it is also possible that these measures genuinely reflect that ΔV is sometimes lower in more

stressful conditions. It is possible to theoretically expect these results with a fitness function with a narrower plateau around the optimum than the Gaussian ($k < 2$, see Fig. 5), although the data available do not globally support this possibility. Of course, it is also possible that the fitness function differs among species and environment, which could then easily account for these observations. In fact, the observation that most data could be interpreted with a single landscape model is the most intriguing result emerging from our survey. However, it seems quite likely that there is variation, even if it is modest, in the shape of the landscape among species and environments, and more data are clearly needed to settle this issue. In line with this discussion, it is worth mentioning the mutation accumulation of Burch and Chao (2004) on RNA bacteriophage Φ6. In this experiment, the authors measure ΔM and ΔV using more or less adapted initial genotypes in a single environment, instead of the same initial genotype in different environments as in our survey. This type of experiment can be interpreted in a similar way as experiments involving different environments (the variation in the distance to the optimum (s_o) is given by the fitness of distinct initial genotypes in a single environment, instead of distinct environments determining the fitness of a single line). Contrary to the main result of our survey, they found that s_{BM} decreased with increasing maladaptation of the initial genotype (s_o). Just as in the present review, ρ_S and ρ_V can be computed from their MA data (not shown), and their mutual relationship across initial genotypes (instead of across environments) can be obtained. This relationship is more consistent with a *linear* log-fitness function ($k = 1$), than with the quadratic function ($k = 2$) suggested by our survey. As far as this interpretation is correct, the discrepancy between the effects of maladaptation generated by environmental change vs. mutation accumulation suggests that a comparison between the two types of experiments would deserve further investigation. Overall, the fact that the fitness function compatible with these data on Φ6 may be quite different from what we found in our survey also indicate that the shape of fitness functions may differ across species (viruses may

show a particular pattern compared to "higher" species) although further experiments are needed in both higher organisms and microbes to assess the generality of this conclusion; the framework we propose in this paper may also be useful for this purpose.

Increased var(s) under stressful conditions

Our results suggest that stressful conditions (at least those considered in our survey) mainly increase the variance of deleterious mutation fitness effects, and have a modest influence (if any) on their average effect (CA hypothesis) or on their total level of expression (CE hypothesis). The finding that stressful conditions tend to increase mutational variance is of course not totally new. For instance, Hermisson and Wagner (2004) proposed a general mechanism for the increase in mutational variance on a quantitative trait (not necessarily fitness related) after an environmental change. However they focused on hidden *standing* genetic variation whereas our study focuses on newly arisen mutational effect, and on their fitness consequence only. An increase in ΔV under stress has also been mentioned previously by several authors (reviewed in Fry and Heinsohn 2002), although without testing between the possible causes of this increase. Our results are also consistent with another study (Remold and Lenski 2001) showing that the variance in fitness among lines carrying a single mutation increased in stressful conditions whereas the average deleterious effect remained unchanged (ΔV and ΔM estimates are however not given in that article). Moreover, the same authors (Remold and Lenski 2004) also studied the fitness effect of single mutations across five genetic backgrounds and in two environments. Among the eighteen mutations studied, some were conditionally neutral (in the sense of Kawecki et al. 1997), but only on specific genetic backgrounds. However, many mutations (7/18) were simply neutral in both environments (unconditionally neutral). Together, although based on a limited amount of studies, these results suggest that stressful conditions do increase the variance of mutation

fitness effects, but not necessarily (and maybe rarely) because of an increase in the expression of deleterious mutations. There are many examples of genes being activated in response to stress (e.g. oxidative, temperature or osmotic stresses), both in the yeast (Toone and Jones 1998) or in mammals (Sonna et al. 2002). Mutations on these genes should indeed be neutral in benign environments, but stress responses are also known to down-regulate some other genes (e.g. in the response to oxidative stress Morel and Barouki 1999). Other stresses may result in the switch to a new resource utilization pathway. In both situations, different genes are being either up or down regulated and the net outcome may thus not necessarily be an increase in the total number of expressed mutations in stressful conditions.

Conditional neutrality, G x E interactions and ecological specialization

Differences in mutation effects across environments are a necessary ingredient for the evolution of ecological specialization (Futuyma and Moreno 1988). Ecological specialization or local adaptation may occur if the direction of selection changes for an allele between environments (antagonistic pleiotropy). It may also occur if the intensity of selection against deleterious mutations at several loci covaries negatively between environments (Fry et al. 1996; Whitlock 1996). The first scenario requires the existence of a trade-off and the occurrence of mutations that are beneficial in at least some environments whereas the second scenario works best when mutations are neutral in one environment but deleterious in another (i.e. under conditional neutrality Kawecki et al. 1997) and accords with the common view that most mutations are deleterious. Many empirical studies document the evolution of ecological specialists in constant environments or of locally adapted genotypes in heterogeneous environments (reviewed in Kassen 2002; Lenormand 2002) and the prevalence of conditional neutrality *vs.* antagonistic pleiotropy has been much debated in this context (cost of specialization Cooper and Lenski 2000; MacLean et al. 2004). There is evidence for the

different types of mutation presented above. In particular, conditionally neutral genetic variation at quantitative trait loci (some of which is revealed by stressful conditions) has been documented in several experiments (reviewed in Hermisson and Wagner 2004). However, some other QTLs with significant fitness effects may show less environment dependence, and evidence for conditional neutrality on traits with very limited impact on fitness (e.g. bristle-number) is not evidence for conditional neutrality for fitness. There is also evidence for antagonistic pleiotropy (Cooper and Lenski 2000; Gazave et al. 2001; MacLean et al. 2004). However, the relative frequency of these different types of mutations is not clearly documented apart from a recent study (Remold and Lenski 2004).

If simple fitness landscape models are a reasonably accurate approximation, as our results suggest, they could provide a rationale to predict the proportion and impact of each of the above type of mutation. Indeed, G x E interactions for mutation fitness effects are inherent to a fitness landscape model in which different environments are characterized by different optimal values for the underlying traits. Fig. 7 sketches the different types of mutation fitness effects (relative to an initial phenotype $\mathbf{z}_o$) that may occur in a simple two-traits landscape with two different optima, O_1 and O_2, determined by two contrasted environments. First, a given mutation **dz** may increase the phenotypic distance from both O_1 and O_2 if the phenotype $\mathbf{z}_o$ + **dz** lies in the white area. Such a mutation would be deleterious in both environments (- / - area) although this deleterious effect may be more severe in one of them. Second, a given mutation may increase the distance from only one of the two optima. Such a mutation would be conditionally neutral or deleterious (- / 0 area). Third, a mutation may not significantly change the distance from either optimum and be neutral (0 / 0 area). Fourth, a mutation may decrease the distance from only one of the optima and be conditionally neutral or beneficial (0 / + area). Fifth, a mutation may decrease the distance from one optimum but increase the distance from the other. Such a mutation would be antagonistic pleiotropic (+ / -

area). Last, a mutation may decrease the distance from both optima and be unconditionally beneficial (+ / + area). The proportion of these different types of mutation depends on the relative position of the optima and the initial phenotype and may be predictable in a given landscape and for a given fitness effect threshold defining a "neutral" mutation.

The evolution of specialization mainly relies on - / 0 mutations (conditional neutrality) and + / - mutations (antagonistic pleiotropy). Our results suggest that whether the environment is stressful or not does not change the total number of expressed deleterious alleles (i.e. of conditionally neutral mutations). However, this does not rule out the potential for ecological specialization by conditional neutrality (in the sense of Kawecki et al. 1997), as different deleterious mutations may be expressed in different environments, even if their total number remains roughly constant across environments. In any case, our survey is mainly concerned with unconditionally deleterious - / - mutations and is not directly relevant to these other contexts, but we note that fitness landscape models may be useful to quantitatively predict the relative prevalence of conditional neutrality and antagonistic pleiotropy.

Implications for the estimation of mutation parameters

Last, and perhaps more importantly, our results suggest that stressful conditions tend to increase the coefficient of variation of mutation fitness effects CV(*s*). This effect could lead to strong underestimation of the mutation rate by the Bateman-Mukai method when fitness is assayed in stressful conditions, sometimes by up to two orders of magnitude. Therefore, it seems a priori wisest to rely on *U* estimates based on fitness assays in an environment to which the control genotype is well adapted. Fortunately, most estimates have been done in this context. Considering that *U* does not vary much across environments compared to CV(*s*), may also be useful when analysing MA data across environments using maximum-likelihood (Keightley 1994; Vassilieva et al. 2000) or minimum distance (Garcia-

Dorado and Marin 1998). In any case, a significant change in *U vs*. CV(*s*) across environments can be tested with these methods, to further infirm or confirm the results of the present study.

Acknowledgments:

We thank P. Jarne, S.F. Elena, D. Waxman and two reviewers for helpful comments on the manuscript, and O. Tenaillon for discussions on our interpretation of empirical results. We also thank J.P. Xu for kindly providing data on his study of *Cryptococcus neoformans*. This work was supported by an ACI grant 0693 from the French Ministry of Research to T. L.

REFERENCES

Bai, Z. D. 1999. Methodologies in spectral analysis of large dimensional random matrices, a review. Statistica Sinica 9:611-662.

Burch, C. L., and L. Chao. 2004. Epistasis and its relationship to canalization in the RNA virus phi 6. Genetics 167:559-567.

Chang, S. M., and R. G. Shaw. 2003. The contribution of spontaneous mutation to variation in environmental response in *Arabidopsis thaliana*: Responses to nutrients. Evolution 57:984-994.

Charlesworth, B., and D. Charlesworth. 1998. Some evolutionary consequences of deleterious mutations. Genetica 103:3-19.

Cooper, V. S., and R. E. Lenski. 2000. The population genetics of ecological specialization in evolving *Escherichia coli* populations. Nature 407:736-739.

Dean, A. M., D. E. Dykhuizen, and D. L. Hartl. 1986. Fitness as a Function of Beta-Galactosidase Activity in Escherichia-Coli. Genet. Res. 48:1-8.

Dean, A. M., D. E. Dykhuizen, and D. L. Hartl. 1988. Fitness Effects of Amino-Acid Replacements in the Beta-Galactosidase of Escherichia-Coli. Mol. Biol. Evol. 5:469-485.

Dykhuizen, D. E., and A. M. Dean. 1990. Enzyme-Activity and Fitness - Evolution in Solution. Tends Ecol. Evol. 5:257-262.

Dykhuizen, D. E., A. M. Dean, and D. L. Hartl. 1987. Metabolic Flux and Fitness. Genetics 115:25-31.

Dykhuizen, D. E., and D. L. Hartl. 1983. Functional-Effects of Pgi Allozymes in Escherichia-Coli. Genetics 105:1-18.

Elena, S. F., and R. E. Lenski. 2003. Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. Nat. Rev. Genet. 4:457-469.

Fernandez, J., and C. Lopez-Fanjul. 1997. Spontaneous mutational genotype-environment interaction for fitness-related traits in *Drosophila melanogaster*. Evolution 51:856-864.

Fisher, R. A. 1930. The genetical theory of natural selection. Oxford University Press, Oxford.

Fry, J. D. 1996. The evolution of host specialization: Are trade-offs overrated? Am. Nat. 148:S84-S107.

Fry, J. D., and S. L. Heinsohn. 2002. Environment dependence of mutational parameters for viability in *Drosophila melanogaster*. Genetics 161:1155-1167.

Fry, J. D., S. L. Heinsohn, and T. F. C. Mackay. 1996. The contribution of new mutations to genotype-environment interaction for fitness in *Drosophila melanogaster*. Evolution 50:2316-2327.

Futuyma, D. J., and G. Moreno. 1988. The Evolution of Ecological Specialization. Annu. Rev. Ecol. Syst. 19:207-233.

Garcia-Dorado, A., C. Lopez-Fanjul, and A. Caballero. 1999. Properties of spontaneous mutations affecting quantitative traits. Genet. Res. 74:341-350.

Garcia-Dorado, A., and J. M. Marin. 1998. Minimum distance estimation of mutational parameters for quantitative traits. Biometrics 54:1097-1114.

Gazave, L., C. Chevillon, T. Lenormand, M. Marquine, and M. Raymond. 2001. Dissecting the cost of insecticide resistance genes during the overwintering period of the mosquito Culex pipiens. Heredity 87:441-448.

Gilligan, D. M., and R. Frankham. 2003. Dynamics of genetic adaptation to captivity. Conserv. Genet. 4:189-197.

Hereford, J., T. F. Hansen, and D. Houle. 2004. Comparing strengths of directional selection: How strong is strong? Evolution 58:2133-2143.

Hermisson, J., and G. P. Wagner. 2004. The population genetic theory of hidden variation and genetic robustness. Genetics 168:2271-2284.

Kassen, R. 2002. The experimental evolution of specialists, generalists, and the maintenance of diversity. J. Evol. Biol. 15:173-190.

Kavanaugh, C. M., and R. G. Shaw. 2005. The contribution of spontaneous mutation to variation in environmental responses of Arabidopsis thaliana: Responses to light. Evolution 59:266-275.

Kawecki, T. J., N. H. Barton, and J. D. Fry. 1997. Mutational collapse of fitness in marginal habitats and the evolution of ecological specialisation. J. Evol. Biol.10:407-429.

Keightley, P. D. 1994. The Distribution of Mutation Effects On Viability in *Drosophila melanogaster*. Genetics 138:1315-1322.

Keightley, P. D. 2004. Comparing analysis methods for mutation-accumulation data. Genetics 167:551-553.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245-261.

Kishony, R., and S. Leibler. 2003. Environmental stresses can alleviate the average deleterious effects of mutations. J. Biol. 2:14.

Kondrashov, A. S., and D. Houle. 1994. Genotype-Environment Interactions and the Estimation of the Genomic Mutation-Rate in *Drosophila melanogaster*. Proc. R. Soc. Lond. [Biol] 258:221-227.

Korona, R. 1999. Genetic load of the yeast *Saccharomyces cerevisiae* under diverse environmental conditions. Evolution 53:1966-1971.

Korona, R. 2004. Experimental studies of deleterious mutation in *Saccharomyces cerevisiae*. Res. Microbiol. 155:301-310.

Lenormand, T. 2002. Gene flow and the limits to natural selection. Trends Ecol. Evol. 17:183-189.

Lynch, M., J. Blanchard, D. Houle, T. Kibota, S. Schultz, L. Vassilieva, and J. Willis. 1999. Perspective: Spontaneous deleterious mutation. Evolution 53:645-663.

MacLean, R. C., G. Bell, and P. B. Rainey. 2004. The evolution of a pleiotropic fitness tradeoff in *Pseudomonas fluorescens*. Proc. Natl. Acad. Sci. USA 101:8072-8077.

Martin, G., and T. Lenormand. 2006. A multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60:893-907.

Morel, Y., and R. Barouki. 1999. Repression of gene expression by oxidative stress. Biochem. J. 342:481-496.

Mukai, T., S. I. Chigusa, L. E. Mettler, and J. F. Crow. 1972. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. Genetics 72:335-355.

Orr, H. A. 2000. Adaptation and the cost of complexity. Evolution 54:13-20.

Parsons, P. A. 1991. Evolutionary Rates - Stress and Species Boundaries. Annu. Rev. Ecol. Syst. 22:1-18.

Remold, S. K., and R. E. Lenski. 2001. Contribution of individual random mutations to genotype-by-environment interactions in *Echerichia coli*. Proc. Natl. Acad. Sci. USA 98:11388-11393.

Remold, S. K., and R. E. Lenski. 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*. Nat. Genet. 36:423-426.

Sokal, R., and F. Rohlf. 1995. Model II regression. Pp. 541-554. Biometry: the principle and practice of statistics in biological research. W.H. Freeman and Company, New York.

Sonna, L. A., J. Fujita, S. L. Gaffin, and C. M. Lilly. 2002. Invited Review: Effects of heat and cold stress on mammalian gene expression. J. Appl. Physiol. 92:1725-1742.

Szafraniec, K., R. H. Borts, and R. Korona. 2001. Environmental stress and mutational load in diploid strains of the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 98:1107-1112.

Toone, W. M., and N. Jones. 1998. Stress-activated signalling pathways in yeast. Genes to Cells 3:485-498.

Vassilieva, L. L., A. M. Hook, and M. Lynch. 2000. The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. Evolution 54:1234-1246.

Welch, J. J., and D. Waxman. 2003. Modularity and the cost of complexity. Evolution 57:1723-1734.

Whitlock, M. C. 1996. The red queen beats the jack-of-all-trades: The limitations on the evolution of phenotypic plasticity and niche breadth. Am. Nat. 148:S65-S77.

Wright, B. E. 2004. Stress-directed adaptive mutations and evolution. Mol. Microbiol. 52:643-650.

Xu, J. P. 2004. Genotype-environment interactions of spontaneous mutations for vegetative fitness in the human pathogenic fungus *Cryptococcus neoformans*. Genetics 168:1177-1188.

Yang, H. P., A. Y. Tanikawa, W. A. Van Voorhies, J. C. Silva, and A. S. Kondrashov. 2001. Whole-genome effects of ethyl methanesulfonate-induced mutation on nine quantitative traits in outbred *Drosophila melanogaster*. Genetics 157:1257-1265.

Figure legend

Fig. 1 Effect of stressful conditions on mutational variance in fitness and Bateman-Mukai estimates of $\bar{s}$. The log-ratio of ΔV estimates ($\log(\rho_V)$, x-axis) and of s_{BM} ($\log(\rho_S)$, y-axis), the Bateman-Mukai estimates of $\bar{s}$, in the stressful *vs*. benign environments are from Table 2 (with s_{BM} computed from Eq. (2)) for various species (indicated on the graph). The black plain line gives the observed linear relationship between estimates (excluding the outlier, top of the graph). Grey dashed lines give the predicted linear relationships for each of the three hypotheses ("CE", "CA", "CV", see text) indicated on the graph.

Fig. 2 Distribution of the relative change in mutational mean and standard deviation of fitness in stressful *vs.* benign environments. Mutation effects on mean (ΔM) and variance (ΔV) in relative fitness are given relative to their values in the benign environment (ΔM^* and ΔV^*). Values on the x-axis refer to the log relative change in mean $\log(\rho_M) = \log(\Delta M/\Delta M^*)$ (28 estimates) or standard deviation $\log(\rho_\sigma) = ½ \log(\Delta V/\Delta V^*)$ (30 estimates). Positive values correspond to an increase in the mean or variance of mutation effects in stressful conditions. All but one (96%) of the $\log(\rho_M)$ estimates fall within the range [-0.5 , 0.5].

Fig. 3 Effect of stress on the mean and variance of *s* with a quadratic fitness function. A fitness landscape for a single phenotypic trait *z* is represented with a Gaussian fitness function (quadratic *log*-fitness function) $\ln W(z) \propto - z^2$. Mutations cause a phenotypic variation around the initial phenotype (x-axis). In grey the initial phenotype is close to the optimum ($z_o = 0$ benign environment), in black, the initial phenotype is maladapted to the environment ($z_o \neq 0$,

stressful environment). The resulting mean $\bar{s}$ and variance V(*s*) of mutation fitness effects are represented in each case on y-axis.

Fig. 4 Variation of $\bar{s}$ and var(*s*) with the fitness distance to the optimum (s_o) with diffrent fitness functions. The fitness functions are of the form $W(\mathbf{z}) = \mathrm{Exp}(-\tfrac{1}{2}\sum_{i=1}^{n}|z_i|^k \lambda_i)$, where k = 1,1.5, 2, 2.5 or 3, as indicated on the graph. The parameter k indicates whether the fitness function has a wider ($k > 2$) or narrower ($k < 2$) plateau around the optimum compared to a Gaussian fitness function ($k = 2$). Several steps are needed to obtain one dot on the figure. First two random variance covariance matrices **S** and **M** of size $n = 50$ are drawn and scaled such that if $k = 2$, $E(\mathrm{Ln}(1+s)) \approx \bar{s}$ would be -0.05. The eigenvalues of the product **S.M** give the λ_i in the expression of $W(\mathbf{z})$. Then the n trait values of the initial phenotype ($\mathbf{z}_o$) are drawn randomly, giving its fitness distance to the optimum $s_o = \log(W_{max}/W(\mathbf{z}_o))$ where $W_{max} = W(\mathbf{0}) = 1$. Then 1000 mutants are drawn around this initial phenotype (i.e. 1000 deviation vectors $\mathbf{dz} = \{dz_i\}_{i\in[1,n]}$, where the dz_i are drawn into independent Gaussians $N(0,1)$). The relative fitness of each mutation is computed as $s(\mathbf{dz}) = W(\mathbf{z}_o + \mathbf{dz})/W(\mathbf{z}_o) - 1$, and used to compute the mean $\bar{s}$ and variance var(*s*) of *s* among mutants. Increasing the number of traits (n) does not change qualitatively the outcome but magnifies the differences among different fitness functions (not shown).

Fig. 5 Variation of log(ρ_S) against log(ρ_V) with different fitness functions. We use the same simulations as in Fig. 4 (with the same grey-level code for each k value). Log(ρ) ratios are computed relative to the case $s_o = 0$. The dashed line corresponds to $\mathrm{Log}(\rho_S) = \mathrm{Log}(\rho_V)$ (CV hypothesis).

Fig. 6 Variation of log(ρ_S) against log(ρ_V) for a Gaussian fitness function (k =2). This figure illustrates the same type of simulations as in Fig. 4 and 5 except that, for each point, only 40 mutants are drawn around the initial phenotype (instead of 1000). Log(ρ) ratios are computed relative to the case s_o = 0. This figure illustrates that drawing a small number of mutations around the initial phenotypes, which is typical of most MA experiments, results in a large uncertainty in the expected relationship between log(ρ_S) and log(ρ_V). The large dark dots represent the observed values surveyed in Table 2. The dashed line corresponds to the observed relationship log(ρ_S) = 0.87 log(ρ_V) (see Table 3, w/o outlier, intercept = 0).

Fig. 7 Mutational G x E interactions for fitness in a fitness landscape model. The figure represents a phenotypic space with two phenotypic traits (z_1 and z_2) with two different optima in two distinct environments ($\mathbf{O}_1$ and $\mathbf{O}_2$). Mutations on a given initial phenotype (black dot) in a given environment may be deleterious (-), neutral (0) or advantageous (+), depending on whether it brings the phenotype away from or closer to the corresponding optimum ($\mathbf{O}_1$ or $\mathbf{O}_2$). The joint effect of a mutation in both environments depends on the position of the mutant in the phenotypic plan. Each coloured area represents a zone in which all mutations have the same qualitative joint effect, encoded as (effect in one environment / effect in the other one). See text for details.

Tables

Table 1. Summary of the predictions. The table summarizes the patterns expected under the three extreme hypothesis: "CE", "CA" and "CV" (see text). As in the text, ΔM and ΔV refer to the mutational change in mean and variance of relative fitness, respectively. The ρ 's are ratios of estimates in stressful *vs*. benign environments: ρ_V for ΔV and ρ_s for Bateman-Mukai estimates of the average fitness effect of mutations. The predicted relationships are also illustrated on Figure 1.

Hypothesis	Prediction across environments	
CE: conditional expression (U varies)	$\Delta M \propto \Delta V$	$\log(\rho_s) = 0$
CA: conditional average ($\bar{s}$ varies)	$\Delta M^2 \propto \Delta V$	$\log(\rho_s) = ½ \log(\rho_V)$
CV: conditional variance (CV(s) varies)	ΔM constant ΔV varies	$\log(\rho_s) = \log(\rho_V)$

Table 2. Environmental variation on mutational mean and variance in fitness. Estimates of mutational mean ΔM and variance ΔV in fitness (relative to the control), given in various environments: st: standard laboratory environment, low (high) D: low (high) density, low (high) T: low (high) temperature, low (high) F: low (high) food quantity or quality (includes diluted media), substance added to the medium when indicated. Fitness traits: *r*: population growth rate, *w*/vial: reproductive output per vial, *dens.*: final population density, *viability*: viability in competition with a reference strain. $\log(\rho_M) = \log(\Delta M/\Delta M^*)$ and $\log(\rho_V) = \log(\Delta V/\Delta V^*)$ and $\log(\rho_U) = 2\log(\rho_M) - \log(\rho_V)$ are the log-ratios of ΔM, ΔV and U_{BM} (Bateman-Mukai estimate of U, $U_{BM} = \Delta M^2/\Delta V$) respectively, relative to their values in the benign environment (indicated in bold). When it was provided in the studies, the control fitness measure is given together with the ratio of control fitness in the stressful *vs.* benign environment, as a standardized measure of stress. Data sources are listed below with the corresponding number of MA generations (into brackets) and with a "*" when ΔV and ΔM are per generation estimates: $1^{(1)}$ and $1^{(2)}$: experiments 1 and 2 in (Fry and Heinsohn 2002)[27-35]*; 2: (Fry et al. 1996)[202]*; 3: (Yang et al. 2001)[mutagenesis ~100]; 4: (Fernandez and Lopez-Fanjul 1997)[104-160]*; 5: (Vassilieva et al. 2000)[214]*; 6: (Szafraniec et al. 2001)[40]; 7: (Korona 1999)[500]; 8: (Xu 2004)[600]; 9: (Kishony and Leibler 2003)[mutagenesis, assumed to be equivalent to 100 generations]. *n* (2*n*): haploid (diploid) strains of *Saccharomyces cerevisiae*. All estimates are for homozygous or haploid effects except (^{H}) heterozygous effects.

species	environment	*trait*	control fitness	% fitness in benign	ΔM	$\log(\rho_M)$	ΔV	$\log(\rho_V)$	$\log(\rho_U)$	source
D. melanogaster	**low D**	***viability***			**-0.0029**		**5.9E-5**			$1^{(1)}$
D. melanogaster	st	*viability*			-0.001	-0.462	3.5E-4	0.78	-1.70	$1^{(1)}$
D. melanogaster	low T	*viability*			-0.004	0.140	1E-3	1.25	-0.97	$1^{(1)}$
D. melanogaster	ethanol	*viability*			-0.0037	0.106	3.8E-4	0.81	-0.59	$1^{(1)}$
D. melanogaster	**low D**	***viability***			**-0.0015**		**5.1E-5**			$1^{(2)}$
D. melanogaster	st	*viability*			-0.0032	0.329	6.9E-4	1.13	-0.47	$1^{(2)}$

D. melanogaster	low T	*viability*			-0.0041	0.437	8.9E-4	1.24	-0.37	1 [(2)]
D. melanogaster	ethanol	*viability*			-0.0039	0.415	1.8E-3	1.55	-0.72	1 [(2)]
D. melanogaster	**st**	***w/vial***	**165**		**-0.0017**		**5.4E-5**			2
D. melanogaster	low T	*w/vial*	157.5	95%	-0.004	0.385	2.9E-5	-0.27	1.04	2
D. melanogaster	tomato	*w/vial*	138.095	84%	-0.0014	-0.082	5.8E-5	0.04	-0.20	2
D. melanogaster	ethanol	*w/vial*	120	73%	-0.0017	0.000	5.4E-5	0.00	0.00	2
D. melanogaster [(H)]	**low D**	***viability***			**-0.0136**		**3.1E-3**			3
D. melanogaster [(H)]	low F+high D	*viability*			-0.0129	-0.023	1.8E-2	0.76	-0.80	3
D. melanogaster [(H)]	low F	*viability*			-0.009	-0.179	7.5E-4	-0.62	0.26	3
D. melanogaster	**st**	***fecundity***			_	_	**1E-4**		_	4
D. melanogaster	high T	*fecundity*			_	_	2.28E-4	0.358	_	4
D. melanogaster	NaCl	*fecundity*			_	_	1.82E-04	0.261	_	4
D. melanogaster	low F	*fecundity*			_	_	1.72E-04	0.235	_	4
D. melanogaster	**st**	***early viab.***			_	_	**9.8E-5**		_	4
D. melanogaster	high T	*early viab.*			_	_	2.79E-4	0.454	_	4
D. melanogaster	NaCl	*early viab.*			_	_	4.62E-5	-0.326	_	4
D. melanogaster	low F	*early viab.*			_	_	6.56E-5	-0.174	_	4
D. melanogaster	**st**	***late viab.***			_	_	**1.6E-5**		_	4
D. melanogaster	high T	*late viab.*			_	_	1.94E-5	0.083	_	4
D. melanogaster	low F	*late viab.*			_	_	1E-4	0.796	_	4
C elegans	**st**	***r***	**1.309**		**-0.0008**		**1.6E-5**			5
C elegans	low T	*r*	0.39	30%	-0.0015	0.273	5.3E-4	1.52	-0.97	5
S. cervisiae 2n [(H)]	**st**	***r***	**0.418**		**-0.0024**		**1.9E-6**			6
S. cervisiae 2n [(H)]	high T	*r*	0.247	59%	-0.19	1.901	9.4E-6	0.63	3.17	6
S. cervisiae 2n [(H)]	**st**	***dens.***	**1.537**		**0**		**2.3E-7**			6
S. cervisiae 2n [(H)]	high T	*dens.*	1.179	77%	-0.083		3.9E-3	4.22		6
S. cervisiae 2n	**st**	***r***	**0.705**		**-0.067**		**1E-2**			7
S. cervisiae 2n	low F	*r*	0.503	71%	-0.0457	-0.164	3.6E-2	0.55	-0.88	7
S. cervisiae n	**st**	***r***	**0.675**		**-0.262**		**7.3E-3**			7
S. cervisiae n	low T	*r*	0.147	22%	-0.236	-0.045	3.4E-2	0.67	-0.76	7
S. cervisiae n	high T	*r*	0.502	74%	-0.5	0.280	2.8E-1	1.58	-1.02	7
S. cervisiae n	low F	*r*	0.534	79%	-0.194	-0.130	1E-2	0.15	-0.41	7
S. cervisiae n	glycerol	*r*	0.272	40%	-0.272	0.016	7.6E-3	0.01	0.02	7
C. neoformans	**st**	***r***	**0.5333**		**-0.3085**		**3.7E-2**			8
C. neoformans	low F+low T	*r*	0.14861	28%	-0.32	0.016	7.3E-2	0.29	-0.26	8
C. neoformans	low T	*r*	0.27083	51%	-0.167	-0.267	1E-2	-0.54	0.00	8
C. neoformans	low F	*r*	0.20625	39%	-0.44	0.155	1.2E-1	0.50	-0.19	8
E. coli	st	*r*		73%	-0.275	0.008	_	_	_	9
E. coli	acidic	*r*		61%	-0.300	0.046	_	_	_	9
E. coli	**high F**	***r***			**-0.270**		**_**	**_**	**_**	9
E. coli	NaCl	*r*		48%	-0.250	-0.033	_	_	_	9
E. coli	redox agent	*r*		54%	-0.210	-0.109	_	_	_	9
E. coli	antibiotic	*r*		43%	-0.100	-0.431	_	_	_	9
E. coli	antibiotic	*r*		48%	-0.150	-0.255	_	_	_	9
E. coli	low T	*r*		12%	-0.150	-0.255	_	_	_	9

Table 3. Summary of statistics for the regression in Fig. 1

Model	Intercept	95% min intercept	95% max intercept	Slope (reduced major axis)	95% min slope	95% max slope	R^2
All data	-0.16	-0.75	0.11	1.03	0.78	1.37	0.55
All data	0	--	--	0.91	0.82	1.06	0.56
w/o outlier	0.02	-0.13	0.13	0.86	0.75	1.00	0.88
w/o outlier	0	--	--	0.87	0.80	0.99	0.88

Fig. 1

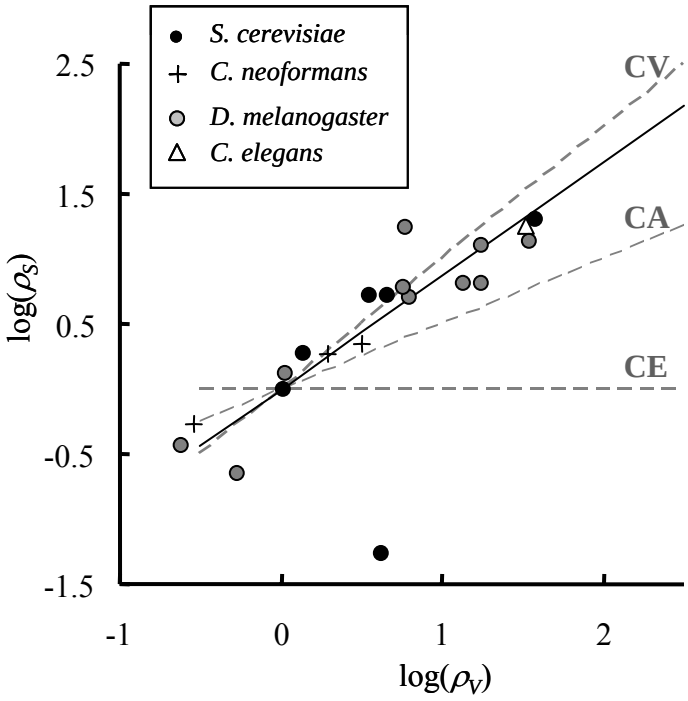

S. cerevisiae
C. neoformans
D. melanogaster
C. elegans
CV
CA
CE
2.5
1.5
0.5
-0.5
-1.5
-1
0
1
2
log(ρS)
log(ρV)

Fig. 2.

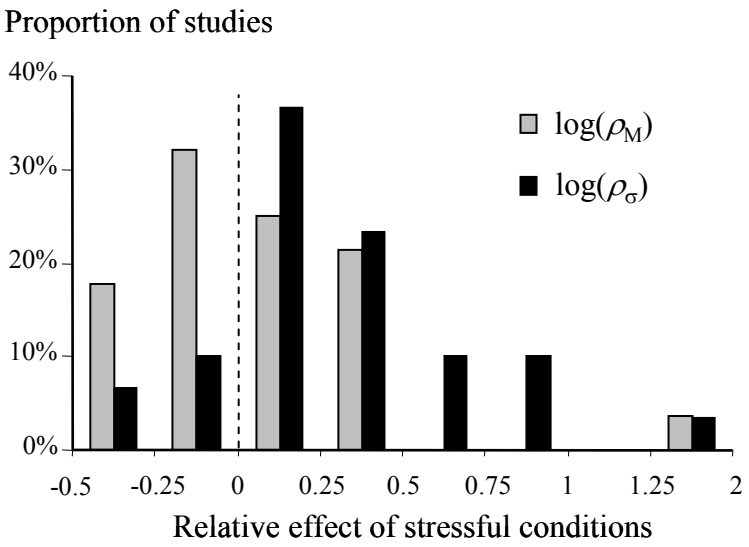

Proportion of studies
40%
30%
20%
10%
0%
log(ρM)
log(ρσ)
-0.5
-0.25
0
0.25
0.5
0.75
1
1.25
2
Relative effect of stressful conditions

Fig. 3

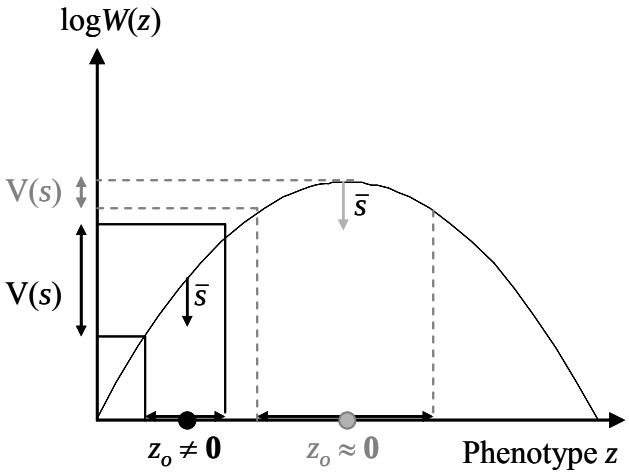


Fig. 4

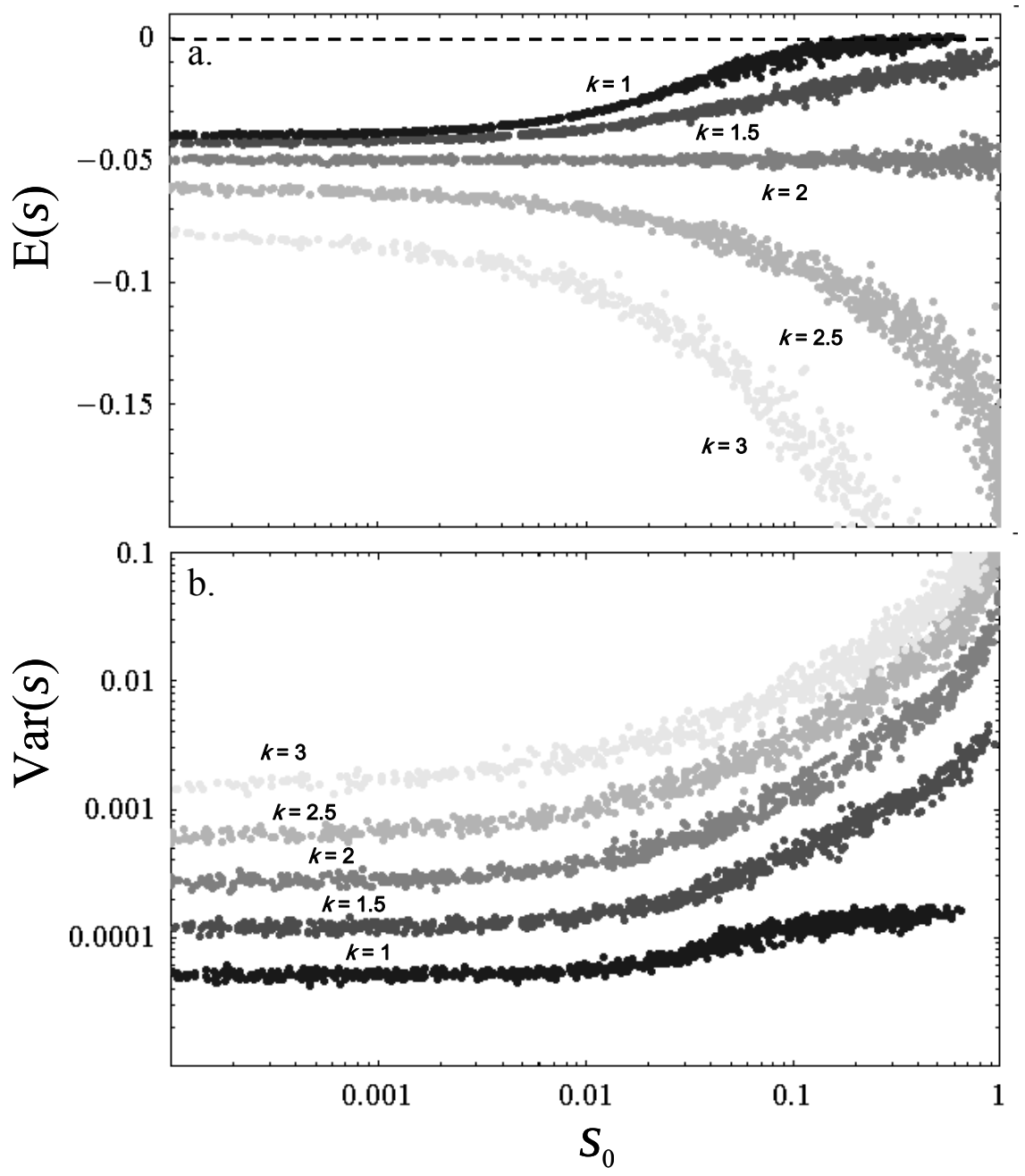

a.
b.
E(s)
Var(s)
S_0
k = 1
k = 1.5
k = 2
k = 2.5
k = 3
0
−0.05
−0.1
−0.15
0.1
0.01
0.001
0.0001
0.001
0.01
0.1
1

Fig. 5

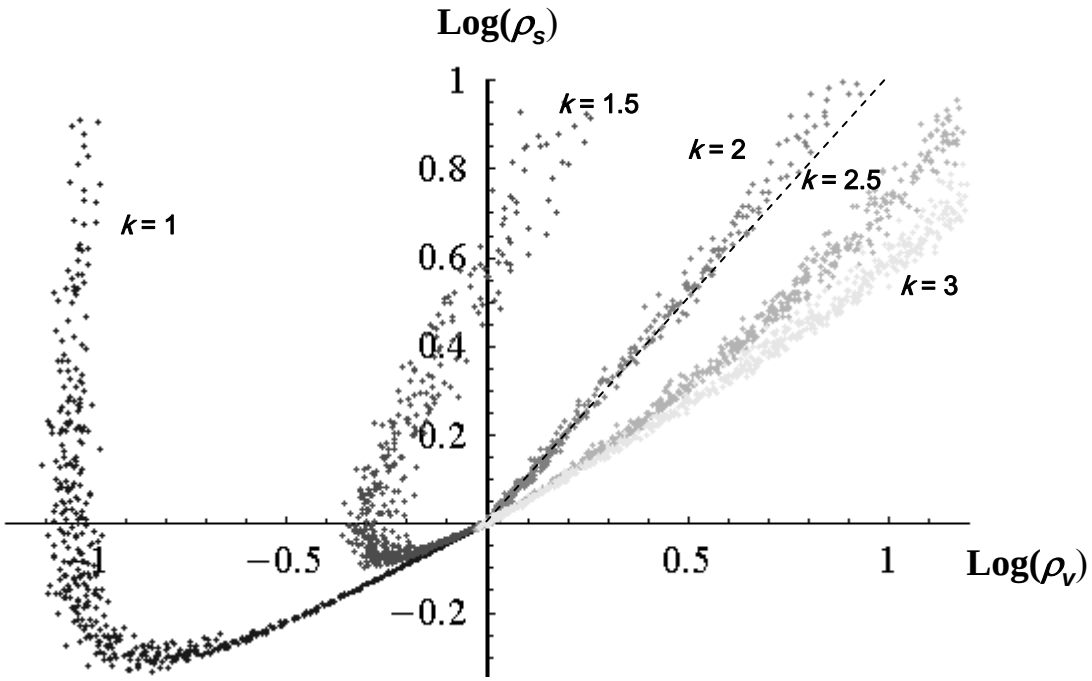


Fig. 6

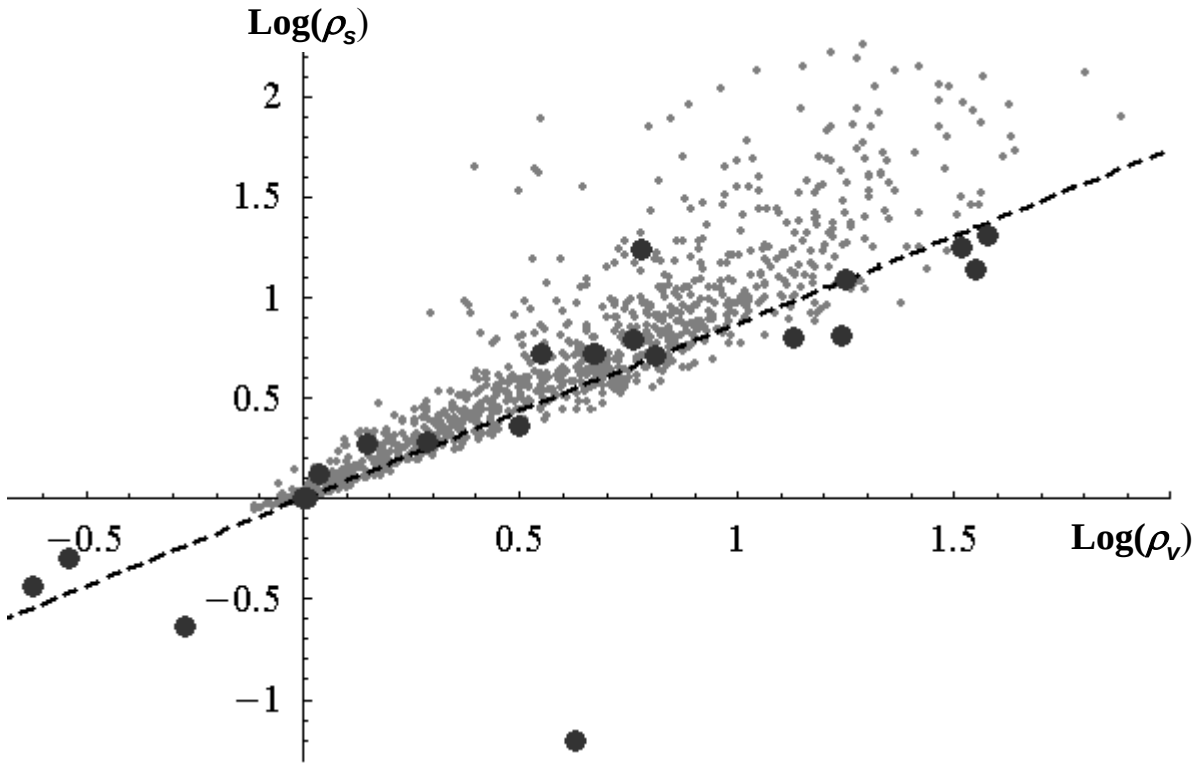
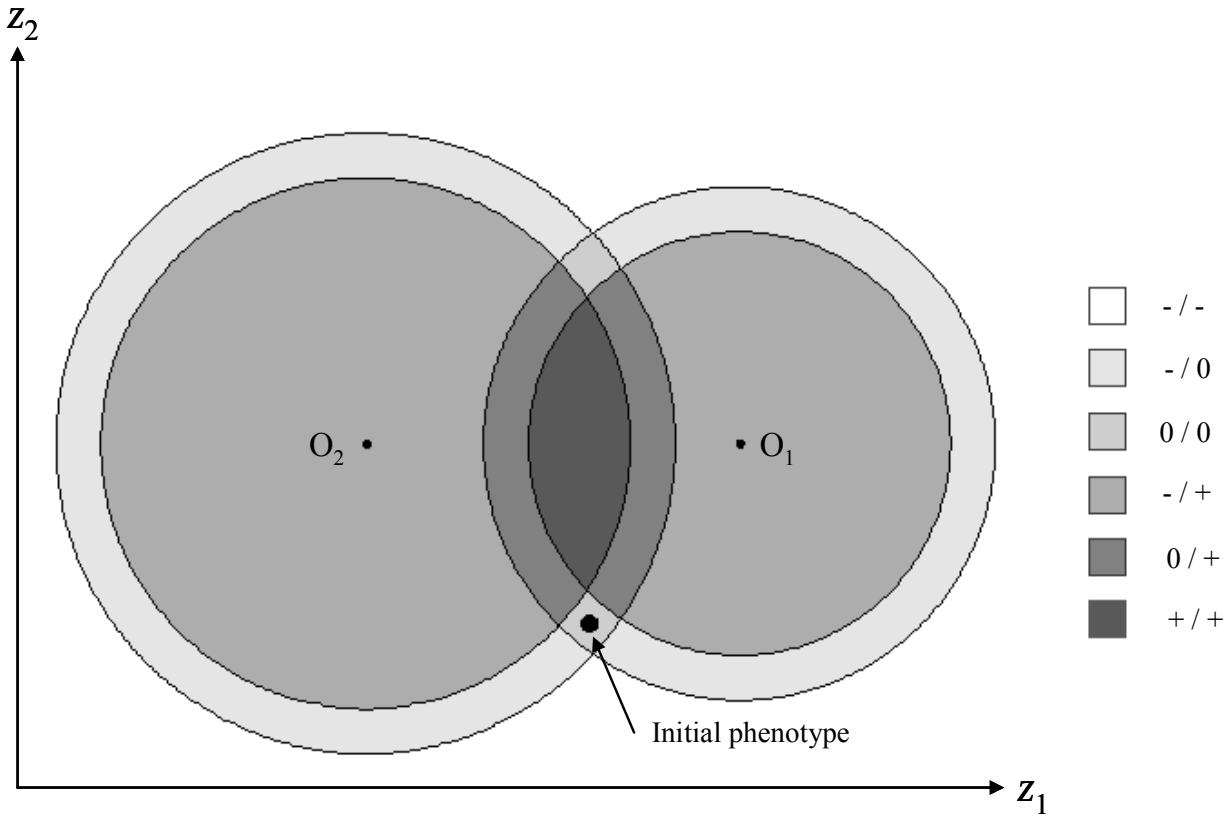


Fig. 7



Predicting fitness trajectories in model organisms

Guillaume Martin and Thomas Lenormand

Centre d'Ecologie Fonctionnelle et Evolutive (CEFE-CNRS UMR 5175)
1919 Rte de Mende, 34 293 Montpellier, France

LRH: Martin G. and Lenormand T.
RRH: Modelling adaptation

Corresponding author : Guillaume Martin

Abstract

The dynamics of adaptation to a new environment is inherently complex, even in the simplest situations such as encountered in experimental evolution. Indeed, the speed with which a population adapts (i.e. the speed of the mean fitness increase) depends on the rate, fitness effects, and fate of beneficial mutations which are hardly known empirically. Recently, several empirical studies have measured the long-term dynamics of adaptation in different model species (mostly micro-organisms), but theory does not, so far, provide quantitative predictions to which such data could be compared.

In this paper, we propose and test such a quantitative approach. Using a model presented in a previous paper, we predict the distribution of mutation fitness effects (including beneficial ones) in a given species, based on parameters of deleterious mutations (the most empirically available). We then parameterize the model, using data from mutation accumulation experiments, for *E. coli* and *D. melanogaster*. Using the prediction and previous results on fixation probabilities, we can predict the long term fitness trajectory of a given population adapting to a new environment. The prediction only requires knowing (i) the demographic regime (effective size, bottlenecks) and (ii) either the maximal fitness eventually reached by the population in this environment, or the initial rate of adaptation. Comparison of our predictions to three long-term empirical fitness trajectories (two studies on *E. coli* and one on *D. melanogaster*) shows that they are accurate. Furthermore, by including or neglecting different factors in the prediction (reproductive system, drift), their impact on rates of adaptation can be inferred. This approach provides a rationale for quantitative tests of adaptation theories using evolution experiments.

Introduction

Modelling the adaptive process requires to understand what favours and what limits the increase with time of mean fitness. Whether a model could quantitatively capture the fitness trajectory of a given population in a given environment is however questionable. First adaptation may be a very stochastic process: the occurrence, characteristics and fate of new mutations seems for instance difficult to predict. Second adaptation may depend on the particular history of a population, for instance on the amount of standing genetic variation that has accumulated. Third, the environment is always changing to some extent in either space or time, making it difficult to predict long term patterns of selection and adaptation.

Even if predicting rates of adaptation is challenging, evolutionary theory provides a framework to address this question. In particular, a first step would be to be able to predict rates of adaptation in the simplest cases: in fixed and controlled environments, with large population sizes and intermediate time scales such that the stochasticity of mutation events is smoothed, yet remaining within micro-evolutionary time scales. Such conditions are typically met in long-term adaptation experiments in model organisms.

The recent years have seen important developments in the study of adaptation, both from theoretical (reviewed in ORR 2005a; ORR 2005b) and empirical (with evolution experiments in micro-organisms, see review in ELENA and LENSKI 2003) standpoints. However so far, "it is unclear whether current theory accomplishes much more than qualitative agreement with the data" (ORR 2005a). So what is needed for a quantitative theory of adaptation? First it is necessary to know the rate of occurrence of new mutations and the distribution of their fitness effect, especially if they are beneficial. Second, it would be critical to know how these parameters vary across various environments and genotypes. Third, it is essential to be able to predict the fate of a new mutation in a given population. This last issue has received considerable attention so that the probability of fixation of advantageous mutations has been worked out for a variety of situations (BARTON and WHITLOCK 1997; CROW and KIMURA 1970; GERRISH and LENSKI 1998 ; ORR 2000; OTTO and WHITLOCK 1997; ROZE *et al.* 2005). However, the mutational parameters and their variation across genotypes and environments are much less known and understood: what is most available empirically is the rate and distribution of effects of deleterious mutations [mostly their average effect, for a review see \Lynch, 1999 #1079; Bataillon, 2000 #1114;]. Indeed, it is very difficult to observe a set of random beneficial mutations because they are often rare. Nonetheless, several studies have used theoretical predictions and empirical results to estimate the rate and effect of beneficial mutations (GERRISH and LENSKI 1998; IMHOF and SCHLOTTERER 2001; MIRALLES *et al.* 1999; ROZEN *et al.* 2002). Although these studies offer a fruitful approach for the study of beneficial mutations, associated with innovative empirical methods, they also present some limitations (GERRISH and LENSKI 1998). In these studies mutational parameters are inferred indirectly from the rate and duration of selective sweeps, i.e. by assuming *a priori* a model of adaptation (as a consequence, it is difficult to evaluate the validity of this particular model of adaptation with this approach). These methods often neglect the interference between deleterious and beneficial mutations (background selection ORR 2000) and assume that the distribution of beneficial effects is exponential. The validity of this last assumption is hard to directly assess empirically (ORR 2003), not mentioning the difficulty to evaluate its generality across species and environments. Typically, the rate of beneficial mutations is estimated to be 10^{-8} or less with these indirect approaches (reviewed in ZEYL 2004). In contrast, direct studies of fitness effects of mutations in maladapted genotypes have revealed up to 4% and 19% of beneficial mutations, in the RNA virus *VSV* (SANJUÀN *et al.* 2004) and *E. coli* (REMOLD and LENSKI 2001), respectively. This discrepancy tends to decrease the confidence that we can

have in indirect estimates of beneficial mutation parameters, or in the possibility to extrapolate from a few experiments.

Alternatively, an elegant approach to overcome our ignorance of the rate and effect of beneficial mutations is to make predictions that do not depend on them. ORR (2003), drawing on the work of GILLESPIE (1984), recently proposed such an approach based on extreme value theory (EVT). He showed that under a limited set of assumptions, the distribution of effects of beneficial mutations on *absolute* fitness is exponential and independent of the fitness rank (the level of adaptation) of the initial genotype in which mutations arise. However, although very few assumptions are needed in this approach, they do limit the range of application of the theory.

First, it assumes that the initial genotype is relatively well adapted (so that beneficial mutations are drawn into the right-tail of the fitness distribution). Simulations (ORR 2003) suggest that the initial genotype must be approximately among the 5% best adapted for the EVT to be accurate, but this threshold may depend on the kind of distribution of fitness assumed. Unfortunately, the cases in which we are most likely to detect adaptation empirically may be precisely those where the initial genotype was not initially so well adapted to its environment. As we have seen above, it is not clear what exactly is the proportion of advantageous mutations in the initial phases of strong adaptation observed in long term experiments on micro-organisms, for example.

Second, the model of adaptation based on the EVT assumes that the distribution of *absolute* fitnesses does not change with time (or with the initial genotype). This assumption may be valid, but it is hard to test. As outlined by Orr (ORR 2005a; ORR 2005b) the assumption does not affect short term predictions (e.g. that the distribution of beneficial effects is exponential at each step of adaptation), but it does affect longer term predictions (i.e. over several generations). Furthermore, it should be noted, as Orr already outlined (ORR 2002; ORR 2003), that most predictions are made on *absolute* fitness effects. Indeed, the distribution of *relative* fitness effects $f(s)$ does depend on the initial genotype's fitness (as the absolute fitness effect of each mutation is standardised by the wild-type fitness to obtain a selection coefficient). As several predictions on fixation probabilities rely on $f(s)$ not on the distribution of absolute fitnesses, knowledge of the wild-type fitness may be needed for predicting adaptation rates. Nevertheless, Orr (ORR 2002) showed that the increase in absolute fitness per adaptive substitution (ΔW in his notations) is only very weakly dependent on the initial genotype's fitness. However, this result has been derived assuming a large sexual population of constant size (such that the fixation probability of a mutation with effect s is $2s$). The derivation of ΔW directly from EVT may be quite difficult in other cases (e.g. asexual populations). Moreover, as it focuses on the right tail of fitness distributions, it cannot account for the interaction of deleterious and beneficial mutations (background selection), that may affect fixation probabilities in asexuals (ORR 2000).

Finally, the predicted fitness increase ΔW (i) does depend on the distribution of beneficial mutation effects on fitness (ORR 2002), which is barely known and (ii) is per adaptive substitution not per generation. To compute the per generation rate of adaptation, one also needs to know the proportion of beneficial mutations not only their effect. Overall, the EVT approach makes more robust predictions on a given step in adaptation than on the long term dynamics of mean fitness (ORR 2005b). The predictions regarding this "first step in adaptive evolution" (BULL and OTTO 2005) have recently been tested empirically (ROKYTA *et al.* 2005): the observations showed a relatively good agreement with the predictions, although corrections had to be made to obtain a good fit to the data.

Overall, although the EVT has the nice property of relying on few assumptions, this approach did not allow, so far, to make quantitative predictions on rates of adaptation in a given species or environment (ORR 2005a; ORR 2005b).

To provide a predictive theory of adaptation, we chose another approach: intending to predict the distribution of beneficial mutations in a new environment (hardly available empirically), based on that of deleterious mutations measured in standard conditions (the most empirically available). To do so, we proposed in a previous paper (MARTIN and LENORMAND 2006b) a model based on continuous phenotypic landscapes as introduced by (FISHER 1930) and later developed by ORR (ORR 1998). To make the model as realistic as possible, the symmetry assumptions of the Fisher model were removed to allow for arbitrary covariation between traits for both selection and mutation effects. The parameters of the model can be estimated from deleterious mutation effect distributions. Then, it allows predicting the change in the distribution of the fitness effects of mutations (including beneficial ones) as a function of the adaptation of the genotype in which they arise. In the present paper, we use this model and previous theoretical results on fixation probabilities, to predict empirical fitness trajectories (i) in a fixed controlled environment and (ii) in model organisms for which deleterious mutation parameters are known (first section). We then compare our predictions with observed trajectories in *E. coli* and *Drosophila* (second section).

Model

We first summarize a previous model (MARTIN and LENORMAND 2006b) for the distribution $f(s)$ of the fitness effect s of mutations arising in a given genotype (hereafter denoted "ancestor"). This model allows predicting the distribution of s in a new environment, based on a measure of this distribution in a standard environment, and on a measure of the fitness of the ancestor relative to a genotype well adapted to the new environment. We then summarize previous theoretical results on fixation probabilities of beneficial mutations according to the demographic regime and reproductive mode. Finally, we use these results and the predicted $f(s)$ to compute the expected rate of adaptation in a sexual or asexual species with known parameters of demographic regime and deleterious mutation rate and effect. We show that the resulting fitness trajectories can be approximated by a simple function of generation time which parameters can be biologically interpreted and compared across species and environments.

Distribution of mutation fitness effects for a given distance to the optimum: To model $f(s)$ we used a fitness surface approach similar to Fisher's model but extended to account for arbitrary covariation between traits for mutation and selection. To do so, we considered a general quantitative genetic framework similar to that in (ZHANG and HILL 2003) but extended to account for beneficial mutations (i.e. phenotypes are not at their optimum). Phenotypes are modelled as a set of n continuous phenotypic traits represented by a column vector $\mathbf{z}$. The fitness $W(\mathbf{z})$ of phenotype $\mathbf{z}$ is a multivariate Gaussian function of the distance between $\mathbf{z}$ and a phenotypic optimum that is set to 0 for all traits, without loss of generality: $W(\mathbf{z}) = \mathrm{Exp}(-\frac{1}{2}\,\mathbf{z}^T\,\mathbf{S}\,\mathbf{z})$, where ^{T} denotes transposition. $\mathbf{S}$ is the $n \times n$ matrix of selective effects on each trait (diagonal terms) and interactions between traits (off-diagonal terms). We consider the effect of mutation on a single genotype, referred to as "ancestor", with phenotype $\mathbf{z}_o$ and we define $s_o = -\log(W(\mathbf{z}_o)/W(\mathbf{0}))$. s_o measures the (log)fitness of the ancestor relative to the best adapted genotype (with phenotype $\mathbf{z} = \mathbf{0}$ by convention), and will be hereafter referred to as "fitness distance to the optimum". We assume that the mutant phenotypic distribution around the ancestor phenotype is multivariate Gaussian with mean $\mathbf{0}$ and arbitrary covariance matrix $\mathbf{M}$ (or at least that each trait can be transformed to get a gaussian distribution of mutant values). As for selection, the model allows for both arbitrary mutational variances on each trait and mutational correlations between traits. Define λ_i the eigenvalues of the product $\mathbf{S}\,\mathbf{M}$ and $\mathrm{CV}(\lambda)$ their coefficient of variation across traits i. Then, when the

ancestor is well adapted to the environment (s_o = 0, only deleterious mutations), the distribution of mutation fitness effects is approximately gamma with shape β_o = ½ n /(1+CV(λ)^2) and scale $\alpha_o = \bar{s} / \beta_o$ where $\bar{s}$ is minus the average fitness effect of mutations (i.e. positive value). These two parameters can be measured from an empirical distribution of deleterious mutation effects. The validity of a gamma approximation seems to be supported by the empirical data on distribution of deleterious mutation fitness effects in standard laboratory (benign) environments (MARTIN and LENORMAND 2006b).

In the general case where the ancestor is not perfectly adapted to the environment ($s_o \geq 0$, i.e. with beneficial mutations), the average effect of mutations $\bar{s}$ remains unchanged, while the variance of s increases (see Appendix 1). This corresponds to an increase in the coefficient of variation of s with increasing maladaptation of the ancestor. This pattern is supported by a review of empirical measures of the impact of stress on mutation fitness effects (MARTIN and LENORMAND 2006a). In these conditions, the distribution of s is approximately a displaced gamma (the sum of a constant plus a gamma) with shape β and scale α and with probability density function

$$f(s) = \frac{e^{-\frac{s_o - s}{\alpha}} (s_o - s)^{\beta-1} \alpha^{-\beta}}{\Gamma(\beta)}, \quad \textbf{(1)}$$

Where $\Gamma(.)$ denotes the gamma function, and where the shape β and scale α are given by

$$\beta = \beta_o \frac{(1 + s_o / \bar{s})^2}{(1 + 2\, q\, s_o / \bar{s})} \text{ and } \alpha = \frac{\bar{s}}{\beta_o} \frac{(1 + 2\, q\, s_o / \bar{s})}{(1 + s_o / \bar{s})}, \quad \textbf{(2)}$$

(see details in Appendix 1). In Eq. (2), q depends on the environment considered. It reflects the fact that even if a given ancestor is at the same fitness distance to the optimum (same s_o) in two different environments, the variance of s (and hence $f(s)$) may vary according to the phenotypic direction to the optimum in each environment (see Appendix 1). All the parameters in Eq. (2) are measurable (and have indeed been measured empirically), except q. Fortunately, assuming that a particular new environment corresponds to a displacement of the optimum in a random phenotypic direction, it can be shown (see Appendix 1) that q = 1 on average, and that its variance across different environments, σ_q^2, can be expected to be small (it is proportional to 1/n the inverse of the number of traits). σ_q^2 depends on the distribution of the eigenvalues λ_i and its order of magnitude of σ_q^2 to can be either (roughly) derived from empirical estimates of β_o or determined under a null model where phenotypic covariances are drawn randomly (see Appendix 1). We found that σ_q should be of the order of $\sigma_q \sim 0.3 - 0.5$, so we considered that q could vary within the 95% confidence interval of a gaussian $N(1,0.3)$ which is $q \in [0.5,1.5]$.

From the probability density of fitness effects given in Eq. (1), the rate of adaptation can be derived. The expected $f(s)$ and corresponding rate of adaptation is obtained by setting q = 1 in Eq. (2), while setting either q = 0.5 or q = 1.5 provides an equivalent of confidence envelope for $f(s)$ and the corresponding rate of adaptation.

Fixation probabilities of beneficial mutations: A fundamental parameter to predict the rate of adaptation is the probability that a given beneficial mutation (assumed to arise in single copy at a given generation) finally reaches fixation. These fixations are the individual steps in

the process of adaptation. The theoretical population genetics literature provides models for the effect of different factors on fixation probabilities, that can thus be taken into account quantitatively.

Probability of escaping drift loss

A beneficial mutation may be initially lost while still at low frequency due to random drift alone. This affects the probability of ultimate fixation and consequently the rate of adaptation. In a sexual diploid, the probability of escaping stochastic loss for a mutation arising in single copy in a population of N diploid individuals with effective size N_e, and with a dominance coefficient h and an homozygous beneficial effect s can be derived from Eq. 8.8.3.21 of (CROW and KIMURA 1970). However, there is no agreement on the value of the dominance coefficient in Drosophila or other species (GARCIA-DORADO *et al.* 1999), particularly among beneficial mutations. We therefore assumed codominance ($h = 1/2$) which gives the probability of escaping drift loss

$$p_{esc}(s) = \frac{1 - Exp(-s\,N_e / N)}{1 - Exp(-2s\,N_e)}. \qquad (3)$$

In a sexual diploid, this probability is assumed to equal the probability of ultimate fixation. Note that this ignores the possible effect of linkage between beneficial mutations sweeping to fixation (Hill-Robertson effect). It thus assumes that the probability that two beneficial mutations segregate simultaneously at linked loci is very small (because linkage decreases with the distance between gene sequences in sexuals). This effect cannot be ignored in asexuals (or strongly selfing species), but it fortunately can be quantified in this case (see below).

In experiments with microbes (e.g. *E.coli*) the total population size is often very large ($N = N_e > 10^6$), so that stochastic loss is unlikely. However, some experimental settings involve strong population bottlenecks by regular dilutions of the medium to maintain exponential growth of the cells. These bottlenecks reduce the fixation probability of beneficial mutations by a factor $D\,\mathrm{Log}(D)^2$ (WAHL *et al.* 2002). Moreover, as reproduction takes place by binary fission in micro-organisms, the probability of escaping drift in a large population is not $2s$ but approximately $p_{esc}(s) = k_f\,s$ where $k_f \approx 2.8$ for a mutation with (haploid) beneficial effect s if mutations appear at random during the cell's life cycle (JOHNSON and GERRISH 2002). Note that in the case of diploid microbes reproducing by binary fission (e.g. diploid yeast) one should set $k_f = 2.8/2 = 1.4$. In the case of studies with micro-organisms (*E. coli*), we did not correct for N_e because of the very large population sizes used, and to obtain analytic expressions for fixation probabilities. Therefore, for these studies, the probability of escaping drift for a mutation with beneficial effect s was set to

$$p_{esc}(s) = D\log(D)^2\,k_f\,s \qquad (4)$$

where $D\,\log(D)^2$ was set to 1 in the absence of dilution bottlenecks. We now turn to factors affecting fixation probability in asexuals only.

Deleterious mutations and background selection

In asexuals, a beneficial mutation may have reduced fixation probability if it appears on a genetic background that is loaded with deleterious mutations, because it cannot "recombine away" from this deleterious background. This effect (background selection) has

been modelled as a reduction in the rate of adaptation by e^{-U/s_H} (ORR 2000; PECK 1994) where s_H is the harmonic mean of deleterious mutations. This quantity may be computed numerically using $f(s)$ given in Eq. (1), but the integral does not converge at the zero bound for s. However, it can reasonably be assumed that nearly neutral deleterious mutations ($N_e\, s < 1$) have no background selection effect. Therefore, the harmonic mean of those deleterious mutations that reduce adaptation rates can be computed numerically using Eq. (1) as

$$s_H = \int_{-\infty}^{-1/Ne} \frac{1}{s} f(s)\, ds\,. \qquad (5)$$

Note that the reduction by e^{-U/s_H} assumes that (i) the frequency of loaded genotypes is at mutation-drift equilibrium (large population) and (ii) beneficial mutations have typically smaller effects than deleterious ones. The latter is not true in our model when away from the optimum (so $\geq \bar{s}$). Therefore, accounting for background selection in this manner may lead to an overestimation of the effect of background selection and subsequent underestimation of adaptation rates in asexuals. A more realistic model has been proposed by (JOHNSON and BARTON 2002) but it cannot be directly implemented with only $f(s)$. In any case, as will be shown in the next section, the predicted effect of background selection in the studies on *E. coli* that we considered is small relative to clonal interference, because of the low mutaiotn rate in this species. Therefore, this potential error is not too problematic.

Beneficial mutations and clonal interference

In an asexual, a given beneficial mutation cannot reach fixation if another mutation with a larger beneficial effect appears in the population. Indeed, as recombination does not occur, beneficial mutations that segregate simultaneously compete for fixation, until ultimately the best mutation removes the other. This effect was modelled by (GERRISH and LENSKI 1998) and termed clonal interference. It reduces the rate of adaptation by an amount $e^{-I(s)}$ where $I(s)$ is the number of beneficial mutations that interfere with a given mutation with selective advantage s. Taking into account the effect of bottlenecks due to the dilution of the medium for microbes (see above), and the possible effect of the mode of reproduction (k_f) the number of interfering mutations is given by (See appendix 2)

$$I(s) = \frac{N \,\mathrm{Log}(N)\, U\, D \,\mathrm{Log}(D)^2 k_f}{s} \times \left(s_o - \alpha\beta + \frac{\alpha\, \Gamma\left(\beta+1, (s_o - s)/\alpha\right) - s_o \Gamma\left(\beta, (s_o - s)/\alpha\right)}{\Gamma(\beta)} \right), \qquad (6)$$

Where β and α are given in Eq. (2).

Rate of adaptation and fitness trajectories

rate of adaptation at a given fitness distance to the optimum

Given the fixation probability of beneficial mutations the rate of adaptation defined as the increase in mean fitness over the generations is computed as the expected fitness effect of

fixed mutations times the number of mutations appearing per generation $N\,U$ (GERRISH and LENSKI 1998; ORR 2000):

$$\frac{d\log(W)}{dt} = N\,U\int_{0}^{s_o} p_{esc}(s)\,s\,f(s)\,e^{-U/s_H}e^{-I(s)}ds\,, \qquad (7)$$

Where $p_{esc}(s)$ is given by Eq. (3) (diploid sexuals) and Eq. (4) (microorganisms) while the exponential terms reflecting background selection (using Eq. (5)) and clonal interference (using Eq. (6)) are only included for asexuals (micro-organisms). Note that mean fitness also increases during the sweep of beneficial mutations (i.e. between successive fixations). Eq. (7) is therefore a continuous time approximation for the discrete process of successive adaptive substitutions. Given empirical values of the mutational parameters (U, β_o and $\bar{s}$) and the demographic parameters (N, N_e, k_f and D) for a given species and experimental design, the rate of adaptation of a population standing at a given fitness distance to the optimum (s_o) can be computed using Eq. (7). We did not find a closed form analytical expression for this integral in the general case, but it can be integrated numerically. However, in the case of a large sexual population, standing away from the optimum (i.e. in the beginning of the boot of adaptation), the rate of adaptation can be expressed as a simple linear function of the fitness distance to the optimum s_o (see appendix 2).

$$\frac{d\log(W)}{dt} \underset{s_o >> \bar{s}}{\approx} N_e\,U\,\frac{\bar{s}}{\beta_o}\,q\,s_o\,, \qquad (8)$$

Where we recall that $q = 1$ on average (Appendix 1). This approximation is only accurate when s_o is large enough relative to $\bar{s}$ but it gives the correct order of magnitude for the rate of adaptation in a wider range of parameters (assuming sexual reproduction), as illustrated on Fig. 1. However, for the fit of empirical trajectories (see next section) we used precise numerical integration of Eq. (7) rather than this approximation.

Fitness trajectory over time

If the fitness of the perfectly adapted genotype is known for a given environment (s_o is known), the expected fitness trajectory of an initially maladapted ancestor in this environment can be predicted by iterating Eq. (7) over the generations. This assumes that all individuals in the population are at the same fitness distance to the optimum $s_o(t)$ at any time t, so that the distribution of fitness effects is the same for mutations arising in any individual. Then, let $\log(W_o)$ and $\log(W_f)$ be the initial and final log mean fitness (when the fitness plateau is reached), and let $\log(W(t))$ be the log mean fitness after t generations. At a given time t, the fitness distance to the optimum is $s_o(t) = \log(W_f / W(t))$, the corresponding rate of adaptation $d\log(W(t))/dt$, can be computed using Eq. (7) with $s_o = s_o(t)$, and the value of $\log(W(t+1))$ at the next generation can be inferred inferred.

The assumption that $s_o(t)$ is approximately the same for all individuals does not seem unrealistic in an asexual population, where each adaptive substitution fixes a single clone in the population. Such genetic homogeneity of evolving experimental populations has been confirmed in *E.coli,* at least until the population gets close to the fitness plateau (DE CRECY-LAGARD *et al.* 2001; PAPADOPOULOS *et al.* 1999; WAHL and KRAKAUER 2000). In a sexual species, however, it can be expected that, even when starting from an initially isogenic line, some genetic variance accumulates in the population over the generations, because a beneficial mutation that fixes may be associated with various genetic backgrounds. However,

if selection is strong enough, it can be assumed that all genotypes are approximately within the same fitness distance to the optimum (i.e. there is little standing genetic variance in fitness). Therefore, in a fitness landscape view, the polymorphism within the population would reflect variation in the direction, rather than the distance, to the optimum. This would correspond to different values of q (as defined in Eq. (2) and Appendix 1) with the same $s_o(t)$. As explained in Appendix 1, we expect that variation in the direction to the optimum translates into little variation in q and that $q = 1$ on average. Therefore we may rely on the approximation of a single distribution of mutation effects (although arising in a genetically variable population in sexuals) to predict average fitness trajectories when selection is initially strong. Note that from this argument (i) we expect more variation among replicate populations in sexuals, and (ii) our assumption may be less realistic in the last phase of adaptation (close to the fitness plateau), as more variance has accumulated and selection is weaker.

The method presented above allows to predict the fitness trajectory over a boot of adaptation by iterating numerical integrations of Eq. (7). However, it does not provide a closed form for $\log(W(t))$ as a function of the number of generations t. Nevertheless, in the case of a large sexual populations, the rate of adaptation is approximately linear with $s_o(t) = \log(W_f/W(t))$ (see Eq. (8)). Similarly, numerical integration of Eq. (7) for various values of s_o showed that this linearity with s_o also holds for asexuals, when not too close to the optimum (e.g. $s_o > 0.05$). Unfortunately, in this case, we did not find any analytic expression for the linear coefficient. However, the fit of the linear regression of $d\log(W)/dt$ on s_o is very good ($R^2 \geq 95\%$) for both sexuals and asexuals and for various values of mutational and demographic parameters (including those given in Table 1 for *E. coli* and Drosophila, see next section). Therefore, we can use this linearity with $s_o(t)$ at any time t to derive a closed form expression for $\log(W(t))$ as a function of t:

$$\begin{cases} \dfrac{d\log\left(W(t)\right)}{dt} \approx a\, s_o(t) = a\log\left(\dfrac{W_f}{W(t)}\right) & \text{A} \\ \log\left(W(t)\right) \approx \log(W_o) + s_o\left(1 - e^{-at}\right) & \text{B} \end{cases} . \qquad (9)$$

The parameter s_o (defined below) gives the log-value of the final fitness W_f relative to the initial W_o, and depends only on the environment to which the population adapts. The parameter a summarizes the effects of the mutational and demographic parameters on the rate of adaptation. For a population of sexual diploids with large effective size N_e, $a \approx N_e\, U\bar{s}\,/\beta_o$ (Eq. (8)), while for asexuals, it is reduced by clonal interference and background selection. The parameter a is a measure of the response to selection imposed by the environment at any time t for a given species and breeding conditions. Consequently, for a given species and empirical design (constant mutational and demographic parameters), a should remain approximately constant for fitness trajectories in different environments, with only variation in s_o.

The function defined in Eq. (9.B) can be easily fitted to empirical trajectories to estimate both parameters. It has a qualitatively (not quantitatively) similar behaviour as the hyperbolic (LENSKI and TRAVISANO 1994) and log-hyperbolic (ELENA *et al.* 1998a) models proposed in previous empirical studies.

Comparing predicted and empirical fitness trajectories

To predict the rate of adaptation, the distance to the optimum (s_o), determined by the level of adaptation of the ancestor to the new environment, has to be known empirically. This can be achieved in two ways. One possibility is to measure the fitness of another genotype

that has long adapted to the new environment considered. s_o can then be estimated as the ratio of the fitness of the well-adapted genotype to that of the focus (ancestor) genotype before it starts to adapt to the new environment. Alternatively, s_o can be measured by following the long term dynamic of a single genotype (the ancestor) until it reaches a fitness plateau. The ratio of this final mean fitness to the initial mean fitness of the ancestor gives an estimate of s_o. We used this second approach to predict initial adaptation rates from known values of final fitness. We first checked the fit of a function of the type defined in Eq. (9.B) on empirical long term fitness trajectories in *E. coli* and Drosophila. Given the value of final log-fitness (s_o from the fit), we can predict the initial rate of adaptation r_o from Eq. (7) and from known values of mutational and demographic parameters in these species and designs. This gives a predicted value of $a = r_o/s_o$ (from Eq (9.A), with $t = 0$, $s_o(t = 0) = s_o$) that can be compared to the fitted value as a test of the model's prediction.

Long-term fitness trajectories in model organisms

Several experiments have sought to measure the change in fitness in a population adapting to a new controlled environment, most studies have used micro-organisms in which rates of adaptation can be measured over thousands of generations (reviewed in ELENA and LENSKI 2003). Some studies have reported fitness trajectories over long enough periods of time that a fitness plateau is reached, so that the value of s_o can be precisely estimated. Note that the model could also be used to predict long term trajectories based on short term dynamics but our aim was to try to validate, rather than use, the model.

GILLIGAN and FRANKHAM (2003) measured the rate of adaptation to captivity of *Drosophila melanogaster* in 20 replicate outbred populations of small size ($N = 1000$, $N_e = 300$) derived from a wild population, over 87 generations. The authors provide the average time at which fitness had reached 25%, 50%, 75% and 95% of its maximum and an estimate of this expected maximum, based on the best fitting exponential model for the fitness trajectory. These values of the average fitness trajectory were used to fit the model in Eq. (9.B). This empirical setting is not the best suited to test our model as adaptation occurs from an initially polymorphic population, hence with potentially substantial standing variation. However, we considered that adaptation to captivity may have mainly relied on newly arisen mutations because (i) the population effective size is small, so that the standing variance in fitness should be reduced, and (ii) the conditions implied by captivity may not be encountered in the wild, so that few beneficial mutations may be segregating in the original population. If this assumption is false, the predicted rate of adaptation should underestimate the observed rate. We did not find any other study reporting long term fitness trajectories in higher organisms with the value of a final fitness plateau explicitly given.

LENSKI and TRAVISANO (1994) provided the first study of long term adaptation in micro-organisms in their famous 10,000 generations experiment with *E. coli*. The value of mean fitness every 500 generation in 12 replicate populations grown in a glucose-limiting medium are available on R. Lenski's website http://www.msu.edu/user/lenski. We used the fitness trajectory on a single population (Ara-1) for the first 2000 generations (available on the same website) to get a higher resolution for this period (as fitness values are given every 100 generations). In this empirical design, regular population bottlenecks (dilution ratio $D = 1:100$) are used to maintain exponential growth

More recently, DE CRECY-LAGARD *et al.*(2001) proposed another study of adaptation of a single population of *E. coli* during 10,000 generations in a thymine-limiting medium. Their empirical design involves a tandem chemostat that allows continuous growth of cells at a constant population size. The trajectory of mean fitness over time was directly read on the Figure 3 of the paper. These two long-term adaptation experiments provide a unique opportunity to compare the dynamics of adaptation in different demographic regimes

(different population sizes, presence or absence of bottlenecks) and in contrasted environments (thymine shortage is much more stressful to *E. coli*, than the glucose-limiting medium).

Other studies (ELENA *et al.* 1998a; ELENA *et al.* 1996) have also measured the dynamics of fitness recovery in debilitated strains of the RNA virus *VSV* previously subjected to strong bottlenecks (reducing its fitness relative to the original strain). In one of these studies, a fitness plateau equal to the original strain's fitness was reached, suggesting that this strain was well adapted to its environment. Such experiment is also well suited for testing the present model. Unfortunately, the estimates of the mean effect of deleterious mutations in *VSV* obtained by two different approaches are very different: $\bar{s} \approx 0.0023$ (ELENA and MOYA 1999) based on Bateman-Mukai estimation and $\bar{s} \approx 0.19$ (SANJUÀN *et al.* 2004) based on the direct effect of single substitutions in a genetically modified *VSV* clone. Considering the high sensitivity of our predictions on empirical estimates of $\bar{s}$ we did not attempt to fit *VSV* fitness trajectories. We believe however, that based on an agreed estimate for $\bar{s}$, comparison of these empirical trajectories with our predictions and with empirical trajectories in *E.coli* would provide a unique opportunity to assess the effect of the very high mutation rate characterizing RNA viruses ($U \sim 1$) on rates of adaptation.

The three empirical fitness trajectories described above were fitted to the exponential model in Eq. (9.B). Least squares estimates of s_o and a for each study are given in Table 1. The model gave a good fit to the data in all three studies. The fit of the model in the *Drosophila* study is very good (R^2 = 99 %) but this only reflects that our model has the same behaviour as the exponential model used to fit the data in (GILLIGAN and FRANKHAM 2003). We did not find any estimate of the R^2 for this study but there clearly is a large variance among replicate trajectories. For the two *E. coli* studies, we give as a comparison the fit of our model and of the log-hyperbolic model (into brackets) used in previous studies (ELENA *et al.* 1998a): *E. coli* in glucose-limiting medium R^2 = 93.7 % (94.6 %), *E. coli* in a thymine-limiting medium R^2 = 89.2 % (82.3 %). Overall, it seems that the fitness dynamics predicted by Eq. (9.B) is as consistent with empirical trajectories as the descriptive models previously used to fit these data.

Estimates of mutational parameters for *D. melanogaster* and *E. coli*

Based on the estimates of final fitness (of s_o), we can predict the initial rate of adaptation in each of these studies using Eq. (7). To do so we use the values of demographic parameters reported in each study: N, N_e, and D (dilution of the medium). We also need to know the value of mutational parameters in each species. The relevant estimates are provided by studies of deleterious mutations: the average deleterious effect $\bar{s}$, the mutation rate U, and the shape of the distribution β_o. We used estimates of these parameters reported in two reviews (GARCIA-DORADO *et al.* 1999; MARTIN and LENORMAND 2006b).

In both species, the estimates of $\bar{s}$ were obtained from Batman-Mukai estimates s_{BM}, reported in Table 1 of (MARTIN and LENORMAND 2006b), and corrected to account for the variation of s, as $\bar{s} = s_{BM} / (1+CV(s)^2)$, where $CV(s)$ is the coefficient of variation of s, reported in Table 2 of (MARTIN and LENORMAND 2006b). For Drosophila, the estimate of $E(s)$ was taken as the average of estimates reviewed in Table 1 of (MARTIN and LENORMAND 2006b). This gives $s_{BM} \approx 0.11$ based on Bateman-Mukai estimation and $E(s) \approx 0.06$ after correcting for variation in s (with CV(s) = 0.92). In *E.coli*, $s_{BM} \approx 0.012$ (KIBOTA and LYNCH 1996) which after correction with CV(s) ≈ 3.09 (from (ELENA *et al.* 1998b)) gives an actual $E(s)$ one order of magnitude smaller $E(s) \approx 0.0011$. The possibility of a strong upward bias in Bateman-Mukai estimates of $E(s)$ in *E.coli* has already been pointed out by (KIBOTA and LYNCH 1996). We did not use the direct $E(s)$ estimate in (ELENA *et al.* 1998b), because the antibiotic marker associated with the transposons used in this experiment has been shown to

have a direct deleterious fitness effect (REMOLD and LENSKI 2001), thus potentially downwardly biasing the estimate of E(s). This might explain why this direct estimate is close to the Bateman-Mukai estimate (which assumes constant s), whereas a strong variance of s was observed in the same study.

Similarly, the estimate of U in both species was taken as the Bateman-Mukai estimate U_{BM} corrected for CV(s). In drosophila, the reported estimates in the review by (GARCIA-DORADO *et al.* 1999) are of the order of $U \approx 0.015$ for viability (without correction as it is based on maximum-distance estimates that account for variation in s), while (KIBOTA and LYNCH 1996) report $U_{BM} \approx 0.0002$ for *E.coli*. After correcting for CV(s) in *E.coli* we get $U \approx 0.0021$ This value is in agreement with the estimate $U = 0.0025$ obtained independently by Drake (DRAKE *et al.* 1998) from mutation rates per gene, which confirms our choice of using the Bateman-Mukai estimate from spontaneous mutations and to correct it by the CV(s) estimate from direct measures by ELENA *et al.* (ELENA *et al.* 1998b). Note that in Drosophila, the estimate of U based on per gene rates of mutation ($U \approx 0.05$) is less reliable than in micro-organisms because it was estimated on a limited number of specific sites (DRAKE *et al.* 1998), so we kept the estimate from MA data reported by Garcia-Dorado (GARCIA-DORADO *et al.* 1999).

The shape parameter β_o of the gamma approximation for $f(s)$ was taken from the fitted gamma distribution shape in (ELENA *et al.* 1998b) for *E.coli*. In Drosophila, the shape estimates provided by statistical analyses of mutation accumulation data tend to have large confidence intervals (KEIGHTLEY 2004), so we estimated β_o based on the moments of $f(s)$ from single transposon insertions (LYMAN *et al.* 1996) reported in (MARTIN and LENORMAND 2006b). The resulting estimate is $\beta_o = 1.2$ or 1.6 when using the second or third moment (respectively), so we took an average $\beta_o = 1.4$.

Note that, for Drosophila, the estimates used here correspond of mutation effects on viability not competitive fitness (the parameter measured in (GILLIGAN and FRANKHAM 2003)). Indeed, the only available estimates of the shape of the distribution of s (i.e. variance and skewness) come from transposon effects on viability (LYMAN *et al.* 1996). It may be assumed however, that a large part of the factor determining fitness is viability. Furthermore, Avila (AVILA and GARCIA-DORADO 2002) recently measured mutational parameters for competitive fitness that were close to those used here ($U_{BM} = 0.03$, $s_{BM} = 0.1$).

Overall, the quality of our predictions may be altered by errors in the assumed values of mutational parameters. The choices made for these estimates are clearly not based on the firmest grounds, particularly as corrections are based on transposon insert effects, which have a direct fitness effect (see above). However, (i) they should give a correct order of magnitude for U and $\bar{s}$, and (ii) there are theoretical reasons to expect little variation in β_o across species (MARTIN and LENORMAND 2006b), which is confirmed by a review of the (precise, albeit scarce) data (MARTIN and LENORMAND 2006b). Furthermore, the fact that several empirical measures are consistent in Drosophila, and that fitness measures in *E coli* can be very precise allows some confidence, at least in the order of magnitudes of these parameters. A summary of the various demographic and mutational parameters used to compute the initial rates of adaptation from Eq. (7) is given in Table 1 (input parameters).

It should also be noted that our approach assumes that the estimates of these mutational parameters obtained in standard laboratory conditions are valid in another environment (that of the adaptation experiment). This assumption is based on several arguments. First, theoretical arguments (MARTIN and LENORMAND 2006b) suggest that β_o may not vary much across environments or species. This was confirmed by a review of empirical estimates showing that it does not vary much across very different species. Second, although there are less reasons to expect that the number of expressed mutations and their

average effect $\bar{s}$ be constant across environments, a review of G×E interactions among deleterious mutations suggests that they are indeed rather constant (MARTIN and LENORMAND 2006a). Finally, we make the assumption that U is also constant across environments. Note that by U we mean the genome-wide *molecular* rate of mutation, not the number of expressed mutations alluded to above. This last assumption may not be valid for micro-organisms under strongly stressful conditions. Indeed, in these very stressful conditions, there is evidence for increased rates of mutations (BJEDOV *et al.* 2003; LOEWE *et al.* 2003), that may be due to e.g. decreased fidelity of DNA polymerases, decreased DNA repair activity or activation of transposons (KIVISAAR 2003). However, the generality of such a phenomenon (known as "stationary phase mutagenesis") is still in debate (SNIEGOWSKI 2004), and some authors suggest that most stress responses only involve increased mutation rates at specific sites not over the whole genome (WRIGHT 2004). In one of the studies on *E. coli* that we consider (LENSKI and TRAVISANO 1994), the populations are clearly not in stationary phase, but this possibility cannot be ruled out in the other study which maintains a constant population size (DE CRECY-LAGARD *et al.* 2001). Overall, in the absence of further information, we relied on extrapolations of U estimates to new environments. This may not be appropriate when considering situations in which adaptive mutators (that increase the mutation rate under stress) may have evolved (TADDEI *et al.* 1997), i.e. (i) in stressful environments often encountered in the wild and (ii) in asexual populations (where the mutator may hitch hike on the beneficial mutations it produces). Fortunately, as we will see in discussion, the predictions for the rate of adaptation in asexuals are rather insensitive to the mutation rate, at least when NU is large, because of clonal interference. This has already been put forward by theoretical approaches (GERRISH and LENSKI 1998) and confirmed empirically (DE VISSER *et al.* 1999).

Predicted fitness trajectories and factors affecting rates of adaptation:

The initial rate of adaptation predicted from Eq (7) with input parameters and estimates of s_o in Table 1 gives the predicted value of $a\, s_o$ (Eq (9.A) with $t = 0$). From this the predicted value of a is directly obtained. The predicted values of a (Table 1) were close to the fitted values, and the latter were always included in the confidence envelope assuming $q \in [0.5,1.5]$. Fig. 2.a (*Drosophila*) and 2.b (*E.coli*) show that the fitted trajectory are close to that predicted using the values of a from the model's prediction.

The most striking result is that the response to selection (parameter a) is more than 20 times larger in Drosophila, whereas the two values (both observed and predicted) are very similar between the two *E. coli* studies in very contrasted conditions. Indeed, a small Drosophila population showed an approximate 1-fold increase in log-fitness in captivity in about 50 generations (Fig 1.a) while much larger *E.coli* populations took more than 2000 generations to reach a comparable fitness plateau. We can use the model to understand which factor (demography, sexual vs. asexual reproduction) generates these differences. We studied the relative impact of demography (parameters N_e and D), background selection and clonal interference on the prediction for each study. In table 1 we include the effect of each of these factors by measuring the relative reduction in adaptation rate predicted when including each of the relevant term into Eq. (7): p_{esc} (instead of $k_f s$) for demography, e^{-U/s_H} for background selection and $e^{-I(s)}$ for clonal interference. The results show that demography (small N_e in Drosophila, dilution bottlenecks for *E. coli* in glucose limiting medium) reduce the predicted rate of adaptation by less than one order of magnitude. Similarly, the predicted effect of background selection in *E. coli* is limited, at least in the initial phase of adaptation (about 30-60 % reduction). On the contrary, the negative effect of clonal interference is very important ($*10^{-6}$) and explains almost completely the large difference between response to selection in *Drosophila* and *E.coli*., despite the latter having $N\,U$ values by several orders of magnitude larger than the former. Clonal interference also explains why the response to selection (a) in

the two *E.coli* studies are very similar despite a 100-fold larger population size and the absence of bottlenecks for the study in the thymine-limiting environment. This "diminishing return" of rates of adaptation on population size (DE VISSER *et al.* 1999; GERRISH and LENSKI 1998) is well established effect of clonal interference. Overall, all the distinct factors taken into account have a significant impact on the prediction, so that neglecting any of them would have resulted in a lessened agreement with the empirical data. As an example, using the large sexual population approximation for the parameter a (Eq. (8)) in drosophila gives $Ne\ U\ \bar{s} / \beta_o = 0.19$ (0.09 - 0.28) instead of 0.11 (0.05-0.16) with the exact numerical integration of Eq. (7) (in Table 1). The approximation therefore gives the correct order of magnitude but is less in agreement with the fitted value $a = 0.074$.

Discussion:

The (somehow surprisingly) good agreement of the predictions with the data tends to support the overall validity of the model, of the previous theoretical results used to compute fixation probabilities and of the estimates of mutational parameters from the empirical literature. Of course, we do not pretend that agreement with such a limited number of empirical studies can validate the approach used here and the various previous results on which it relies. Rather, we hope it is an illustration that a quantitative test of adaptation theories is possible, based only on fitness trajectories over time, and on measures of deleterious mutation parameters (the most empirically available). Our aim was to propose a rationale for fitting empirical fitness trajectories with biologically meaningful parameters, and to propose a way to test theoretical predictions based on these fits. Our model also provides a way to quantify the relative influence of various factors affecting rates of adaptation (demography, background selection, clonal interference etc..), as illustrated in Table 1, by exploring which factor can be neglected without removing the agreement with the empirical data. Our results suggest that in the two studies of *E. coli*, clonal interference is the main limit to adaptation, explaining the much slower rate of adaptation in the bacteria than in Drosophila despite a much larger mutation supply.

Our approach is open to further empirical tests, for example by studying fitness trajectories with similar empirical designs and model species, using different environmental challenges. In such a case we expect that only the final fitness (s_o) should change and not the response to selection (parameter a). The model could also be tested in other organisms for which the relevant mutational parameters are known: the yeast *S. cerevisiae*, the nematode *C. elegans* and the virus *VSV* (with the limit that there is no agreement on $\bar{s}$ estimates in this species).

The requirements to test the model (in organisms with known mutational parameters) is to measure changes in fitness or fitness components (i) in a new constant environment, (ii) in a population with limited initial standing variance in fitness, and (ii) over a sufficiently long period of time that a fitness plateau is neared (providing a precise estimate of s_o). Given such data, the whole fitness trajectory could be predicted, and the predicted value of a could be compared with its observed (fitted) value. The sensitivity of the predictions on empirical estimates depends on the reproductive mode. In sexuals, $a \approx N_e\ U\ \bar{s} / \beta_o$ (Eq. (8)) so that precise estimates are needed for each parameter. In asexuals however, the impact of N_e estimates is limited due to clonal interference (GERRISH and LENSKI 1998) as observed here and in a previous empirical study (DE VISSER *et al.* 1999). The sensitivity on U estimates may be larger, as U determines the relative importance of clonal interference and background selection. Whether this effect is captured by the model could be tested in the *VSV* which has a very high mutation rate. This is all the more important as we can expect that the assumption of smaller beneficial than deleterious effects may be unrealistic when the ancestor is maladapted, as observed by REMOLD and LENSKI (REMOLD and LENSKI 2001). In this case,

modelling background selection in terms the factor e^{-U/s_H} may not give a good approximation, and more realistic approaches should be used (e.g. as in JOHNSON and BARTON 2002). Sensitivity on the parameter β_o may be important, and it is the most difficult to measure empirically. However, a survey of distributions of *s* across species suggests (yet based on few studies) that this parameter may not vary much, and that in most species the distribution is rather skewed. Such limited variation is expected theoretically if mutational and selective covariations between phenotypic traits are common (MARTIN and LENORMAND 2006b). This is interesting as it suggests that we may rely on some quantitative invariance of the shape of fitness effects distributions β_o of the order of, say, 1 to 4. On the contrary, the sensitivity of predictions on $\bar{s}$ is important as this parameter can vary by several orders of magnitude across species (ELENA and MOYA 1999; KIBOTA and LYNCH 1996; LYNCH *et al.* 1999; MARTIN and LENORMAND 2006b). Numerical integration of Eq. (7) suggested that the rate of adaptation also scales with $\bar{s}$ in asexuals (not shown). Therefore, the validity of the predictions may depend a lot on accurate empirical estimates of $\bar{s}$. This illustrates the importance of correcting the classic Bateman-Mukai estimates by variation in *s*, e.g. through maximum-likelihood (KEIGHTLEY 1994) or minimum-distance (GARCIA-DORADO and MARIN 1998) methods.

Another consequence of the predicted scaling with $\bar{s}$ is that, contrary to the first intuition, the present model suggests that a larger average deleterious effect may go with increased responses to new environments. This result comes from the landscape approach used in our model: $\bar{s}$ gives a measure of the average (fitness) size of random "mutational steps" in the landscape. The scale of deleterious effects measured in standard conditions (where all steps are towards decreased fitness as the genotype is well-adapted), is also the scale of potential steps towards a new optimum, in a new environment. One may argue that when background selection has a strong impact on fixation probabilities (*U* large) increased effect of deleterious mutations may limit adaptation. However, mutations of larger deleterious effects are also expected to have a lower frequency in the population, so that beneficial mutations are less likely to be "trapped" in strongly deleterious backgrounds. As a result, increased average deleterious effect of mutation is not expected to result in more background selection under our model, at least in the simplest situation. Indeed, integrating Eq. (4) for s_o close to zero and large N_e, gives $e^{-U/s_H} = Exp\left(-U\beta_o/((\beta_o - 1)\bar{s})\right)$ so that larger $\bar{s}$ values correspond to decreased background selection. Comparisons across empirical estimates reveals a large increase in $\bar{s}$ from lower to higher organisms. If our model is valid, this suggests that higher organisms may adapt more rapidly, all else being equal than lower ones (with time measured in number of generations). This might compensate for their lower generation time. Such predictions (*all else being equal*) could be tested by comparing the parameter *a* given here with those obtained from diploid yeast *S. cerevisiae* (lower sexual), and *C. elegans* (higher asexual).

To predict adaptation rates in *E. coli* we assumed a value of $\bar{s}$ one order of magnitude lower than the Bateman-Mukai estimates (KIBOTA and LYNCH 1996) and the direct measure in (ELENA *et al.* 1998b). This was done to account for a possibly large effect of the antibiotic marker used in the latter study. Such a large effect may seem implausible. However, we recall that, as explained above, taking into account such direct marker effect allows to reconcile (biased) estimates from spontaneous mutation accumulation (KIBOTA and LYNCH 1996) and (unbiased) direct measures (i) of $\bar{s}$ from insertion mutagenesis (ELENA *et al.* 1998b) and of *U* from mutation rates per gene (DRAKE *et al.* 1998). Furthermore, as our predictions are sensitive to errors in the value assumed for $\bar{s}$ (see above), the quality of the predictions for two independent studies on *E. coli*, suggest that an estimate of $\bar{s} \approx 0.001$ may be more

realistic than $\bar{s} \approx 0.01$. Transposon insertion is a powerful method to produce single mutants, and thus measure $f(s)$ directly, which is not possible with spontaneous mutation accumulation. It is also probably the method that can most readily be extended to other model organisms (and it has already been used in yeast, drosophila and *E. coli*). However, we believe that some corrections must be made on the resulting estimates of $f(s)$ parameters to account for potential direct effects of transposition.

Our model is an attempt to propose quantitative predictions on rates of adaptation, based on empirically available quantities. It provides simple predictions that seem consistent with the three empirical studies analysed here. The model also provides a rationale for assessing quantitatively the relative impact of various factors limiting adaptation (demography, sex, etc.). It would have to be tested in other species and, in Drosophila and *E. coli*, under other environmental challenges. If empirical tests confirm the present results, the model could be used to predict long term fitness trajectories from only initial adaptation rates. Indeed the initial rate is $r_o = r_o(s_o, a) \approx s_o a$, where a can be predicted so that both parameters (s_o and a) can be known from initial fitness trajectories. This could have important applications for the study of the dynamics of adaptation in economically or medically important species.

Figure Legend

Figure 1: Predicted rate of adaptation (y-axis, log-scale) as a function of the fitness distance to the optimum (s_o x-axis)
Plain lines give the exact numerical integration of Eq. (7) for each value of s_o. Dashed lines give the linerar approximation in Eq. (8) $d\log(W)/dt = N_e\ U\ \bar{s}\ /\beta_o\ s_o$. The mutational and demographic parameters used are those for Drosophila given in Table 1 except that three distinct values of β_o (indicated on the graph) were assumed.

Figure 2: Predicted, fitted and observed fitness trajectories
Log relative fitness ($\log(W(t)/W_o)$) is given as a function of the number of generations since the initiation of the adaptation experiment. Squares: observed trajectories, the grey plain line: fitted trajectory (using the exponential model in Eq. (9.B) and fitted s_o and a in Table 1), thick dashed line: predicted trajectory (idem but using the fitted s_o and predicted a for q = 1 in Table 1), thin dashed lines: envelope (idem but using the fitted s_o and predicted a for q = 0.5 or 1.5 in Table 1. **a.** Adaptation to captivity conditions in *Drosophila melanogaster* from (GILLIGAN and FRANKHAM 2003) **b.** Long-term adaptation in *E. coli*. Black squares: thymine-limiting medium from (DE CRECY-LAGARD *et al.* 2001), white squares: glucose-limiting medium from (LENSKI and TRAVISANO 1994) (scaled as $5*\log(W(t)/W_o)$ for comparisons).

Appendix 1: change in the variance of *s* with s_o, and moments of *q*

Let $\mathbf{\Lambda} = \mathbf{diag}(\lambda_i)$, be the diagonal matrix with diagonal elements λ_i. Then, assuming s is small enough that $\log(1+s) \approx s$, the fitness effect s of a given mutation can be expressed as a quadratic form in standard gaussian vectors

$$s \approx \log(1+s) = -\frac{1}{2}\mathbf{dx}^{\mathrm{T}}\mathbf{\Lambda}\,\mathbf{dx} - \mathbf{x}_{\mathrm{o}}^{\mathrm{T}}\mathbf{\Lambda}\,\mathbf{dx} \qquad (10)$$

Where $\mathbf{dx}$ follows a standard multivariate gaussian distribution with mean $\mathbf{0}$ and covariance matrix identity $\mathbf{I}$. The vector $\mathbf{x}_o$ is obtained by a linear transformation of the phenotype of the ancestor $\mathbf{z}_o$ and is such that $s_o = ½\ \mathbf{x}_o\ \mathbf{\Lambda}\ \mathbf{x}_o$ (see MARTIN and LENORMAND 2006b). The mean and variance of s defined in Eq. (10) are given by

$$\begin{aligned} E(s) &= -\bar{s} = -\frac{1}{2}Tr(\mathbf{\Lambda}) \\ V(s) &= \frac{1}{2}Tr(\mathbf{\Lambda}^2) + \mathbf{x}_{\mathrm{o}}^{\mathrm{T}}\mathbf{\Lambda}^2\,\mathbf{x}_{\mathrm{o}} \end{aligned}, \qquad (11)$$

Where $Tr(.)$ denotes matrix trace. It can be seen from Eq. (11) that the average s does not depend on the ancestor phenotype (on $\mathbf{x}_o$), which is due to the quadratic approximation for the fitness function $W(\mathbf{z})$ (see MARTIN and LENORMAND 2006b). However, the variance of s increases with increasing maladaptation of the ancestor (with $|\mathbf{x}_o|$). Denote $V(s)^* = ½\ \mathrm{Tr}(\mathbf{\Lambda}^2) = \bar{s}^2 / \beta_o$, the variance of s when the ancestor is perfectly adapted ($\mathbf{x}_o = \mathbf{0}$, $s_o = 0$). Then from Eq. (11), the variance of s for any s_o ($\mathbf{x}_o$) can be expressed as

$$V(s) = V(s)^*\left(1 + 2\,q\,\frac{s_o}{\bar{s}}\right), \qquad (12)$$

Where, because $s_o = ½\ \mathbf{x}_o\ \mathbf{\Lambda}\ \mathbf{x}_o$, q is a ratio of quadratic forms in $\mathbf{x}_o$

$$q = \frac{Tr(\mathbf{\Lambda})}{Tr(\mathbf{\Lambda}^2)}\frac{\mathbf{x_o}\ \mathbf{\Lambda}^2\ \mathbf{x_o}}{\mathbf{x_o}\ \mathbf{\Lambda}\ \mathbf{x_o}}. \qquad (13)$$

The displaced gamma approximation in Eq. (1) is obtained by fitting the distribution of s to $s_o - \gamma$ where γ is gamma distributed with parameters α and β chosen so that s_o - γ has the same mean and variance as given in Eqs. (11) and (12). This gives the value of α and β in Eq. (2). Assume that a new environment sets a new phenotypic optimum, then a new vector $\mathbf{x}_o$ connects the ancestor phenotype to this new optimum. In other words, $\mathbf{x}_o$ changes with the environment (even though the ancestor phenotype remains unchanged) because it is defined relative to the optimum in this environment. Note that the scaling of $\mathbf{x_o}$ is irrelevant as q is a ratio: the value of q does not depend on the norm of $\mathbf{x}_o$ but on its direction which is a random variable with respect to different environments. As q is a ratio of quadratic forms in random vectors ($\mathbf{x}_o$), we can use the fact that q is independent of its denominator (CONNIFFE and SPENCER 2001; GUPTA and KABE 1998) so that, for any order, the raw moment of the ratio (q) equals the ratio of the raw moments of its numerator by its denominator. Then, assuming that

$E(\mathbf{x}_o) = \mathbf{0}$ (no bias towards a given direction), the mean of the numerator and denominator of q can be derived using $E(\mathbf{x}_o\ \mathbf{\Lambda}\ \mathbf{x}_o) = Tr(\mathbf{\Lambda})$ and $E(\mathbf{x}_o\ \mathbf{\Lambda}^2\ \mathbf{x}_o) = Tr(\mathbf{\Lambda}^2)$, so that it is obvious that $E(q) = 1$.

We can also use this property of ratios to derive the second raw moment $E(q^2)$, provided some assumptions on the distribution of $\mathbf{x}_o$ for different environments. Assuming that $\mathbf{x}_o$ may point in any direction across environments, we can consider that $\mathbf{x_o}$ follows any circular distribution in space, for example a standard multivariate gaussian distribution with mean $\mathbf{0}$ and covariance matrix identity $\mathbf{I}$. We can then derive the 2nd moments of the numerator and denominator: $E((\mathbf{x}_o\ \mathbf{\Lambda}\ \mathbf{x}_o)^2) = 2Tr(\mathbf{\Lambda}^2) + Tr(\mathbf{\Lambda})^2$ and $E((\mathbf{x}_o\ \mathbf{\Lambda}^2\ \mathbf{x}_o)^2) = 2Tr(\mathbf{\Lambda}^4) + Tr(\mathbf{\Lambda}^2)^2$. Let us define two parameters determined by the distribution of the λ_i across traits i: $v_2 = \overline{\lambda^2} / \overline{\lambda}^2$ and $v_4 = \overline{\lambda^4} / \overline{\lambda^2}^2$ where bars denote the average across eigenvalues λ_i. Then the variance of q can be written

$$\sigma_q^2 = E(q^2) - 1 = \frac{1}{n}\ \frac{2(v_4 - v_2)}{1 + 2\,v_2 / n}. \qquad (14)$$

Note that σ_q^2 is proportional to $1/n$ so that in the limit of a large number of traits it should be small. However, it has been shown that the parameter v_2 might also be very large if e.g. matrices **S** and **M** are drawn randomly (MARTIN and LENORMAND 2006b). One way to get an order of magnitude for σ_q^2 is to notice that $v_2/n = (1+\mathrm{CV}(\lambda)^2)/n = 2/\beta_o$ (see the definition in Eq. (2)). Therefore, assuming that v_4/n is of the same order as v_2/n (even if both are not negligible) we get that

$$\sigma_q^2 \sim \frac{2\,v_2 / n}{1 + 2\,v_2 / n} \sim \frac{1}{1 + \beta_o}. \qquad (15)$$

Based on the observed shape parameters of distributions of mutation fitness effects in model organisms $\beta_o \sim 2$ (see Table 1 and MARTIN and LENORMAND 2006b), we get a (very rough) estimate for σ_q of the order of 0.5.

Another approach to determine σ_q is to consider that **S** and **M** are drawn randomly as large Wishart matrices (a classic model of random covariance matrices, see (MARTIN and LENORMAND 2006b). Under this null model the values of v_2 and v_4 can be derived using Random Matrix Theory (BAI 1999), according to n and the strength of phenotypic correlations. This gives a more precise estimate for σ_q that is always ≤ 0.3 for n sufficiently large ($n > 15$). Overall we chose to set the range of variation for q to the 95% confidence interval of the normal with mean 1 and standard deviation $\sigma_q = 0.3$ which gives $q \in [0.5,1.5]$.

Appendix 2: analytic expressions for large populations

The probability of beneficial mutations when $f(s)$ follows the displaced gamma defined in Eq. (1) is given by

$$p_b = \int_o^{s_o} f(s)\,ds = 1 - \frac{\gamma\left(\beta, s_o/\alpha\right)}{\Gamma(\beta)}. \qquad (16)$$

Where $\gamma(.)$ denotes the lower incomplete gamma function and α and β are given by Eq. (2). This probability increases with increasing distance to the optimum (s_o). In the case of a large

diploid sexual population, the probability of fixation of an advantageous mutation is $p_{esc} = N_e/N\, s$ (limit of Eq. (3) for s<<N,N_e). In this case defining $x = s_o/\alpha$, and using properties of the incomplete gamma function the rate of adaptation can be expressed as

$$\frac{d\,\log(W)}{dt} = N\,U \int_0^{s_o} p_{esc}\; s\; f(s)\; ds = N_e\,U\,\bar{s}^2 \left(\frac{e^{-x} x^{\beta} (x - 1 - \beta)}{\Gamma(\beta)} + ((x - \beta)^2 + \beta)\; p_b \right). \qquad (17)$$

Finding a linear approximation for this rate as a function of s_o is not straightforward. However, we can use the fact that when s_o is large relative to $\bar{s}$, the shape of the displaced gamma (β in Eq. (2)) becomes large so that $f(s)$ is approximately gaussian with mean and variance given in Eq. (11) and (12) above. Therefore, when away from the optimum, the distribution of s can be approximated by a gaussian with mean $-\bar{s}$ and variance $\bar{s}^2 / \beta_o\,(1 + 2\,q\,s_o / \bar{s})$. Using the probability density function $g(s)$ of this gaussian instead of the displaced gamma $f(s)$ and simplifying assuming $\bar{s} << s_o$ gives a linear expression for the rate of adaptation in a large sexual population away from the optimum

$$\frac{d\,\log(W)}{dt} = N\,U \int_0^{\infty} \frac{N_e}{N}\, s^2\; g(s)\; ds \underset{s_o >> \bar{s}}{\approx} N_e\,U\,\frac{\bar{s}}{\beta_o}\, q\, s_o\,. \qquad (18)$$

In the case of asexuals, clonal interference decreases the fixation probability of beneficial mutations. This effect depends on the number of mutations $I(s)$ interfering with a given mutation of beneficial effect s. $I(s)$ is computed as the number of beneficial mutations with effect $s_i > s$ that appear during the selective sweep of the focal mutation (with effect s) and that escape drift loss (GERRISH and LENSKI 1998). Indeed, in a very large population, any of these interfering mutations of effect s_i (arising on a different background) will ultimately eliminate the focal mutation. Using the probability density in Eq. (1), the number of interfering mutations can be derived through a change of variable from s_i to $s_i' = s_i - s$ as

$$I(\mathrm{s}) = \frac{N\,U\,\log(N)}{s} \int_{s_i'=0}^{s_o - s} p_{esc}(s + s_i')\;\; f(s + s_i')\; ds_i'\,. \qquad (19)$$

Which yields the result in Eq. (6)

References:

AVILA, V., and A. GARCIA-DORADO, 2002 The effects of spontaneous mutation on competitive fitness in *Drosophila melanogaster*. Journal of Evolutionary Biology 15: 561-566.

BAI, Z. D., 1999 Methodologies in spectral analysis of large dimensional random matrices, a review. Statistica Sinica 9: 611-662.

BARTON, N. H., and M. C. WHITLOCK, 1997 The evolution of metapopulations, pp. 183-210 in *Metapopulation Biology*, edited by H. I.A. and G. M.E. Academic Press.

BJEDOV, I., O. TENAILLON, B. GERARD, V. SOUZA, E. DENAMUR *et al.*, 2003 Stress-induced mutagenesis in bacteria. Science 300: 1404-1409.

BULL, J. J., and S. P. OTTO, 2005 The first steps in adaptive evolution. Nature Genetics 37: 342-343.

CONNIFFE, D., and J. E. SPENCER, 2001 When moments of ratios are ratios of moments. Journal of the Royal Statistical Society Series D-the Statistician 50: 161-168.

CROW, J., and M. KIMURA, 1970 *An introduction to population genetics theory*. Alpha Editions, Minneapolis.

DE CRECY-LAGARD, V., J. BELLALOU, R. MUTZEL and P. MARLIERE, 2001 Long term adaptation of a microbial population to a permanent metabolic constraint: overcoming thymineless death by experimental evolution of Escherichia coli. BMC Biotechnology 1: 10.

DE VISSER, J., C. W. ZEYL, P. J. GERRISH, J. L. BLANCHARD and R. E. LENSKI, 1999 Diminishing returns from mutation supply rate in asexual populations. Science 283: 404-406.

DRAKE, J. W., B. CHARLESWORTH, D. CHARLESWORTH and J. F. CROW, 1998 Rates of spontaneous mutation. Genetics 148: 1667-1686.

ELENA, S. F., M. DAVILA, I. S. NOVELLA, J. J. HOLLAND, E. DOMINGO *et al.*, 1998a Evolutionary dynamics of fitness recovery from the debilitating effects of Muller's ratchet. Evolution 52: 309-314.

ELENA, S. F., L. EKUNWE, N. HAJELA, S. A. ODEN and R. E. LENSKI, 1998b Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica 103: 349-358.

ELENA, S. F., F. GONZALEZCANDELAS, I. S. NOVELLA, E. A. DUARTE, D. K. CLARKE *et al.*, 1996 Evolution of fitness in experimental populations of vesicular stomatitis virus. Genetics 142: 673-679.

ELENA, S. F., and R. E. LENSKI, 2003 Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. Nature Reviews Genetics 4: 457-469.

ELENA, S. F., and A. MOYA, 1999 Rate of deleterious mutation and the distribution of its effects on fitness in vesicular stomatitis virus. Journal of Evolutionary Biology 12: 1078-1088.

FISHER, R. A., 1930 *The genetical theory of natural selection*. Oxford University Press, Oxford.

GARCIA-DORADO, A., C. LOPEZ-FANJUL and A. CABALLERO, 1999 Properties of spontaneous mutations affecting quantitative traits. Genetical Research 74: 341-350.

GARCIA-DORADO, A., and J. M. MARIN, 1998 Minimum distance estimation of mutational parameters for quantitative traits. Biometrics 54: 1097-1114.

GERRISH, P. J., and R. E. LENSKI, 1998 The fate of competing beneficial mutations in an asexual population. Genetica 103: 127-144.

GILLESPIE, J. H., 1984 Molecular Evolution over the Mutational Landscape. Evolution 38: 1116-1129.

GILLIGAN, D. M., and R. FRANKHAM, 2003 Dynamics of genetic adaptation to captivity. Conservation Genetics 4: 189-197.

GUPTA, A. K., and D. G. KABE, 1998 Moments of ratios of quadratic forms. Statistics & Probability Letters 38: 69-71.

IMHOF, M., and C. SCHLOTTERER, 2001 Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. Proceedings of the National Academy of Sciences of the United States of America 98: 1113-1117.

JOHNSON, T., and N. H. BARTON, 2002 The effect of deleterious alleles on adaptation in asexual populations. Genetics 162: 395-411.

JOHNSON, T., and P. J. GERRISH, 2002 The fixation probability of a beneficial allele in a population dividing by binary fission. Genetica 115: 283-287.

KEIGHTLEY, P. D., 1994 The Distribution of Mutation Effects On Viability in *Drosophila melanogaster*. Genetics 138: 1315-1322.

KEIGHTLEY, P. D., 2004 Comparing analysis methods for mutation-accumulation data. Genetics 167: 551-553.

KIBOTA, T. T., and M. LYNCH, 1996 Estimate of the genomic mutation rate deleterious to overall fitness in *E-coli*. Nature 381: 694-696.

KIVISAAR, M., 2003 Stationary phase mutagenesis: mechanisms that accelerate adaptation of microbial populations under environmental stress. Environmental Microbiology 5: 814-827.

LENSKI, R. E., and M. TRAVISANO, 1994 Dynamics of Adaptation and Diversification - a 10,000-Generation Experiment with Bacterial-Populations. Proceedings of the National Academy of Sciences of the United States of America 91: 6808-6814.

LOEWE, L., V. TEXTOR and S. SCHERER, 2003 High deleterious genomic mutation rate in stationary phase of *Escherichia coli*. Science 302: 1558-1560.

LYMAN, R. F., F. LAWRENCE, S. V. NUZHDIN and T. F. C. MACKAY, 1996 Effects of single P-element insertions on bristle number and viability in *Drosophila melanogaster*. Genetics 143: 277-292.

LYNCH, M., J. BLANCHARD, D. HOULE, T. KIBOTA, S. SCHULTZ *et al.*, 1999 Perspective: Spontaneous deleterious mutation. Evolution 53: 645-663.

MARTIN, G., and T. LENORMAND, 2006a The fitness effect of mutations in stressful environments: a survey in the light of fitness landscape models. Evolution.

MARTIN, G., and T. LENORMAND, 2006b A multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60: 893-907.

MIRALLES, R., P. J. GERRISH, A. MOYA and S. F. ELENA, 1999 Clonal interference and the evolution of RNA viruses. Science 285: 1745-1747.

ORR, H. A., 1998 The population genetics of adaptation: The distribution
of factors fixed during adaptive evolution. Evolution 52: 935-949.

ORR, H. A., 2000 The rate of adaptation in asexuals. Genetics 155: 961-968.

ORR, H. A., 2002 The population genetics of adaptation: The adaptation of DNA sequences. Evolution 56: 1317-1330.

ORR, H. A., 2003 The distribution of fitness effects among beneficial mutations. Genetics 163: 1519-1526.

ORR, H. A., 2005a The genetic theory of adaptation: A brief history. Nature Reviews Genetics 6: 119-127.

ORR, H. A., 2005b Theories of adaptation: what they do and don't say. Genetica 123: 3-13.

OTTO, S. P., and M. C. WHITLOCK, 1997 The probability of fixation in populations of changing size. Genetics 146: 723-733.

PAPADOPOULOS, D., D. SCHNEIDER, J. MEIER-EISS, W. ARBER, R. E. LENSKI *et al.*, 1999 Genomic evolution during a 10,000-generation experiment with bacteria. Proceedings of the National Academy of Sciences of the United States of America 96: 3807-3812.

PECK, J. R., 1994 A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137: 597-606.

REMOLD, S. K., and R. E. LENSKI, 2001 Contribution of individual random mutations to genotype-by-environment interactions in *Echerichia coli*. Proceedings of the National Academy of Science 98: 11388-11393.

ROKYTA, D. R., P. JOYCE, S. B. CAUDLE and H. A. WICHMAN, 2005 An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nature Genetics 37: 441-444.

ROZE, D., F. ROUSSET and Y. MICHALAKIS, 2005 Germline bottlenecks, biparental inheritance and selection on mitochondrial variants: A two-level selection model. Genetics 170: 1385-1399.

ROZEN, D. E., J. DE VISSER and P. J. GERRISH, 2002 Fitness effects of fixed beneficial mutations in microbial populations. Current Biology 12: 1040-1045.

SANJUÀN, R., A. MOYA and S. F. ELENA, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences of the United States of America 101: 8396-8401.

SNIEGOWSKI, P., 2004 Evolution: Bacterial mutation in stationary phase. Current Biology 14: R245-R246.

TADDEI, F., M. RADMAN, J. MAYNARDSMITH, B. TOUPANCE, P. H. GOUYON *et al.*, 1997 Role of mutator alleles in adaptive evolution. Nature 387: 700-702.

WAHL, L. M., P. J. GERRISH and I. SAIKA-VOIVOD, 2002 Evaluating the impact of population bottlenecks in experimental evolution. Genetics 162: 961-971.

WAHL, L. M., and D. C. KRAKAUER, 2000 Models of experimental evolution: The role of genetic chance and selective necessity. Genetics 156: 1437-1448.

WRIGHT, B. E., 2004 Stress-directed adaptive mutations and evolution. Molecular Microbiology 52: 643-650.

ZEYL, C., 2004 Capturing the adaptive mutation in yeast. Research in Microbiology 155: 217-223.

ZHANG, X. S., and W. G. HILL, 2003 Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. Evolution 57: 1761-1775.

Table 1: Fitted and predicted fitness trajectories

Species	input parameters							trajectory parameters			Limits to adaptation		
	mutation			demography / reproduction				fitted		predicted (s_o)	Background selection	Clonal interference	drift
	β_o	$\bar{s}$	U	N	N_e	k_f	D	s_o	a	a if q =1 (+/- 0.5)			
Drosophila melanogaster [1]	1.25	0.06	0.015	1000	300	1	no	1.186	0.074	0.11 (0.053 - 0.16)	no	no	14%
E. coli in glucose limiting medium [2]	2.81	0.0011	0.0022	$5\ 10^8$	$5\ 10^8$	2.8	0.01	0.37	0.00073	0.0012 (0.00045 - 0.0019)	46% - 5.6%	$4.1\ 10^{-6}$	21%
E. coli in thymine limiting medium [3]	2.81	0.0011	0.0022	10^{10}	10^{10}	2.8	no	3.04	0.0023	0.0024 (0.0012 - 0.0037)	71% - 5.6%	$5.2\ 10^{-8}$	100%

The input parameters for mutation where taken from the literature (see 2nd section) and the demographic parameters were given in each study. The trajectory parameters are (i) the values of s_o and a from the fit of an exponential model as given in (9.B) and (ii) the predicted value of a based on the fitted s_o value (= $1/s_o$ * predicted initial rate of adaptation using Eq. (7)). Values into brackets give the "envelope" of a for q = 0.5 or 1.5. The last part of the Table gives the influence of the inclusion of each adaptation limiting factor, computed as the ratio of the rate of adaptation with the factor included to that of a large sexual population ($p_{esc} = k_f s$, no clonal interference, no background selection).
Sources are [1] : (GILLIGAN and FRANKHAM 2003), [2] : (LENSKI and TRAVISANO 1994), [3] : (DE CRECY-LAGARD *et al.* 2001).

Figure 1

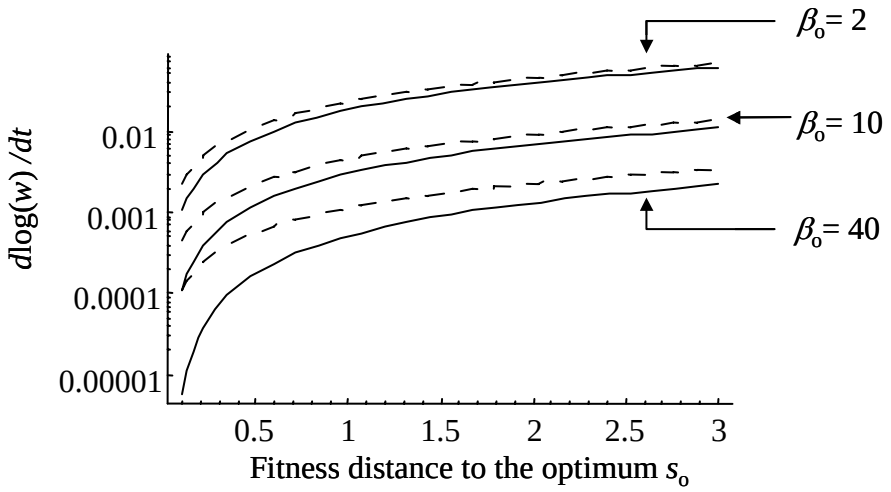
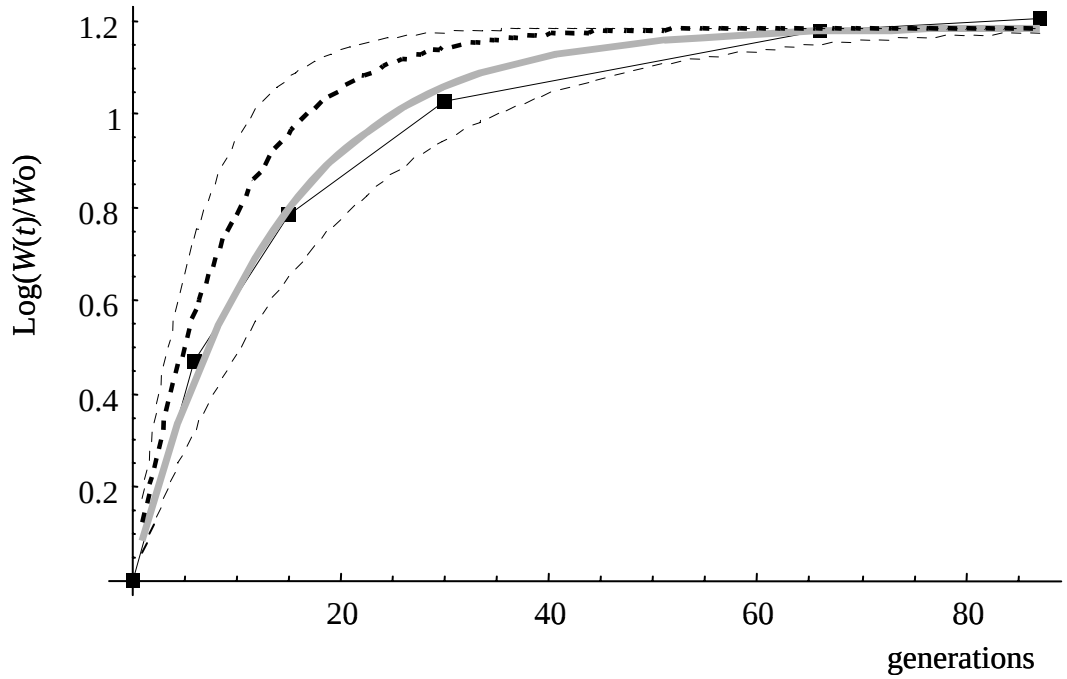
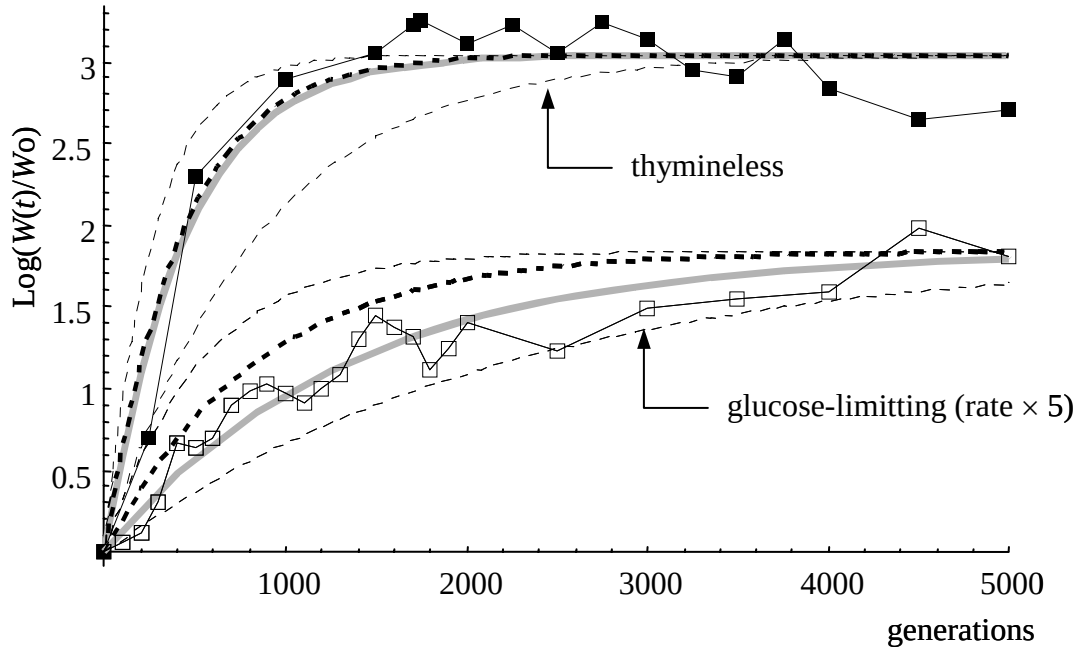

β_o= 2
β_o= 10
β_o= 40
$d\log(w)/dt$
0.01
0.001
0.0001
0.00001
0.5
1
1.5
2
2.5
3
Fitness distance to the optimum s_o

Figure 2

a.


1.2
1
0.8
0.6
0.4
0.2
Log(W(t)/Wo)
20
40
60
80
generations

b.


3
2.5
2
1.5
1
0.5
Log(W(t)/Wo)
thymineless
glucose-limitting (rate × 5)
1000
2000
3000
4000
5000
generations

DOI: 10.1534/genetics.104.039982

Selection for Recombination in Structured Populations

Guillaume Martin,* Sarah P. Otto*,† and Thomas Lenormand*,[1]

**CEFE-CNRS, 34293 Montpellier, France and †Zoology Department, University of British Columbia, Vancouver, V6T 1Z4 British Columbia, Canada*

Manuscript received December 20, 2004
Accepted for publication May 1, 2005

ABSTRACT

In finite populations, linkage disequilibria generated by the interaction of drift and directional selection (Hill-Robertson effect) can select for sex and recombination, even in the absence of epistasis. Previous models of this process predict very little advantage to recombination in large panmictic populations. In this article we demonstrate that substantial levels of linkage disequilibria can accumulate by drift in the presence of selection in populations of any size, provided that the population is subdivided. We quantify (i) the linkage disequilibrium produced by the interaction of drift and selection during the selective sweep of beneficial alleles at two loci in a subdivided population and (ii) the selection for recombination generated by these disequilibria. We show that, in a population subdivided into n demes of large size N, both the disequilibrium and the selection for recombination are equivalent to that expected in a single population of a size intermediate between the size of each deme (N) and the total size (nN), depending on the rate of migration among demes, m. We also show by simulations that, with small demes, the selection for recombination is stronger than both that expected in an unstructured population ($m = 1 - 1/n$) and that expected in a set of isolated demes ($m = 0$). Indeed, migration maintains polymorphisms that would otherwise be lost rapidly from small demes, while population structure maintains enough local stochasticity to generate linkage disequilibria. These effects are also strong enough to overcome the twofold cost of sex under strong selection when sex is initially rare. Overall, our results show that the stochastic theories of the evolution of sex apply to a much broader range of conditions than previously expected.

WHY sex and recombination are so widespread in nature is an age-old debate in evolutionary biology. While some theories invoke mechanistic advantages for sex (*e.g.*, DNA repair), most theories account for sex on the basis of its effects on multilocus allelic combinations (Kondrashov 1993). These hypotheses can be termed *generative hypotheses*, because they focus on the effects of sex and recombination on the array of genotypes generated within a population. According to several generative hypotheses, sex and recombination are advantageous because they facilitate the response to selection by reducing negative linkage disequilibria, whereby beneficial alleles are found on low-fitness genetic backgrounds, thereby increasing the genetic variance in fitness (Maynard Smith 1971; Felsenstein 1974). This class of explanations is consistent with experiments showing that higher rates of recombination often evolve as a pleiotropic response to artificial selection for other traits (see Otto and Barton 2001 for a review).

Generative hypotheses can be classified according to the force generating linkage disequilibria (LD) within a population (Kondrashov 1993; Barton 1995a; Otto and Lenormand 2002). LD can be produced by selection involving epistasis (Feldman *et al.* 1980) or can result from the interaction of drift with directional selection (Hill and Robertson 1966). In this article we focus on drift-based explanations for LD. Because drift in the presence of selection causes an accumulation of negative linkage disequilibria, linkage imposes a limit on the efficacy of natural selection in finite populations (Hill and Robertson 1966; Felsenstein 1974; Barton 1995b). Sex and recombination release populations from this limit, by allowing beneficial alleles within different individuals to come together, as recognized early on by both Fisher (1930) and Muller (1932). The accumulation of negative disequilibria due to drift in the presence of selection on linked loci is often referred to as the Hill-Robertson effect (HRE). The HRE results in selective interference among loci, which reduces the probability of fixation of beneficial alleles as the population size gets smaller or as linkage among loci tightens (Hill and Robertson 1966; Barton 1995b). In the extreme case of an asexual population, the HRE affects the entire genome and is then referred to as "clonal interference" (Gerrish and Lenski 1998).

The HRE imposes an important limit on the response to selection on linked sets of loci in sexual species

[1]*Corresponding author:* CEFE -CNRS, 1919 Route de Mende, 34293 Montpellier Cedex 5, France. E-mail: thomas.lenormand@cefe.cnrs.fr

and on the whole genome of asexuals (Barton 1995b; Barton and Partridge 2000). The HRE operates whenever several alleles are segregating simultaneously within a population, regardless of whether these alleles are favorable and spreading [as in the Fisher-Muller model (Fisher 1930; Muller 1932)], deleterious mutations [as in Muller's ratchet (Muller 1932)], or both (Peck 1994). Under all of these scenarios, the HRE selects for sex and recombination to reduce negative associations among the most fit alleles generated by drift in the presence of selection (Felsenstein and Yokoyama 1976; Otto and Barton 2001; Otto and Lenormand 2002).

This "stochastic theory" (Kondrashov 1993) for the advantage of sex has the seductive property of being widely applicable because all populations are finite. Moreover, it does not require any particular form of epistasis, provided that epistasis is typically small (Otto and Barton 1997, 2001). The LD produced by the HRE is, however, inversely proportional to the size of the population and can be very small in a large sexual population, except when many loci undergo selection simultaneously (Iles *et al.* 2003). Thus, it would appear at first that the HRE cannot provide a compelling explanation for the maintenance of sex in large populations. At the other extreme, the HRE can also fail as an explanation in very small populations (Otto and Barton 2001) or in very stable environments because too few beneficial alleles will segregate simultaneously. Indeed, a reduction of the advantage of sex in small populations has been confirmed experimentally (Colegrave 2002), although it is not clear whether the advantage for sex observed in the larger populations was due to the HRE or to weak synergistic epistasis. Therefore, the generality of the HRE as an explanation for the ubiquity of sex among eukaryotes is not straightforward; its main restrictive requirement is that populations must be of intermediate size—large enough for several mutations to segregate and yet small enough for significant linkage disequilibria to develop.

The theory above assumes, however, that populations are unstructured. With the discovery of genetic markers, a considerable amount of data have accumulated, showing that most populations exhibit spatial structure, at least through isolation by distance. However, the effect of spatial structure on multilocus adaptation and on the evolution of recombination has received relatively little attention. The effect of population subdivision on the maintenance of sex has been examined in a few studies. In particular, it has been shown that population structure can enhance the advantage of sexual over asexual lineages under Muller's ratchet (Peck *et al.* 1999). Furthermore, by increasing the frequency of homozygotes, population structure can impart an advantage to sex in diploids by reducing the mutation load (Agrawal and Chasnov 2001) and improving the efficacy of selection (Otto 2003), although these advantages arise from segregation rather than from recombination and are not directly related to the HRE. These models show that sex can be maintained without the need for synergistic epistasis, provided that the population is subdivided. Nevertheless, we lack a general analytical framework in which to understand the role of drift on linkage disequilibrium and on the evolution of sex and recombination in structured populations.

In this article, we explore the interaction of drift and selection in subdivided populations. Using an island model of selection in the absence of epistasis, we develop an analytical model to quantify the average LD generated during the selective sweep of beneficial alleles at two linked loci. The analytical model assumes that drift is weak within each deme, and we use simulations for the case of smaller deme sizes (for both a sex and a recombination modifier). We demonstrate that negative associations develop among selected alleles in subdivided populations of any size. We also show that these associations reduce the rate of spread of favorable alleles, although this effect is substantial only when selection is strong relative to recombination. These negative associations select for increased rates of sex and recombination, even in very large populations to a level that can overcome the twofold cost of sex under strong selection. Rates of migration and deme size are shown to play a critical role in determining the strength of selection for sex and recombination.

In the first part of this article, we summarize, in compact vector notation, the model introduced by Barton and Otto (2005) to predict the expected linkage disequilibrium generated between two linked loci exposed to directional selection and drift in a single large population. In the second part, we extend this model to a population subdivided into a number of large demes. We derive recursion equations for the expectation and variance of the mean linkage disequilibrium generated between selected loci by the HRE. We also give a simplified approximate expression for the LD under weak selection and loose linkage (quasi-linkage equilibrium). We then give the expected frequency change at a modifier locus, changing the recombination rate between the selected loci. We supplement this analysis with simulation results for the case of smaller demes. We also give simulation results for the case of a modifier of sex arising in an asexual population (with or without a twofold cost). The reader should keep in mind that our analytical model is intended to quantify the LD generated in a metapopulation and the subsequent selection for recombination and that it does not fully capture the limits to adaptation imposed by the HRE. Indeed, the analysis assumes that beneficial alleles are sufficiently common in the whole population that they always fix, which ignores the influence of the HRE on the fixation probability of beneficial alleles. We end by discussing the implications of our findings for the validity of the stochastic theory for the evolution of sex and its empirical tests.

MODEL

Our model builds on the single-population analysis of Barton and Otto (2005), describing the dynamics of linkage disequilibria in a single finite population. We analyze an island model with an arbitrary number of demes for the development of disequilibria among selected loci. We assume that selection is multiplicative and homogeneous over space, so that random genetic drift is the ultimate source of disequilibria among loci. To extend the method to subdivided populations, we introduce a compact vector notation. We begin with a general description of this model. We then summarize the results for a single population and finally turn to the case of a structured population.

Genetic setting: We model a population consisting of n demes, each containing $2N$ chromosomes. Initially, we keep track of two alleles at each of two loci, j and k, separated by r_{jk} units of recombination. Later, we add a third locus that modifies the rate of recombination. We assume multiplicative viability selection, so that no LD is generated by epistasis. Because of our assumption that selection is multiplicative both within and among loci, this model describes a population of either $2Nn$ haploid individuals or Nn diploid individuals. We follow the frequency x_j (x_k) of a beneficial allele with selective advantage s_j (s_k) at locus j (k), as well as the linkage disequilibrium x_{jk}. The three variables characterizing a given population $\{x_j, x_k, x_{jk}\}$ at time t are written as elements of a vector $\mathbf{x}$, where we use the set of subscripts $U = \{j, k, jk\}$ to denote the elements in $\mathbf{x}$. In the following analysis, it is useful to refer to one of these subscripts without specifying which one, which we do by using the subscripts a, b, or c. For instance, the definition of $\mathbf{x}$ is $\mathbf{x} \equiv \{x_a\}_{a\in U}$. To distinguish among demes when there is more than one deme, we add $[i]$ to denote the value of a variable or of a vector in deme i.

Life cycle: The life cycle consists of either selection in the haploid phase followed by random mating or random mating followed by selection in the diploid phase, after which meiosis occurs to produce an effectively infinite population of haploid juveniles. At this stage, population regulation occurs, such that a finite population of individuals is sampled in each deme, followed by migration in the haploid phase. We chose this life cycle because it allows direct comparison with an infinite unstructured population in the limit as migration rates increase. An alternative life cycle in which haploid migration was followed by syngamy and then by random sampling of diploid individuals was also studied. The results were qualitatively similar but do not reduce to the case of a single unstructured population as migration rates increase because there is always one generation of drift followed by selection within each deme, causing a small Hill-Robertson effect. At each locus, beneficial alleles start in linkage equilibrium ($x_{jk} = 0$ at $t = 0$) and sweep from a low initial frequency toward fixation.

Stochastic fluctuations around the deterministic trajectory: Because of drift, allele frequencies at both loci and linkage disequilibrium deviate from the trajectory they would follow in an infinite population (or deterministic trajectory). We denote this deterministic trajectory at time t by the vector $\mathbf{x}^* = \{x_j^*, x_k^*, x_{jk}^*\}$. Following Barton and Otto (2005), we focus on the deviations from $\mathbf{x}^*$, which occur in the presence of drift. We let $\mathbf{dx} = \{dx_j, dx_k, dx_{jk}\}$ describe the vector of these deviations. Thus, at any time t, the vector of allele frequencies and LD can be written as the sum of their deterministic values and the stochastic deviations, $\mathbf{x} = \mathbf{x}^* + \mathbf{dx}$. We begin by deriving a recursion for $\mathbf{dx}$ from one generation to the next along a given stochastic trajectory. We then compute the recursions for the expected deviations over all possible trajectories, $E[\mathbf{dx}]$. It is this expected deviation that is of greatest interest to us, as it describes the expected effect, over all possible stochastic outcomes, of drift and selection on allele frequencies and LD during selective sweeps.

Single population: To begin, we describe the case of a single deme ($n = 1$), where the deterministic trajectory is determined only by recombination and selection. For given values of parameters s_j, s_k, and r_{jk}, let us write the vector of recursions as $\mathbf{f} \equiv \{f_a\}_{a\in U} = \{f_j, f_k, f_{jk}\}$, which are three functions that determine the values of the allele frequencies (x_j, x_k) and linkage disequilibrium (x_{jk}) after selection and recombination (expressions for f_j, f_k, and f_{jk} are given in Equation A1 in appendix a). After one generation, the deterministic trajectory vector becomes $\mathbf{f}(\mathbf{x}^*)$. In an infinite population with no initial LD and no epistasis (as assumed here), x_{jk} remains zero, and each locus evolves independently. Consequently, the deterministic trajectory is described by the recursions obtained by setting x_{jk} to zero in $\mathbf{f}$,

$$\begin{aligned} x_j^{*\prime} &= f_j(\mathbf{x}) = x_j + \frac{s_j x_j(1-x_j)}{\phi_j} \\ x_k^{*\prime} &= f_k(\mathbf{x}) = x_k + \frac{s_k x_k(1-x_k)}{\phi_k} \\ x_{jk}^{*\prime} &= 0, \end{aligned} \tag{1}$$

where $\phi_j = 1 + s_j(x_j - \frac{1}{2})$ and $\phi_k = 1 + s_k(x_k - \frac{1}{2})$.

For a given trajectory of allele frequency and LD in a finite population, drift creates deviations from the deterministic trajectory that accumulate over time. At a given time t, the finite population is characterized by the vector $\mathbf{x} = \mathbf{x}^* + \mathbf{dx}$, where $\mathbf{x}^*$ is given by (1). The recursion for the stochastic trajectory $\mathbf{x}$ is similar to that for $\mathbf{x}^*$ except that drift occurs each generation. After selection and recombination, $\mathbf{x}$ becomes $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}^* + \mathbf{dx})$. Drift then occurs, which corresponds to multinomial sampling from the pool of four haplotypes after selection (this is true even in the diploid case under strict multiplicative selection). Sampling adds a random vector of perturbations $\boldsymbol{\zeta} = \{\zeta_j, \zeta_k, \zeta_{jk}\}$ to $\mathbf{f}(\mathbf{x})$. These perturbations are small as long as the population size is

large, and their moments can be found from the multinomial distribution.

After one generation in a finite population (*i.e.*, recombination, selection, and drift), the vector $\mathbf{x}$ becomes

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}^* + \mathbf{dx}) + \boldsymbol{\zeta} = \mathbf{f}(\mathbf{x}^*) + \mathbf{dx}', \tag{2}$$

where $\mathbf{f}(\mathbf{x}^*)$ is the value of the deterministic trajectory and $\mathbf{dx}' \equiv \{dx'_a\}_{a\in U}$ is the deviation from the deterministic trajectory in the next generation. From (2), we can write the recursion for deviations from the deterministic trajectory as

$$\mathbf{dx}' = \mathbf{f}(\mathbf{x}^* + \mathbf{dx}) - \mathbf{f}(\mathbf{x}^*) + \boldsymbol{\zeta} = \mathbf{dx}_s + \boldsymbol{\zeta}, \tag{3}$$

where $\mathbf{dx}_s = \mathbf{f}(\mathbf{x}^* + \mathbf{dx}) - \mathbf{f}(\mathbf{x}^*)$ represents the value of the vector of deviations after selection and meiosis (before drift).

Approximation in a large population: Assuming that populations are large enough that all deviations $\mathbf{dx}$ remain small (say, of order dx), we can obtain an approximate expression for $\mathbf{dx}'$ by performing a Taylor series expansion of (3) of $\mathbf{f}(\mathbf{x}^* + \mathbf{dx}) - \mathbf{f}(\mathbf{x}^*)$ around the deterministic trajectory $\mathbf{x}^*$ for each of the three recursions f_j, f_k, and f_{jk} in $\mathbf{f}$. Because the main effect of drift is to introduce variance around the deterministic trajectory, we must keep terms to second order in the deviations in the Taylor Series (see Barton and Otto 2005), yielding

$$dx'_a = \sum_{b\in U} \frac{\partial f_a}{\partial x_b}(\mathbf{x}^*)\,dx_b + \frac{1}{2}\sum_{(b,c)\in U} \frac{\partial^2 f_a}{\partial x_b \partial x_c}(\mathbf{x}^*)\,dx_b dx_c + \zeta_a + o(dx^2). \tag{4}$$

Because we need a compact notation before analyzing the case of multiple demes, we introduce a vector notation describing each of the deviation terms in (4). We have already described the vector $\mathbf{dx}$, whose terms occur in the first sum of (4). In addition, we require a vector describing the products of deviations, $dx_b dx_c$, which are $O(dx^2)$ terms. Ignoring the order in which the product is taken, there are six elements of this vector, corresponding to the deviations of a pair of variables (x_b, x_c) with $b \le c \in U^2$ (ordering the set U by $j < k < jk$). For convenience we refer to each of the six pairs of subscripts (b, c) as elements of the set

$$V \equiv \{(b,c)\}_{b\le c\in U^2} = \{(j,j),(j,k),(j,jk),(k,k),(k,jk),(jk,jk)\}. \tag{5}$$

The 1×6 vector of the products of deviations is thus defined by

$$\mathbf{dx^2} \equiv \{dx_b dx_c\}_{(b\le c)\in U^2} = \{dx_j^2, dx_j dx_k, dx_j dx_{jk}, dx_k^2, dx_k dx_{jk}, dx_{jk}^2\}. \tag{6}$$

We can then rewrite recursion (4) for the whole system using only matrix notation as

$$\mathbf{dx}' = \mathbf{D_1 dx} + \mathbf{D_2 dx^2} + \boldsymbol{\zeta} + o(\mathbf{dx^2}), \tag{7}$$

where $\mathbf{D_1}$ is the 3×3 matrix containing the partial derivatives for the first-order terms in the Taylor series ($\mathbf{D_1}$ represents the gradient of $\mathbf{f}$ at point $\mathbf{x}^*$), and where $\mathbf{D_2}$ is the 3×6 matrix containing the different coefficients of the second-order terms in the series. Matrices $\mathbf{D_1}$ and $\mathbf{D_2}$ are given explicitly in appendix a.

Because the recursions for the deviations $\mathbf{dx}$ depend on $\mathbf{dx^2}$, we must also describe recursions for $\mathbf{dx^2}$. These recursions are obtained by taking the expectation of the products of deviations after one generation, $dx'_a dx'_b$, $(a, b) \in V$, and approximating the result to second order in the deviations as in (4). The recursion for the expected value of $\mathbf{dx^2}$ after one generation (recombination, selection, and drift) can then be written as

$$\mathbf{dx^{2\prime}} = \mathbf{D_3 dx^2} + \boldsymbol{\zeta^2} + o(\mathbf{dx^2}), \tag{8}$$

where $\boldsymbol{\zeta^2} \equiv \{\zeta_{a,b}\}_{(a,b)\in V} = \{\zeta_a \zeta_b\}_{a\le b\in U^2}$ is defined in the same manner as $\mathbf{dx^2}$ in (6) but with the corresponding stochastic perturbation terms, and where $\mathbf{D_3}$ is the 6×6 matrix containing products of first partial derivatives of $\mathbf{f}$ at point $\mathbf{x}^*$ and is obtained by identification in a similar way as $\mathbf{D_1}$ and $\mathbf{D_2}$ ($\mathbf{D_3}$ is also given in appendix a).

Next, we give the distribution of the multinomial perturbation vector $\boldsymbol{\zeta}$ under the assumption of a large population size. In the following, we refer to $\mathbf{dx}$ and $\mathbf{dx^2}$ as first- and second-order moments of deviations.

Moments of the multinomial distribution: The exact expectations of the perturbations introduced by sampling, $E[\zeta_a]$ and $E[\zeta_a \zeta_b]$, are computed from the multinomial distribution and are given in appendix b. To order $1/2N$, the effect of sampling on first-order moments simplifies to $E[\boldsymbol{\zeta}] \simeq_{N\gg 1} \mathbf{0}$ (one round of drift produces negligible deviation). However, drift does produce variance in the deviations: the effect of sampling on second-order moments is, to order $1/2N$,

$$E[\boldsymbol{\zeta^2}] \underset{N\gg 1}{\simeq} \frac{1}{2N}\mathbf{c}, \tag{9}$$

where $\mathbf{c} = \{x_j^*(1-x_j^*),\ 0,\ 0,\ x_k^*(1-x_k^*),\ 0,\ x_j^*(1-x_j^*)\cdot x_k^*(1-x_k^*)\}$ is a 1×6 vector with nonzero terms equal to the genetic variances of x_j, x_k, and x_{jk}, evaluated along the deterministic trajectory. We use the same vector notation as the one defined for $\mathbf{dx^2}$ in (6).

Because we are interested in evaluating the expected trajectory for the different possible stochastic outcomes, we want to compute the expectations of $\mathbf{dx}'$ and $\mathbf{dx^{2\prime}}$, which are obtained by taking the expectations of recursions (7) and (8). Note that the elements in the three matrices $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{D_3}$ are partial derivatives of $\mathbf{f}$ evaluated along the deterministic trajectory $\mathbf{x}^*$; consequently, they are independent of $\mathbf{dx}$ and are not random variables. Thus, the recursion for the expected deviations and product of deviations over one generation is given by

$$E[\mathbf{dx}'] = \mathbf{D_1}E[\mathbf{dx}] + \mathbf{D_2}E[\mathbf{dx^2}] + o(1/2N) \quad (10a)$$

$$E[\mathbf{dx^2}'] = \mathbf{D_3}E[\mathbf{dx^2}] + \frac{\mathbf{c}}{2N} + o(1/2N). \quad (10b)$$

Recursion (10a) summarizes recursions (4a) and (4b) in Barton and Otto (2005), while recursion (10b) summarizes recursions (5a), (5b), and (5c) in Barton and Otto (2005).

As drift is the initial source of variation and introduces variances of order $1/2N$, recursions (10a) and (10b) are of order $1/2N$. As long as the stochastic perturbations in **dx** are small relative to the allele frequencies (*i.e.*, as long as alleles are not close to fixation), they can be approximated by a Gaussian distribution with mean and variance given by recursions (10a) and (10b), respectively (Barton and Otto 2005). These approximate recursions are valid for large populations (*i.e.*, for $1/2N$ small) and are not valid if the selective sweeps start from a very low allele frequency.

Production of negative linkage disequilibrium: Here we describe the development of the elements of $E[\mathbf{dx}]$, namely the expected deviation from the deterministic trajectory for allele frequencies ($E[dx_j]$ and $E[dx_k]$) and for the LD ($E[dx_{jk}]$). These are the quantities of greatest evolutionary relevance, because they describe the effects of drift on the spread of beneficial alleles and on linkage disequilibria. In particular, $E[dx_{jk}]$ determines the amount and sign of linkage disequilibrium within the population, because no LD is generated along the deterministic trajectory under multiplicative selection ($x^*_{jk} = 0$, see Equation 1).

By inspecting Equation 10, note that random genetic drift generates variance around the trajectory [the term $\mathbf{c}/2N$ in (10b)], but it does not directly bias the allele frequencies or LD [*i.e.*, drift does not contribute directly to (10a)]. Because $\mathbf{D_3}$ in (10b) contains only zero or positive terms, this variance is converted into positive covariances between deviations in allele frequencies and in LD by the action of selection. Because $\mathbf{D_2}$ in (10a) contains only zero or negative terms, however, this positive covariance between deviations causes, on average, negative deviations in the allele frequency trajectory as well as negative LD. A more detailed interpretation of this process can be found in Barton and Otto (2005).

In short, the interaction of drift and selection generates negative deviations, on average, for both the allele frequencies and LD, relative to their expected values in the absence of drift (deterministic trajectory). In other words, negative genetic associations build up among selected loci ($E[dx_{jk}] < 0$) and the selective sweep of beneficial alleles is delayed relative to the time course of selection in an infinite population ($E[dx_j] < 0$). Because the ultimate source of negative deviations is the variance introduced by drift, the expected deviations are inversely proportional to the population size N and become exceedingly small in very large populations. We now focus on how this process is modified in a subdivided population.

Subdivided population: *Fluctuations within demes around the deterministic trajectory:* We make the key assumption that selection is homogeneous in space so that no linkage disequilibria can be produced deterministically, as would be the case if selection coefficients at the selected loci covaried across demes (Lenormand and Otto 2000). We further assume that all demes start at linkage equilibrium and at the same allele frequencies. Consequently, the initial conditions and deterministic forces are homogeneous, so the deterministic trajectory $\mathbf{x}^*$ is the same for all demes at any time, and the only difference among demes is due to the stochastic deviations that build up during the selective sweeps occurring in different demes. This homogeneous deterministic trajectory $\mathbf{x}^*$ equals that of a single population, given by (1). Deviations will differ, however, from one deme to another. We denote $\mathbf{dx}[i] = \{dx_a[i]\}_{a\in U}$ as the vector of deviations from $\mathbf{x}^*$, along a given stochastic trajectory in deme *i*. Our aim is to compute the expected value of the vector of average deviations across all demes, which we denote as

$$\overline{\mathbf{dx}} \equiv \{\overline{dx}_a\}_{a\in U}, \quad \text{where } \overline{dx}_a = \frac{1}{n}\sum_{i=1}^{n} dx_a[i]. \quad (11)$$

For any variable or vector, we denote the mean taken across all demes with a bar.

As in the single-population model, we also need to compute the recursion for the mean of second-order moments taken across all demes. Using the notation introduced in (5) and (6), we define the vector of the second-order moments, averaged across demes, as

$$\overline{\mathbf{dx^2}} \equiv \{\overline{dx_a dx_b}\}_{a\le b\in U}, \quad (12)$$

where, for any couple of variables (x_a, x_b), $(a, b) \in V$, $\overline{dx_a dx_b} = (1/n)\sum_{i=1}^{n} dx_a[i]dx_b[i]$. In our calculations, we also need the product of the average deviations,

$$\overline{\mathbf{dx}}^{\mathbf{2}} \equiv \{\overline{dx}_a \overline{dx}_b\}_{a\le b\in U}, \quad (13)$$

where $\overline{dx}_a\overline{dx}_b = (1/n^2)\left(\sum_{i=1}^{n} dx_a[i]\right)\left(\sum_{i=1}^{n} dx_b[i]\right)$. For simplicity, we describe the three moments defined in (11), (12), and (13) as the first moment, the within-deme second moment, and the among-deme second moment, respectively.

As in the single-population model, we must compute the recursion over one generation for the expectation of the three moments ($\overline{\mathbf{dx}}$, $\overline{\mathbf{dx^2}}$, and $\overline{\mathbf{dx}}^{\mathbf{2}}$) taken across all the possible stochastic trajectories in each deme. To calculate these recursions, we first compute the joint effect of recombination, selection, and drift on the moments, using the results of the single-population model, and then we add the effect of migration. Finally, taking the expectation over all possible trajectories, we derive the recursion for the expected value of the

moments over a complete generation in a subdivided population.

Effect of recombination, selection, and drift on the moments in a subdivided population: Because meiosis, selection, and drift occur independently in each deme, the recursion for the deviation vector $\mathbf{dx}[i]$ in any deme i before migration occurs is similar to that given in the single-population model. Consequently, assuming that all demes are large enough, the recursion (7) describes the value of $\mathbf{dx}[i]$ along a given stochastic trajectory before migration,

$$\mathbf{dx}[i]' = \mathbf{D_1}\mathbf{dx}[i] + \mathbf{D_2}\mathbf{dx^2}[i] + \boldsymbol{\zeta}[i] + o(\mathbf{dx^2}), \qquad (14)$$

where $\mathbf{dx}^2[i]$ is the vector of products of deviations in deme i defined as in (6) and $\boldsymbol{\zeta}[i]$ is the perturbation introduced by sampling in deme i on the local vector of allele frequencies and LD. The coefficients in matrices $\mathbf{D_1}$ and $\mathbf{D_2}$ in (14) are evaluated along the deterministic trajectory, common to all demes. As a consequence, the recursion is the same for all demes. Using this fact, it is easy to deduce from (14) the value of the three moments after recombination, selection, and drift, following their definitions given in (11), (12), and (13). Taking the expectation over all possible trajectories, we obtain the expected value of the three moments before migration,

$$\begin{aligned} E[\overline{\mathbf{dx}}'] &= \mathbf{D_1}E[\overline{\mathbf{dx}}] + \mathbf{D_2}E[\overline{\mathbf{dx^2}}] + E[\overline{\boldsymbol{\zeta}}] + o(\mathbf{dx^2}) \\ E[\overline{\mathbf{dx^2}}'] &= \mathbf{D_3}E[\overline{\mathbf{dx^2}}] + E[\overline{\boldsymbol{\zeta}^2}] + o(\mathbf{dx^2}) \\ E[\overline{\mathbf{dx}}^{\mathbf{2}\prime}] &= \mathbf{D_3}E[\overline{\mathbf{dx}}^{\mathbf{2}}] + E[\overline{\boldsymbol{\zeta}}^{\mathbf{2}}] + o(\mathbf{dx^2}), \end{aligned} \qquad (15)$$

where $\overline{\boldsymbol{\zeta}} \equiv \{\overline{\zeta}_a\}_{a\in U}$ is the average across all demes, of the stochastic perturbations $\boldsymbol{\zeta}[i]$ that are introduced by drift in each deme i, $\overline{\boldsymbol{\zeta}^2} \equiv \{\overline{\zeta_a\zeta_b}\}_{(a,b)\in U^2}$ is the average of the products of these perturbations, and $\overline{\boldsymbol{\zeta}}^{\mathbf{2}} \equiv \{\overline{\zeta}_a\overline{\zeta}_b\}_{(a,b)\in U^2}$ is the product of average perturbations.

Moments of the perturbation vectors: We now compute the expectations for the effect of n independent multinomial samplings in the n demes on the three moments: $E[\overline{\boldsymbol{\zeta}}]$, $E[\overline{\boldsymbol{\zeta}^2}]$, and $E[\overline{\boldsymbol{\zeta}}^{\mathbf{2}}]$. The expectations of the sampling vectors $\boldsymbol{\zeta}[i]$ in a given deme i are computed from the position along the deterministic trajectory (common to all demes) and from the deme size $2N$. Because deme sizes are assumed to be large, we deduce from (9) that, to order $1/2N$, $E[\overline{\boldsymbol{\zeta}}] = \mathbf{0} + o(1/2N)$, and

$$E[\overline{\boldsymbol{\zeta}^2}] = \overline{E[\boldsymbol{\zeta}^2]} \underset{N\gg 1}{\simeq} \frac{1}{2N}\mathbf{c} + o(1/2N). \qquad (16)$$

The effect of drift on the within-deme second moment is thus inversely proportional to the local deme size $2N$, whether the demes are isolated or connected by migration. This point is important; it ensures that some stochasticity is present even in an infinite population, provided that the population is subdivided into demes of finite size. The among-deme second moments are the products of average deviations by themselves. As drift occurs independently in each deme, each random vector $\boldsymbol{\zeta}[i_1]$ is independent of $\boldsymbol{\zeta}[i_2]$ when $i_1 \neq i_2$. Using this independence and the fact that for any i_1, $E[\boldsymbol{\zeta}[i_1]] = \mathbf{0} + o(1/2N)$, we obtain, to order $1/2N$,

$$E[\overline{\boldsymbol{\zeta}}^2] = \frac{1}{n^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} E[\boldsymbol{\zeta}[i_1]\boldsymbol{\zeta}[i_2]] = \frac{1}{n^2}\sum_{i=1}^{n} E[\boldsymbol{\zeta}[i]^2] = \frac{1}{2nN}\mathbf{c}. \qquad (17)$$

Sampling has an equivalent effect on the among-deme second moments as it would have on the second moments of a single population of the same total size (*i.e.*, of size $2nN$). Consequently, the among-deme second moments will be much smaller than the within-deme second moments in a population composed of a large number of demes ($n \gg 1$).

Effect of migration on allele frequencies and linkage disequilibrium in the *n*-island model: We next give an exact recursion for the effect of migration on allele frequencies and linkage disequilibrium in an n-island model and the change by migration of the three moments defined in (11), (12), and (13). Details of the derivation are given in APPENDIX B.

We first note that the n-island model can be reduced to a two-island model. Indeed, migration changes haplotype frequencies within a deme, as if this focal deme exchanged migrants with a migrant pool at a rate $m_e = mn/(n-1)$. Consequently, the effect of migration on allelic frequencies and LD can be derived for any deme, using a two-demes recursion (see, *e.g.*, BARTON and GALE 1993) and the values of allele frequencies and LD in the migrant pool (see APPENDIX B). The recursion for the change in allele frequencies and LD averaged across demes (*i.e.*, on $\overline{\mathbf{x}} = \{\overline{x}_j, \overline{x}_k, \overline{x}_{jk}\}$) is given by

$$\delta_m[\overline{\mathbf{x}}] = \left\{ \begin{array}{c} 0 \\ 0 \\ m_e(2-m_e)\overline{\Delta_j\Delta_k} \end{array} \right\} = m_e(2-m_e)\overline{\Delta_j\Delta_k}\,\mathbf{u_{jk}}, \qquad (18)$$

where $\mathbf{u}_{jk} = \{0, 0, 1\}$ is the unit vector representing the linkage disequilibrium, and where $\overline{\Delta_j\Delta_k} = \overline{x_jx_k} - \overline{x}_j\overline{x}_k$ is the covariance between allele frequencies at loci j and k, taken across demes, *i.e.*, the spatial covariance between allele frequencies in the whole population. Equation 18 shows that migration (i) does not affect the metapopulation allele frequencies, as expected, and (ii) increases the average linkage disequilibrium per deme by a quantity $m_e(2-m_e)\overline{\Delta_j\Delta_k}$ each generation. Thus, migration transforms a proportion $m_e(2-m_e)$ of the spatial covariance between allele frequencies at loci j and k into local linkage disequilibrium between these loci.

We now use recursions for the effect of migration on local (B5) and average (18) allele frequencies and LD to compute the effect of migration on the deviation

moments given by (11), (12), and (13). Under homogeneous selection, the effect of selection and recombination on the deterministic trajectory is identical in all demes (differences between demes are only due to stochastic deviations). As a consequence, migration does not affect the deterministic trajectory and changes only the deviations $\mathbf{dx}[i]$ in each deme i ($\delta_m[x_a[i]] = \delta_m[dx_a[i]]$ for any variable a). Using this fact we directly obtain the effect of migration on the first moments $\overline{\mathbf{dx}}$,

$$\delta_m[\overline{\mathbf{dx}}] = m_e(2 - m_e)\overline{\Delta_j\Delta_k}\,\mathbf{u_{jk}}, \tag{19}$$

where $\overline{\Delta_j\Delta_k} = \overline{dx_j dx_k} - \overline{dx_j}\,\overline{dx_k}$. The effect of migration on each product of local deviations $dx_a[i]dx_b[i]$ in deme i is also computed from (B5). We then take the average of these products across demes to obtain the effect of migration on the within-deme second moments, $\overline{\mathbf{dx}^2}$. The resulting expression is simplified by dropping $O(dx^3)$ terms (large deme approximation). We then obtain

$$\delta_m[\overline{\mathbf{dx}^2}] = -m_e(2 - m_e)\overline{\mathbf{\Delta}^2} + o(\mathbf{dx}^2), \tag{20}$$

where $\overline{\mathbf{\Delta}^2} = \overline{\mathbf{dx}^2} - \overline{\mathbf{dx}}^2$ can be interpreted as the vector of spatial variances and covariances between all variables. Finally, we similarly compute the effect of migration on the among-deme second moments, using the product of migration effects on average deviations $\overline{dx_a}\,\overline{dx_b}$, which is given in (18). Migration has no or negligible effect, $o(dx^2)$, on this moment, for allele frequency and LD, respectively: $\delta_m[\overline{\mathbf{dx}}^2] = \mathbf{0} + o(\mathbf{dx}^2)$.

Recursions over one generation: We can now compute the recursion for the expected value of the three moments describing deviations in a population with a life cycle where migration occurs after selection, recombination, and drift. We obtain the overall changes by combining the changes on the three moments due to recombination and selection (15), drift [(16) and (17)], and migration [(19) and (20)]. We obtain a closed recursion system for the expected value of the three moments over one generation [dropping the $o(1/2N)$ for simplicity]:

$$E[\overline{\mathbf{dx}}'] = \mathbf{D_1}E[\overline{\mathbf{dx}}] + \mathbf{D_2}E[\overline{\mathbf{dx}^2}] + \frac{m_e(2 - m_e)}{(1 - m_e)^2}E[\overline{\Delta_j''\,\Delta_k''}]\mathbf{u_{jk}} \tag{21a}$$

$$E[\overline{\mathbf{dx}^2}'] = \mathbf{D_3}E[\overline{\mathbf{dx}^2}] + \frac{\mathbf{c}}{2N} - \frac{m_e(2 - m_e)}{(1 - m_e)^2}E[\overline{\mathbf{\Delta}^2}''] \tag{21b}$$

$$E[\overline{\mathbf{dx}^2}''] = \mathbf{D_3}E[\overline{\mathbf{dx}^2}] + \frac{\mathbf{c}}{2nN}. \tag{21c}$$

In (21b), the vector of spatial variances and covariances $E[\overline{\mathbf{\Delta}^2}] = E[\overline{\mathbf{dx}^2}] - E[\overline{\mathbf{dx}}^2]$ follows the recursion

$$E[\overline{\mathbf{\Delta}^2}'' = (1 - m_e)^2\left(\mathbf{D_3}E[\overline{\mathbf{\Delta}^2}] + \frac{\mathbf{c}}{2N}(1 - 1/n)\right) \tag{22}$$

over one generation. $E[\overline{\Delta_j''\Delta_k''}]$ in (21a) is the second element in this vector. Because selection occurs independently in each deme, the evolution of recombination depends on local LD between the selected loci. The expectation of this local LD is given by the third element of $E[\overline{\mathbf{dx}}]$ and its variance across demes is given by the sixth element in $E[\overline{\mathbf{\Delta}^2}]$. System (21) extends the single-population model (10) to a subdivided population for any migration rate, number of demes, recombination rate, and selection coefficients, provided that the demes remain large and alleles are not close to fixation.

Selection for recombination: To quantify how recombination evolves in response to the disequilibria generated by the Hill-Robertson effect, we introduce a third locus i modifying r, the recombination rate between the selected loci. As in Barton and Otto (2005), allele 1 at locus i corresponds to a higher recombination rate between loci j and k than allele 0. More precisely, genotypes $\{0, 0\}$,$\{1, 0\}$, and $\{1, 1\}$ at locus i correspond to recombination rates $r - dr$, r, and $r + dr$, respectively. We assume that the three loci are in the order i, j, k and that the recombination rate between i and j is R. We study the change in the frequency x_i at the modifier locus. To include this third locus, we have to keep track of four new variables in our vector recursions: the modifier allele frequency (x_i), the two-locus linkage disequilibria x_{ij} and x_{ik}, and the three-locus LD x_{ijk}. The recursions for the effect of recombination and selection on the seven variables (deterministic function $\mathbf{f}$) are given by Equations A2 in Barton and Otto (2005), for a weak modifier ($dr \ll r$).

Extending the method described above for a subdivided population to three loci and computing the effect of migration on the four new variables (see appendix b), we compute a new system of vector recursions that is similar to system (21) but with vectors and matrices of higher dimension (see appendix a), with qualitatively similar effects of migration (described in appendix b).

Comparison with exact simulations: Simulations were performed to check the analysis and to obtain results for small deme sizes. The simulations followed the same life cycle, using exact recursions for the effects of selection, random mating, meiosis, and migration on haplotype frequencies. Drift was simulated by multinomial sampling within each deme. To study the evolution of recombination, a recombination modifier (third locus) was included with the same effect as described above. We also performed simulations with a sex modifier, in which case individuals $\{0, 0\}$,$\{1, 0\}$, and $\{1, 1\}$ at locus i were supposed to have sex with probability σ_1, σ_2, and σ_3, respectively. We also introduced the possibility that individuals reproducing sexually produced, *e.g.*, half as many daughters as individuals reproducing asexually (*i.e.*, a twofold cost).

RESULTS

General effect of population subdivision: We now give a general interpretation of the effect of structure on

the system compared to the extreme cases: $m_e = 0$ (isolated demes) and $m_e = 1$ (panmictic population). The among-deme second moments [see (21c)] are equivalent to the second moments of deviations in a single population (10b) of size $2nN$ (the total size of the population); these moments can be interpreted as the variances and covariances of deviations in the migrant pool. The within-deme second moments [see (21b)] are also closely related to the second moments of deviations of a single population. Indeed, using (22), (21b) can be written

$$E[\overline{\mathbf{dx}^{2\prime\prime}}] = a_m^2\left(\mathbf{D_3}E[\overline{\mathbf{dx}^2}] + \frac{\mathbf{c}}{2N}\right) + (1 - a_m^2)\left(\mathbf{D_3}E[\overline{\mathbf{dx}}^2] + \frac{\mathbf{c}}{2nN}\right), \tag{23}$$

where $a_m = 1 - m_e$ ranges between 0 and 1 as m_e ranges between 0 and 1. Consequently, migration tends to buffer the variances and covariances of deviations produced locally [first term in (23)], bringing them closer to the lower variance produced in a population of total size $2nN$ [second term in (23)]. Consequently, $E[\overline{\mathbf{dx}^2}]$ ranges between the value expected for a single population of size $2N$ and that for a population of size $2nN$. Finally, the first moments (21a) are produced by local variances and covariances, $E[\mathbf{dx}^2]$, in the same way as in a single population (see 10a), except that migration also directly favors positive linkage disequilibrium by the admixture of populations with different allele frequencies [contributing the term $m_e(2 - m_e)E[\overline{\Delta_j\Delta_k}]$. Indeed, recursion (22) indicates that any element in $E[\overline{\mathbf{\Delta}^2}]$ (including $E[\Delta_j\Delta_k]$) is always positive (all the elements in $\mathbf{D_3}$ are positive). This is true when demes are large (*i.e.*, under our model's assumptions) because the selected alleles spread faster in those demes in which, by chance, positive disequilibrium arises. We see below that this result does not hold for small demes.

As a check, the results for a subdivided population converge upon the results for a single population for extreme values of the migration rate. When $m = m_e = 0$, recursions (21a) and (21b) reduce to recursions (10a) and (10b) for a single population of size $2N$. Conversely, when $m = 1 - 1/n$ ($m_e = 1$), recursions (21a) and (21b) reduce to recursions (10a) and (10b) for a single population of size $2Nn$.

Overall, in a subdivided population of any total size, but with large demes, migration always opposes the creation of negative linkage disequilibrium by drift in the presence of selection. This occurs because (i) the effect of drift is buffered locally by migration and (ii) migration is a direct source of positive LD by admixture. Nevertheless, neither of these effects tends to be large enough to alter the expectation that LD becomes negative.

Infinite subdivided population: Interestingly, the effects of drift do not disappear even in a population with infinite total size as long as the size of each deme ($2N$) is finite. Indeed, in the limit as n increases to infinity, the among-deme second moments, scaled to $1/2nN$ (see 21c), become negligible, so that $E[\overline{\mathbf{dx}}^2] = E[\overline{\mathbf{\Delta}^2}]$ and recursion (21) simplifies to

$$E[\overline{\mathbf{dx}}''] = \mathbf{D_1}E[\overline{\mathbf{dx}}] + \mathbf{D_2}E[\overline{\mathbf{dx}^2}] + \frac{m(2-m)}{(1-m)^2}E[\overline{dx_j'' dx_k''}]\mathbf{u_{jk}} \tag{24a}$$

$$\mathrm{E}[\overline{\mathbf{dx}^{2\prime\prime}}] = (1 - m)^2\left(\mathbf{D_3}E[\overline{\mathbf{dx}^2}] + \frac{\mathbf{c}}{2N}\right). \tag{24b}$$

Thus, even in an infinite metapopulation, some variance in deviations is produced by drift within each deme, which causes the expected average disequilibrium across demes to become negative [because of the negative elements of $\mathbf{D_2}$ in (21b)].

We illustrate this result in Figure 1, where we give the maximum absolute value of the average LD per deme in a very large population ($2nN \geq 5 \times 10^6$) under weak selection, for various migration and recombination rates. Figure 1 shows that even in a weakly structured population ($Nm \geq 1$), a substantial LD can build up in a very large population to a level similar to that expected if the demes were completely isolated as long as gene flow is not too strong (note the steep decrease for large m). Figure 1 also illustrates that, as in a panmictic population, the LD is very low when linkage is loose. Note also that results illustrated in Figure 1 are a lower bound for

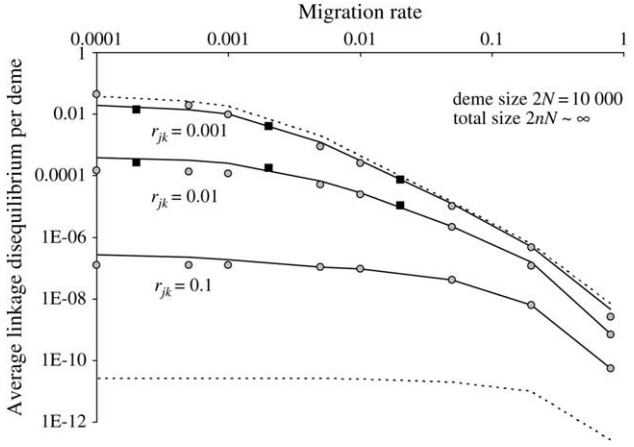


FIGURE 1.—Log-log plot of the maximum absolute value of the average LD between the selected loci, that is reached during the sweeps, for varying migration rate (x-axis) and for three recombination rates (indicated on the graph). The values were obtained from iteration of recursion (21) (lines), from simulations (squares), or from the weak selection approximation (28) (shaded circles). Dashed curves indicate the two limits for the recombination rates $r = 0$ (top line) and $r = \frac{1}{2}$ (bottom line). For each value of m and r, the number of demes n was chosen large enough (always >500) that further increasing n had very little effect on the result (infinite population limit). Simulation results were averaged over 1800 ($m = 0.0002$ and 0.002) and 10,000 ($m = 0.02$) sweeps. Other parameters are $s_j = s_k = 0.005$, with an initial beneficial allele frequency of 0.1 at both loci and deme size $2N = 10{,}000$.

the LD produced when the number of demes is smaller or when selection is stronger.

LD under weak selection: *Loose linkage:* When the processes that reduce linkage disequilibrium are large relative to the processes generating disequilibrium, it is possible to derive an analytical solution for the steady-state level of linkage disequilibrium. This steady-state level depends on the current allele frequencies and is known as a "quasi-linkage equilibrium" (QLE) (Barton and Turelli 1991). When recombination rates are large relative to selection and drift [$r \gg 1/(2N)$, s_j, s_k], QLE values can be determined, from (21b) and (21c), for the variance in LD within demes, $E[\overline{dx_{jk}^2}]$, and among demes, $E[\overline{dx}_{jk}^2]$, as well as for the covariances between LD and allele frequencies within demes, $E[\overline{dx_{jk}\,dx_j}]$, and among demes, $E[\overline{dx}_{jk}\overline{dx}_j]$ (details not shown). For the spatial variances and covariances between allele frequencies, $E[\overline{\Delta_j\Delta_k}]$, to reach a steady state, however, it is also required that the migration rate be large relative to selection and drift [$m_e \gg 1/(2N)$, s_j, s_k], as appears from recursion (22). All spatial covariances in $E[\overline{\mathbf{\Delta}^2}]$ are produced by drift and selection within demes and are reduced by migration. Assuming that the migration rate is large enough, the equilibrium value of these covariances, noted $E[\hat{\mathbf{\Delta}}^{\mathbf{2}}]$, can be obtained by solving the matrix equation $E[\mathbf{\Delta}^{2\prime\prime}] = E[\mathbf{\Delta}^2]$ and using recursion (22).

Once the QLE values have been calculated for the second moments, the steady-state level of linkage disequilibrium can be determined from (21a) by setting $E[\overline{dx}''_{jk}] = E[\overline{dx}_{jk}]$. To denote this QLE approximation in a population subdivided into n demes of size $2N$, we use a hat, $E[\widehat{dx}_{jk}]_{2N,n}$. For a single unstructured population of size $2N$, Barton and Otto (2005) found that $E[\widehat{dx}_{jk}]_{2N} = -2s_j s_k x_j(1-x_j)x_k(1-x_k)(1-r)/(2Nr^3)$. In a structured population, we find that the linkage disequilibrium falls between the expected LD in a single population of size $2N$ (the size of the deme) and of size $2nN$ (the total size of the population) and can be written as

$$E[\widehat{dx}_{jk}]_{2N,n} = (\alpha E[\widehat{dx}_{jk}]_{2N} + (1-\alpha)E[\widehat{dx}_{jk}]_{2nN}), \quad (25)$$

where

$$0 \le \alpha \equiv a_m^2 \frac{(1-a_r)^2(a_r - 1/2)\,(1+a_m^2 a_r)}{a_r(1-a_m^2 a_r)\,(1-a_m^2 a_r^2)} < 1, \quad (26)$$

where $a_m = 1 - m_e$ as in (23) and $a_r = 1 - r$. As the migration rate increases, α decreases, and the linkage disequilibrium becomes increasingly similar to that expected in a single unstructured population of size $2nN$.

Using (26), we can define a QLE population size, N_{QLE}, according to the population size of an unstructured population that leads to the same expected amount of linkage disequilibrium as that in a structured population. From (25), this equivalent population size is

$$N_{\mathrm{QLE}} = \frac{nN}{1+(n-1)\alpha} \underset{n\to\infty}{\to} \frac{N}{\alpha}. \quad (27)$$

Infinite population: In a population with a very large number of demes, the within-demes and between-demes variances and covariances are equal, $E[\overline{\mathbf{dx}^2}] = E[\overline{\mathbf{\Delta}^2}]$ [see (24b)]. Assuming that migration is strong enough, these variances and covariances reach an equilibrium $E[\hat{\mathbf{\Delta}}^{\mathbf{2}}]$. It is then possible, for any recombination rate r, to solve the differential equation for $E[\overline{dx}_{jk}]$ by a continuous time approximation (*i.e.*, under weak selection), using the method presented in Barton and Otto (2005, Equations B4a and B4b). We obtain the average LD per deme after t generations of the selective sweep,

$$E[\overline{dx}_{jk}]_{2N,\infty} = \alpha E[\widehat{dx}_{jk}]_{2N}(1 - e^{-rt}), \quad (28)$$

where α is defined above in (26) and $E[d\hat{x}_{jk}]_{2N}$ is defined above as the QLE for a panmictic population of the size of the deme ($2N$). This approximation makes no assumptions on the recombination rate provided that the population has a very large number of demes. As with the QLE approximation in an infinite population (27), this approximation corresponds to the LD produced in a single panmictic population of a finite size $2N/\alpha$. The agreement between this approximation and both simulations and recursion (21) is illustrated in Figure 1. The approximation is less accurate with very low migration ($m \le 0.0001$) when the weak structure assumption is no longer met.

QLE for the modifier frequency: Using the three-locus version of recursion (21), we can compute the expected change in the frequency of a modifier at QLE, assuming that migration and recombination rates are large relative to selection and drift (Figure 2). The result is a complicated function of the parameters describing the population (m, n, $2N$) and the genetic map (r and R).

When there is no migration among demes, the predicted change in the modifier collapses down to the results presented in Barton and Otto (2005) for a single unstructured population. With migration, we present results for the special case in which the loci are equidistant ($R = r$). When migration is weak [but still assuming that m, $r \ll 1$, $r \gg 1/(2N)$, s_j, s_k], the predicted change in the modifier at QLE is to leading order in m and r:

$$E[dx_i] \simeq \frac{dr\,s_j^2 s_k^2 x_i(1-x_i)x_j(1-x_j)x_k(1-x_k)}{r^3(r+m_e)(r+2m_e)^2(3r+2m_e)^2} \times \left(\frac{(1-m_e)^2(48m_e^4 + 264m_e^3 r + 534m_e^2 r^2 + 455m_e r^3 + 134r^4)}{16N(r+m_e)} + \frac{30m_e^5 + 149m_e^4 r + 281m_e^3 r^2 + 240m_e^2 r^3 + 87m_e r^4 + 8r^5}{Nnr^2} \right). \quad (29)$$

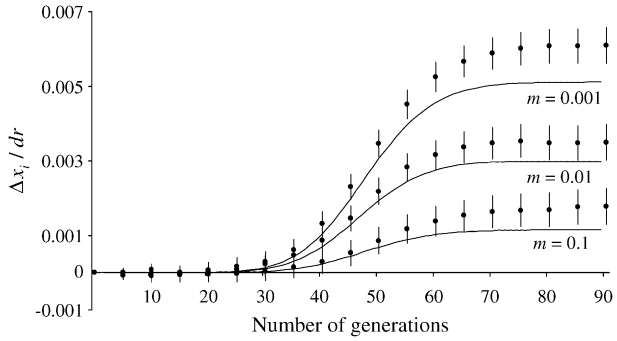


FIGURE 2.—Value of the average modifier frequency change over time Δx_i, scaled to the modifier effect ($dr = 0.03$) for three values of the migration rate m indicated on the graph. Lines indicate the values obtained from recursion (21) for three loci and dots indicate the results of simulations with 95% confidence intervals averaged over 10^7 sweeps. Other parameters are $n = 5$, $2N = 10{,}000$, $r_{jk} = r_{ij} = s_j = s_k = 0.1$, an initial beneficial allele frequency of 0.01, and an initial modifier frequency $x_i = 0.5$.

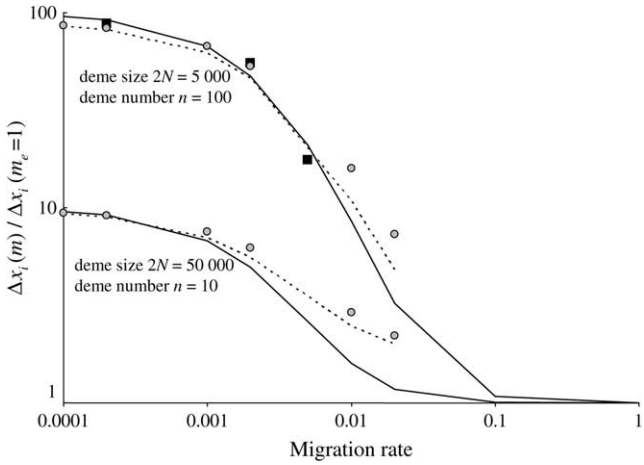


FIGURE 3.—Ratio between the cumulative modifier frequency change, over the selective sweeps, in a structured population, $\Delta x_i(m)$, and the same frequency change in the absence of structure, $\Delta x_i(m_e = 1)$, for different values of the migration rate m (x-axis, on log scale). The total population size is kept constant $2nN = 500{,}000$ with either 100 demes of size $2N = 5000$ (top curve) or 10 demes of size $2N = 50{,}000$ (bottom curve). The values are obtained with iteration of recursion (21) (solid lines), with the QLE approximation (29) (dashed lines), or with the single-population QLE approximation (29) with a population size $2N_{QLE}$ given in (27) (shaded circles). Simulation results averaged over 10,000 sweeps are indicated for the case $2N = 5000$ (solid squares). Other parameters are $s_j = s_k = r_{ij} = r_{jk} = 0.01$, with an initial beneficial allele frequency of 0.1, and a modifier effect $dr = 0.005$ with initial frequency $x_i = 0.5$.

At the other extreme, in an unstructured population ($m_e = 1$) with equidistant loci, we retrieve the result presented in Equation 7a of BARTON and OTTO (2005) for a population of total size $2Nn$:

$$E[dx_i] \simeq \frac{1.868\,dr\,s_j^2 s_k^2 x_i(1-x_i)x_j(1-x_j)x_k(1-x_k)}{Nnr^5}. \quad (30)$$

Integrating the QLE frequency change over the selective sweeps yields the cumulative frequency change and the average per-generation selection coefficient at the modifier locus. Indeed, because the beneficial alleles rise from an initial frequency p_0 to fixation, the cumulative frequency change is obtained by integrating $x_j(1-x_j)x_k(1-x_k)$ over time, yielding $(1-p_0)^2(1+2p_0)/6s$.

Figure 3 shows that in a subdivided population with large demes, the frequency change at the modifier locus can be orders of magnitude larger than that in the corresponding panmictic population even for Nm values >1. Figure 3 also illustrates that the QLE approximation captures this behavior under weak selection. As might be expected intuitively, Equation 30 with Nn replaced by N_{QLE} (27) also provides a reasonable approximation for the frequency change at the modifier locus, although it is less accurate than (29) (see Figure 3). However, these approximations work best in a parameter range where the selection for recombination is weak (for instance, the maximum selection illustrated in Figure 3 is $0.001\,dr$).

Overall, we observe similar properties for the rate of change of an allele modifying recombination rates and for the linkage disequilibrium in a subdivided population. In both cases, the predictions fall between those expected in an undivided population whose size is that of the deme ($2N$) and those in a panmictic population of size $2nN$. Furthermore, both the change in the modifier and the linkage disequilibria can be substantial in large, even infinitely large, populations, as long as the population is sufficiently structured.

Smaller deme sizes: For smaller deme sizes, the deviations from the deterministic trajectory can no longer be assumed small, and our analysis breaks down. We thus turned to simulations to study the development of LD and selection for recombination. We used the same simulations as presented above and each simulation was run until the polymorphism was lost at both selected loci or at the modifier locus (so that no further change in the frequency of the modifier could be expected). For a large population ($2nN = 100{,}000$), the effect of deme size on the average per-generation selection coefficient for recombination (scaled to the modifier effect) is illustrated in Figure 4. With realistic values of selection coefficients ($s = 0.1$) and tight linkage ($r = 0.01$), the selection coefficient for recombination can be substantial (of the order of $0.1\,dr$). It also shows that our model is a good approximation as long as the deme size $2N$ is not less than a few thousand. Indeed with smaller demes, the beneficial alleles are often lost temporarily from a deme due to drift, which reduces the amount of local LD. In this context, a small amount of migration favors negative linkage disequilibria directly by admixture (as appeared in

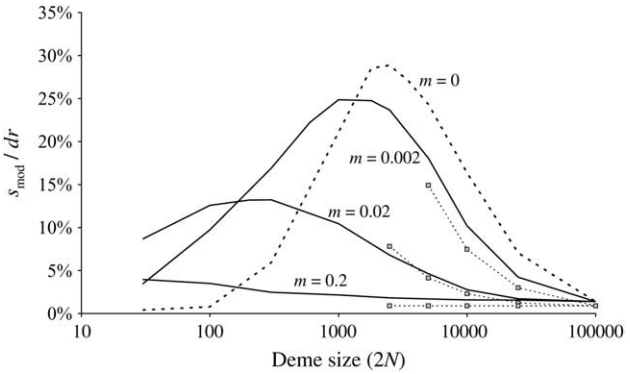


FIGURE 4.—Effect of population structure on the per generation selection coefficient on the recombination modifier, s_{mod}, averaged over the t generations of the selective sweep and scaled by the modifier effect dr. $s_{mod} = \Delta x_i(t)/(tx_i(1 - x_i))$, where $\Delta x_i(t)$ is the average cumulative modifier frequency change over the selective sweep. The value is given for different deme sizes $2N$ (x-axis) and migration rates (indicated). Lines show simulation results and dots indicate the prediction from recursion (21) iterated over $t = 100$ generations (the expected time taken by the selective sweep along the deterministic trajectory). Other parameters are $s_j = s_k = 0.1$, $r_{ij} = r_{jk} = 0.01$, with an initial beneficial allele frequency of 0.02, and a modifier effect $dr = 0.005$ with initial frequency $x_i = 0.5$.

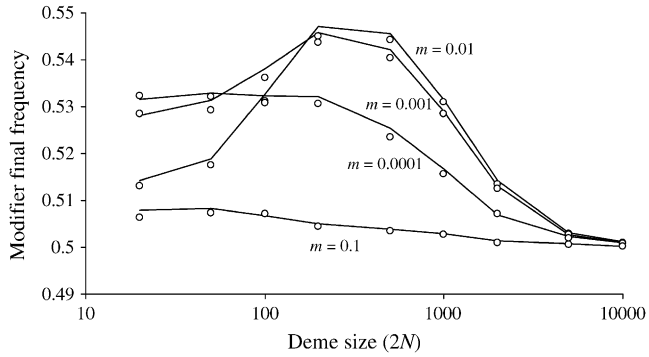


FIGURE 5.—Effect of population structure on modifier final frequency at the end of the sweeps (the initial frequency is 0.5) for different deme sizes $2N$ (x-axis) and migration rates (indicated) under strong selection ($s_j = s_k = 1$). The total population size is kept constant, $2nN = 10{,}000$. Lines correspond to a sex modifier with the probability to reproduce sexually (with recombination rate set to $\frac{1}{2}$) $\sigma_1 = 0.02$, $\sigma_2 = 0.03$, and $\sigma_3 = 0.04$ for individuals carrying zero, one, or two copies of the modifier, respectively. Dots correspond to a recombination modifier with $dr = 0.005$ and $r_{ij} = r_{jk} = 0.015$. Initial frequency of selected alleles is 0.01.

simulations, not shown), because it restores polymorphism to individual demes. This can be interpreted more precisely using recursion (18) because this recursion makes no assumption on the deme size so that the average amount of LD per deme produced by admixture is always $m_e(2 - m_e)E[\overline{\Delta_j \Delta_k}]$, even in small demes. When migration is infrequent and demes are small enough that alleles can be locally lost, the Hill-Robertson effect within each deme makes it more likely that the beneficial allele at one locus is lost while the beneficial allele at the other remains, particularly when the selection coefficients at each locus are of the same order. This generates a negative $E[\Delta_j \Delta_k]$, so that contrary to the large demes case, the effect of admixture, when the population structure is substantial, is to favor negative LD. Overall, the LD produced in a population subdivided into small demes is maximum for an intermediate rate of migration, whereas it is maximum for $m = 0$ when demes are large. These results are illustrated in Figure 4 (compare deme size above or <1000). When considering small demes, selection for recombination is more efficient in a subdivided population than it would be if demes were either isolated or completely connected.

Sex modifiers: We also performed simulations in which the locus i was a sex modifier. Figure 5 illustrates the effect of population structure as above but with strong selection. Figure 5 also shows that a sex or a recombination modifier has the same behavior. In Figure 6, we also show how the LD generated by the Hill-Robertson effect in a subdivided population selects for increased sex/recombination at a level sufficient to overcome the twofold cost of sex. Note in Figure 6 that increased sex would not be favored in the absence of structure ($m_e = 1$). These conclusions hold only for a weak modifier effect under very strong selection

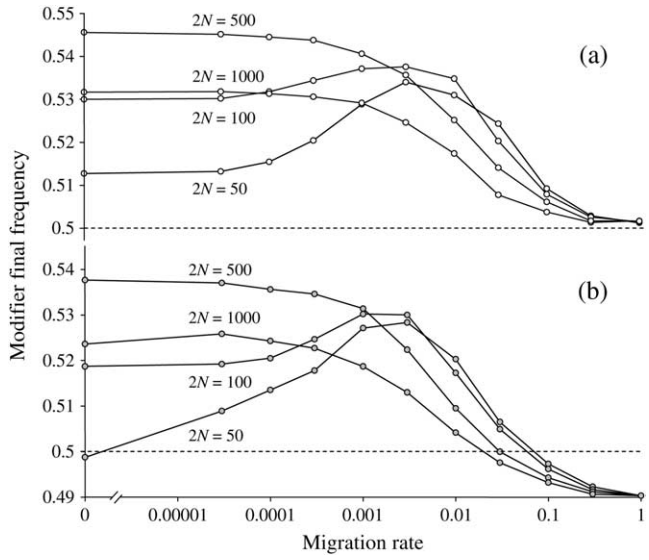


FIGURE 6.—Effect of population structure on a sex modifier final frequency at the end of the sweeps (the initial frequency is 0.5) for different deme sizes $2N$ (indicated) and migration rates (x-axis, note the log-scale and the value for $m = 0$) under strong selection ($s_j = s_k = 1$). The total population size is kept constant, $2nN = 10{,}000$. As in Figure 5, the probability to reproduce sexually (with recombination rate set to $\frac{1}{2}$) is $\sigma_1 = 0.02$, $\sigma_2 = 0.03$, and $\sigma_3 = 0.04$ for individuals carrying zero, one, or two copies of the modifier, respectively. In a, there is no cost of sex whereas in b individuals who reproduce sexually produce half as many daughters compared to individuals reproducing asexually (twofold cost). Initial frequency of selected alleles is 0.01.

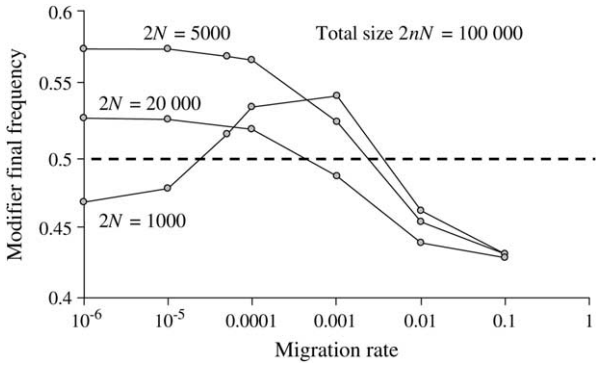


FIGURE 7.—Effect of population structure on a sex modifier final frequency at the end of the sweeps. The same as that in Figure 6b is shown (with a twofold cost of sex) but for asexuals *vs.* weakly sexual organisms ($\sigma_1 = 0$, $\sigma_2 = 0.01$, and $\sigma_3 = 0.02$), weaker selection ($s_j = s_k = 0.1$), and larger total population size ($2nN = 100{,}000$).

($s_j = s_k = 1$). Indeed, with $s_j = s_k = 0.1$ and the same parameters as those in Figure 6, a modifier increasing sex does not invade but is less disfavored for intermediate levels of population structure (not shown). However, a weak sex modifier can overcome the twofold cost and invade a structured asexual population of large total size ($2nN = 100{,}000$) under weaker selection (for instance, with $s = 0.1$, see Figure 7) under a wide range of population structures.

DISCUSSION

Drift influences the response to selection of a set of linked loci in a way that is not predicted by the dynamics of each locus considered separately. The interaction of drift and selection tends to build up negative associations between favorable alleles at linked loci (*i.e.*, negative linkage disequilibria), a process known as the HRE. This process generates negative LD in the absence of epistasis, but the latter will also contribute to the development of LD. The negative linkage disequilibrium created by drift in the presence of selection causes beneficial alleles to be associated with deleterious alleles at other loci. Metaphorically, negative LD stores genetic variance in fitness by "hiding" good alleles on bad genetic backgrounds. This variance is restored by the action of recombination. Consequently, modifiers of recombination increase in frequency because they help regenerate good combinations of alleles and rise in frequency along with these combinations.

This stochastic advantage to recombination works in a single population but its magnitude decreases with population size and vanishes when population size tends toward infinity (BARTON and OTTO 2005). Thus, the stochastic theory for the evolution of sex provides a poor explanation for the maintenance of sex in species with large and unstructured populations. The stochastic theory for sex also fails at the other extreme, in very small populations, because several mutations are unlikely to segregate simultaneously in small populations. Of course, real populations are spatially structured to some extent, and thus we set out to determine the effect of population structure on the stochastic theory for sex. We reached two main conclusions: (i) substantial linkage disequilibrium and selection for recombination can occur in a large—even infinite—population provided that it is subdivided, and (ii) substantial linkage disequilibrium and selection for recombination can also occur in very small demes that are connected by migration, because the polymorphism at the selected loci is maintained at the metapopulation level.

Linkage disequilibria generated by drift with large demes: When the subpopulations (demes) are large, the LD generated by drift with selection in a subdivided population falls between the LD expected in an isolated deme of size $2N$ (when $m = 0$) and that in a single deme of size $2nN$ (when $m = 1$). More precisely, when selection is weak relative to migration and recombination, the average LD per deme equals that expected in an unstructured population of size N_{QLE}, with

$$N_{\mathrm{QLE}} = \frac{nN}{1 + (n-1)\alpha} \underset{n\to\infty}{\to} \frac{N}{\alpha},$$

where α varies between 0 and 1 [see (26)]. This result holds for any recombination rate in a very large metapopulation [see (28)] and remains accurate for small m-values (see Figure 1).

This apparently simple result summarizes a complex underlying process. In a subdivided population, we can distinguish two sources of LD. The first source of LD is drift with selection, which produces a negative LD on average, as in a single population of size $2N$. This process relies upon the creation of variance in LD, which is produced within each deme by drift but is destroyed by migration among demes (see Equation 21b). These antagonistic effects imply that a subdivided population exhibits an intermediate level of variance in LD between a completely structured ($m_e = 0$) and an unstructured population ($m_e = 1$). The second source of disequilibrium is migration (admixture), which favors positive LD whenever allele frequencies covary positively across demes (Equation 21a). Indeed, with large demes, we expect this spatial covariance to be positive, because both alleles sweep faster than average in those demes with a positive LD and both alleles sweep slower than average in those demes with a negative LD. Consequently, allele frequencies covary positively across demes, and admixture generates a small amount of positive disequilibrium. However, when selection is homogeneous over space, the effect of admixture is small because it relies on the variance among demes in LD. Indeed, our results indicate that when migration is weak, the variance among demes in LD may be large but a small proportion is converted into positive LD, while

when migration is strong, this variance is reduced. Overall, for any level of structure, admixture never overwhelms the Hill-Robertson effect, so that the linkage disequilibrium remains negative, on average.

Overall, the net effect of these two sources of linkage disequilibrium (*i.e.*, drift with selection and admixture) is a negative linkage disequilibrium with a value intermediate between that expected in an unstructured population of size $2N$ and that of size $2nN$. With large demes, gene flow limits the production of negative disequilibrium (i) because the effect of drift is buffered locally by migration and (ii) because migration is a direct source of positive LD through admixture. As in a single population, the resulting LD is larger when selection is strong and equal at both loci and when linkage is tight [in an infinite metapopulation (see 28) or at QLE (see 25) in a finite metapopulation, x_{jk} is proportional to $s_j s_k / r^3$].

Linkage disequilibria generated by drift with small demes: In a structured population with small demes, the simultaneous fixation of beneficial alleles at both loci is impeded by drift, and one of the beneficial alleles is often lost locally, at least transiently. Therefore, in small isolated populations, there is less scope for the Hill-Robertson effect because of a lack of polymorphism. However, our simulations revealed that a small amount of gene flow among subpopulations is enough to restore the polymorphism at selected loci and allow the Hill-Robertson effect to occur. Indeed, the amount of linkage disequilibrium generated is much higher when small demes are connected by migration than when they are not (see Figure 6, left side). In contrast to the case of large demes, allele frequencies often covary negatively across demes, because the spread of beneficial alleles interferes with the local fixation of other beneficial alleles (Hill and Robertson 1966). Consequently, admixture can itself generate negative disequilibrium, causing the average LD per deme to be larger than that expected either for a set of isolated populations of small size or for a large unstructured population.

Evolution of recombination in a subdivided population: In the absence of epistasis, increased recombination is favored only if selected loci are negatively associated. Our model shows that, as in a single population, a Hill-Robertson effect occurs in subdivided populations, which generates negative LD and therefore selects for higher rates of recombination. Our results show that population structure has qualitatively similar effects on the frequency of alleles that increase sex or recombination and on the LD between selected loci. When demes are large, the frequency change at the modifier locus is intermediate between the value expected in a single population of size $2N$ and that in an unstructured population of size $2nN$, whereas when demes are small (simulation results), the frequency change at the sex or recombination modifier locus is larger than expected for both migration limits. Unlike a single unstructured population, the genetic associations generated by drift with selection do not vanish when the total population size gets very large or when local deme size gets very small. As a consequence, selection for sex and recombination is effective in a structured population under a broad range of conditions.

Limits of the approach and perspective: The analytic model developed by Barton and Otto (2005) and extended here to structured populations assumes that within-deme drift is weak enough that beneficial alleles sweep at both loci (*i.e.*, that $Ns \gg 1$). Furthermore, both our simulations and analysis assume that initial beneficial allele frequencies are relatively high ($p_0 \geq 1\%$) and do not vary substantially among demes. These conditions might be met in a weakly structured population undergoing an environmental change with selection on standing variation. In addition, our approximations assume that (i) selection is not too strong relative to migration (and relative to recombination in the case of the QLE approximation) and that (ii) the effect of the modifier on the recombination rate is weak. More theoretical work is needed to relax these assumptions and describe the full spectrum of effects that population structure can have on the development of disequilibria, the spread and fixation of beneficial alleles, and the evolution of recombination. In addition to being often subdivided, natural populations may also experience heterogeneous selection across habitats or epistatic selection across loci. Both factors can generate LD and influence the evolution of recombination (Lenormand and Otto 2000). Including weak epistasis in our model should be possible [by modifying function **f** in (1)], but modeling heterogeneous selection might be complicated by the fact that we consider constant deterministic trajectories across demes. In any case, the interaction of these factors with population structure remains to be fully explored.

Implications for the theories of the evolution of sex and empirical tests: Our results suggest that drift could be an important factor favoring sexual reproduction, even in infinite populations, provided that these populations are subdivided into demes of finite size. This could be relatively common in natural populations, which almost always exhibit some level of structure. However, the advantage of sex or recombination due to the HRE can be weak if selection is too weak and the linkage is not tight enough. Consequently, particularly when some cost of sex is included (*e.g.*, increased duration of cell division in isogamous species or twofold cost in anisogamous species), sex/recombination may evolve only in populations of intermediate size and under rapid environmental change with strong selection (Otto and Barton 2001; Otto and Lenormand 2002). We showed that in a metapopulation, a weak sex modifier can overcome the twofold cost over a much broader range of population size and for weaker—but still substantial—selection. Similarly, population structure can increase the advantage of segregation (Agrawal and Chasnov 2001; Otto 2003) and contribute to the maintenance of sexual reproduction. Whether the rate of environmental

change and the strength of selection are sufficient in nature for beneficial mutations to drive the evolution of sex remains, however, an open question. Nevertheless, all experimental evidence demonstrating an advantage to recombination relied on either strong artificial selection (see Otto and Lenormand 2002) or abrupt environmental change (Colegrave 2002; Goddard *et al.* 2005).

Most of the benefit of recombination is gained by a modest amount of sex whereas the twofold cost of sex is proportional to the rate of sex. As a consequence, the evolution of high rates of sex remains difficult to explain. Our results show that given substantial directional multiplicative selection and population structure, a low rate of sex (with the twofold cost) is stable against complete asexuality even in very large populations. The evolution of higher rates of sex seems unlikely in our two-locus study. However, when considering the evolution of sex *vs.* asex instead of recombination (*i.e.*, when a twofold cost applies), modeling many loci is particularly important as a sex modifier changes recombination rates over the whole genome. As suggested by Barton and Otto (2005), our model could be extended to several loci by summing over pairwise LD. The general matrix recursions [(21) and (24)] that describe the interplay of drift and migration on metapopulation moments should remain unchanged in this context. Although they did not consider a twofold cost, simulations by Iles *et al.* (2003) showed that adding more loci for a given additive fitness variance resulted in a greater advantage to recombination in a panmictic finite population and in a larger range of population sizes where this advantage is substantial. More work is needed to determine quantitatively the magnitude of the HRE in structured populations with numerous loci and to determine the amount of sex and recombination that is ultimately favored.

An empirical prediction from our analysis is that there should be a positive correlation between levels of population structure and recombination rates. However, using the usual F_{st} to measure population structure may be misleading. As shown in our model, the effect of structure on linkage disequilibria and on selection for recombination is not simply determined by F_{st} [see (26), (27), and (29)] but depends in a complicated way on recombination rates, migration rates, and the number and size of demes. Moreover, the power of this approach is weakened by the fact that species will differ in their genomic maps, their history of selection, and their total population size.

Our analysis also predicts how linkage disequilibria should vary across a genome in the presence of selection and drift but in the absence of epistasis. In a weakly structured population, with weak multiplicative selection and loose linkage, using the QLE approximations (26), we can find a simple relationship between F_{st}, average LD, and the spatial covariance between allele frequencies $\overline{\Delta_j\Delta_k}$ for any pair of genes separated by r recombination units:

$$\overline{x}_{jk} \underset{n\to\infty}{\to} -\frac{\overline{\Delta_j\Delta_k}}{F_{st}}\frac{1-2r}{2Nr}.$$

Keeping in mind the various assumptions made in the QLE analysis, there should be a linear relationship between LD and $\overline{\Delta_j\Delta_k}(\frac{1}{2}-r)/r$ measured for different pairs of loci if the Hill-Robertson effect is an important mechanism shaping the disequilibria. This prediction has the nice property that it does not depend on the strength of selection, because $\overline{\Delta_j\Delta_k}$is measured, not estimated. However, this spatial covariance might often be too small ($\ll F_{st}$) to be correctly measured and one needs to know which allele is favored at each locus. This prediction illustrates that the effect of the HRE on LD may be more readily detected in a structured than in a panmictic population.

Summary: In this article we develop explicit recursions for the effect of drift, selection, and migration in a three-locus system under the island model. These recursions allow us to quantify the effect of structure on the production of linkage disequilibrium between two selected loci by drift in the presence of selection (the Hill-Robertson effect) when deme size is large. We find that, on average, negative disequilibria develop among selected loci. Because of these negative associations among favored alleles, modifier alleles that increase the rate of recombination spread. The rate of this spread is much more substantial in a structured population, contributing to a plausible explanation for why sex and recombination are so ubiquitous.

The authors gratefully acknowledge Nick Barton, Sylvain Billiard, and two anonymous reviewers for helpful comments on the manuscript. This work was supported by grant Action Concertée Incitative jeune chercheur (no. 0693) from the French ministry of research to T.L. and by a National Sciences and Engineering Research Council grant from Canada and a poste rouge from Centre National de la Recherche Scientifique to S.P.O. G.M. benefited from a fellowship from the French ministry of research.

LITERATURE CITED

Agrawal, A. F., and J. R. Chasnov, 2001 Recessive mutations and the maintenance of sex in structured populations. Genetics **158:** 913–917.

Barton, N. H., 1995a A general model for the evolution of recombination. Genet. Res. **65:** 123–144.

Barton, N. H., 1995b Linkage and the limits to natural selection. Genetics **140:** 821–841.

Barton, N. H., and K. S. Gale, 1993 Genetic analysis of hybrid zones, pp. 13–45 in *Hybrid Zones and the Evolutionary Process*, edited by R. G. Harrison. Oxford University Press, Oxford.

Barton, N., and S. P. Otto, 2005 Evolution of recombination due to random drift. Genetics **169:** 2353–2370.

Barton, N. H., and L. Partridge, 2000 Limits to natural selection. BioEssays **22:** 1075–1084.

Barton, N. H., and M. Turelli, 1991 Natural and sexual selection on many loci. Genetics **127:** 229–255.

Colegrave, N., 2002 Sex releases the speed limit on evolution. Nature **420:** 664–666.

Feldman, M. W., F. B. Christiansen and L. D. Brooks, 1980 Evolution of recombination in a constant environment. Proc. Natl. Acad. Sci. USA **77:** 4838–4841.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

Felsenstein, J., and S. Yokoyama, 1976 Evolutionary advantage of recombination. 2. Individual selection for recombination. Genetics **83:** 845–859.
Fisher, R. A., 1930 *The Genetical Theory of Natural Selection.* Oxford University Press, Oxford.
Gerrish, P. J., and R. E. Lenski, 1998 The fate of competing beneficial mutations in an asexual population. Genetica **103:** 127–144.
Goddard, M. R., H. Charles, J. Godfray and A. Burt, 2005 Sex increases the efficacy of natural selection in experimental yeast populations. Nature **434:** 636–640.
Hill, W. G., and A. Robertson, 1966 The effect of linkage on the limits to artificial selection. Genet. Res. **8:** 269–294.
Iles, M. M., K. Walters and C. Cannings, 2003 Recombination can evolve in finite populations given selection on sufficient loci. Genetics **165:** 2249–2258.
Kondrashov, A. S., 1993 Classification of hypotheses on the advantage of amphimixis. J. Hered. **84:** 372–387.
Lenormand, T., and S. P. Otto, 2000 The evolution of recombination in a heterogeneous environment. Genetics **156:** 423–438.
Maynard Smith, J., 1971 What use is sex? J. Theor. Biol. **30:** 319–335.
Muller, H. J., 1932 Some genetic aspects of sex. Am. Nat. **66:** 118–138.
Nei, M., and W. H. Li, 1973 Linkage disequilibrium in subdivided populations. Genetics **75:** 213–219.
Otto, S. P., 2003 The advantages of segregation and the evolution of sex. Genetics **164:** 1099–1118.
Otto, S. P., and N. H. Barton, 1997 The evolution of recombination: removing the limits to natural selection. Genetics **147:** 879–906.
Otto, S. P., and N. Barton, 2001 Selection for recombination in small populations. Evolution **55:** 1921–1931.
Otto, S. P., and T. Lenormand, 2002 Resolving the paradox of sex and recombination. Nat. Genet. **3:** 252–261.
Peck, J. R., 1994 A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics **137:** 597–606.
Peck, J. R., J. Yearsley and G. Barreau, 1999 The maintenance of sexual reproduction in a structured population. Proc. R. Soc. Lond. Ser. B Biol. Sci. **266:** 1857–1863.
Wolfram, S., 1991 *Mathematica.* Addison-Wesley, New York.

Communicating editor: M. Uyenoyama

APPENDIX A: MATRIX NOTATIONS

Explicit expression of the vector function f: The deterministic changes in allele frequencies and linkage disequilibrium after multiplicative selection and recombination are given by the vector function $\mathbf{f} = \{f_j, f_k, f_{jk}\}$, from Barton and Otto (2005),

$$x'_j = f_j(\mathbf{x}) = x_j + \frac{s_j x_j(1-x_j)\phi_k + s_k x_{jk}(1-s_j(x_j-1/2))}{W} \tag{A1a}$$

$$x'_k = f_k(\mathbf{x}) = x_k + \frac{s_k x_k(1-x_k)\phi_j + s_j x_{jk}(1-s_k(x_k-1/2))}{W} \tag{A1b}$$

$$x'_{jk} = f_{jk}(\mathbf{x}) = \frac{x_{jk}(1-r)(1-s_j^2/4)(1-s_k^2/4)}{W^2}, \tag{A1c}$$

where $\phi_j = 1+s_j(x_j-1/2)$, $\phi_k = 1+s_k(x_k-1/2)$, and $\bar{W} = \phi_j\phi_k + s_j s_k x_{jk}$ are the mean fitnesses of the population, and $\mathbf{x} = \{x_j, x_k, x_{jk}\}$ is the vector of allele frequencies and LD at the previous generation. (A1c) shows that multiplicative selection alone cannot produce but may change the linkage disequilibrium (x'_{jk} is proportional to x_{jk}).

Exact expressions for the matrices $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{D_3}$: The first and second partial derivatives of the vector function **f** with respect to the three variables x_j, x_k, and x_{jk}, evaluated along the deterministic trajectory **x***, are the elements in matrices $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{D_3}$. These derivatives can be computed directly from (A1) and are also given in appendix b of Barton and Otto (2005). For each variable in **dx** (resp. $\mathbf{dx^2}$), the corresponding row in matrix $\mathbf{D_1}$ (resp. $\mathbf{D_2}$) is directly computed by identification to the coefficients of the first- (resp. second-) order Taylor series expansion of $\mathbf{dx_s} = \mathbf{f}(\mathbf{x}+\mathbf{dx}) - \mathbf{f}(\mathbf{x})$, the difference between the stochastic and deterministic trajectories after selection [see (3)]. The 3×3 matrix $\mathbf{D_1}$ contains the first partial derivatives of **f**, which multiply the elements of **dx** in (4), and equals the gradient of **f** at point **x***,

$$\mathbf{D_1} = \mathbf{grad}(\mathbf{f}(\mathbf{x}^*)) = \begin{bmatrix} \frac{a_j}{\phi_j^2} & 0 & \frac{s_k a_j}{\phi_j^2\phi_k} \\ 0 & \frac{a_k}{\phi_k^2} & \frac{s_j a_k}{\phi_j\phi_k^2} \\ 0 & 0 & \frac{(1-r)a_j a_k}{\phi_j^2\phi_k^2} \end{bmatrix}, \tag{A2}$$

where $a_j = 1 - s_j^2/4$ and $a_k = 1 - s_k^2/4$. Similarly, the 3×6 matrix $\mathbf{D_2}$ contains the second partial derivatives of **f**, which multiply the elements of $\mathbf{dx^2}$ in (4), as well as the coefficient $\frac{1}{2}$:

$$\mathbf{D_2} = \begin{bmatrix} \frac{-s_j a_j}{\phi_j^3} & 0 & \frac{-2s_j s_k a_j}{\phi_j^3\phi_k} & 0 & \frac{-s_k^2 a_j}{\phi_j^2\phi_k^2} & \frac{-s_j s_k^2 a_j}{\phi_j^3\phi_k^2} \\ 0 & 0 & \frac{-s_j^2 a_k}{\phi_j^2\phi_k^2} & \frac{-s_k a_k}{\phi_j^3} & \frac{-2s_j s_k a_k}{\phi_j\phi_k^3} & \frac{-s_j^2 s_k a_j}{\phi_j^2\phi_k^3} \\ 0 & 0 & \frac{-2(1-r)s_j a_j a_k}{\phi_j^3\phi_k^2} & 0 & \frac{-2(1-r)s_k a_j a_k}{\phi_j^2\phi_k^3} & \frac{-2(1-r)s_j s_k a_j a_k}{\phi_j^3\phi_k^3} \end{bmatrix}. \tag{A3}$$

Note that each term in $\mathbf{D_2}$ is negative, which demonstrates that all variances and covariances of deviations will tend to favor negative deviations with the term $\mathbf{D_2 dx^2}$. Similarly, for each variable in $\mathbf{dx^2}$, *i.e.*, for each $\{dx_a dx_b\}_{(a,b)\in V}$, the corresponding row in the 6×6 matrix $\mathbf{D_3}$ is obtained by identification to the coefficients of the Taylor series expansion of the products of deviations after selection and recombination $(f_a(\mathbf{x}+\mathbf{dx}) - f_a(\mathbf{x}))(f_b(\mathbf{x}+\mathbf{dx}) - f_b(\mathbf{x}))$. We then obtain, for the two-locus system:

$$\mathbf{D_3} = \begin{bmatrix} \frac{s_j a_j^2}{\phi_j^4} & 0 & \frac{2s_k a_j^2}{\phi_j^4\phi_k} & 0 & 0 & \frac{s_k^2 a_j^2}{\phi_j^4\phi_k^2} \\ 0 & \frac{a_j a_k}{\phi_j^2\phi_k^2} & \frac{s_j a_j a_k}{\phi_j^3\phi_k^2} & 0 & \frac{s_k a_j a_k}{\phi_j^2\phi_k^3} & \frac{s_j s_k a_j a_k}{\phi_j^3\phi_k^3} \\ 0 & 0 & \frac{(1-r)a_j^2 a_k}{\phi_j^4\phi_k^2} & 0 & 0 & \frac{(1-r)s_k a_j^2 a_k}{\phi_j^4\phi_k^3} \\ 0 & 0 & 0 & \frac{a_k^2}{\phi_k^4} & \frac{2s_j a_k^2}{\phi_j\phi_k^4} & \frac{s_j^2 a_k^2}{\phi_j^2\phi_k^4} \\ 0 & 0 & 0 & 0 & \frac{(1-r)a_j a_k^2}{\phi_j^2\phi_k^4} & \frac{(1-r)s_j a_j a_k^2}{\phi_j^3\phi_k^4} \\ 0 & 0 & 0 & 0 & 0 & \frac{(1-r)^2 a_j^2 a_k^2}{\phi_j^4\phi_k^4} \end{bmatrix}. \tag{A4}$$

Matrices $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{D_3}$ were computed using Mathematica (Wolfram 1991) and are available upon request. The values of x_j and x_k in ϕ_j and ϕ_k are evaluated along the deterministic trajectory ($x_j = x_j^*$, $x_k = x_k^*$, and $x_{jk} = x_{jk}^* = 0$).

Exact moments introduced by the multinomial sampling: The adult population is sampled from the surviving juveniles according to a multinomial distribution, as in the standard Wright-Fisher model. Following Barton and Otto (2005), the moments of the multinomial distribution are used to determine the expected values of the perturbations:

$$E[\boldsymbol{\zeta}] = \begin{Bmatrix} \zeta_j \\ \zeta_k \\ \zeta_{jk} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ -x_{jk}/2N \end{Bmatrix}. \tag{A5}$$

The variances and covariances of perturbations are given by

$$E[\boldsymbol{\zeta}^2] = \begin{Bmatrix} E[\zeta_j^2] \\ E[\zeta_j\zeta_k] \\ E[\zeta_j\zeta_{jk}] \\ E[\zeta_k^2] \\ E[\zeta_k\zeta_{jk}] \\ E[\zeta_{jk}^2] \end{Bmatrix} = \frac{1}{2N} \begin{Bmatrix} x_j(1-x_j) \\ x_{jk} \\ -x_{jk}(2x_j-1) \\ x_k(1-x_k) \\ -x_{jk}(2x_k-1) \\ x_j(1-x_j)x_k(1-x_k) + (2x_j-1)(2x_k-1)x_{jk} - x_{jk}^2 \end{Bmatrix}. \tag{A6}$$

Large deme size approximation: The sources of negative deviations $E[dx_a]$ are the variances and covariances of deviations, which are of order $O(dx^2)$. Consequently, in our approximation for large population size, terms in $dx_a/2N$ are $o(dx^2)$ and are negligible. Therefore, although the actual sampling is made from populations following stochastic trajectories ($x_a = x_a^* + dx_a$), the values of $E[\zeta_a]$ and $E[\zeta_a\zeta_b]$ are approximately independent of the actual values of the deviations, dx_a. Thus, the perturbations caused by drift within a generation are determined by the population size and the allele frequencies on the deterministic trajectory. Note that this result explains why terms in $E[\zeta_a dx_b]$ were dropped from (8). The approximation for large population size of the exact expressions of $E[\boldsymbol{\zeta}]$ and $E[\boldsymbol{\zeta}^2]$ is thus obtained by replacing any x_a by x_a^* in (A5) and (A6), so that

$$E[\boldsymbol{\zeta}^2] = \frac{\mathbf{c}}{2N} + o(N^{-1}) \quad \text{and} \quad E[\boldsymbol{\zeta}] = \mathbf{0} + o(N^{-1}), \tag{A7}$$

where $\mathbf{c} = \{x_j^*(1-x_j^*), 0, 0, x_k^*(1-x_k^*), 0, x_j^*(1-x_j^*)\, x_k^*(1-x_k^*)\}$ is a 1×6 vector with the nonzero terms equal to the genetic variances of x_j, x_k, and x_{jk}, evaluated along the deterministic trajectory.

Three-locus recursions: When including a modifier locus, i, that modifies the recombination rate between loci j and k, four additional variables are needed to describe the system: the allele frequency at the modifier locus (x_i) and the three additional LD that are defined when including locus i (x_{ij}, x_{ik}, and x_{ijk}). We define new deviation vectors including these variables: the 1×7 vector $\mathbf{dx}$ of first-order deviations, the 1×28 vector $\mathbf{dx^2}$ of second-order deviations, excluding the repeated products, and the three corresponding metapopulation moments $\overline{\mathbf{dx}}$, $\overline{\mathbf{dx^2}}$, and $\overline{\mathbf{dx}}^{\mathbf{2}}$. We then follow the same method as that described for the two-locus model. The recursions for the deterministic change, after one round of recombination and selection, for the four additional variables can be found in (A2c)–(A2e) of Barton and Otto (2005). From these recursions, and in the same way as that for the two-locus model, we generate the 7×7 matrix $\mathbf{D_1}$, the 7×28 matrix $\mathbf{D_2}$, and the 28×28 matrix $\mathbf{D_3}$ (available upon request). As in the two-locus model, the multinomial sampling effect is negligible on the vector $E[\overline{\mathbf{dx}}]$, while it introduces variance in the vectors $E[\mathbf{dx^2}]$ and $E[\overline{\mathbf{dx}}^{\mathbf{2}}]$. Finally, the effect of migration on the moments in a subdivided population is exactly the same as that for the two-locus model except for the three-locus linkage disequilibrium dx_{ijk} (see appendix b).

APPENDIX B: EFFECT OF MIGRATION ON ALLELE FREQUENCIES AND LINKAGE DISEQUILIBRIA IN THE n-ISLAND MODEL

Let us consider a focal deme i, from which a fraction m of individuals emigrate, and into which a comparable number of individuals immigrate from all other demes. Let $\mathbf{v}[i]$ be the vector containing the frequencies of multilocus haplotypes in deme i. After migration, the new genotype frequency vector is given by

$$\mathbf{v}[i]' = (1-m)\mathbf{v}[i] + m\frac{1}{n-1}\sum_{\substack{i_1=1 \\ i_1 \neq i}}^{n} \mathbf{v}[i_1], \tag{B1}$$

which can also be written

$$\mathbf{v}[i]' = (1-m_e)\mathbf{v}[i] + m_e\bar{\mathbf{v}}, \tag{B2}$$

where $m_e = mn/(n-1)$ and $\bar{\mathbf{v}} = (1/n)\sum_{i=1}^{n} \mathbf{v}[i]$ is the vector giving the haplotype frequencies in the whole population (or equivalently in the migrant pool). This is exactly the recursion for a two-island system where one deme is the focal deme i and the other is the migrant pool (with haplotype frequency vector $\bar{\mathbf{v}}$).

This result is valid for any number of loci, but let us first consider the two-locus case. For any variable $\{x_a\}_{a\in U}$ at a given time, let us denote the difference between the value of x_a in deme i and the mean of x_a over all demes by $\Delta_a[i] = x_a[i] - \overline{x}_a$. The allele frequencies and linkage disequilibrium in the migrant pool (denoted by the index $i = mp$) are given by

$$\begin{aligned} x_j[mp] &= \overline{x}_j \\ x_k[mp] &= \overline{x}_k \\ x_{jk}[mp] &= \overline{x}_{jk} + \overline{\Delta_j\Delta_k}, \end{aligned} \tag{B3}$$

where

$$\overline{\Delta_j\Delta_k} = \frac{1}{n}\sum_{i=1}^{n} \Delta_j[i]\Delta_k[i] = \overline{x_j x_k} - \overline{x}_j\overline{x}_k \tag{B4}$$

is the covariance between allele frequencies at loci j and k, taken across demes, *i.e.*, the spatial covariance between allele frequencies in the whole population (*cf.* also Nei and Li 1973). The recursion for the effect of migration on allele frequencies and LD in the focal deme i is the same as that for the two-island system (given, *e.g.*, in Barton and Gale 1993), where we use the values of $x_a[mp]$ given in (B3) for the other deme (migrant pool). The change due to migration $\delta_m[\mathbf{x}[i]]$ on the vector $\mathbf{x}[i]$ of allele frequencies and LD in the focal deme i is thus given by

$$\delta_m[\mathbf{x}[i]] = \begin{Bmatrix} \delta_m[x_j[i]] \\ \delta_m[x_k[i]] \\ \delta_m[x_{jk}[i]] \end{Bmatrix} = \begin{Bmatrix} -m_e\Delta_j[i] \\ -m_e\Delta_k[i] \\ -m_e(\Delta_{jk}[i] - \overline{\Delta_j\Delta_k}) + m_e(1-m_e)\Delta_j[i]\Delta_k[i] \end{Bmatrix}. \tag{B5}$$

Taking the average across demes of $\delta_m[\mathbf{x}[i]]$ in (B5), the effect of migration on the average allele frequencies and LD in the whole population (*i.e.*, on $\overline{\mathbf{x}} = \{\overline{x}_j, \overline{x}_k, \overline{x}_{jk}\}$) gives recursion (18). For the three-locus system, recursion (18) has to be changed to include the effect of migration on the other two-locus linkage disequilibria (x_{ij} and x_{ik}), which is obtained simply by switching indexes: for example, the change in the linkage disequilibrium dx_{ij} is $m_e(2-m_e)\overline{\Delta_i\Delta_j}$. However, the effect of migration on the three-locus linkage disequilibrium x_{ijk} has to be computed. Following Barton and Turelli (1991) we define the three-locus linkage disequilibrium x_{ijk} by

$$x_{ijk} = \mathrm{cov}(X_i, X_j, X_k) = \sum_X v_X(X_i - E[X_i])(X_j - E[X_j])(X_k - E[X_k]), \tag{B6}$$

where, for any diallelic locus l, X_l is a binary variable with value 1 for one of the alleles and 0 for the other, and v_X is the frequency of a given three-locus haplotype $\{X_i, X_j, X_k\}$ in the population considered (*i.e.*, either the focal deme i or the migrant pool mp). The change in x_{ijk} for a given deme in the n-island model can be computed as in a two-island system with migration between the deme considered and the migrant pool (deme mp). The value of $x_{ijk}[\mathrm{mp}]$, the three-locus LD in the migrant pool relative to its average across demes x_{ijk}, is

$$x_{ijk}[\mathrm{mp}] = \overline{x}_{ijk} - (\overline{\Delta_i\Delta_j\Delta_k} + \overline{\Delta_i\Delta_{jk}} + \overline{\Delta_j\Delta_{ik}} + \overline{\Delta_k\Delta_{ij}}). \tag{B7}$$

Then, from (B2), the change in the average x_{ijk} by migration is

$$\begin{aligned} \delta_m[\overline{x}_{ijk}] = {} & m_e(2-m_e)\overline{(\Delta_i\Delta_{jk} + \overline{\Delta_k\Delta_{ij}} + \overline{\Delta_j\Delta_{ik}})} \\ & - m_e(3 - 2m_e(3-m_e))\overline{\Delta_i\Delta_j\Delta_k}. \end{aligned} \tag{B8}$$

Taking into account the fact that any $\Delta_a = dx_a - \overline{dx_a}$ is of the order of deviations dx_a, and removing $O(dx^3)$ terms, we finally obtain the large-deme approximation

$$\delta_m[\overline{x}_{ijk}] = m_e(2-m_e)(\overline{\Delta_i\Delta_{jk}} + \overline{\Delta_k\Delta_{ij}} + \overline{\Delta_j\Delta_{ik}}) + o(dx^2). \tag{B9}$$