Year : 2023

A systems biology approach to bacterial gene essentiality

Martins Bravo Afonso

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Department of Fundamental Microbiology

A systems biology approach to bacterial gene essentiality

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Afonso MARTINS BRAVO

Master de l'Université NOVA de Lisboa

Jury

Prof. Denys Alban, Président
Prof. Jan-Willem Veening, Directeur de thèse
Dr. Athanasios Typas, Co-directeur de thèse (EMBL)
Prof. Alexandre Persat, Expert
Prof. Lars Dietrich, Expert

Lausanne
(2023)

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Department of Fundamental Microbiology

A systems biology approach to bacterial gene essentiality

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Afonso MARTINS BRAVO

Master de l'Université NOVA de Lisboa

Jury

Prof. Denys Alban, Président
Prof. Jan-Willem Veening, Directeur de thèse
Dr. Athanasios Typas, Co-directeur de thèse (EMBL)
Prof. Alexandre Persat, Expert
Prof. Lars Dietrich, Expert

Lausanne
(2023)

UNIL | Université de Lausanne

Faculté de biologie et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Alban	**Denys**
Directeur·trice de thèse	Monsieur	Prof.	Jan-Willem	**Veening**
Co-directeur·trice	Monsieur	Dr	Athanasios	**Typas**
Expert·e·s	Monsieur	Prof.	Alexandre	**Persat**
	Monsieur	Prof.	Lars	**Dietrich**

le Conseil de Faculté autorise l'impression de la thèse de

Afonso Martins Bravo

Master degree in Medical microbiology, Universidade Nova de Lisboa, Portugal

intitulée

A systems biology approach to bacterial gene essentiality

Lausanne, le 6 octobre 2023

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Alban Denys

The Zen of Python, by Tim Peters

Beautiful is better than ugly.

Explicit is better than implicit.

Simple is better than complex.

Complex is better than complicated.

Flat is better than nested.

Sparse is better than dense.

Readability counts.

Special cases aren't special enough to break the rules.

Although practicality beats purity.

Errors should never pass silently.

Unless explicitly silenced.

In the face of ambiguity, refuse the temptation to guess.

There should be one-- and preferably only one --obvious way to do it.

Although that way may not be obvious at first unless you're Dutch.

Now is better than never.

*Although never is often better than *right* now.*

If the implementation is hard to explain, it's a bad idea.

If the implementation is easy to explain, it may be a good idea.

Namespaces are one honking great idea -- let's do more of those!

This page is intentionally left blank.

What are you looking for?

All my data…

Xue Liu

That's a problem…

Julien Dénéréaz

This page is intentionally left blank.

Academic Summary

Under any given environment, there are genetic *loci* required for sustaining life, and others which are not. Such defines both gene essentiality, if a gene is required in all conditions, and conditional essentiality, if a gene is needed in only some environments.

As high-throughput essentiality determining methods continue to become widespread, so have the conditions and organisms being assayed. Indeed, several works have now demonstrated the context dependence of gene essentiality, varying not only in regards to specific environments, but also at the strain level. Expectedly, such results have highlighted the caveats of primarily screening model organisms under standardized laboratory conditions, often dissimilar from their natural niche. Insight into these species/strain dependent differences could potentially yield novel targetable essential pathways and functions, allowing for greater control of the targeted organisms.

In this thesis we develop new computational pipelines for independently processing data from transposon mutagenesis (Tn-seq), and CRISPR interference sequencing (CRISPRi-seq) techniques. Both Tn-seq and CRISPRi-seq use pooled libraries of mutants to infer gene fitness, however, while the first assesses a cell´s ability to survive any individual gene disruption originating from a transposon insertion event, the second determines gene fitness based on a cell´s ability to survive gene transcriptional repression.

We respectively apply Tn-seq and CRISPRi-seq to explore gene essentiality in *Escherichia coli* and *Streptococcus pneumoniae*. We then dwell into the problem of lack of phylogenetic conservation of essential genes, and how conditional essentiality can emerge in highly specific environments.

In chapter 2 we present 2FAST2Q, a stand-alone easy-to-use program for counting features in FASTQ files. 2FAST2Q solves a recent issue in bioinformatics, the lack of an unified versatile program that can filter and extract data from raw sequencing files. 2FAST2Q is particularly useful for CRISPRi-seq data, where the sequences corresponding to all the individual mutant strains can be counted and filtered according to the user needs. 2FAST2Q is also capable of *de novo* finding unknown sequences based on known anchor sequences, a useful feature when performing experiments with barcoded mutant strains.

In chapter 3 we introduce TnSeeker. TnSeeker is a linux based Python pipeline capable of inferring gene essentiality from fastq files originating from Tn-seq experiments. TnSeeker uses both the relative transposon distribution, and local GC content biases, to determine if any given genetic feature, such as a gene sub-domain region, is essential. Such is performed using a conservative approach based on a self-optimizing threshold defining algorithm: Genes known to be essential across a multitude of organisms are used to determine the optimal essentiality defining cutoff for the dataset being analyzed, with the threshold being defined as the significance value that most accurately recapitulates the known data.

Using TnSeeker, and Tn-seq, we determined the core-essentialome of 8 *E. coli* strains. We demonstrated that only genes related with protein biosynthesis are significantly enriched when comparing essential genes across all *E. coli* strains. Conversely, enrichment for all expected life sustaining functions (DNA and central metabolism) is seen when either considering the essentialome of strains by themselves, or the pan-essentialome. We discuss that such effect could derive from non-orthologous gene displacement, arising from the action of mobile genetic elements such as viruses and plasmids.

In chapter 4 we adapt an arraying technique known as SUDOKU to deconvolute the built transposon *E. coli* UTI89 library. By plating any pooled library, randomly picking enough single colonies, arraying them into compartmentalized units, and sequencing all compartments as a series of mixed pools, it is possible to backtrack any found mutants into their original compartment. Such arrayed libraries allow for the permanent storage of individual mutants, and facilitate the massive screening (for example for chemogenomics) of new microorganisms. We further enhance SUDOKU by allowing it to also function with CRISPRi libraries.

Finally, in chapter 5, we explore the gene essentiality of the human nasopharynx inhabiting opportunistic pathogen *S. pneumoniae. S. pneumoniae* has been shown to be negatively associated with carriage of another common pathogen, *Staphylococcus aureus*. Using CRISPRi-seq, we observe that one general purpose efflux pump, SPV 686/7/8 (here renamed as ArpABC), capable of exporting several described antibiotics and antimicrobial peptides, is essential at pH 6 only when *S. aureus* is present. We further explore how *S. pneumoniae* and *S. aureus* interact when present in a human cell matrix, where typical cell wall essential genes of *S. pneumoniae* become not as essential.

Such results highlight the impact of studying conditional essentiality, or how essentiality changes in regards to different conditions, often drastically different from the ones typically used in laboratories. Indeed, throughout this thesis we explore how essentiality is a function of the environment, where each lineage of organism is both optimally and uniquely adapted to all different scenarios it encounters in its natural environment.

The techniques used in this thesis are only one dimension of the unprecedented worldwide ongoing characterization of bacteria. As unparalleled amounts of data continue to be generated, we are now closer than ever into fully integrating our collective knowledge, and finally take the next step in biology: the full understanding and modeling of bacteria, perhaps the simplest most successful organisms in the planet.

Résumé Académique

Dans un environnement donné, il existe des locus génétiques nécessaires au maintien de la vie, et d'autres qui ne le sont pas. Cela définit à la fois l'essentialité du gène, si un gène est requis dans toutes les conditions, et l'essentialité conditionnelle, si un gène n'est nécessaire que dans certains environnements.

Alors que les méthodes de détermination de l'essentialité à haut débit continuent de se répandre, les conditions et les organismes testés se sont également répandus. En effet, plusieurs études ont maintenant démontré la dépendance au contexte de l'essentialité des gènes, variant non seulement en fonction des environnements spécifiques, mais également au niveau de la souche. De tels résultats ont alors mis en évidence les dangers de cribler principalement des organismes modèles dans des conditions de laboratoire standardisées, souvent différentes de leur niche naturelle. Un aperçu de ces différences dépendantes des espèces ou des souches pourrait potentiellement produire de nouvelles voies et fonctions essentielles ciblables, permettant un meilleur contrôle des organismes ciblés.

Dans cette thèse, nous développons de nouveaux pipelines pour le traitement indépendant des données issues des techniques de mutagenèse par transposon (Tn-seq) et de séquençage d'interférence CRISPR (CRISPRi-seq). Tn-seq et CRISPRi-seq utilisent des bibliothèques regroupées de mutants pour déduire la forme physique des gènes, cependant, alors que la première évalue la capacité d'une cellule à survivre à toute perturbation génétique individuelle provenant d'un événement d'insertion de transposon, la seconde détermine l'importance des gènes en fonction de la capacité des cellules à survivre à la répression transcriptionnelle des gènes.

Nous appliquons respectivement Tn-seq et CRISPRi-seq pour explorer l'essentialité des gènes chez *Escherichia coli* et *Streptococcus pneumoniae*. Nous nous attardons ensuite sur le problème du manque de conservation phylogénétique des gènes essentiels, et comment l'essentialité conditionnelle peut émerger dans des environnements hautement spécifiques.

Dans le chapitre 2, nous présentons 2FAST2Q, un programme autonome facile à utiliser pour compter les fonctionnalités dans les fichiers fastq. 2FAST2Q résout un problème récent en bioinformatique, l'absence d'un programme polyvalent unifié capable de filtrer et d'extraire des données à partir de fichiers de séquençage bruts. 2FAST2Q est particulièrement utile pour les données CRISPRi-seq, où les séquences correspondant à toutes les souches mutantes individuelles peuvent être comptées et

filtrées en fonction des besoins de l'utilisateur. 2FAST2Q est également capable de trouver de novo des séquences inconnues basées sur des séquences d'ancrage connues, une fonctionnalité utile lors de la réalisation d'expériences avec des souches mutantes à code-barres.

Dans le chapitre 3, nous présentons TnSeeker. TnSeeker est un pipeline Python basé sur Linux capable de déduire l'essentialité des gènes à partir de fichiers fastq provenant d'expériences Tn-seq. TnSeeker utilise à la fois la distribution relative des transposons et les biais du contenu GC locaux pour déterminer si un trait génétique donné, telle qu'une région de sous-domaine génique, est essentiel. Ceci est effectué à l'aide d'une approche conservatrice basée sur un algorithme de définition de seuil d'auto-optimisation: les gènes connus pour être essentiels dans une multitude d'organismes sont utilisés pour déterminer le seuil optimal définissant l'essentialité pour l'ensemble de données analysées, le seuil étant défini comme la signification statistique, valeur qui récapitule le plus fidèlement les données connues.

À l'aide de TnSeeker et de Tn-seq, nous avons déterminé le noyau essentiel de 8 souches d'*E. coli*. Nous avons démontré que seuls les gènes liés à la biosynthèse des protéines sont considérablement enrichis lors de la comparaison des gènes essentiels dans toutes les souches d'E. coli. À l'inverse, l'enrichissement pour toutes les fonctions de maintien de la vie attendues (ADN et métabolisme central) est observé lorsque l'on considère soit l'essentialome des souches elles-mêmes, soit le pan-essentialome. Nous discutons du fait qu'un tel effet pourrait dériver d'un déplacement de gène non orthologue, résultant de l'action d'éléments génétiques mobiles tels que des virus et des plasmides.

Dans le chapitre 4, nous adaptons une technique de mise en réseau connue sous le nom de SUDOKU pour déconvoluer la bibliothèque de transposon construite *E. coli* UTI89. En étalant sur plaque n'importe quelle bibliothèque regroupée, en choisissant au hasard suffisamment de colonies individuelles, en les répartissant en unités compartimentées et en séquençant tous les compartiments comme une série de *pools* mixtes, il est possible de remonter tous les mutants trouvés dans leur compartiment d'origine. De telles bibliothèques en réseau permettent le stockage permanent de mutants individuels et facilitent le criblage massif (par exemple pour la chimiogénomique) de nouveaux micro-organismes. Nous améliorons encore SUDOKU en lui permettant de fonctionner également avec les bibliothèques CRISPRi.

Enfin, au chapitre 5, nous explorons l'essentialité génétique du pathogène opportuniste *Streptococcus pneumoniae* habitant le nasopharynx humain. Il a été démontré que *S. pneumoniae* est négativement associé au portage d'un autre agent pathogène courant, *Staphylococcus aureus*. En utilisant CRISPRi-seq, nous observons qu'une pompe à efflux à usage général, SPV 686/7/8 (rebaptisée ici ArpABC), capable d'exporter plusieurs antibiotiques et peptides antimicrobiens décrits, est essentielle à pH 6 uniquement lorsque *S. aureus* est présent. Nous explorons plus en détail comment *S. pneumoniae* et *S. aureus* interagissent lorsqu'ils sont présents dans une matrice cellulaire humaine, et observons que *S. pneumoniae* affiche un besoin réduit de gènes essentiels liés à la paroi cellulaire.

Ces résultats mettent en évidence l'impact de l'étude de l'essentialité conditionnelle, ou comment l'essentialité change en fonction de différentes conditions, souvent radicalement différentes de celles généralement utilisées dans les laboratoires. En effet, tout au long de cette thèse, nous explorons comment l'essentialité est une fonction de l'environnement, où chaque lignée d'organisme est à la fois adaptée de manière optimale et unique à tous les différents scénarios qu'elle rencontre dans son environnement naturel. Ensemble, ces données peuvent être utilisées pour tester de nouvelles cibles antibiotiques, ou même pour déterminer des molécules de ciblage spécifiques à une souche/espèce.

Les techniques utilisées dans cette thèse ne sont qu'une dimension de la caractérisation mondiale sans précédent des bactéries en cours. Alors que des quantités inégalées de données continuent d'être générées, nous sommes maintenant plus proches que jamais d'intégrer pleinement nos connaissances collectives et de franchir enfin la prochaine étape de la biologie : la compréhension et la modélisation complètes des bactéries, peut-être les organismes les plus simples et les plus performants de la planète.

Non-Academic Summary

Under any given environment, there are genetic features required for sustaining life, and others which are not. Such definition could characterize what is known as an essential, or lethal, gene. Indeed, gene essentiality has been an ongoing topic in genetics since its formal humble beginnings in the XX century, when mutations started being systematically categorized. Since then, mutations have unfolded the inner workings of cells and organisms, and remain, to this day, at the core of most biology related fields. Nowadays, mutations can be induced on a genome wide scale, and their respective outcomes assayed. Currently, such high-throughput measurements can only be efficiently achieved by using variations of either the Tn-seq, or CRISPRi-seq, methods. The first relies on randomly integrating transposons into an organism´s DNA in such a frequency that all genetic features are disrupted, and thus, non-functional. The second is based on using an engineered protein, dCas9, to block any chosen genomic feature from becoming active, ultimately resulting in that feature's non-functionality. Both methods are done *en masse* in the way that any genomic feature disruption exists within a single cell, in turn existing within a larger pool of all the individually disrupted cells: mutants. By simultaneously submitting such different mutants to distinct conditions, it becomes possible to determine what features are required for survival under any tested environment, as any mutations in any important locations will slowly be eliminated from the mutant pool population. The determination of what mutations are/not present at the end of the experiment indicates which genes are/not essential. Such is accomplished by sequencing all the mutants present in the pool and comparing their relative abundance to the same pool of mutants at the start of the experiment.

On chapter 2 we approached how this sequencing data can be demultiplexed from the millions of sequences corresponding to the mutants in the pool, into human readable data. To this end we developed 2FAST2Q, an intuitive user-friendly computer program that can count the abundances of any sequences, and thus of the mutants in the pool, using several user-defined filtering steps. We exemplify how 2FAST2Q can be used, and how different usage parameters impact downstream analysis.

In chapter 3 we developed a computational pipeline capable of determining essential genes without the need to compare the starting and end populations of mutants. Such is possible when using the Tn-seq method by determining where all transposons have inserted into in the genome. In a pool of transposon mutants, a gene

without transposons in any given gene, but with transposons at a high enough frequency in its surroundings, would imply that that gene should also be carrying transposons, and that not observing such would mean that cells carrying those insertions did not survive, and that that gene is essential. The developed pipeline individually determines, for all these cases, and using a self-optimizing algorithm based on known essential genes present in most organisms, whether a gene is essential or not. Using this method, termed TnSeeker, we determined all the common essential genes in a collection of 8 different strains of *Escherichia coli*. We observed that all strains exhibit a reduced overlap in common essentials, with most of these being related to protein synthesis. Conversely, other known essential cell functions such as DNA maintenance were only present when strains were either considered by themselves, and/or as the sum of all. Such phenomenon could potentially be related with the evolutionary mobility of essential genes via mobile elements, such as viruses. Due to the shuffling of genes across organisms, it is possible for species and strains to have essential cell functions performed by different genes, thus reducing any gene similarity, and confusing the analysis of shared essential genes.

Upon creating a pool of mutants, it might be desirable to individually sort them and create an individual collection, where single characterized mutants can be used in isolation for specific experiments. In chapter 4 we adapted a method, known as SUDOKU, to array unknown pooled mutants into known locations within a plate matrix. We effectively arrayed one *E. coli* mutant pooled transposon library into its curated isolated single mutant form.

Finally, in chapter 5, we used the CRISPRi-seq method to explore how the human nasopharynx inhabiting opportunistic pathogen *Streptococcus pneumoniae* interacts with another common pathogen, *Staphylococcus aureus*. We observed that one efflux pump, SPV 686/7/8 (here named as ArpABC), previously implicated in the detoxification of bactericions (molecules with bactericidal activity), becomes essential at pH 6 only when *S. aureus* is present. Such highlights the impact of studying conditional essentiality, or how essentiality changes in regards to different conditions, often drastically different from the ones typically used in laboratories. Indeed, throughout this thesis we explore how essentiality is a function of the environment, where each lineage of organism is both optimally and uniquely adapted to all different scenarios it encounters in its natural environment. Together, such data can be used

for assaying novel antibiotic targets, which often disrupt cellular essential functions, or even be used to determine strain/species specific targeting molecules.

The techniques used in this thesis are only one dimension of the unprecedented worldwide ongoing characterization of bacteria. As unparalleled amounts of data continue to be generated, we are now closer than ever to fully integrating our collective knowledge and finally take the next step in biology: the full understanding and modeling of bacteria, perhaps the simplest most successful organisms in the planet.

Résumé non académique

Dans un environnement donné, il existe des traits génétiques nécessaires au maintien de la vie, et d'autres qui ne le sont pas. Une telle définition pourrait caractériser ce que l'on appelle un gène essentiel, ou mortel. En effet, l'essentialité des gènes est un sujet récurrent en génétique depuis ses modestes débuts formels au XXe siècle, lorsque les mutations ont commencé à être systématiquement catégorisées. Depuis lors, les mutations ont dévoilé le fonctionnement interne des cellules et des organismes et restent, à ce jour, au cœur de la plupart des domaines liés à la biologie. De nos jours, des mutations peuvent être induites à l'échelle du génome et leurs résultats respectifs mesurés. Actuellement, de telles mesures à haut débit ne peuvent être réalisées efficacement qu'en utilisant des variantes des méthodes Tn-seq ou CRISPRi-seq. La première repose sur l'intégration aléatoire de transposons dans l'ADN d'un organisme à une fréquence telle que tous les traits génétiques sont perturbés et donc non fonctionnels. La seconde est basée sur l'utilisation d'une protéine modifiée, dCas9, pour empêcher tout gène ciblé de devenir actif, et d'entraîner ainsi la non-fonctionnalité de ce gène. Les deux méthodes sont effectuées de la manière dont toute perturbation des traits génétiques existe dans une seule cellule, demeurant à son tour dans un plus grand *pool* contenant toutes les cellules individuellement perturbées: les mutants. En soumettant simultanément ces différents mutants à des conditions distinctes, il devient possible de déterminer quels traits génétiques sont nécessaires à la survie dans n'importe quel environnement testé, car toute mutation dans n'importe quel endroit important du génome sera lentement éliminée de la population du *pool* de mutants. La détermination des mutations présentes ou non présentes à la fin de l'expérience indique quels gènes sont essentiels ou non. Ceci est accompli en séquençant tous les mutants présents dans le *pool* et en comparant leur abondance relative au même *pool* de mutants présents au départ.

Au chapitre 2, nous avons abordé la manière dont ces données de séquençage peuvent être démultiplexées à partir des millions de séquences correspondant aux mutants du *pool*, en données lisibles par l'homme. À cette fin, nous avons développé 2FAST2Q, un programme informatique intuitif qui peut compter les abondances de n'importe quelle séquence, et donc des mutants dans le *pool*, en utilisant plusieurs étapes de filtrage de séquence définies par l'utilisateur. Nous illustrons comment

2FAST2Q peut être utilisé et comment différents paramètres d'utilisation impactent l'analyse en aval.

Dans le chapitre 3, nous avons développé un pipeline informatique capable de déterminer les gènes essentiels sans avoir besoin de comparer les populations de départ et d'arrivée des mutants. Cela est possible lors de l'utilisation de la méthode Tn-seq en déterminant où tous les transposons se sont insérés dans le génome. Dans un *pool* de mutants transposons, un gène sans transposons dans un gène donné, mais avec des transposons à une fréquence suffisamment élevée dans son entourage, impliquerait que ce gène devrait également porter des transposons, et que ne pas en observer signifierait que les cellules portant ces insertions n'ont pas survécu et que ce gène est essentiel. Le pipeline développé détermine individuellement, pour tous ces cas, si un gène est essentiel ou non à l'aide d'un algorithme d'auto-optimisation basé sur des gènes essentiels connus présents dans la plupart des organismes. En utilisant cette méthode appelée TnSeeker, nous avons déterminé tous les gènes essentiels communs dans une collection de 8 souches différentes d'*Escherichia coli*. Nous avons observé que toutes les souches présentent un chevauchement réduit dans les éléments essentiels communs, la plupart d'entre eux étant liés à la synthèse des protéines. À l'inverse, d'autres fonctions cellulaires essentielles connues telles que la maintenance de l'ADN n'étaient présentes que lorsque les souches étaient soit considérées par elles-mêmes, soit comme la somme de toutes. Un tel phénomène pourrait potentiellement être lié à la mobilité évolutive des gènes essentiels via des éléments mobiles, tels que les virus. En raison du brassage des gènes entre les organismes, il est possible que les espèces et les souches aient des fonctions cellulaires essentielles exécutées par différents gènes, réduisant ainsi toute similitude génétique et perturbant l'analyse des gènes essentiels partagés.

Lors de la création d'un *pool* de mutants, il peut être souhaitable de les trier individuellement et de créer une collection individuelle, où des mutants caractérisés uniques peuvent être utilisés isolément pour des expériences spécifiques. Dans le chapitre 4, nous avons adapté une méthode, connue sous le nom de SUDOKU, pour organiser des mutants inconnus dans des emplacements connus au sein d'une matrice de plaques. Nous avons efficacement classé une bibliothèque de transposons de mutants *E. coli* mélangé dans un *pool* dans sa forme de mutant unique isolé et organisé, créant ainsi une collection de mutants individuels.

Enfin, au chapitre 5, nous avons utilisé la méthode CRISPRi-seq pour explorer comment le pathogène opportuniste *Streptococcus pneumoniae* habitant le nasopharynx humain interagit avec un autre pathogène commun, *Staphylococcus aureus*. Nous avons observé qu'une pompe à efflux, SPV 686/7/8 (rebaptisée ici ArpABC), auparavant impliquée dans la détoxification des bactéricides (molécules à activité bactéricide), ne devient indispensable à pH 6 qu'en présence de *S. aureus*. Cela met en évidence l'impact de l'étude de l'essentialité conditionnelle, ou comment l'essentialité change en fonction de différentes conditions, souvent radicalement différentes de celles généralement utilisées dans les laboratoires. En effet, tout au long de cette thèse, nous explorons comment l'essentialité est une fonction de l'environnement, où chaque lignée d'organisme est à la fois adaptée de manière optimale et unique à tous les différents scénarios qu'elle rencontre dans son environnement naturel. Ensemble, ces données peuvent être utilisées pour tester de nouvelles cibles antibiotiques, qui perturbent souvent les fonctions cellulaires essentielles, ou même être utilisées pour déterminer des molécules de ciblage spécifiques à une souche ou une espèce.

Les techniques utilisées dans cette thèse ne sont qu'une dimension de la caractérisation mondiale sans précédent des bactéries en cours. Alors que des quantités inégalées de données continuent d'être générées, nous sommes maintenant plus proches que jamais d'intégrer pleinement nos connaissances collectives et de franchir enfin la prochaine étape de la biologie : la compréhension et la modélisation complètes des bactéries, peut-être les organismes les plus simples et les plus performants de la planète.

Acknowledgements (and a short personal story)

My Ph.D. journey had several official starts. Depending on who you ask, people might say it started in either January 2017, 2018, or 2019. At least the month will be correct, and all will agree on when it ended, 2023.

In truth, the story began on the 16th of June 2016, when I was accepted to integrate the XXth edition of the GABBA program at Porto University, in Porto, Portugal. During my interview process there I meet the GABBA coordinator, António Amorim. He enthusiastically argued against all my Master´s research project, and inadvertently taught me my first Ph.D. lesson: Don´t be afraid to stand up for your data, or what you think is correct. I hated that interview overall, but it was one of the most enlightening experiences I ever had. Thank you **António**, for being such a strong example, and pushing your students into their own paths.

In January 2017 I moved from Lisbon to Porto. In there, I joined the 8 other GABBA fellowship holders and started what would be 7 months of continuous classes developing both soft and hard skills. During this period, the 9 of us spent the better part of most days clumped together in a small stuffy room, working on assignments, and practicing presentations. We would then often continue and party until late on Wednesdays and Fridays. At some points we were spending more time together than apart, and we quickly became a family. I do not think I ever had as much fun in 6 months as when I was with them, there. Thank you so much, **Abel**, **Catarina**, **Márcia**, **Marta**, **Rita**, **Rui**, **Tiago**, and **Vanessa**. We will always have our ´sopa´ & Gin Fridays at the ´Pinguim´, and all those clubbing nights. In more detail, I would personally like to thank **Abel**, for being the most stylish GABBA; **Catarina**, for being one of the most down to earth people I ever met; **Márcia**, for not letting anyone forget about *Drosophila*; **Marta**, the best drunk contemporary dancing partner anyone could ever ask for; **Rita**, the grown up of the group; **Rui**, for always being up for anything; **Tiago**, for all the amazing conversations and insights at ´Pinguim´; **Vanessa**, for all the hugs, and always taking care of all of us. To all of you: Once a GABBA, always a GABBA.

The GABBA fellowships, now discontinued due to governmental changes, were given to people, not institutions. In practice this allowed for the Ph.D. project to be carried out in any lab of the student´s choosing. As the fellowship followed the person, it was expected that the students would take the initiative and arrange some kind of research venture for themselves. And so, it was. After a couple of months of prospective lab visits, I met Athanasios Typas, aka Nassos, at his EMBL lab in

Heidelberg, Germany. Nassos agreed to accept me in his lab by offering me an exciting dual supervised Ph.D. project. I would thus start my research at EMBL, and then continue it at UNIL, in Lausanne, Switzerland, as part of a multi lab collaboration project. At that point, the Lausanne lab, directed by Jan-Willem Veening, would then become my main lab, and UNIL my Ph.D. university. This is the reason this thesis has two co-supervisors, and why most people reading this in Lausanne might be, until this point in the story, confused by the lack of a connection between these described events, and my life as a UNIL Ph.D. candidate. I hope the link is clearer now.

In January 2018 I moved to Heidelberg, Germany, to start working on what would eventually be the 3rd and 4th chapter of this thesis. In there I saw more snow than I had ever seen before, but also saw the sun less than I had ever seen before. Such a transition was rough, but I quickly adapted to the warm EMBL and Typas lab culture, especially due to **Philipp**, the best friend any expat could ask for. Philipp introduced me to all EMBL people he could find, showed me around the Heidelberg region, included me in all the Ph.D. parties, and, most importantly, listened to my weekly frustrations. Thank you, Philipp, without you my experience in Germany would have been drastically different. At the Typas lab I also met loads of other amazing people: **Anna**, simultaneously the lab´s youngest Ph.D., and senior Ph.D. Thank you for always touring the Christmas markets, having ´feuerzangenbowle´(s), and going out for ramen in Frankfurt; **Elisabetta**, always busy everywhere, but always around for a nice chat; **Jacob**, a very stubborn Greek with a heart of gold; **Karin**, for all the super fun and easy going chats, and cool hikes around the Neckar; **Mathilda**, for looking out for all of us; **Sarela** and **Chris**, for their lightheaded Spanish mood. **Bede**, for his hilariously inappropriate jokes delivered with a deep Kiwi accent. **Alexandra**, for always having my back, and supporting me when needed. **Nazgul**, the most successful crypto girl around, and the best for existential late night Instagram conversations; And finally to all the other Typas lab members I have meet over the years. You all always lived by the motto ´work hard, party hard´, and I love you for it.

I would like to give a special mention to the people at the **EMBL Salsa club**, it was always my favorite night of the week, great dancers, and loads of fun; the **EMBL Swing dancing club**, for not being scared of shouting “frei machen” when needed; and my **German teacher**, which I saw twice a week for 1 year, and was the best language teacher I ever had. *Danke schön, ich habe etwas Deutsch gelernt*.

After having spent 1 year living in Heidelberg, I permanently moved to Lausanne in the first or second week of January 2019, and officially started (again) my Ph.D. in Jan-Willem´s lab at UNIL. In order to continue my research projects, and analyze the resulting data, I started digging into probably the most life changing professional skill I ever had the opportunity to learn: Python programing. Indeed, prior to 2019 I hardly knew more than how to open a blank windows terminal, and now I am the author of a published Python module. I consider obtaining fluency in Python my biggest and most useful Ph.D. achievement. Thank you Guido van Rossum for inventing such an intuitive programing language: Python.

The first person I properly met at UNIL was **Vincent**, with whom I shared a small windowless corner conference room for 1 year. Vincent was, and still is, my go to person when I want to discuss research ideas or ask about statistical approaches to problems. Despite preferring R over Python, he is a very chilled dude, and I appreciate his around the clock availability to always take time to help someone out.

During 2019 I was mostly obsessed with learning Python and trying to solve my research problems (chapter 3 and 4) in this isolated office and was hardly ever in the lab. I confess this might have made me look a bit isolated, but the easy-going people of the Veening lab continued to reach out and soon enough I was feeling right at home. For this, I would like to give a special thanks to **Julien**, for supporting all the lab´s nerdiest needs and conversations. Never change your humor type, and thank you for all the casual geekiness; **Clément**, for being one of the main lab party instigators, and always being up for anything risky; **Xue**, for her point-blank roasting humor, having my back in the lab, and always being supportive in her unique way. Two things are certain: you will be a great PI, and I will always have my fur; **Lance**, for being the unbeatable lord of the rings; **Jun**, for his quiet but sharp demeanor; **Paddy**, the best hiking/camping/skiing/chatting buddy around. You rekindled my wanderlust, and made me discover parts of the world I would have never seen, and experiment sports I would have never tried. Thank you for teaching me how to ski, and always taking me along for your mountain hiking expeditions. You are always in a good mood, and we always have nice easy-going chats; **Dimitra**, for always being the pre- and -after party/clubbing organizer. The group will always remember how some nights were forgotten... Thank you for dragging me into that Covid winter running session with Bob that the both of us mildly hated. That set me off in the course to eventually running the marathon, and becoming a healthier person. You also made me seriously think into

my post Ph.D. future at the right time, and for that, I am also very grateful; **Hammam**, he never misses a beat, and is always ready and willing to drag us wherever there is music and dancing; **Rita**, for always being funny, and a great dancer; **Ophélie**, for managing to find overnight sleeping trains against all odds, and being 2gangsta; **Liselot**, for all the tiktoks, and just putting up with me; **Nádia**, who accompanied me into the world of stand-up-paddleboarding in the crystal waters of lac Léman; **Nina**, and her extremely joyful mood, and obsession with my flowy red pants; **Vik**, the starboy. Never stop doing your thing and being a trailblazer; **Axel**, for keeping the order; **Jesús**, for his willingness to fuck up our enemies; **Jessica,** and her eager hyper-activeness; **Monica**, for helping us all, and not being too harsh on lab coat regulations; **Bevika**, for steadily overcoming her fears; **Maria**, for being born a mere 2h before I was, and therefore, being a cool person; **Johann**, for always trying to join and never having an excuse to not have a drink; **Amelieke**, for being so fun to talk to; To all other past and current Veening lab members, thank you for your help.

I would also like to say a few words to **Christopher**, my first master student: You were a great dedicated student, and I am proud of what you accomplished. I hope I helped you along in your path, wherever it may lead you.

To **Nassos** and **Jan-Willem**, the 3 of us embarked on this somewhat bumpy journey together, but what project isn´t. You both always accepted me and provided me with all the space and time needed to not only pull this together, but also see different parts of the world in the process. If it weren´t for you, I wouldn't have met most of the here mentioned people, or developed myself into what I am today. For that, I am forever grateful. Thank you.

Now, progressively going back in time, I would like to give a special thank you to some old friends. **Jorge**, for being the eternal bike touring partner, and overall chill dude that he is. **Alicia**, for unwillingly setting this Ph.D. story in motion. I hope you are all right, wherever it may be. The BCM lads, **André**, **Daniel**, and **Diogo**. We might have become a bit distant, but when we are together it's like no time has passed at all, and we are all back at the FCT canteen, eating those disgusting desserts. In particular, **André**, for having a heart of gold and always being there no matter what; **Daniel**, for having the guts to continue chasing his dreams; **Diogo**, for entrusting me with the honor of being the official translator at his wedding; Lastly, to my oldest friend, **João Miguel**, may you always be on the light-side.

A deep thank you to all the friends and acquaintances not here mentioned, current and past, may you always have good fortune. To all the haters, may you politely fuck off.

A special thank you to my aunt **Clara**, for caring.

To my **brother**, may you find your path.

To my **father**, for always encouraging me.

To my **mother**, I think this thesis is for you.

Table of Contents

This page is intentionally left blank.

Chapter 1

General Introduction

The quest for artificial mutagenesis

To understand what essential genes are, and how they are determined, is to explore how the discipline of genetics came to be. The landmark paper by Gregor Mendel linking hereditarily with ‘particulate units’, and the later rediscovery of his work by Hugo de Vries, Carl Correns, and Erich von Tschermak, prompted early XXth century biological experimentalists to ponder how external factors could influence the natural genetic state of an organism (DeMarini, 2020; Wassom, 1989). Indeed, de Vries seems to have predicted the current course of molecular biology when he wrote, in 1901:

> *Knowledge of the principles of mutation will certainly sometime in the future enable a fully planned artificial induction of mutations, i.e., the creation of new properties in plants and animals. Moreover, man will likely be able to produce superior varieties of cultivated plants and animals by commanding the origin of mutations. (Vries, 1901)*

Initial genetics works progressed slowly as mutation studies relied on visually assessing phenotypic variations occurring by natural mutagenesis, an inefficient, serendipitous, and slow process. Indeed, it was not until 1927 that the first mutagen, the X-ray, was undoubtedly confirmed to increase the mutation rate in *Drosophila melanogaster*, the first widespread genetic model (Muller, 1927). The discovery of the mutagenic effects of UV radiation, and the first chemical mutagen, mustard gas, soon followed. Artificial mutagenesis was thus achievable, increasing the throughput of mutational studies by several orders of magnitude, and ultimately boosting genetic research by providing insight into mutation types, genetic recombination, and the nature of the ‘particulate units’ – genes themselves (Wassom, 1989).

Notably, despite the concept of a ‘transforming principle’: the notion that bacteria are capable of transferring genetic information (Griffith, 1928); and the acceptance that genetic information, and thus genes, reside on the chromosomes (Morgan, 1911); many of such works proceeded without the idea of DNA as the cellular material controlling inheritance. Indeed, such was only concisely demonstrated in 1944, when Oswald Avery, Colin MacLeod, and Maclyn McCarty reported the

transformation of unencapsulated Pneumococcus cells into encapsulated cells, by the sole addition of DNA (Avery *et al.*, 1944).

Transposon mutagenesis

By the late 1940's, X-ray based genetic characterization work in several organisms, most notably on *D. melanogaster*, had led to the description of the 'position-effect'[1]: the concept of a gene's effect being dependent on its relative position in the chromosome (Lewis, 1950). Barbara McClintock, then studying the genetic composition of the short arm of chromosome 9 in maize, and intrigued by this non-organism-specific phenomenon, characterized a new type of 'position-effect': mutations similar to the ones produced by known mutagens were consistently being observed in the same specific *locus*. Moreover, such events were associated with the insertion of 'chromatin' adjacent to the *locus* showing 'position-effect' (Mcclintock, 1950). The first notion of transposon had thus appeared, but it was not until several decades later, in the 1970's, that the word, or concept, would start being widely used and understood, particularly when describing the translocation of drug-resistance elements from prokaryotic plasmids and bacteriophages (Berg *et al.*, 1975; Dougan & Sherratt, 1977; Heffron *et al.*, 1975; Kleckner *et al.*, 1975; Kleckner *et al.*, 1977; Ptashne & Cohen, 1975).

Currently, transposons are defined as mobile DNA delimited by terminal inverted repeats, which are used by transposases to mediate their own transposition between nonhomologous insertion sites. Transposases are thus enzymes capable of mediating the excision and reintegration of a transposon. Transposon replication can either be 'cut-and-paste', if the sequence is removed to a new *locus*, or 'copy-paste', if a new sequence is copied to a new place, while maintaining the original transposon insertion site (Craig, 1997; Reznikoff, 1993; Sandoval-Villegas *et al.*, 2021).

Of particular early importance are the transposable element 10 (Tn10), encoding a tetracycline resistance marker, from which the *tetR* inducible system

[1] At the time, 'position-effect' seemed to encompass several different phenomena, however, the described effect in regards to the white/red phenotype in *D. melanogaster* are now known to be mostly derived from gene silencing due to a change in the position of a gene due to chromosome translocation (position-effect variegation) (Elgin & Reuter, 2013)

originates (Beck *et al.*, 1982; Kleckner *et al.*, 1975); and Tn5, containing a kanamycin resistance marker (Berg *et al.*, 1975; Berg *et al.*, 1982).

Groundbreaking work by Nancy Kleckner *et al.* was perhaps the first described instance of a transposon based whole-genome mutagenesis assay. Using a bacteriophage as the delivery vector of Tn10 into *Salmonella*, several independent transductants were isolated, collected, and screened for different auxotrophic mutations. The resulting auxotrophic mutants were in line with those obtained when performing the standard chemical mutagenesis assay, proving that transposons are capable of randomly inserting into many different genomic sites, and of causing gene loss of function phenotypes. However, the biggest fundamental difference between transposon and chemical mutagenesis was the relative easier method of transposon insertion mapping, and thus of gene characterization using at-the-time methods. The existence of known constant transposon sequences, and the carriage of an antibiotic marker, implied that homology, restriction mapping, and selection could now be used to more easily further study, move or delete known/unknown *loci* (Kleckner *et al.*, 1975; Kleckner *et al.*, 1977). Tn10 insertion sites, however, were readily shown to not be as randomly distributed into the genome as the Tn5 generated ones. The same was also observed for other contenders: Tn3, Tn9, and phage Mu. Tn5, also due to its early characterization and high insertion frequency, thus became the preferred transposon when performing transposon mutagenesis. Indeed, by the mid 1980's, Tn5 mutagenesis, aided by the recently invented nucleotide sequencing (Maxam & Gilbert, 1977; Sanger *et al.*, 1977), had already been performed in several different bacterial species, and helped map dozens of new genes and functions (de Bruijn & Lupski, 1984).

However, despite the existence of several well characterized transposons, random transposon mutagenesis was still a technique not easily transposed to non-model bacteria. In the case of Mycobacteria and several Gram-positive bacteria, the use of species-specific transposons, often not available and not having the same manipulation conveniences as the now easily engineered Gram negative transposons, was usually required. The long expected breakthrough arrived in the late 1990's, with the development of *in vitro* transposition using the horn fly *Haematobia irritans* transposon *Himar1* (Lampe *et al.*, 1996). *Himar1*, inserting only into TA sites, had previously been demonstrated to be able to transpose in both insects and protozona,

suggesting a broad spectrum of use. Moreover, *Himar1 in vitro* transposition requires the presence of only a single protein. These factors were quickly adapted for the development of an improved transposon mutagenesis assay, GAMBIT, firstly applied in *Haemophilus influenzae* and *Streptococcus pneumoniae* (Akerley *et al.*, 1998), and then in Mycobacteria (Rubin *et al.*, 1999). This assay was probably the first where a transposon high insertion saturation was required for gene essentiality inference, and the first resembling modern transposon mutagenesis assays. GAMBIT allowed for the easier mapping of essential *loci* by leveraging PCR with an anchored chromosomal primer and a transposon specific primer, with the length of all fragments being determined by electrophoresis. If the insertion of transposons did not cause loss of viability in a particular *locus*, a sequential increase in PCR fragment length, corresponding to the continuous detection of transposons increasingly further from the anchor primer, should be seen. If the transposon landed on a non-viable *locus*, a gap in this continuously increasing PCR length map would be detected, and the *locus* would be deemed essential. This contrasted with at-the-time transposon methods for detecting genes with non-viable mutations, which relied on the selection and isolation of conditional essential mutants, or required wild type complementation (Akerley *et al.*, 1998).

Essential genes

At this point in time, gene essentiality inference across all domains of life using different methods greatly accelerated, so a clearer definition on this topic is required before proceeding further. As initially described in 1961 by Salome Gluecksohn-Waelsch:

> *In bacteria, a mutation is considered lethal if growth and survival of the mutant organism cannot be achieved by manipulation of the environment. If, on the other hand, it is possible to devise environmental conditions which permit the organism to survive, the mutation becomes 'viable' and therefore may be called a 'conditional lethal' mutation.* (Gluecksohn-Waelsch, 1961, p. 2)

60 years later the concept remains mostly unchanged: An essential (lethal) gene is a gene whose disturbance, in any given environment, causes either serious growth impairment or cell death. Or, in a broader sense: Under any given environment, there are genetic *loci* required for sustaining life, and others which are not.

The next-generation sequencing booster

The early 2000's brought about major efforts in full genome sequencing, and in the generation of complete gene-deletion mutant collections/libraries in model organisms like *Escherichia coli* and *S. cerevisiae* (Baba *et al.*, 2006; Giaever *et al.*, 2002; Tong *et al.*, 2001). The assembly of double gene-deletion libraries quickly followed, permitting the high-throughput analysis of genetic interactions such as synthetic lethality (Butland *et al.*, 2008; Tong *et al.*, 2001; Tong *et al.*, 2004; Typas A, 2008). These and similar methodologies allowed a peak of unprecedented detail into gene essentiality. However, to this day, mass application of these techniques remains unfeasible for most organisms.

Contemporary to the first gene-deletion mutant collections, microarrays, a method by which the amount of specific DNA sequences can be quantified based on fluorescence, had start to gain popularity (Schena *et al.*, 1995). By hybridizing the fluorescently labelled transposon bordering sequences against known sequences, normally from specific chromosomal *loci*, microarrays allow for the relative differential quantification of said sequences between different conditions. More abundant

transposon mutants would result in higher intensity fluorescent signals, and vice versa. By using known sequences from *loci* of interest where each sequence was nested within a micro array format, the simultaneous individual assessment of the relative fitness of hundreds of genetic *loci*, from transposon mutagenesis experiments, was finally achievable.

Such breakthrough had thus allowed, for the first time using what could be considered a high-throughput format, the genome wide level characterization of condition specific bacterial adaptations. Eventually, several variations of this combination of saturating transposon mutagenesis with a microarray readout were developed. In 2001, TraSH (transposon site hybridization) appeared (Sassetti *et al.*, 2001), in 2004, MATT (Microarray tracking of transposon mutants) (Salama *et al.*, 2004), and in 2007, GAF (genomic array footprinting) (Bijlsma *et al.*, 2007), with several works subsequently leveraging these, or other method variations, for novel gene essentiality and pathway discovery (Chan *et al.*, 2005; Girgis *et al.*, 2007; Joshi *et al.*, 2006).

The first decade of the XXIst century brought about new sequencing technologies. Indeed, the advent of next-generation sequencing (NGS), following the 'first-generation' Sanger method, has brought significant experimental paradigm shifts in biology. For the first time, millions of nucleotide base pairs could be automatically sequenced in a few days, generating what is currently known as 'big-data'. To handle such developments, bioinformatics was pushed to the forefront of biology. Indeed, by providing general sequencing analysis pipelines, or by handling 'big-data' via custom made software, bioinformatics is increasingly becoming more and more entwined with biological advances (Barba *et al.*, 2014). In **chapter 2**, we explore these methods further by presenting 2FAST2Q, a sequencing data counting program (Bravo *et al.*, 2022).

Despite the relative success of microarrays, the ever decreasing cost of NGS soon enabled another revolution by allowing the microarray part be skipped, and further increasing throughput. Indeed, several NGS based transposon mutagenesis methods, collectively known as Tn-seq, emerged in 2009: High-throughput INSeq (Goodman *et al.*, 2009); High-throughput insertion tracking by deep sequencing' (HITS) (Gawronskia *et al.*, 2009); Transposon directed insertion-site sequencing (TraDIS) (Langridge *et al.*, 2009); and Transposon sequencing (Tn-seq) (Gawronskia

et al., 2009; Goodman *et al.*, 2009; Langridge *et al.*, 2009; van Opijnen *et al.*, 2009). Similar to the microarrays, these allowed for unparalleled genome-wide gene essentiality and fitness studies in a wide range of species, under virtually infinite different conditions, but eventually in a more streamlined and cheaper format (Chao *et al.*, 2016; Deutschbauer *et al.*, 2014). Tn-seq is described in detail in **chapter 3**, but the overall procedure is similar to the previously mentioned GAMBIT method, although with the PCR fragments undergoing NGS, instead of being assessed by electrophoresis (figure 1).

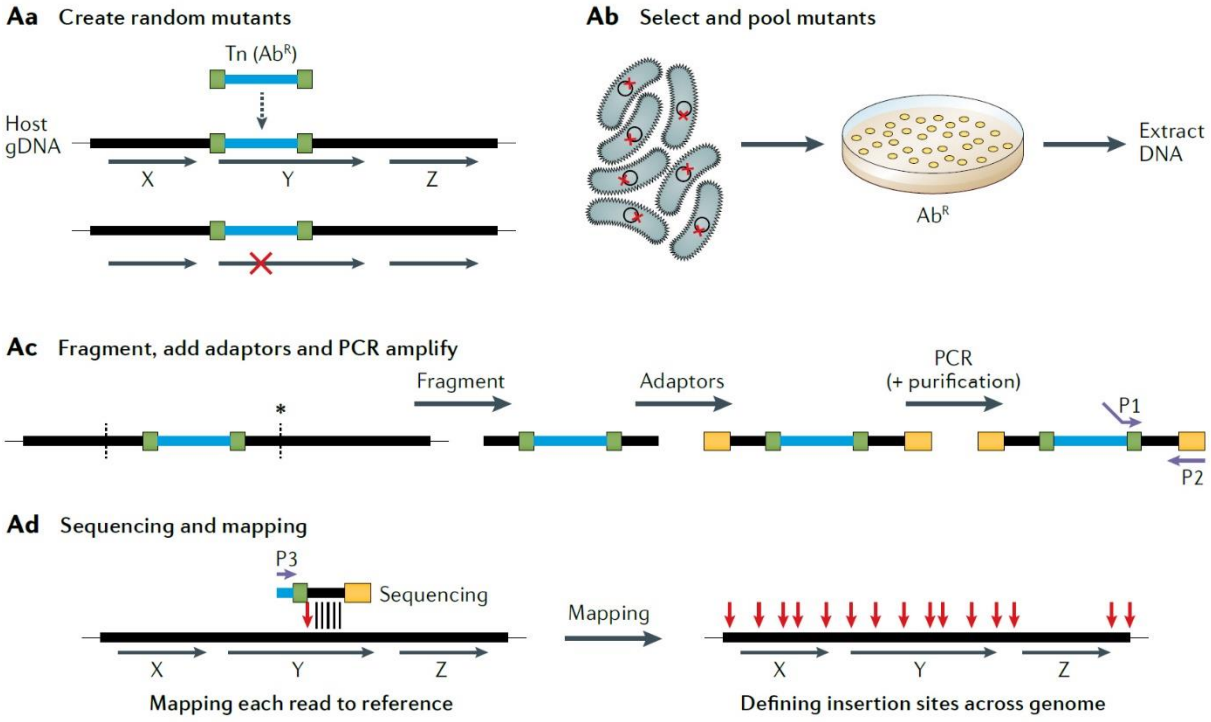


Figure 1 | Basic Tn-seq method schematic.

Adapted from Cain *et al.* (Cain *et al.*, 2020)

The CRISPR revolution

It has been a long-standing goal in biology to induce precise targeted mutations easily and reliably in the genome of any organism. The basis for such a method was serendipitously discovered in 1987 in *E. coli* (Ishino *et al.*, 1987), and independently found in several different bacteria/archaea species in subsequent years. However, it was not until 2002 that it was recognized as a distinct class of element, and appropriately named as Clustered regularly interspaced short palindromic repeats, CRISPR (Jansen *et al.*, 2002).

Further works soon discovered that some regions (spacer regions) of CRISPR had similarity with some sequences of bacteriophages/plasmids, and that several genes in the system, the CRISPR-associated (Cas) genes, were involved in CRISPR activity. Later comparative genomics analyses demonstrated that CRISPR-Cas is a Prokaryotic acquired immunity system against invading viruses and plasmids, similar to the Eukaryotic RNA interference system (Ishino *et al.*, 2018; Makarova *et al.*, 2006). It is now known that CRISPR-Cas can recognize and cleave DNA via the concerted action of double-strand break inducing Cas protein(s), and the homology-based targeting of RNAs. Upon infection of a host by foreign DNA, sequences known as protospacers are acquired from this DNA and introduced into the bacterial genome at the CRISPR locus as a spacer sequence. These protospacers are sourced from regions flanked by protospacer adjacent motifs (PAM), a 2-5bp long sequence that varies across bacteria. Upon subsequent infections, the host can use the expressed sequence from these spacers as guidance for Cas mediated DNA cleavage, thus avoiding re-infection. As the host does not display a PAM sequence adjacent to the spacer, activity against self, and thus cell death, is bypassed. CRISPR-Cas is therefore considered to be a form of inherited bacterial acquired immunity. Several types of CRISPR-Cas systems exist, however, the type II CRISPRi-Cas9 from *Streptococcus pyogenes* is probably one of the simplest. It consists of a three-component system between Cas9, a single multi-functional effector, and two RNAs: CRISPR RNA (crRNA), the spacer; and the trans-activating crRNA (tracrRNA), that links Cas9 with the crRNA (Ishino *et al.*, 2018; Jinek *et al.*, 2012a; Kozovska *et al.*, 2021; Zhang *et al.*, 2021).

In 2012, the mentioned CRISPR-Cas9 system was for the first time adapted in short succession by 3 independent groups for gene engineering (Cong *et al.*, 2013;

Jinek *et al.*, 2012b; Mali *et al.*, 2013). In this case the system was further simplified into a 2 component system by linking both crRNA and the tracrRNA into a single molecule, named single guide RNA (sgRNA). By modifying the sequence of this sgRNA, it is possible to direct the Cas9 into cleaving the DNA at any given permissible *locus*, creating double-strand-breaks (DSB), and inducing the cellular repair mechanisms. Depending on whether a donor DNA template is provided (artificially/naturally) or not, DSB are either repaired by non-homologous end joining (NHEJ), resulting in genomic deletions, or homology directed repair (HDR), leading to precise substitutions (Kozovska *et al.*, 2021). Despite revolutionary in the Eukaryotic field, where gene manipulation and editing remained troublesome, such accomplishments were not as groundbreaking in bacteria, where simple homologous based methods were already in use. Moreover, most bacteria lack the NHEJ system, with Cas9 chromosome cleavage often resulting in cell death (Cui & Bikard, 2016).

CRISPR(i-Seq): Another mutagenesis method, or a worthy successor?

Observations that mutations in the Cas9 nuclease domains resulted in inactivation of DNA cleavage whilst maintaining DNA binding (Jinek *et al.*, 2012b), quickly prompted the development of CRISPR interference (CRISPRi) (Bikard *et al.*, 2013; Qi *et al.*, 2013).

CRISPRi uses a nuclease inactivated (dead) Cas9 (dCas9), in conjunction with a designed sgRNA, to repress the expression of targeted genes by blocking the binding of DNA binding proteins such as RNA polymerases. By having dCas9 expression and/or the sgRNA under the control of an inducible promoter, the obtained knockdown is both inducible and reversible. Through modulation of the time at which a knockdown is created, CRISPRi can in this way enable the study of essential genes, whose knockout is lethal, and thus difficult to study using other methods.

It has recently become feasible to synthesize large sgRNA libraries that target all permissible *loci* in a genome. Upon introducing these sgRNA *en masse* into an organism expressing dCas9, a pooled CRISPRi library is obtained. Theoretically, each individual cell in this pool will then display a different sgRNA, and thus be able of having a knockdown for the sgRNA targeted *locus* (figure 2). When submitting the CRISPRi mutant pool to a selective pressure, such as sub-inhibitory concentrations of antibiotic, cells will compete locally based on the fitness effect of the knockdowned

locus. Over generations, a worse cell fitness will result in lower amounts of that same mutant and thus of the sgRNA, and vice-versa. At the end of the challenge, all the sgRNAs in the pool are sequenced by NGS and counted. In **chapter 2**, we explore the sgRNA counting process in detail with the python program 2FAST2Q. Finally, differential analysis is performed comparing different conditions of the same CRISPRi mutant pool, including induced vs. non induced, resulting in a relative fitness value for every mutant in the library. (Bock *et al.*, 2022; de Bakker *et al.*, 2022; Liu *et al.*, 2017; Peters *et al.*, 2016; Rousset *et al.*, 2021). In **chapter 5**, the method CRISPRi-seq (CRISPRi coupled with NGS, as described (de Bakker *et al.*, 2022)) is used to examine the interactions between *Streptococcus pneumoniae* and competing species.

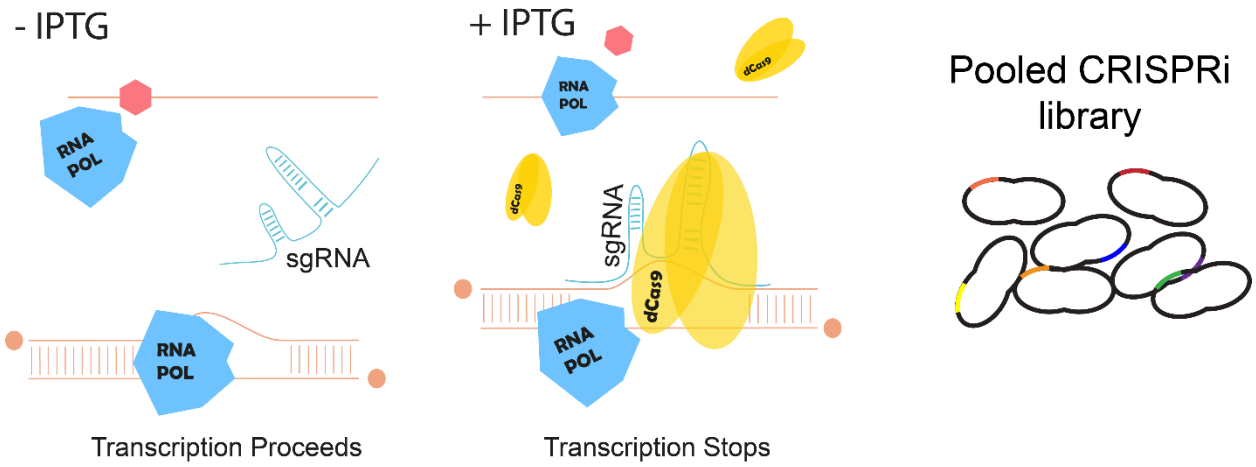


Figure 2 | CRISPRi-seq library.

dCas9 is used to prevent RNA polymerase DNA binding at the sgRNA binding place in the presence of an inducer (IPTG). By having multiple sgRNAs, it is possible to create a pooled library where each mutant targets a different *locus*.

CRISPRi-seq Vs. Tn-seq: Different tools for different jobs

Several methods (or variations of such) currently exist for assaying essentiality and genetic interactions: the creation of pure gene knock-out libraries such as the KEIO collection (Baba *et al.*, 2006); knock-down strategies such as CRISPRi-seq (Bikard *et al.*, 2013; de Bakker *et al.*, 2022); and high-density transposon mutagenesis (Tn-seq) (Gawronskia *et al.*, 2009; Goodman *et al.*, 2009; Langridge *et al.*, 2009; van Opijnen *et al.*, 2009). Among these, the two latter leverage pooled mutant libraries in conjunction with next generation sequencing (NGS) to determine *loci* essentiality in a high-throughput manner. Depending on the setup, both these techniques are able to detect changes in mutant fitness, and track gene essentiality shifts over time for any given experimental condition, and thus conditional essentiality (de Bakker *et al.*, 2022; Gallagher *et al.*, 2020; Liu *et al.*, 2017). However, unlike Tn-seq, where essential insertions are random and disappear from the initial population over time not allowing their relative fitness to be easily measured, CRISPRi-seq is inducible and reversible by default. This behavior permits the systematic measurement of all target genes fitness, essential or not, without loss. Moreover, due to requiring *de novo* sgRNA design, CRISPRi-seq can be used to create tilling libraries, targeting only genes of interest and thus reducing costs. However, when performed using a high enough saturation library, Tn-seq has the potential to assay essentiality over smaller *loci.* Indeed, the essentiality of small genome areas such as promoters, intergenic regions, and protein domains can be determined this way using a properly engineered transposon cassette, at the cost of requiring more transposon insertions and thus a larger sequencing capacity (Cain *et al.*, 2020). Such a screen is harder to implement using CRISPRi due to the requirement for large libraries without off-target sgRNAs (careful design is needed), a PAM sequence, and the possible existence of cofounding polar effects. This latter arises from dCas9 blocking the transcription of multiple genes that might be transcribed as single transcripts, such as operons (Liu *et al.*, 2017; Qi *et al.*, 2013; Zhang *et al.*, 2021).

Due to being based on the absence/presence of random insertions, Tn-seq data analysis is often case dependent, requiring convoluted data analysis. These pitfalls and current state-of-art are described in detail in **chapter 3 and 4.** CRISPRi-seq, due to being more systematic, especially when using reduced libraries with one sgRNA

per gene, have a (by comparison) simplified analysis procedure (see **chapter 2** and **5**).

The transposon intrinsic ability to target all cellular DNA, including extra chromosomal DNA, together with its reduced cost and easiness of implementation, places Tn-seq as an ideal first line screen tool in the study of novel genes across multiple uncharacterized species and strains. The CRISPRi-seq more expensive setup can then be used as a follow-up technique in selected organisms, or for selected pan-genome genes, to study essential gene fitness or as a cross-validation method.

As described, gene essentiality/lethality assaying techniques have defined genetics since its inception. Now, as biology moves past model organisms, embraces high-throughput condition testing, and enables the analysis of large interaction networks, such assays continue to be adapted and relevant. Indeed, an increasing body of literature demonstrates the plasticity and context dependence of gene essentiality, dramatically varying not only between different species, but also at the strain level, for any same condition (Carey *et al.*, 2018; Coe *et al.*, 2019; Poulsen *et al.*, 2019; Rancati *et al.*, 2018; Rosconi *et al.*, 2022; Rousset *et al.*, 2021). Ultimately, in-depth insight into how essentiality shifts across different conditions and organisms enables not only the identification of novel biotechnologically relevant gene functions and antimicrobial targets, but also helps further the quest for finding a minimal life sustaining genome (for any given condition).

Thesis Outline

From the early days of relying of natural mutation, to the current era of large-scale custom made mutant libraries, gene essentiality continues to be a relevant biological puzzle. Throughout this thesis, we explore this issue using complementary state-of-the-art methods: Tn-seq and CRISPRi-seq. With both approaches relying on NGS, we first present a sequencing data analysis program: 2FAST2Q. 2FAST2Q is a versatile and intuitive standalone program capable of extracting and counting feature occurrences, such as sgRNAs and bar-coded transposons, from sequencing files (**Chapter 2**). In **chapter 3**, we built pooled transposon mutant libraries for 8 *E. coli* strains, and developed a novel self-optimizing non-transposon-specific pipeline capable of inferring gene essentiality from transposon insertions. We present the respective benchmarking, the resulting *E. coli* pan-essentialome, and strain-specific essential genes. We further utilized one of these transposon libraries, from strain UTI89, and adapted a pooled library arraying technique known as SUDOKU to deconvolute this library into its arrayed format (**chapter 4**) (Anzai *et al.*, 2017; Erlich *et al.*, 2009). Moreover, we modified the technique in conjunction with 2FAST2Q to work with CRISPRi libraries, and randomly bar-coded transposons. Regarding this latter, we also demonstrate the uses and pitfalls of using such method to infer gene fitness, in our previously built libraries. Finally, in **chapter 5**, we leverage CRISPRi-seq methods to study *Streptococcus pneumoniae* gene essentiality in a competing environment with both *Staphylococcus aureus*, and/or human nasopharynx cells. We demonstrate genes that are beneficial in such conditions, and elaborate on their possible modes of action.

References

Akerley, B. J., Rubin, E. J., Camilli, A., Lampe, D. J., Robertson, H. M., & Mekalanos, J. J. (1998). Systematic identification of essential genes by in vitro mariner mutagenesis. *Proc. Natl. Acad. Sci, 95*, 8927-8932.

Anzai, I. A., Shaket, L., Adesina, O., Baym, M., & Barstow, B. (2017). Rapid curation of gene disruption collections using Knockout Sudoku. *Nat Protoc, 12*(10), 2110-2137. doi:10.1038/nprot.2017.073

Avery, O., MacLeod, C., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med, 79*(2), 137–158.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol, 2*, 2006 0008. doi:10.1038/msb4100050

Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses, 6*(1), 106-136. doi:10.3390/v6010106

Beck, C., Mutzel, R., Barbé, J., & Muller, W. (1982). A multifunctional gene (tetR) controls Tn10-encoded tetracycline resistance. *Journal of Bacteriology, 150*(2), 633-642.

Berg, D. E., Davies, J., Allet, B., & Rochaix, J.-D. (1975). Transposition of R factor genes to bacteriophage λ. *Proc. Nat. Acad. Sci., 72*(9), 3628-3632.

Berg, D. E., Johnsrud, L., Mcdivitt, L., Ramabhadran, R., & Hirschel, B. J. (1982). Inverted repeats of Tn5 are transposable elements. *Proc. Natl. Acad. Sci, 79*, 2632-2635.

Bijlsma, J. J., Burghout, P., Kloosterman, T. G., Bootsma, H. J., de Jong, A., Hermans, P. W., & Kuipers, O. P. (2007). Development of genomic array footprinting for identification of conditionally essential genes in Streptococcus pneumoniae. *Appl Environ Microbiol, 73*(5), 1514-1524. doi:10.1128/AEM.01900-06

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., & Marraffini, L. A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res, 41*(15), 7429-7437. doi:10.1093/nar/gkt520

Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., Stanley, G., Chen, S., Garnett, M., Li, W., Moffat, J., Qi, L. S., Shapiro, R. S., Shendure, J., Weissman, J. S., & Zhuang, X. (2022). High-content CRISPR screening. *Nature Reviews Methods Primers, 2*(1). doi:10.1038/s43586-021-00093-4

Bravo, A. M., Typas, A., & Veening, J.-W. (2022). 2FAST2Q: a general-purpose sequence search and counting program for FASTQ files. *PeerJ, 10*. doi:10.7717/peerj.14041

Butland, G., Babu, M., Diaz-Mejia, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O., Mori, H., Wanner, B. L., Lo, H., Wasniewski, J., Christopolous, C., Ali, M., Venn, P., Safavi-Naini, A., Sourour, N., Caron, S., Choi, J. Y., Laigle, L., Nazarians-Armavil, A., Deshpande, A., Joe, S., Datsenko, K. A., Yamamoto, N., Andrews, B. J., Boone, C., Ding, H., Sheikh, B., Moreno-Hagelseib, G., Greenblatt, J. F., & Emili, A. (2008). eSGA: E. coli synthetic genetic array analysis. *Nat Methods, 5*(9), 789-795. doi:10.1038/nmeth.1239

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & Opijnen, T. v. (2020). A decade of advances

in transposon-insertion sequencing. *Nat Rev Genet, 21*(9). doi:10.1038/s41576-020-0244-

Carey, A. F., Rock, J. M., Krieger, I. V., Chase, M. R., Fernandez-Suarez, M., Gagneux, S., Sacchettini, J. C., Ioerger, T. R., & Fortune, S. M. (2018). TnSeq of Mycobacterium tuberculosis clinical isolates reveals strain-specific antibiotic liabilities. *PLoS Pathog, 14*(3), e1006939. doi:10.1371/journal.ppat.1006939

Chan, K., Kim, C. C., & Falkow, S. (2005). Microarray-based detection of Salmonella enterica serovar Typhimurium transposon mutants that cannot survive in macrophages and mice. *Infect Immun, 73*(9), 5438-5449. doi:10.1128/IAI.73.9.5438-5449.2005

Chao, M. C., Abel, S., Davis, B. M., & Waldor, M. K. (2016). The design and analysis of transposon insertion sequencing experiments. *Nat Rev Microbiol, 14*(2), 119-128. doi:10.1038/nrmicro.2015.7

Coe, K. A., Lee, W., Stone, M. C., Komazin-Meredith, G., Meredith, T. C., Grad, Y. H., & Walker, S. (2019). Multi-strain Tn-Seq reveals common daptomycin resistance determinants in Staphylococcus aureus. *PLoS Pathog, 15*(11), e1007862. doi:10.1371/journal.ppat.1007862

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex Genome EngineeringUsing CRISPR/Cas Systems. *Science, 339*, 819-823.

Craig, N. L. (1997). Target site selection in transposition. *Annu. Rev. biochm., 66*, 437-474.

Cui, L., & Bikard, D. (2016). Consequences of Cas9 cleavage in the chromosome of Escherichia coli. *Nucleic Acids Res, 44*(9), 4243-4251. doi:10.1093/nar/gkw223

de Bakker, V., Liu, X., Bravo, A. M., & Veening, J.-W. (2022). CRISPRi-seq for genome-wide fitness quantification in bacteria. *Nature Protocols, 17*, 252–281

de Bruijn, F. J., & Lupski, J. R. (1984). The use of transposon Tn5 mutagenesis in the rapid generation of correlated physical and genetic maps of DNA segments cloned into multicopy plasmids — a review. *Gene, 27*, 131-149.

DeMarini, D. M. (2020). The mutagenesis moonshot: The propitious beginnings of the environmental mutagenesis and genomics society. *Environ Mol Mutagen, 61*(1), 8-24. doi:10.1002/em.22313

Deutschbauer, A., Price, M. N., Wetmore, K. M., Tarjan, D. R., Xu, Z., Shao, W., Leon, D., Arkin, A. P., & Skerker, J. M. (2014). Towards an informative mutant phenotype for every bacterial gene. *J Bacteriol, 196*(20), 3643-3655. doi:10.1128/JB.01836-14

Dougan, G., & Sherratt, D. (1977). The Transposon Tn 1 as a Probe for Studying ColE1 Structure and Function. *Molec. gen. Genet., 151*, 151-160.

Elgin, S. C., & Reuter, G. (2013). Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harb Perspect Biol, 5*(8), a017780. doi:10.1101/cshperspect.a017780

Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M., & Hannon, G. J. (2009). DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res, 19*(7), 1243-1253. doi:10.1101/gr.092957.109

Gallagher, L. A., Bailey, J., & Manoil, C. (2020). Ranking essential bacterial processes by speed of mutant death. *Proc Natl Acad Sci U S A, 117*(30), 18010-18017. doi:10.1073/pnas.2001507117

Gawronskia, J. D., Wonga, S. M. S., Giannoukosb, G., Wardb, D. V., & Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-

wide screen for Haemophilus genes required in the lung. *Proc. Natl. Acad. Sci., 106*(38), 16422-16427.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Ve´ronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., Bakkoury, M. E., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., ldener, U. G., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kötter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., & Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature, 418*, 387-391.

Girgis, H. S., Liu, Y., Ryu, W. S., & Tavazoie, S. (2007). A comprehensive genetic characterization of bacterial motility. *PLoS Genet, 3*(9), 1644-1660. doi:10.1371/journal.pgen.0030154

Gluecksohn-Waelsch, S. (1961). Lethal Genes and analysis of differentiation. *Science, 142*(3597), 1269-1276.

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., & Gordon, J. I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe, 6*(3), 279-289. doi:10.1016/j.chom.2009.08.003

Griffith, F. (1928). The significance of pneumococcal types. *J Hyg, 227*(2), 114-159.

Heffron, F., Rubens, C., & Falkow, S. (1975). Translocation of a plasmid DNA sequence which mediates ampicillin resistance: Molecular nature and specificity of insertion. *Proc.Nat.Acad.Sci., 72*(9), 3623-3627.

Ishino, Y., Krupovic, M., & Forterre, P. (2018). History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J Bacteriol, 200*(7). doi:10.1128/JB.00580-17

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., & Nakata, D. (1987). Nucleotide Sequence of the iap Gene, Responsible for Alkaline Phosphatase Isozyme Conversion in Escherichia coli, and Identification of the Gene Pqroduct. *Journal of Bacteriology, 169*(12), 5429-5433.

Jansen, R., Embden, J. D. A. v., Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology, 43*(6), 1565-1575.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012a). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science, 337*(6096), 816 - 821.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012b). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science, 337*, 816-821.

Joshi, S. M., Pandey, A. K., Capite, N., Fortune, S. M., Rubin, E. J., & Sassetti, C. M. (2006). Characterization of mycobacterial virulence genes through genetic interaction mapping. *PNAS, 103*(31), 11760-11765.

Kleckner, N., Chan, R. K., Tye, B.-K., & Botstein, D. (1975). Mutagenesis by

insertion of a drug-resistance element carrying an inverted repetition. *Journal of Molecular Biology, 97*(4), 561-575.

Kleckner, N., Roth, J., & Botstein, D. (1977). Genetic Engineering in Viva Using Translocatable Drug-resistance Elements. *J. Mol. Biol., 116*, 125-159.

Kozovska, Z., Rajcaniova, S., Munteanu, P., Dzacovska, S., & Demkova, L. (2021). CRISPR: History and perspectives to the future. *Biomed Pharmacother, 141*, 111917. doi:10.1016/j.biopha.2021.111917

Lampe, D. J., Churchill, M. E. A., & Robertson, H. M. (1996). A purified mariner transposase is sufficient to mediate transposition in vitro. *The EMBO Journal, 15*(19), 5470-5479.

Langridge, G. C., Phan, M. D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., & Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res, 19*(12), 2308-2316. doi:10.1101/gr.097097.109

Lewis, E. B. (1950). The Phenomenon of Position Effect. In (pp. 73-115).

Liu, X., Gallay, C., Kjos, M., Domenech, A., Slager, J., Kessel, S. P. v., Knoops, K., Sorg, R. A., Zhang, J.-R., & Veening, J.-W. (2017). High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. *Mol Syst Biol, 13*(5), 931. doi:10.15252/msb.20167449

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., & Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct, 1*, 7. doi:10.1186/1745-6150-1-7

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science, 339*(6121), 823-826. doi:10.1126/science.1232033

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Nati. Acad. Sci., 74*(2), 560-564.

Mcclintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A, 36*, 344-355.

Morgan, T. H. (1911). Random Segregation Versus Coupling in Mendelian Inheritance. *Science, 34*(873), 384.

Muller, H. J. (1927). Artificial transmutation of the gene. *Science, 66*(1699), 84-87.

Peters, J. M., Colavin, A., Shi, H., Czarny, T. L., Larson, M. H., Wong, S., Hawkins, J. S., Lu, C. H. S., Koo, B. M., Marta, E., Shiver, A. L., Whitehead, E. H., Weissman, J. S., Brown, E. D., Qi, L. S., Huang, K. C., & Gross, C. A. (2016). A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell, 165*(6), 1493-1506. doi:10.1016/j.cell.2016.05.003

Poulsen, B. E., Yang, R., Clatworthy, A. E., White, T., Osmulski, S. J., Li, L., Penaranda, C., Lander, E. S., Shoresh, N., & Hung, D. T. (2019). Defining the core essential genome of Pseudomonas aeruginosa. *Proc Natl Acad Sci U S A, 116*(20), 10072-10080. doi:10.1073/pnas.1900570116

Ptashne, K., & Cohen, s. N. (1975). Occurrence of insertion sequence (IS) regions on plasmid deoxyribonucleic acid as direct and inverted nucleotide sequence duplications. *J. Bact, 122*(2), 776-781.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-

guided platform for sequence-specific control of gene expression. *Cell, 152*(5), 1173-1183. doi:10.1016/j.cell.2013.02.022

Rancati, G., Moffat, J., Typas, A., & Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. *Nat Rev Genet, 19*(1), 34-49. doi:10.1038/nrg.2017.74

Reznikoff, W. S. (1993). The tn5 transposon. *Annu. Rev. Microbiol, 47*, 945-963.

Rosconi, F., Rudmann, E., Li, J., Surujon, D., Anthony, J., Frank, M., Jones, D. S., Rock, C., Rosch, J. W., Johnston, C. D., & van Opijnen, T. (2022). A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat Microbiol*. doi:10.1038/s41564-022-01208-7

Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernandez-Rodriguez, J., Clermont, O., Denamur, E., Rocha, E. P. C., & Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the Escherichia coli species. *Nat Microbiol, 6*(3), 301-312. doi:10.1038/s41564-020-00839-y

Rubin, E. J., Akerley, B. J., Novik, V. N., Lampe, D. J., Husson, R. N., & Mekalanos, J. J. (1999). In vivotransposition of mariner-based elements in entericbacteria and mycobacteria. *Proc. Natl. Acad. Sci., 96*, 1645-1650.

Salama, N. R., Shepherd, B., & Falkow, S. (2004). Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. *J Bacteriol, 186*(23), 7926-7935. doi:10.1128/JB.186.23.7926-7935.2004

Sandoval-Villegas, N., Nurieva, W., Amberger, M., & Ivics, Z. (2021). Contemporary Transposon Tools: A Review and Guide through Mechanisms and Applications of Sleeping Beauty, piggyBac and Tol2 for Genome Engineering. *Int J Mol Sci, 22*(10). doi:10.3390/ijms22105084

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Nati. Acad. Sci., 74*(23), 5463-5467.

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2001). Comprehensive identification of conditionally essential genes in mycobacteria. *PNAS, 98*(22), 12712-12717.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science, 270*, 467-470.

Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W. V., Bussey, H., Andrews, B., Tyers, M., & Boone, C. (2001). Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science, 294*, 2364-2368.

Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Ménard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.-M., Shapiro, J., Sheikh, B., Suter, B., Wong, S. L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H., & Boone, C. (2004). Global Mapping of the Yeast Genetic Interaction Network. *Science, 303*, 808-813.

Typas A, N. R., Siegele DA, Shales M, Collins SR, Lim B, Braberg H, Yamamoto N, Takeuchi R, Wanner BL, Mori H, Weissman JS, Krogan NJ, Gross CA. (2008). A tool-kit for high-throughput, quantitative

analyses of genetic interactions in E. coli. *Nat Methods, 5*(9), 781-787.

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods, 6*(10), 767-772. doi:10.1038/nmeth.1377

Vries, H. d. (1901). *Die Mutationstheorie*. Leipzig: Verlag von Veit und comp.

Wassom, J. S. (1989). Origins of Genetic Toxicology and the Environmental Mutagen Society. *Environmental and Molecular Mutagenesis, 14*, 1-6.

Zhang, R., Xu, W., Shao, S., & Wang, Q. (2021). Gene Silencing Through CRISPR Interference in Bacteria: Current Advances and Future Prospects. *Front Microbiol, 12*, 635227. doi:10.3389/fmicb.2021.635227

Chapter 2

2FAST2Q: A general-purpose sequence search and counting program for FASTQ files

Afonso M. Bravo[1], Athanasios Typas[2], Jan-Willem Veening[1]

[1]Department of Fundamental Microbiology, Faculty of Biology and Medicine, University of Lausanne, Biophore Building, Lausanne 1015, Switzerland.
[2]Genome Biology Unit, EMBL, Heidelberg, Germany

This chapter is currently published as:

Bravo AM, Typas A, Veening J. 2022. 2FAST2Q: a general-purpose sequence search and counting program for FASTQ files. PeerJ, 10e14041.

DOI: *10.7717/peerj.14041*

Afonso M. Bravo conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft. Athanasios Typas conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft. Jan-Willem Veening conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Abstract

The increasingly widespread use of next generation sequencing protocols has brought the need for the development of user-friendly raw data processing tools. Here, we explore 2FAST2Q, a versatile and intuitive standalone program capable of extracting and counting feature occurrences in FASTQ files. Despite 2FAST2Q being previously described as part of a CRISPRi-seq analysis pipeline, in here we explore in detail the program's functionality, and its broader applicability and functions.

2FAST2Q is built in Python, with published standalone executables in Windows MS, MacOS, and Linux. It has a familiar user interface, and uses an advanced custom sequence searching algorithm.

Using published CRISPRi datasets in which *Escherichia coli* and *Mycobacterium tuberculosis* gene essentiality, as well as host-cell sensitivity towards SARS-CoV2 infectivity were tested, we demonstrate that 2FAST2Q efficiently recapitulates published output in read counts per provided feature. We further show that 2FAST2Q can be used in any experimental setup that requires feature extraction from raw reads, being able to quickly handle Hamming distance-based mismatch alignments, nucleotide wise Phred score filtering, custom read trimming, and sequence searching within a single program. Moreover, we exemplify how different FASTQ read filtering parameters impact downstream analysis, and suggest a default usage protocol.

2FAST2Q combines easiness of use and versatility to offer an all-around single-step FASTQ sequence extraction tool. It efficiently processes not only CRISPRi-seq / random-barcode sequencing datasets on any up-to-date laptop, but also handles the advanced extraction of *de novo* features from FASTQ files. We expect that 2FAST2Q will not only be useful for people working in microbiology but also for other fields in which amplicon sequencing data is generated. 2FAST2Q is available as an executable file for all current operating systems without installation and as a Python3 module on the PyPI repository (also available at https://veeninglab.com/2fast2q).

Introduction

Next generation sequencing (NGS) has drastically changed the landscape of experimental biology, not only by helping to characterize cellular networks to an unprecedented level, but also by generating vast quantities of data. Typical NGS data generated by Illumina sequencing is delivered in the form of a so-called FASTQ file: a text file that contains the inferred DNA sequences with their respective quality scores, typically existing in a compressed form with the extension *.fastq.gz. However, as newer sequencing platforms become available, so do the sequencing file types. For example, Oxford Nanopore sequencers store their data in FAST5 format, which requires conversion to FASTQ before any traditional downstream sequence analysis can be performed. More than the file itself, the compression format can also vary: DRAGEN ORA (.ora) is currently being rolled out by Illumina as an alternative to the standard .gz format. Nonetheless, despite these constant advancements, FASTQ remains the standard format, in large part probably due to the current convergence of NGS analysis programs to mainly accept .FASTQ as first input. In time, however, data requirements might change and lead to the need of either further pre-analysis format-exchange programs, or the rewrite of current bioinformatics core programs.

Currently, dozens of tools exist for FASTQ file analysis, however, as big data handling becomes an increasingly needed skill in biology, so does the demand for versatile user-friendly applications. With NGS becoming simpler and widespread, so must its respective data processing. There is, therefore, a need for intuitive, reproducible, and versatile tools that can handle the sometimes overwhelming initial raw data processing steps.

NGS applications often require features to be extracted and counted from FASTQ files for downstream analysis. Several analysis tools and scripts exist for systematic reverse genetic screens, such as CRISPRi-seq (Liu *et al.*, 2021), and random-barcode sequencing (RB-Seq) (Cain *et al.*, 2020; Wetmore *et al.*, 2015). However, at the moment, such pipelines tend to overspecialize into CRISPR/Cas9 workflows, be complex, or require informatics skills beyond the average user (Li *et al.*, 2014; Liao *et al.*, 2019; Winter *et al.*, 2016; Winter *et al.*, 2017). A notable example, MAGeCK, allows for both feature counting and downstream feature differential analysis (Li *et al.*, 2014). MAGeCK, due to being primarily optimized for this latter, has some caveats regarding more complex feature extractions procedures. Indeed, when

dealing with mismatches or dynamic read trimming/feature extraction it requires the installation of 3rd party command line only software such as bowtie2 and/or cutadapt. Current more user friendly approaches such as CRISPRAnalyzeR and PinAPL-Py are also limiting in throughput in regards to searching and returning reads with specific sequences, especially when considering sequence mismatches, nucleotide wise Phred score filtering, and dynamic sequence search using multiple sequences of variable length (Spahn *et al.*, 2017; Winter *et al.*, 2017). This is particularly important in the cases where a user wants to control these processing parameters to easily extract and count know/unknown variable location/length sequences from their experimental setup. As such, when such advanced requirements are needed, it is the current standard to create custom made pipelines for handling the specifics of the experiment, normally in conjunction with bioinformatics tools such as Trimmomatic, cutadapt, and/or Bowtie2 (Bolger *et al.*, 2014; Langmead & Salzberg, 2012; Martin, 2011). This assumes bioinformatics, sequencing, and programing knowledge, requiring weeks or months of time to implement from scratch for the average user, with the alternative being outsourcing the data processing. This latter requiring either extra funds, or the right willing colleague.

Here, we explore 2FAST2Q, a fast and versatile FASTQ file processor for extracting and counting sequence occurrences from raw reads. 2FAST2Q requires no installation by default (when using the executables), and works in all common operative systems. 2FAST2Q has been previously published as part of a CRISPRI-Seq protocol, however, in this work we further elaborate on the program's functionalities (de Bakker *et al.*, 2022). We demonstrate novel applications and provide an in-depth description of 2FAST2Q. As a proof of concept, we show that 2FAST2Q efficiently and reliably counts single guide RNA (sgRNA) features in FASTQ files originating from published prokaryotic and eukaryotic CRISPRi-seq experiments. Moreover, we explore alternative 2FAST2Q functions, and how these can be used for any *de novo* sequence searching, or for extracting and counting any kind of sequences from FASTQ files using advanced search and filtering methods.

Results

Developing 2FAST2Q

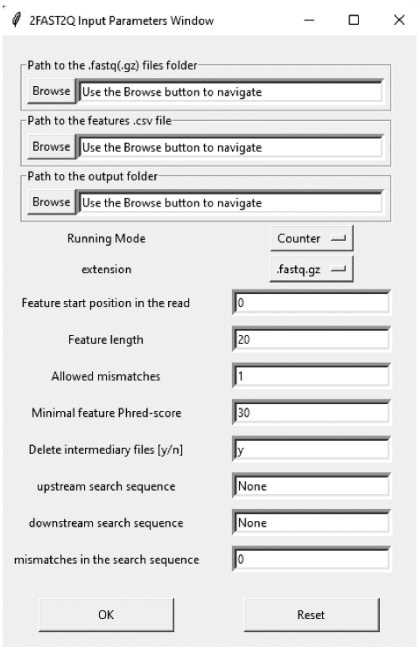
A major goal when doing targeted (amplicon) sequencing is to know the abundance of each target within a sample. To that end, we wrote the Python-based tool called 2FAST2Q (Figure 1). 2FAST2Q is able to efficiently extract, align, filter, and count DNA sequences from standard FASTQ files in a single step. 2FAST2Q also performs mismatch sequence searching, nucleotide Phred score quality filtering, dynamic sequence search and trimming (including double sequence search), and automatically loads and detects FASTQ (.gz compressed or not) files. The program also exists as an easy-to-use intuitive executable version for MS Windows, macOS, and Linux, requiring no installation. Alternatively, 2FAST2Q is also available as a Python3 package in the PyPI repository, and can be installed with the “pip install fast2q” command. As input, 2FAST2Q requires only a FASTQ (.gz compressed or not) file, and, when reference feature sequences exist (i.e.: sgRNAs, barcodes), a .csv file with all the lookup DNA sequences. As an output, 2FAST2Q returns an ordered .csv file with all the raw feature counts per condition, as well as quality control statistics (Figure 1B-C). 2FAST2Q contrasts with other current methods by being easy to setup and intuitive to use (Figure 1A), while simultaneously maintaining advanced configuration settings such as efficient mismatched sequence searching, and quality filtering. 2FAST2Q is thus able of going beyond traditional CRISPRi experimental setups, handling any kind of feature extraction, know or unknown, from FASTQ files.

Counting features using 2FAST2Q

An important feature of performing CRISPRi-seq or RB-seq is to obtain reliable counts of each sgRNA or barcode, for any experimental condition. When using 2FAST2Q in “counting mode”, (i.e.: for CRISPRi-seq, or sequence barcode counting), it can be used to quickly obtain an absolute feature sequence count from FASTQ files. Moreover, it might also be of interest to extract all features existing before/in-between/after a given sequence. 2FAST2Q has an “extract and count mode” for this occasion, where the program doesn’t require the input of any feature sequences, and will retrieve the count of all found read sequences. In both instances, the program can search for any feature by either specifying a starting read position, or by providing upstream and/or downstream constant search sequences. The feature length must be

specified, except in the latter, where variable sized sequences can be retrieved and/or aligned to (Figure 2).

A



B

	A	B	C	D	E	F	G
1	#Feature	41-1Ti9_G	41-1Ti9_G	41-1Ti9_LE	41-1Ti9_LE	41-1Ti9_M	41-1Ti9_M
2	1	283	318	300	367	290	322
3	2	255	234	443	485	367	384
4	3	165	179	286	331	296	323
5	4	164	166	215	236	192	198
6	5	32	43	88	85	75	67
7	6	2	0	0	0	0	0
8	7	64	55	99	125	140	140
9	8	79	102	173	170	151	143
10	9	275	279	286	314	297	315
11	10	389	387	297	356	300	346
12	11	207	207	333	360	286	333
13	12	125	153	215	205	183	194
14	13	35	58	83	87	86	66
15	14	0	0	0	0	0	0

C

	A	B	C	D	E	F	G
1	#2FAST2Q version: 2.3.4						
2	#Mismatch: 0						
3	#Phred Score: 30						
4	#Feature Length: 20						
5	#Feature start position in the read: 0						
6	#Running mode: C						
7	#Upstream search sequence: None						
8	#Downstream search sequence: None						
9	#Mismatches in search sequence: 0						
10	#Sample r	Running T	Running T	Total num	Total num	Number o	Number o
11	41-1Ti9_G	32.37	seconds	3946258	2128978	2128978	0
12	41-1Ti9_G	34.07	seconds	4140906	2191905	2191905	0

Figure 1 | 2FAST2Q interface, and outputs.

A) All program parameters are given by interacting with 2FAST2Q user interface. 2FAST2Q outputs two .csv files; a raw read count file for all samples (**B**), and a file with each sample statistics (**C**). Each independent file is considered to be a sample, and the file name the sample name.

Benchmarking 2FAST2Q

2FAST2Q was initially benchmarked against a published CRISPRi-seq dataset comprising 479M reads dispersed over 118 FASTQ files (Rousset *et al.*, 2021). In this study, Rousset *et al.* examined which genes are essential in *Escherichia coli* under different environmental conditions using CRISPRi-seq (Rousset *et al.*, 2021). 2FAST2Q was used to find an alignment and count the occurrence of each feature from a list of 11,629 sgRNAs across all the 118 files. When only considering perfect alignments between a feature and a read, 2FAST2Q was able to output the final compiled sgRNA count table in 7 min on a personal desktop computer (33s per sample distributed over 10 parallel cores). For comparison, the same files and parameters were also input into MAGeCK. Despite its faster individual file processing speed, its lack of inbuilt sample multiprocessing resulted in a total run time of 23 min. Moreover, MAGeCK fails to return an organized file for all combined samples, leaving the user

with the individual count files for each sample (118 in this case). MAGeCK also requires explicit indication of all the FASTQ files to be processed, a time consuming step which 2FAST2Q performs automatically, unless indicated. When comparing the read counts returned from both 2FAST2Q and MAGeCK, a perfect correlation (r=1) was observed for all features (supplementary figure 3), indicating similar read counting accuracy.

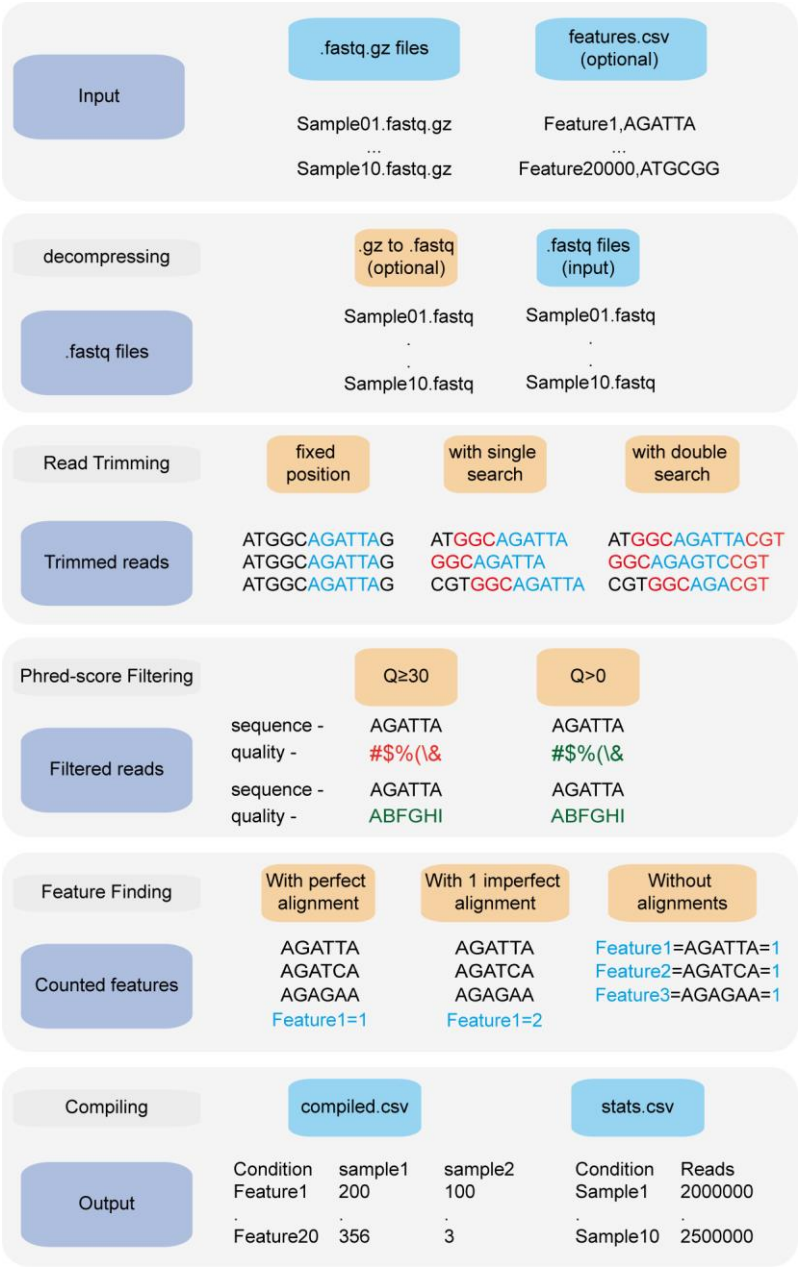


Figure 2 | 2FAST2Q pipeline.

2FAST2Q requires only .fastq.gz (or .fastq) files as input. When in alignment mode, a csv file with all the features must also be provided. 2FAST2Q performs all described steps automatically and without requiring external software. Trimming parameters, filtering scores, and mismatch tolerances can be easily adjusted using 2FAST2Q graphic interface.

When allowing for 1 mismatch in the sgRNA search count, the total 2FAST2Q run time only increased by 2min, to 9min. Under these program conditions, this corresponds to a more than 40x speed improvement over the use of similar purpose standard search functions, such as the Python regex module match function ("Python Software Foundation, Python Language Reference, version 3.7, Available at http://www.python.org,"). For mismatch searching, MAGeCK requires the use of Bowtie2, and respective setup, and was thus not used for further benchmarking.

Using the same dataset published by Rousset, F. *et al.* (Rousset *et al.*, 2021), we assessed the impact of different initial 2FAST2Q parameters on both absolute feature counts, and on downstream data analysis. When not using any Phred-score filtering (Q≥0), and not allowing for any mismatches, we were able to fully recapitulate the reported total read counts/sgRNA for all conditions (Figure 3E) (supplementary tables 1 and 2). However, high-quality read length has been reported to improve Illumina sequencing results interpretation (Bokulich *et al.*, 2013). We therefore implemented a filtering for nucleotide wise Phred-scores (Q), where all the sequenced nucleotide scores corresponding to the found feature read location are required to be above an indicated threshold. As expected, filtering using Q≥30, indicating a 0.1% probability of a nucleotide sequencing mistake, lowers the amount of reads/sgRNA. In some cases, by more than 1 order of magnitude (Figure 3G). However, when considering the millions of reads generated by a typical sequencing experiment, the presence of mismatches in high quality reads is a likely event (any length of 20 nucleotides with Q≥30 have, at most, a 2% chance of having a mismatch: 0.001 * 20 = 0.02). We then also added a feature mismatch search where a read is considered valid if it unambiguously aligns to a single feature for any number of considered mismatches, thus retrieving more high quality reads, especially from lower overall quality sequencing runs. Allowing for mismatches expectedly increased the number of reads/feature (Figure 3A, 3C, 3I and 4A), without significantly sacrificing total run time (Figure 4B) (supplementary tables 3 and 4). As an extreme benchmark case, we allowed for the same number of mismatches as the feature length (20bp) (Figure 3I-J). In practice, these parameters mapped any read to its closest feature, meaning the sequence that unambiguously differs the least from the read. This is performed by an inbuilt safety mechanism, where if more than one feature possible matches the read at the lowest amount of allowed mismatches (i.e. 1), the read is always discarded, but otherwise kept. In regards to the Rousset, F. *et al.* dataset, which is on average of high

quality, these parameters recovered on average 3% more reads/sgRNA (Fig 4A). However, it is conceivable that the use and outcome of these parameters varies depending on the experimental setup and user requirements, requiring careful consideration before proceeding to downstream data analysis. In here, we report only on the possibilities of 2FAST2Q functionalities.

Higher stringency parameters can aid in biological discovery

We used the Jupyter notebook analysis pipeline published by Rousset, F. *et al.* (Rousset *et al.*, 2021) to assay how these different read processing scenarios impact downstream analysis. Using the different read count tables directly outputted by 2FAST2Q, we calculated and compared the median gene scores as defined by Rousset *et al.* (essentially, the median of the $\log_2$ fold change for each feature in all experimental replicates) for the LB medium and gut microbiota medium (GMM) conditions. Using more stringent criteria than Rousset, F. *et al.* (a gene is considered significant if it has an absolute gene fold change ≥ 4, instead of ≥ 3.5), we compared how different Phred-scores and mismatch filtering criteria influenced downstream analysis, namely how these criteria influence gene score calculations, and thus gene essentiality (Figure 3B, 3D, 3F, 3H, and 3J).

We observed a higher stringency for the 2FAST2Q parameters of 1 mismatch, and base pair quality filtering of ≥ 30 (Fig 3B), with fewer genes being considered essential for any given condition with these criteria than with the criteria that recapitulate the published data (0 mismatches allowed, and no Q consideration) (Figure 3F). As expected, different read filtering criteria resulted in fold change differences, and consequently in differences in the genes considered essential for these conditions. What criteria to use would depend on the specifics of each individual experiment. The default 2FAST2Q parameters uses Q ≥ 30, while allowing up to 1 mismatch (representing for any 20 basepair (bp) sequence, a 5% bp deviation error with, at maximum, a 2% chance of any nucleotide being wrongly sequenced). As shown in Figure 3, although the default setting of 2FAST2Q give slightly fewer significant hits, they were all also reported by Rousset *et al*. It is also conceivable for a user to be interested in aligning all reads to their closest matching feature. Like mentioned before, this is possible by setting the total amount of mismatches to the same length of the feature. Once again, we intend only to demonstrate the range of

uses of 2FAST2Q. Ultimately, the biological relevance of which parameters to choose is left upon the user.

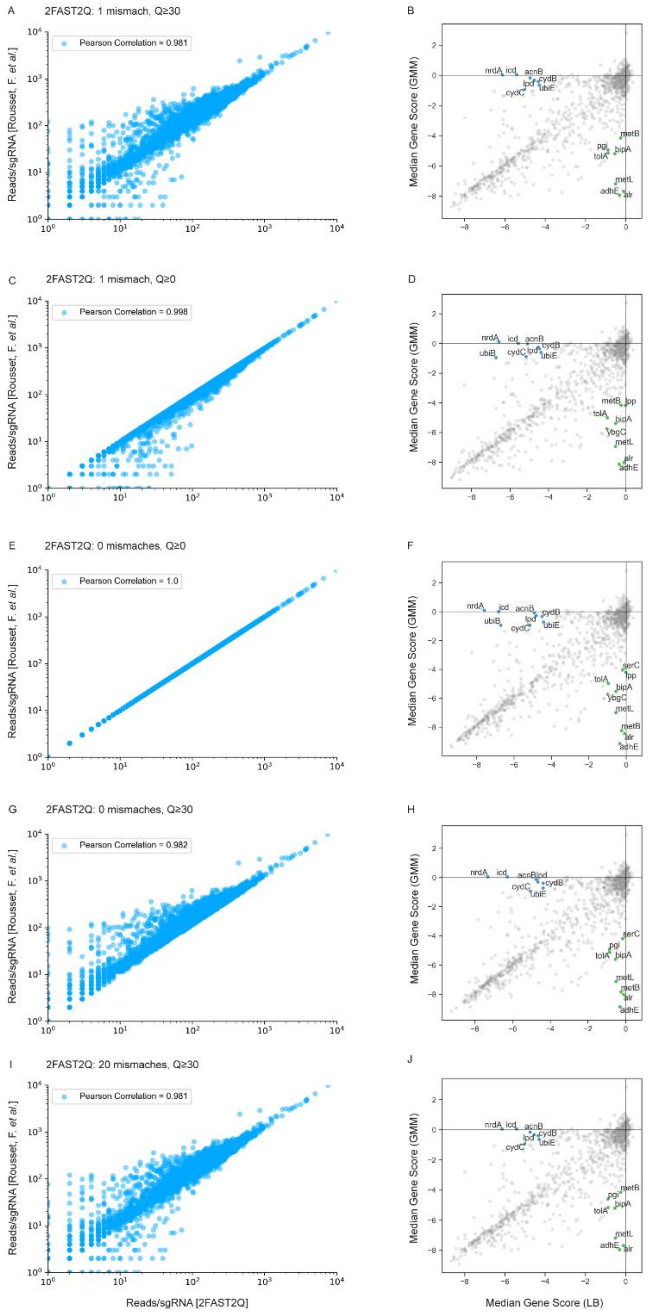


Figure 3 | Absolute read counts/sgRNA for the Rousset, F. *et al.* dataset MG1655 LB 1 condition (Rousset *et al.*, 2021).

The total read counts using different 2FAST2Q mismatch and/or quality filtering inputs are plotted against those reported by Rousset, F. *et al*. Pearson correlation for each plot is also shown. Plots **B**, **D**, **F**, **H** and **J** were generated using an adaptation of the published Jupyter notebook analysis pipeline, and highlight the significant genes (absolute fold change ≥ 4 in one condition and ≤ 1 in the other) when using different 2FAST2Q input parameters. green: fold change < 4 in GMM media, and > -1 in LB; blue: fold change < 4 in LB media, and > -1 in gut microbiota medium (GMM);

2FAST2Q dynamically performs FASTQ feature extraction

Under certain experimental setups, the extraction of features from FASTQ files might require the use of a dynamic trimming and search function (i.e.: when the location and/or size of the feature differs from read to read) (Fig 2). In this case, a delimiting search sequence of any length (up and/or downstream of the feature) can be provided. Similar to feature mismatch search, an arbitrary number of mismatches can also be indicated for the search sequence-based trimming, as well as a minimum Phred-score. 2FAST2Q will search each read for the indicated sequences, returning the correctly trimmed read for further processing, and bypassing the need for more complex tools such as Trimmomatic and Bowtie2. As a proof of concept, we used a published CRISPRi-seq dataset by Wei *et al.* (Wei *et al.*, 2021), where dynamic read trimming was required. In this study, a CRISPRi screen was performed using Vero-E6 cells (kidney epithelial cells from an African green monkey) infected with SARS-CoV-2 to identify host genes important for viral replication (Wei *et al.*, 2021). In this dataset, the location of each feature was at a variable location within the read. 2FAST2Q dynamic trimming allowed each read to be independently trimmed based on the relative location of the found search sequences, thus always returning the correct feature location. Using this method, we submitted 6 FASTQ files (SRR14668185 - SRR14668190) for 2FAST2Q processing. As search sequence we used a 10bp upstream constant sequence (CGAAACACCG), allowing for 1 mismatch search error in this sequence. We used the provided list of 84,953 sgRNA sequence features, and ran 2FAST2Q (Q≥30, 0mismatches). 2FAST2Q simultaneously processed all 6 samples, comprising 324M reads, within 8 minutes on a standard desktop PC (data not shown). This result corresponds to a slowdown of only 22% (speed comparisons were determined using processed reads/second) when compared with the non-dynamic feature extraction process, such as the one we used for the same parameter 2FAST2Q run with the Rousset, F. *et al.* dataset (Rousset *et al.*, 2021) (Q≥30, 0mismatches).

Recently, Bosch *et al.* published a CRISPRi-seq experimental setup with variable length sgRNAs (Bosch *et al.*, 2021). In this case, both the trimming of each read and the length of each sgRNA need to be considered read by read. This is a feature, to our knowledge, beyond easy implementation in any of the programs mentioned in this work. Once again, 2FAST2Q was also able to extract, count, and

align all the found features in a *Mycobacterium tuberculosis* dataset (SRR13734827), to the provided 96,700 long sgRNA file, this time using 2 delimiting constant search sequences (upstream: GTACAAAAAC; downstream: TCCCAGATTA), while allowing for 1 mismatch in each. The returned variable length sequence between the two constant search sequences was used for perfect match alignment against the sgRNAs (data not shown). When compared with the non-dynamic extraction process used by 2FAST2Q in the Rousset, F. *et al.* dataset, a slowdown of 44% was observed, also in line with what was observed for the Wei *et al.* dataset.

As the sequence search algorithm uses a similar process to the one used for feature alignment mismatch, a similar speed improvement over standard Python functions is also obtained. Together, these benchmarks demonstrate that 2FAST2Q is a versatile and quick computational tool that can extract relevant features and counts from FASTQ files.

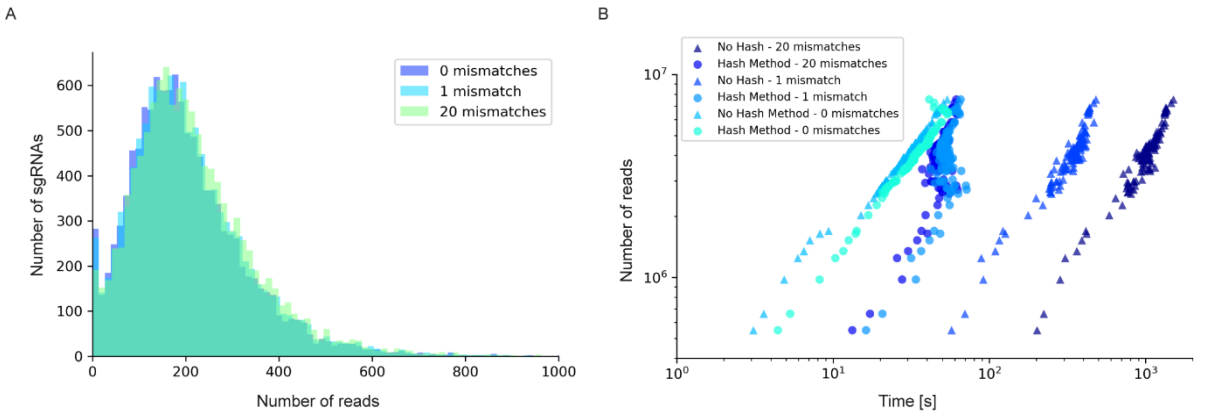


Figure 4 | Read/sgRNA distribution and runtime analysis of 2FAST2Q with different mismatch parameters and algorithms.

Data analysis was performed on the Rousset *et al.* "UTI89_T0" fastq sample (Rousset *et al.*, 2021) when submitted to 2FAST2Q analysis with either 0, 1, or 20 mismatches (and Phred score ≥ 30). Increasing mismatches allows for greater read recovery by matching a given read to its closest matching (and thus most likely) feature. **A**) The median reads/sgRNA increased from 182 to 187, and then to 196, when considering 0,1, and 20 mismatches, respectively. **B**) 2FAST2Q runtime analysis demonstrates the efficiency of real time creation of pre-processed failed/passed read hash tables (see methods) vs. the "no hash" method, where each read is always processed *de novo* for mismatches.

Discussion

FASTQ files are the current standard sequencing output file format. Considering that new sequencing based differential analysis based experimental techniques emerge on an almost monthly basis, the need for easy-to-use, versatile and efficient programs specifically designed for extracting and counting features form FASTQ files is pressing. We thus developed a fast and intuitive tool for counting sequence occurrences in FASTQ files. We have recently implemented 2FAST2Q in our CRISPRi-seq pipeline and have found it useful in the first step of data analysis (de Bakker *et al.*, 2022). However, in here we describe novel 2FAST2Q functionalities and explore the program's parameter versatility, which cover most current user applications that require the extraction and counting of specific feature sequences, such as CRISPRi-seq, RB-Seq, and general amplicon-seq. Despite only handling single-ended FASTQ files at the moment, the processing of paired-ended files is possible by running two separate instances of 2FAST2Q. The program will automatically compile all samples at the end if all intermediary files of the first run are copied to the output folder of the second instance while processing. However, if a feature spans both reads, and both reads from the paired-ended are to be analyzed as a single contiguous read, a pre-process step of read merging (for example using PEAR) is recommended. The resulting merged reads can be input into 2FAST2Q as normal.

Depending on the desired output, current methods might require users to handle several different software pipelines in order to extract and filter relevant data from FASTQ files. However, 2FAST2Q is a standalone program that can, in a single step, efficiently and quickly perform nucleotide wise quality filtering, mismatch sequence searching, *de novo* feature extraction, and sequence occurrence counting. 2FAST2Q outputs an individually compiled, easy to interpret, excel readable .csv file with all the ordered feature counts per sample, alongside a file with relevant sample statistics.

2FAST2Q fully recapitulated the feature counts independently returned by MAGeCK, and reported by Rousset, F. *et al.*, for all conditions when using the same filtering criteria. 2FAST2Q was also successful at extracting features starting at different positions per read when using a published dataset of a CRISPRi screen on eukaryotic cells that were infected with SARS-CoV-2 (Wei *et al.*, 2021). 2FAST2Q

inbuilt search functions also allow for more complex experimental setups. For example, recent work by Bosch *et. al* applied CRISPRi-seq with variable length sgRNAs to identify conditionally essential genes in *M. tuberculosis* (Bosch *et al.*, 2021). By providing up and downstream search sequences, 2FAST2Q was able to extract and count these sgRNAs in a single-step. In the case of experiments with more than one feature per read, such as with dual barcode sequencing, or dual CRISPRi-seq, it is conceivable that 2FAST2Q could also be used, taking into account that the parameters need to be adjusted to capture different features per read each time, and by compiling the data at the end.

Besides being able to align and count provided features in FASTQ files, 2FAST2Q is also able to extract and count all unique read sequences when in "extract and count mode". In this case, all different sequences that fulfill the required parameters are returned, with any possible mismatches being accounted as distinct sequences.

As experiments that produce large datasets (>1GB) become more widespread, the need for versatile, fast, and easy to use software that handles raw data becomes more sought out. It is thus our hope that 2FAST2Q can contribute to facilitate the processing of the large amounts of sequencing data originating from NGS studies.

Here, we explored and benchmarked 2FAST2Q, a tunable novel Python3-based program capable of single-step quality filtering, read feature searching, extraction, and feature counting in FASTQ files. 2FAST2Q exists as a standalone program, not requiring any installation whatsoever when using the executable files, and as a Python module available at the PyPI depository. We demonstrated how 2FAST2Q can be used for the processing of FASTQ files originating from different experimental setups, and how it handles different input parameters to adapt to most conceivable datasets requiring feature counting. 2FAST2Q is an intuitive program, that we believe can streamline sequencing data feature extraction for most users, without the need for advanced bioinformatics setups, or the use of multi-step complex pipelines.

Methods

Installation and code availability:

All 2FAST2Q executable files can be downloaded from zenodo: https://zenodo.org/record/5410822. The code, usage instructions, and test datasets are available on GitHub: https://github.com/veeninglab/2FAST2Q. 2FAST2Q is also a Python package, and can be accessed on PyPI: https://pypi.org/project/fast2q/. When using the executable version on MS Windows or MacOS, no further installation is required and a double click on the executable should suffice. For a more in depth description, please see the online tutorial on https://veeninglab.com/2fast2q. 2FAST2Q is fully implemented in Python3.

Usage considerations:

All indicated 2FAST2Q running times were performed on a desktop PC with a 12 core 3.7GHz processor, and 32GB of RAM. However, 2FAST2Q runs on any up-to-date desktop or laptop. When using 2FAST2Q without mismatch search (perfect alignment only), sample processing should be in the order of seconds or minutes. When using the mismatch search, it is possible for 2FAST2Q analysis to take several minutes per sample. When processing more than one sample, 2FAST2Q will automatically parallelize all analyses by distributing each sample per available processor core.

2FAST2Q fast sequence mismatch search function was possible due to the use of Python numpy (Harris *et al.*, 2020) and numba (Lam *et al.*, 2015) modules. An advanced and in-depth tutorial on 2FAST2Q parameters is available on GitHub and PyPI.

2FAST2Q algorithm

When initialized in standard feature count mode, 2FAST2Q will automatically handle all compressed or uncompressed FASTQ files, and create a hash table for all supplied sequence features. 2FAST2Q will then forward all samples for parallel processing, which can be monitored via a progress bars (supplementary figure 1). Each FASTQ file is sequentially read, saving RAM space. The individually loaded reads are submitted for trimming based on the indicated parameters, either using a

fixed position, or a dynamic search. The first assumes the presence of a fixed feature length in the same location for all reads. The second requires one or two search sequences. When one sequence (either up or downstream) is provided, 2FAST2Q will search the read until the sequence is found, and return the adjacent predetermined sized feature (again, either up or downstream). When two sequences are used, 2FAST2Q will return any feature within the found search sequences. The location and feature length parameter can thus be ignored in this latter scenario. A sequence mismatch search can also be performed.

Following read trimming, the Phred-score corresponding to each nucleotide of the trimmed sequence is considered. If any of the scores is below the indicated parameter threshold, the read is discarded.

If the read passes quality control, depending on the user input, the found feature is either retuned, or an alignment against the input features is attempted. Feature alignment is performed using either mismatch search or not. By default, 2FAST2Q will always first check for a perfect match. Perfect matching uses hashing, directly comparing all features to the read sequence using hashing runtime complexity. When dealing with mismatches, 2FAST2Q will perform sequence search based on a faster custom made search algorithm. At first, all feature/search sequences are converted to their numerical binary form, subsequently reducing them to integer8 format using numpy. Sequence mismatches are counted by tracking the non-zero result positions of subtracting both sequences. 2FAST2Q mismatch search is therefore based on a Hamming distance calculation. As simple numpy constructs, arithmetic operations can be easily processed using the Python Numba module njit decorator. Therefore, all 2FAST2Q search functions are pre-compiled and effectively run at much faster speeds (supplementary figure 2). All read sequences searches, and features mismatch alignments are performed using this approach, allowing all search operations to run faster than standard Python code. Moreover, reads that fail to safely align, within the given parameters, to any of the provided features, are stored and used for quick hashed based comparison. The same is performed for reads that align with mismatches. By performing the much faster hashed comparison, this feature avoids the slower *de novo* mismatch search for previously seen same sequence reads. Runtime is thus decreased, paradoxically maintaining sample processing time as file size increase. “Already seen read” hashing is especially useful with datasets comprising dozens of different independent samples from the same sequencing run

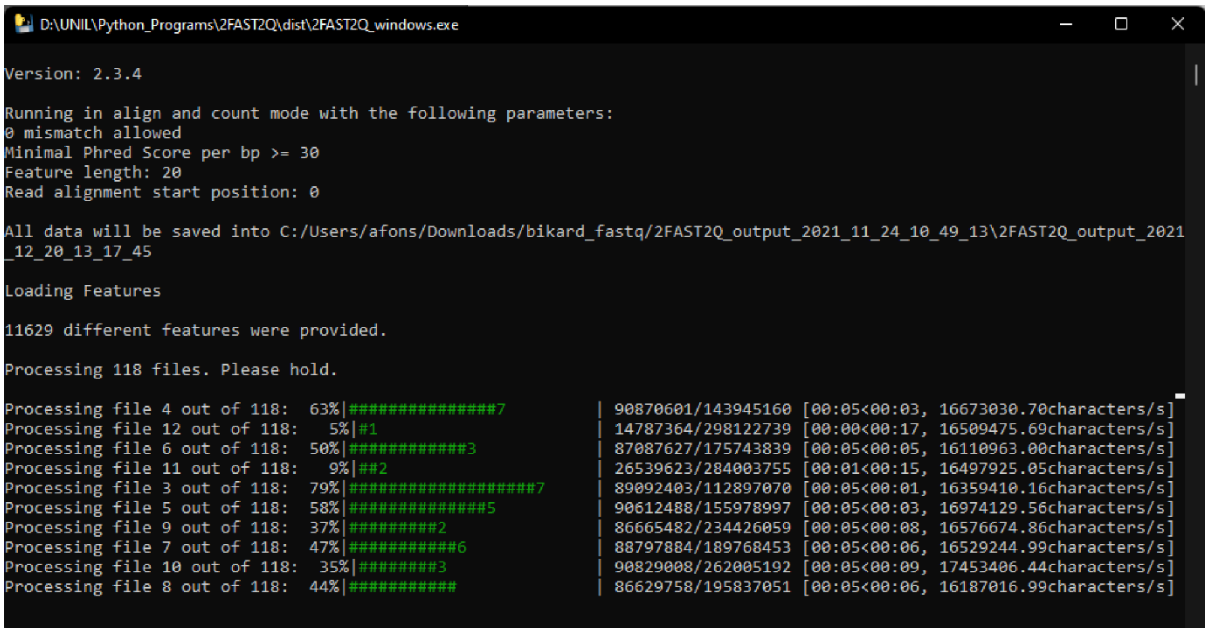
(see results). In this case, the generated failed/passed read hash tables for each sample are compiled and used as a seed to the next batch of samples. Each new sample thus takes advantage of the already processed reads in a previous sample, avoiding reprocessing the exact same read several times.

A Python dictionary with a class feature count is used to keep track of all found aligned sequences. When no feature file is provided (i.e. when running in “Extractor+Counter” mode), all found read sequences are returned and counted. Each FASTQ file will originate a unique output file. At the end of the analysis, all samples’ files are compiled into a single file, which can be readily used for downstream applications.

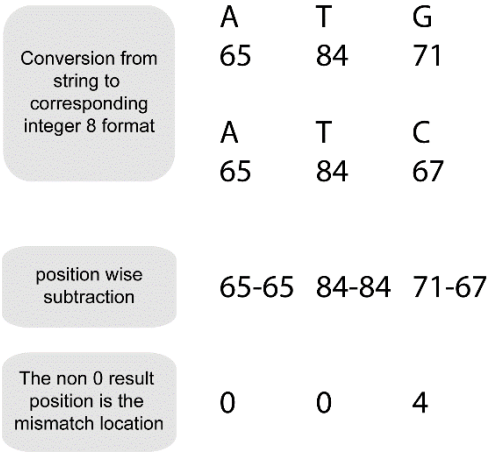
Acknowledgements

We thank Julien Dénéréaz and Vincent de Bakker for their software tests, and all members of the Veening lab for helpful discussions.

Supplementary

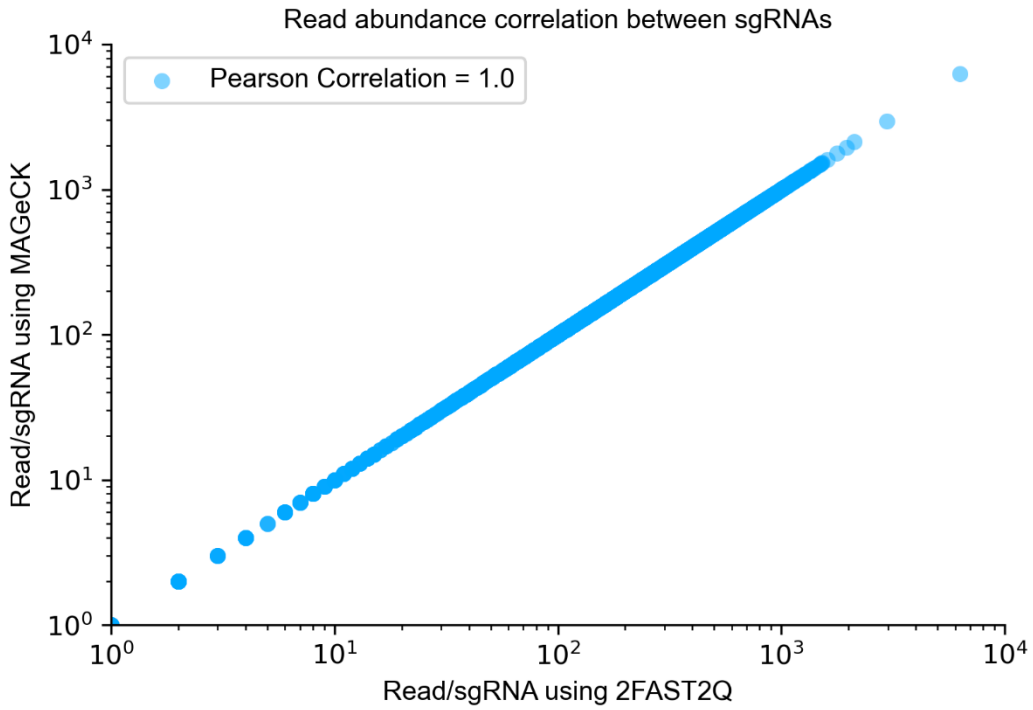


Supplementary figure 1 | 2FAST2Q parallel sample processing screen.



Supplementary figure 2 | 2FAST2Q Hamming distance-based mismatch search algorithm.

When dealing with mismatches, 2FAST2Q will preemptively convert all the input feature sequences into their respective binary integer format using 8bits encoding. This step ensues faster downstream processing, decreases RAM usage, and allows mismatches to be calculated by a simple subtraction performed at machine code speed.



Supplementary figure 3 | Absolute read counts/sgRNA for the Rousset, F. *et al.* dataset UTI89 T0 condition (Rousset *et al.*, 2021).

The total read counts were obtained using MAGeCK (y axis) and 2FAST2Q (x axis). Pearson correlation is also shown.

References

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *nature methods, 10*(1), 57-60.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bosch, B., DeJesus, M. A., Poulton, N. C., Zhang, W., Engelhart, C. A., Zaveri, A., Lavalette, S., Ruecker, N., Trujillo, C., Wallach, J. B., Li, S., Ehrt, S., Chait, B. T., Schnappinger, D., & Rock, J. M. (2021). Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. *Cell, 184*(17), 4579-4592 e4524. doi:10.1016/j.cell.2021.06.033

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & Opijnen, T. v. (2020). A decade of advances in transposon-insertion sequencing. *Nat Rev Genet, 21*(9). doi:10.1038/s41576-020-0244-

de Bakker, V., Liu, X., Bravo, A. M., & Veening, J.-W. (2022). CRISPRi-seq for genome-wide fitness quantification in bacteria. *Nature Protocols, 17*, 252–281

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Rio, J. F., Wiebe, M., Peterson, P., Gerard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357-362. doi:10.1038/s41586-020-2649-2

Lam, S. K., Pitrou, A., & Seibert, S. (2015). *Numba*. Paper presented at the Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology, 15*(554).

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res, 47*(8), e47. doi:10.1093/nar/gkz114

Liu, X., Kimmey, J. M., Matarazzo, L., Bakker, V. d., Maele, L. V., Sirard, J.-C., Nizet, V., & Veening, J.-W. (2021). Exploration of Bacterial Bottlenecks and Streptococcus pneumoniae Pathogenesis by CRISPRi-Seq *Cell Host & Microbe, 29*, 107-120.

Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet Journal, 17*, 10-12.

Python Software Foundation, Python Language Reference, version 3.7, Available at http://www.python.org.

Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernandez-Rodriguez, J., Clermont, O., Denamur, E., Rocha, E. P. C., & Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the Escherichia coli species. *Nat Microbiol, 6*(3), 301-312. doi:10.1038/s41564-020-00839-y

Spahn, P. N., Bath, T., Weiss, R. J., Kim, J., Esko, J. D., Lewis, N. E., & Harismendy, O. (2017). PinAPL-Py:

A comprehensive web application for the analysis of CRISPR/Cas9 screens. *Scientific Reports, 7*.

Wei, J., Alfajaro, M. M., DeWeirdt, P. C., Hanna, R. E., Lu-Culligan, W. J., Cai, W. L., Strine, M. S., Zhang, S. M., Graziano, V. R., Schmitz, C. O., Chen, J. S., Mankowski, M. C., Filler, R. B., Ravindra, N. G., Gasque, V., de Miguel, F. J., Patil, A., Chen, H., Oguntuyo, K. Y., Abriola, L., Surovtseva, Y. V., Orchard, R. C., Lee, B., Lindenbach, B. D., Politi, K., van Dijk, D., Kadoch, C., Simon, M. D., Yan, Q., Doench, J. G., & Wilen, C. B. (2021). Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. *Cell, 184*(1), 76-91 e13. doi:10.1016/j.cell.2020.10.028

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio, 6*(3), e00306-00315. doi:10.1128/mBio.00306-15

Winter, J., Breinig, M., Heigwer, F., Brügemann, D., Leible, S., Pelz, O., Zhan, T., & Boutros, M. (2016). caRpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. *Bioinformatics, 32*(4), 632-634. doi:10.1093/bioinformatics/btv617

Winter, J., Schwering, M., Pelz, O., Rauscher, B., Zhan, T., Heigwer, F., & Boutros, M. (2017). CRISPRAnalyzeR: Interactive analysis, annotation and documentation of pooled CRISPR screens. *bioRxiv*. doi:10.1101/109967

Chapter 3

TnSeeker: A self-optimizing Tn-seq gene domain essentiality prediction program

Afonso M. Bravo[1], Alexandra Koumoutsi[2], Jan-Willem Veening[1], Athanasios Typas[2]

[1]Department of Fundamental Microbiology, Faculty of Biology and Medicine, University of Lausanne, Biophore Building, Lausanne 1015, Switzerland.

[2]Genome Biology Unit, EMBL, Heidelberg, Germany

Afonso M. Bravo conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the chapter, and approved the final draft. Alexandra Koumoutsi conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the chapter, and approved the final draft. Athanasios Typas conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft. Jan-Willem Veening conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Abstract

In this work we explore how different transposon sequencing (Tn-seq) data analysis approaches and methodologies influence gene essentiality inference. We performed Tn-seq in 8 different *E. coli* strains and developed a novel general purpose bioinformatics pipeline, termed TnSeeker. TnSeeker operates at the gene domain level, determining gene essentiality from any kind of Tn-seq dataset, while automatically compensating for transposon-insertion sequence biases, and self-optimizing essentiality calling thresholds from a user defined set of essential genes. TnSeeker is also able to determine strand-specific differential transposon insertions, enabling the study of opposite orientation expression effects. We demonstrate that TnSeeker uses a stringent approach, avoiding classifying genes that are too small to be statistically evaluated for any given library saturation. TnSeeker is also capable of determining essentiality at the sub gene level, being capable of self-arranging gene domain sizes according to their respective transposon insertion frequencies, and inferring essentiality regions.

Besides using the built transposon libraries, we further benchmarked TnSeeker with independent *Streptococcus pneumoniae*, *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa* Tn-seq datasets, and compared its performance against a naïve method and other popular tools in the field. Finally, we used TnSeeker to evaluate the essentialome of the 8 *E. coli* strain collection, exploring the positive correlation of pan-essentialome size with number of analyzed strains, and the convergent tendency of the core-essentialome.

Introduction

The construction of a saturated transposon mutant library is normally initiated by introducing a randomly inserting engineered transposon cassette into the genome of an organism of interest, and thus disrupting the insertion site via the creation of a truncated product. If such disruption proves lethal, no insertion is observed at that location and the gene is inferred to be essential. An antibiotic resistance marker is usually carried in the transposon cassette for selection. However, a transcription terminator might not be included to allow transcription to read-through, reducing polar effects (Wetmore *et al.*, 2015). In some cases, the transposon marker promoter can drive the over-expression of downstream genes. If these genes are essential or display fitness advantages, the insertion is then selected for. Such observations are relatively rare, apparently being mostly associated with toxin-antitoxin systems (Hutchison *et al.*, 2019). Similarly, transposons with outward-facing promoters (inducible or not) have also been used, creating a system capable of examining gain-of-function mechanisms by overexpression of nearby genes (Christen *et al.*, 2011; Coe *et al.*, 2019).

In prokaryotes, the Tn5 (Berg *et al.*, 1975) or Mariner transposases (Rubin *et al.*, 1999) are frequently used for library building as both are capable of efficient “cut-and-paste” transposition, while simultaneously displaying low insertion specificity (Larivière *et al.*, 2021). Mariner transposons differ from Tn5 by their ability to target TA/AT dinucleotides only, with Tn5 displaying no sequence specificity, despite exhibiting some preference for GC-rich sites (Brian Green *et al.*, 2012). In literature, Mariner based strategies are normally preferred over Tn5 as saturation of all TA sites is more easily achieved with fewer mutants, while maintaining an equal distribution of reads across all insertion sites. Tn5 preference for GC hotspots might also lead to biased insertion patterns and, consequently, biased read distribution across the genome, resulting in larger transposon libraries being needed.

Transposon-based gene essentiality inference is ultimately determined by the number of insertions per genetic site of interest. The higher the number of expected insertions, the smaller a *locus* needs to be to accurately determine essentiality. *Mariner* transposon methods are consequently at a disadvantage in regards to assaying the essentiality of smaller genes. For example, in *Escherichia coli* there is on average >30 TA sites per 1kb of genome (on average, only 3% of the genome can be directly evaluated), with high GC genes having lesser putative sites. Under these

circumstances statistical power is an issue, with Mariner transposons capping at the *loci* resolution of 200bp for *Vibrio cholerea* in one example (Chao *et al.*, 2016; Chao *et al.*, 2013). Tn5 transposons are therefore advantageous over their Mariner counterpart when higher essentiality resolution is required, albeit at the cost of requiring larger libraries and increased sequencing capacity.

Depending on the target species and available tools, the transposition process is usually done by either conjugation or transformation. In the first, a suicide plasmid carrying a transposase and transposon system is commonly used. When conjugated into a non-permissible host (i.e. not able to replicate the plasmid), the plasmid-expressed transposase will transpose the transposon into the host genome without copying itself. Plasmid loss over generations, together with transposon-positive selection, will then ideally lead to the creation of a pooled single transposon insertion carrying population (Bouhenni *et al.*, 2005; Rubin *et al.*, 1999). Regarding transformation, purified transposase is often used to integrate, *in vitro,* a transposon cassette into the purified DNA of the target organism. The transposon carrying DNA is then transformed back into the organism, and selected for (van Opijnen *et al.*, 2009). Whichever the method, during the selection phase, genes required for survival that have been disrupted by the transposon insertion will start to be depleted from the library: creating the baseline essential gene pool. Recently, Gallagher *et al*. have developed 'transformation transposon insertion mutant sequencing' (TFNseq), enabling the tracking of transposon insertions in genes immediately after transformation, effectively following their loss during subsequent growth, and monitoring the speed at which baseline essential genes are formed (Gallagher *et al.*, 2020). This phenomenon therefore implies the existence of a 3rd class of essentiality: fitness genes. These are genes whose classification (essential or non-essential) depends on the amount of elapsed generations, and on their ability to compete in the pooled population, in any given environment (Langridge *et al.*, 2009; Lluch-Senar *et al.*, 2015; Miravet-Verde *et al.*, 2020). Maria Lluch-Senar *et al.* have previously demonstrated this effect in *Mycoplasma pneumoniae,* having determined the optimum number of cell generations for essentiality analysis in the specified laboratory conditions (Lluch-Senar *et al.*, 2015).

Following selection, the pooled transposon library is ready to use and can be submitted to different conditions for assaying conditional essentiality. After DNA extraction and sequencing library building, next-generation sequencing (NGS) of the

transposon border yields the flanking DNA sequence, allowing for transposon mapping onto the organism's genome. Based on literature, Tn-seq analysis can then be subdivided into two major methods: comparative Tn-seq, and 'snapshot' Tn-seq.

When doing comparative Tn-seq-based assays, the reads originating from all transposon insertions are differentially compared across conditions, and a gene is deemed essential/fitness if it has significantly less reads than in the control condition (Helmann *et al.*, 2019; van Opijnen *et al.*, 2009). Sufficient coverage in all conditions, and in all insertion sites, is therefore paramount to achieve statistical significance. Considering as example a prokaryotic transposon library with 100,000 insertions, obtaining enough read coverage to overcome inherent biases such as genes with a low frequency of insertion sites or reduced read numbers might require extensive amplicon sequencing, and thus be outside the scope of usability for most users (for an average read coverage of 150 reads, 15M reads would be needed for one sample alone).

With 'snapshot' Tn-seq, insertions can be binary classified into absent or not, with essentiality being determined based on the relative local frequency of these. Non-essential features typically display a significantly higher insertion ratio than essential ones, and these latter exhibit less insertions than what is expected by chance alone (Cain *et al.*, 2020). Significantly less sequencing coverage is thus needed than comparative Tn-seq, allowing for increased throughput. 'Snapshot' Tn-seq, however, poses its own issues, requiring more complex data analysis methods to discern essentiality from a static map of transposon insertions. To a certain degree, several published programs/methods have taken different approaches to tackling some of these problems. Indeed, although based on the simple concept of presence/absence of transposon insertions, Tn-seq data analysis remains elusive without any "one-fits-all" pipeline. Despite the existence of multiple Tn-seq normalizations, methods, and tools, there is little understanding of when to use one over another (Barquist *et al.*, 2016; Chao *et al.*, 2016; DeJesus *et al.*, 2015; DeJesus & Ioerger, 2013, 2016; Emily C. A. Goodall *et al.*, 2018; Larivière *et al.*, 2021; Miravet-Verde *et al.*, 2020; Nlebedim *et al.*, 2021; Pritchard *et al.*, 2014; Rahman *et al.*, 2022; Solaimanpour *et al.*, 2015; Zomer *et al.*, 2012). It is also unclear how different data treatments and approaches influence essentiality determination, especially when considering different biological and experimental conditions (Miravet-Verde *et al.*, 2020). Firstly, there is the combined

effect of the transposase and the organism themselves, with possible biases being introduced based on local genome factors, such as GC content, or on transposase insertion hotspots. Secondly, one must consider the existence of fitness genes, with essentiality determination being dependent on several factors: the number of generations elapsed since library inception, and thus differential mutant clearance time from the population; the longevity of any given gene product, and its cellular abundance; and the possible detection of dead cells with deleterious insertions, increasing the chance of false positives. Thirdly, when preparing the sequencing library it is possible for biased and/or chimeric PCRs to happen, respectively falsely increasing the relative abundance of any given insertion, or leading to the miss-mapping of transposon insertions (Miravet-Verde *et al.*, 2020; Wetmore *et al.*, 2015). Lastly, further biases can also be introduced into the system by the user itself, by either being over or under stringent with the sequencing quality control, genome alignment parameters, and essentiality determination method (Laehnemann *et al.*, 2016; Nicholas A Bokulich *et al.*, 2013). Moreover, further confusion is added when considering essential (and non-essential) genes typically display insertions in the N- and C- terminal regions, while also possibly having non-essential gene/protein domains, and/or yet unresolved lethal transposon insertions.

Considering all these factors, how is it then possible to confidently label a gene as essential? A naïve approach would be to determine if the insertions appear in a frequency lower than the one observed across the genome, and if so, call the gene essential. In this case, there would then be a skew towards smaller insertion-free genes. A more complex method could normalize such insertions to gene length, or even ignore insertions at both ends of a gene. Such considerations, however, still leave the question of how to proceed in genes that display “some” insertions, and what and where these “some” insertions are. Does the frequency of these in the population matter (i.e. is an insertion with more reads than another relevant)? Does transposon orientation relative to gene influence essentiality? Also, more than genes, is it possible to examine essentiality as a function of genomic *loci* (i.e. gene domains, small non-coding regions, or intragenic regions)? Several works have examined different parts of these questions over the years. For example, in 2011 Beat Christen *et al.* used their own custom made approach to determine the essentiality of *Caulobacter crescentus* non-coding elements, alongside non-essential domains of essential *loci* (Christen *et al.*, 2011). This was based on finding significant insertion free gaps in genes, and was

tailored to their own experimental data. More recently, A. S. M. Zisanur Rahman defined “essential domain-containing” genes in *Burkholderia cenocepacia,* where the essentiality of gene ‘sub domains’ could also be assayed (Rahman *et al.*, 2022), albeit also using a homemade approach. Indeed, custom approaches to Tn-seq analysis are the norm in the field, but several generalizing programs do exist. A noteworthy mention is TRANSIT, a multi-option program capable of inferring essential genes directly from sequencing data (DeJesus *et al.*, 2015). TRANSIT offers several parameters and claims to handle both Himar and Tn5 data, however, the question remains on the best method to choose, and how to define a threshold for calling a certain gene essential, especially in the ‘fitness gene’ cases?

Here, we developed TnSeeker, a novel ‘one-fits-all’ pipeline towards general Tn-seq data analysis. TnSeeker is able to automatically define an essentiality threshold from “gold set” genes, and can adjust its calculation to any kind of transposon and genome content bias. Moreover, it can also assess possible preferences in transposon orientation relative to gene and evaluate *loci* at the sub domain level. To benchmark TnSeeker, we generated saturating Tn5 libraries for 8 different strains of *Escherichia coli* and determined their respective pan-essentialome (shared essential genome).

Results

The pKMW7 Tn5 vector successfully creates random transposon libraries in *E. coli* natural isolates

Using the randomly barcoded pKMW7 Tn5 transposon suicide vector library (Wetmore *et al.*, 2015), we performed a pilot transposon mutagenesis assay in 34 natural isolate strains, and determined the respective transformation efficiency by CFU counting. The strains for which transposon library saturation could be efficiently achieved (22 out of 35 tested had a conjugation/electroporation transformation efficiency >= 10^{-6}) were short-listed into 8 in such a way that phylogeny distribution, pathogenic/commensal diversity, and environmental niche differences were maximized (Supplementary table 1).

Table 1 | Summary of the built pilot transposon libraries.

The transposon library corresponding to BW25113 was sequenced twice. Library size corresponds to the total number of collected mutants following selection by plating.

Strain	Unique insertions (MAPQ≥40, Phred≥10)	Unique insertions (MAPQ≥0, Phred≥0)	Library size (CFUs)
BW25113 #1	38,906	44,353	180,000
BW25113 #2	14,649	17,354	180,000
IAI 16	23,689	27,894	110,000
IAI 33	16,205	18,977	96,000

Based on the calculated transposon mutagenesis efficiencies, initial transposon libraries were built for 3 strains to obtain a total number of mutants in the 100,000 – 200,000 range (theoretically, an insertion every ~31bp). Despite obtaining such CFUs, we discovered the real number of unique transposon insertions (same chromosome position and orientation) to be only around 20% of the total library size (table 1). Furthermore, only 4,784 insertions were in common between the 2 sequencing runs of the same BW25113 library, suggesting suboptimal sequencing coverage. As expected, mapping parameters and read quality filtering influenced the number of found unique transposon insertions, with no quality control (MAPQ≥0, Phred≥0) returning on average 15% more unique insertions than the chosen filtering parameters (table 1; see methods). To better limit sequence and alignment errors, and thus

possibly bias downstream insertion location analysis, we opted for the more stringent filtering method. It is noteworthy that inherent differences in strain growth dynamics probably influenced Tn5 insertion distribution due to a distinct number of total generations, thus introducing biases into any downstream comparisons (supplementary figure 1).

In order to better control the number of elapsed generations upon library building, and therefore minimize any arising biases from differential transposon site clearance across all transposon libraries, a new transposon mutagenesis method was adapted from the one employed by Wetmore *et al.* (Wetmore *et al.*, 2015). Essentially, following conjugation, transposon mutant selection is directly performed in liquid media, instead of overnight on a solid agar surface, under continuous exponential growth until 30 generations pass, upon which the library is frozen. Using this method, new libraries with total mutant CFU > 1M were constructed (table 2).

Despite the larger library sizes, the total number of recovered unique insertions was still 10-15% of total CFUs. In the case of the BW25113 strain (the larger library) an insertion was observed every 18bp (every 1.4bp was the theoretical expectation).

Read and insertion distribution analysis showed Tn5 integration over the entire genome, albeit with low coverage (median < 1 read per million [RPM] per insertion), with sporadic insertions capturing large amounts of reads (12,545 RPM as the extreme value). No major differences in insertion frequency were observed between the positive and negative DNA strands across all libraries, but an enrichment in both was observed around the origin of replication, probably arising from multifork DNA replication (figure 1 and supplementary figure 2). Tn5 nucleotide insertion bias analysis of all libraries also revealed no strain or Tn5 mutagenesis method differences. In fact, independently of the used transposon library building experimental method (plated Vs. pooled), Tn5 displayed significantly more affinity for dinucleotides starting with G/C than expected by random chance (figure 1B).

Table 2 | Summary of the transposon libraries built during this study.

Total and unique genes available in the *E. coli* 909 strain panel were determined based on the work of Galardini *et al*. (Galardini *et al.*, 2017). In here, unique genes are genes that do not exist in the other used strains. Unique essential genes correspond to genes that are uniquely essential in only 1 of the libraries, having explicitly been deemed non-essential in the remaining 7 libraries (or not existing) (supplementary table 2). Genes too small for statistical inference ("Genes too small for assaying") in any of the libraries were not considered. These correspond to genes for which there is no confidence in essentiality prediction (see methods). Only read alignments with a nucleotide Phred-score ≥10, and mapping quality (MAPQ) ≥ 40 were considered to be valid. When appropriate, the insertion sites of the same library, independently sequenced several times, were merged together for essentiality calculations. All libraries were built using antibiotic selection in liquid media for exactly 30 generations.

Strain	**Total genes**	**Unique Genes**	**Essential genes** (% total w/o N/A)	**Unique Essential Genes** (w/o N/A)	**N/A** (Genes too small for significance)	**Unique insertions** (MAPQ>=40, Phred>=10)	**Library size (CFUs)**
BW25113	4,313	432	267 (6.5%)	45	202	257,952	3,240,000
UTI 89	4,839	665	257 (5.5%)	21	131	251,437	1,396,000
IAI 33	4,562	645	219 (5.2%)	27	472	58,071	2,065,000
Nissle 1917	4,589	497	220 (5.5%)	55	561	93,436	288,300
IAI 16	4,481	661	253 (6%)	14	264	70,096	5,369,000
IAI 13	4,391	449	159 (3.9%)	2	321	98,372	8,850,000
NRG 857 C	4,542	531	103 (2.4%)	5	302	157,533	2,460,000
EcoR-63	4,573	481	211 (5.9%)	29	1,022	32,917	1,947,000

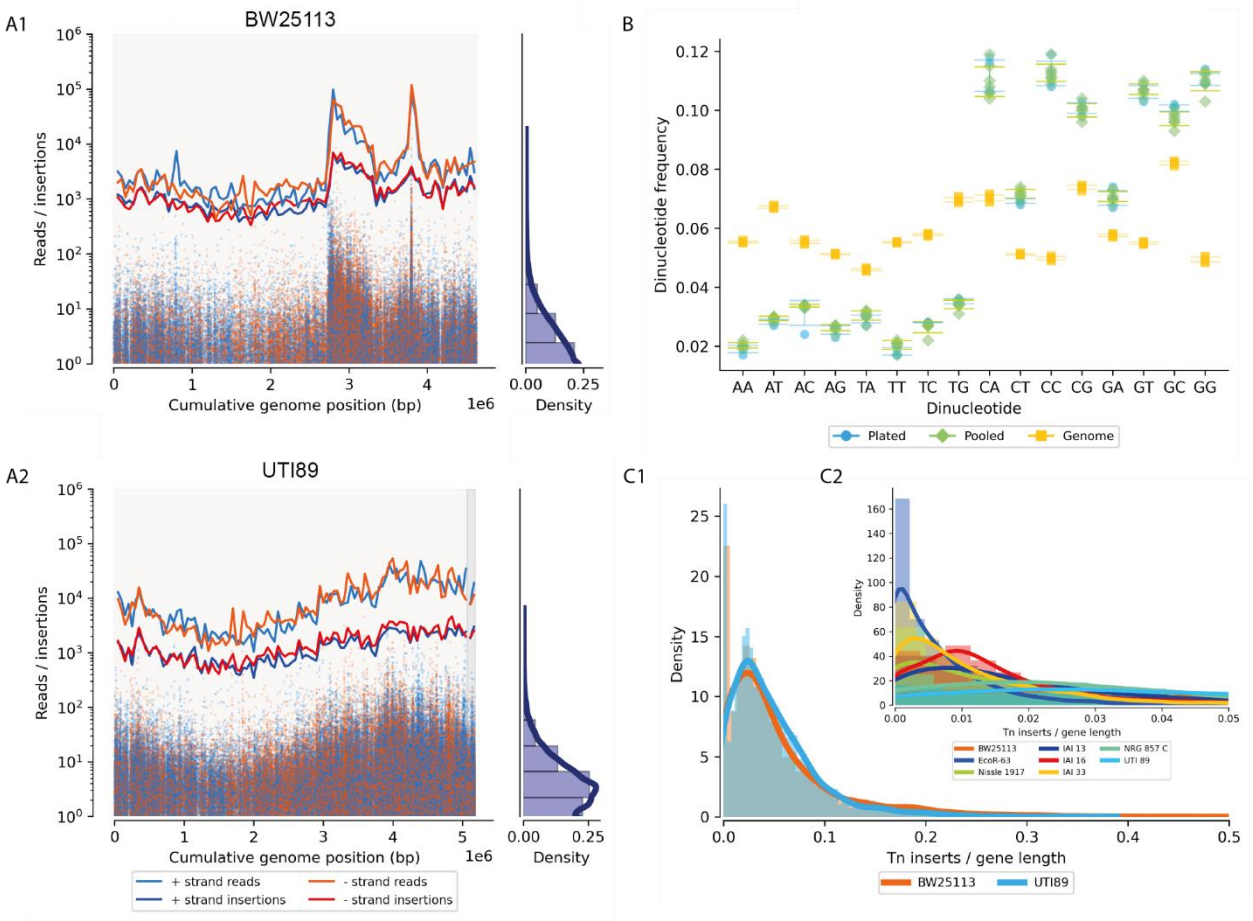


Figure 1 | Tn5 library bias analysis.

A) Read distribution (reads per million [RPM]) across the chromosome. Different contigs are indicated by different background shades. In the case of **A2**, such correspond to the chromosomal DNA and a plasmid. Number of reads and insertions in the positive and negative DNA strand, per bin of 50,000 bp, are indicated by the red/blue lines. Density plot represents the read abundance per insertion. Origin of replication is located at position 4,052,525 for UTI89 and 3,936,725 for BW25113. **A1**) BW25113 strain. **A2**) UTI89 strain. **B**) Chromosome dinucleotide Tn5 insertion bias in the current experimental setup, compared to genome bias. The first 2 bp for all libraries following the Tn5 insertion were used for calculation (plated method: table 1; pooled method: table 2). **C**) Total insertions per gene per gene length were calculated (disregarding insertions in the first and last 10% of a gene), and their distribution plotted. Density (y-axis) refers to kernel density estimation, and is therefore dependent on the units of the used data. A probability can be obtained by integrating the density over a given range (x-axis), which sums to 1. **C1**) Only the density for the BW25113 and UTI89 libraries is shown. **C2**) Density comparison between libraries with different Tn5 saturations.

The misleading issue of genes without transposon insertions: why experimental and local genomic context matters.

Insertion normalization per gene length demonstrates the existence of 2 distinct groups of genes: those without insertions, and those with. The range of the number of normalized insertions in genes with insertions decreases the less insertions a library has, whilst increasing the number of zero-insertion genes (figure 1C1 and 1C2). We used this latter phenomenon as a naïve statistical control attempt at inferring essential genes, where all genes without insertions between the first and last 10% of a gene were deemed essential. Using this method, the number of inferred essential genes is shown to be inversely related with library size, rendering this approach unreliable for most of the built libraries where hundreds or thousands of essential genes were returned (supplementary figure 3A). Indeed, despite recapitulating previously described BW25113 essentials to a great degree (>90% cumulative overlap with 3 independent datasets) (supplementary figure 3B), such approach requires large transposon libraries to be accurate, while still seemingly still missing several essentials (lower precision). This is due to the assumption that all genes without any insertion are essential, thus not considering the role of mixed genes (genes with essential and non-essential domains); unresolved essential insertions; relative transposon orientation biases; the absence of normalization for gene GC content, overall transposon insertion frequency; and the lack of an adequate definition of a statistically significant threshold for deeming a feature essential. Such problems seem to persist in our datasets even when using more sophisticated alternative methods like the TRANSIT program, or ANUBIS. Indeed, both returned more than 600 essential genes (DeJesus *et al.*, 2015; Miravet-Verde *et al.*, 2020) (supplementary figure 3C and 3D). To address these issues, we developed a novel Tn-seq analysis pipeline, named TnSeeker.

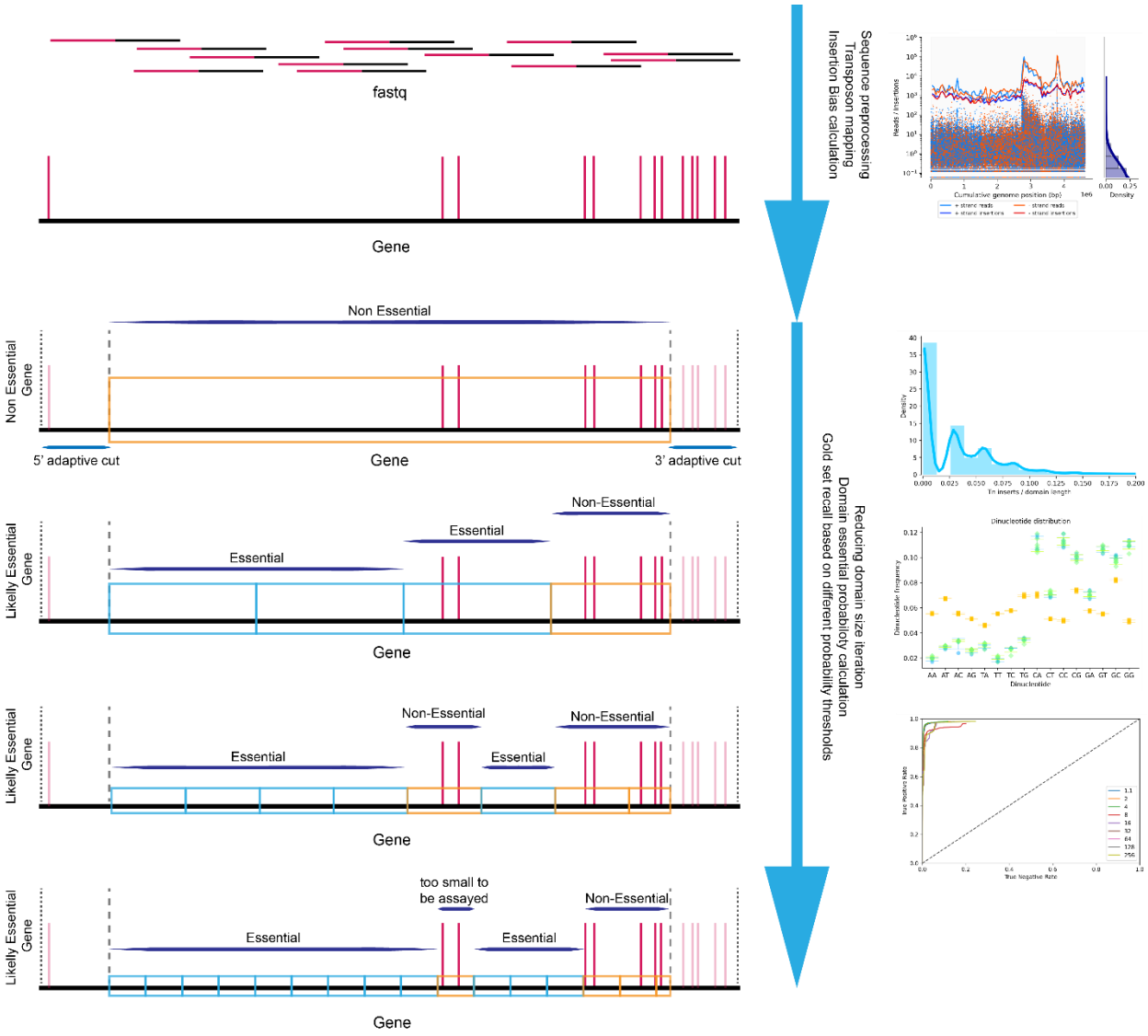


Figure 2 | TnSeeker pipeline.

Reads are processed based on the existence of an input transposon sequence using a custom-made algorithm similar to 2FAST2Q (chapter 1), and mapped to the bacterial genome using Bowtie2. Iterative sub-feature subdividing then proceeds, with significance being determined at each stage. Ultimately, an optimal domain size and significance threshold is chosen based on the recapitulation of the essential and non-essential genes of a reference gold set.

TnSeeker uses a high-confidence conservative approach for inferring essential domains

TnSeeker infers essentiality by combining linear density (insertions per unit of genome) with a modified gene specific sliding window approach, while performing self-optimizing thresholding based on a gene gold set. Transposon sequence insertion biases are also automatically compensated according to the local nucleotide distribution.

TnSeeker exists as a Python3 package pipeline, requiring Bowtie2 to be installed (and callable) for sequence alignments (Langmead & Salzberg, 2012). The program consists of two main parts: 1) Read trimming based on the existence of a transposon sequence, followed by read alignment which ultimately results in the creation of a TnSeeker-specific formatted insertion table used as input for; 2) The essentiality inference program that returns a fully annotated table with the essentiality classification of every feature, sub-domain, and domain-specific transposon direction bias (figure 2) (see methods). Moreover, TnSeeker is also capable of simultaneously extracting transposon barcode sequences and associating them with specific transposon locations, while determining their uniqueness and abundance in the overall pool, a feature explored in detail in chapter 4.

TnSeeker's key method is its ability to automatically determine an optimal genomic window size to subdivide all the genomic features into. This is performed for each gene by iterating through increasingly larger windows, calculating the probability of the observed transposon insertions being smaller than expected by chance for each such domain (whilst adjusting for transposon nucleotide bias), and then iterating again through different essentiality probability thresholds until an optimum is found (see methods). This latter is calculated from the true positive/negative gene recall of every such parameter combination, and then the iteration restarts. These true positive/negative essential genes are here defined as the 'gold set' genes, and consist of an user defined set of genes that are supposed to be either essential (positive), or mostly non-essential (negative), in any bacterial species (see methods). This thresholding process can be visualized via receiver-operator curves (ROC), where, for each domain size and probability threshold, the optimal point that maximizes true positive and minimizes true negative genes can be inferred from. Indeed, the optimal curves for all the built libraries in this study, plus 3 others from different studies, are

shown in figure 3B, with figure 3A displaying the linear density distribution of the optimal domain window size for strains BW25113 and UTI89. By comparing with the naïve approach distribution shown in figure 1C, it's possible to observe how subdividing each gene creates a more discrete linear density distribution. As this latter is closely related with the abovementioned probability of transposon insertion, it is thus easier to automatically iterate and determine a threshold at which a given domain with a given linear density is deemed essential of not. Conversely, it is also possible to determine genes or domains that either are, or become too small to be statistically assayed. These consist of features that, for any given linear density, no statistical significance can possibly be determined considering the reduced size of the feature relative to the observed library transposon saturation. Such features are deemed by TnSeeker as limbo genes (or 'too small to be assayed'), as they cannot be confidently classified as either essential or non-essential. Most of the current Tn-seq methods seem to not take into consideration such possibility, thus skewing essentiality in favor of smaller genes, especially at lower library saturations. TnSeeker is in this regard conservative, as it will avoid classification of ambiguously essential features, thus minimizing false positives at the cost of increased limbo genes. Such can be observed in figure 3C, where the effect of library saturation on the percentage of returned essential genes is much smaller for TnSeeker than for the naïve method (where every gene with 0 transposon insertions is considered essential). This is further highlighted in figure 3F, where the different combinations of essentials overlap returned only by other programs/methods (TRANSIT, ANUBIS, and naïve) for the same BW25113 Tn5 library is increased when the limbo genes are included. A similar increase is also observed in the interaction overlap between all programs except TnSeeker and the independent dataset by Koo BM. *et al.* (Koo *et al.*, 2017). Both these factors indicate that TnSeeker indeed avoids classification of a certain category of genes, and that such genes tend to be overrepresented essentials in other programs, while not necessarily being true essentials due to their lower overlap with other methods. For example, TRANSIT returned an excess of method-specific essentials, whilst not resulting in greater accuracy (compared with the remaining methods consensus). In fact, ROC analysis interestingly demonstrates that the true-negative rate was on par with the stringent naïve approach, despite this latter's better sensitivity (figure 3D). A similar effect is also seen when comparing TnSeeker with essential genes published

in other datasets (figure 3E), with the differences between these different methods and libraries also being evident in the ROC analysis (figure 3D). In the case of the Goodall ECA *et al.* dataset (Emily C. A. Goodall *et al.*, 2018), the ratio of true negatives is higher than the true positives, further reinforcing the effect that data analysis, in this case the choice of the gold set genes, can have on essentiality inference (see methods).

Interestingly, only 12 genes were uniquely assigned as essential in BW25113, while being labelled as non-essential in the remaining datasets (figure 3E). Among these, 4 of them are directly related with anaerobic and nitrogen metabolism (ydgN, fnr, glnG and glnD) (supplementary table 3), possibly more revealing of the conditions the library was incubated on, and not deeming any further in-depth study.

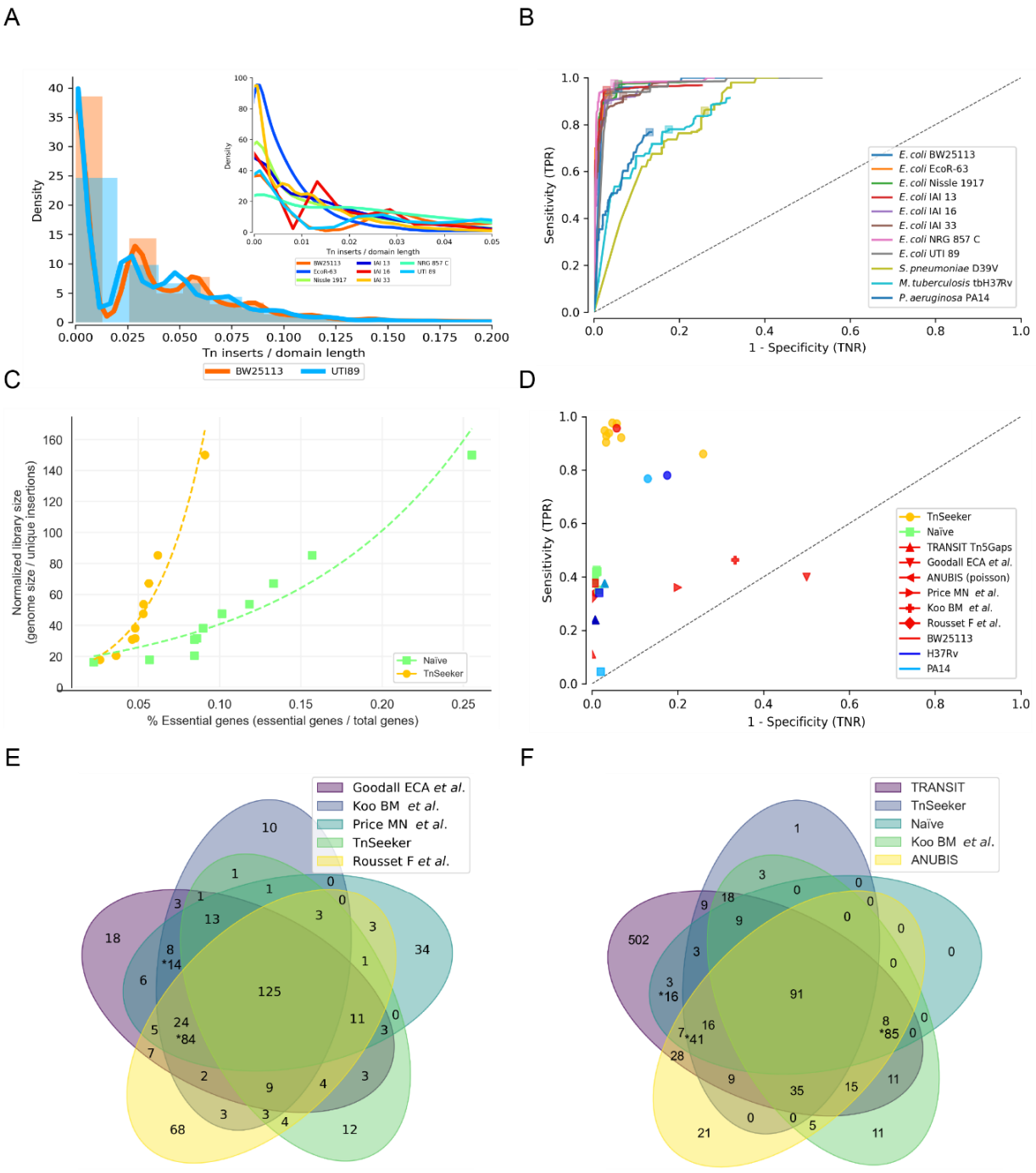


Figure 3 | Outcomes and comparisons for distinct essentiality methods.

A) Optimal TnSeeker divider domain linear density distribution for all the built Tn5 libraries strains, highlighting BW25113 and UTI89, the highest saturation strains. **B**) Optimal ROC for all analyzed transposon libraries in this study. For each strain, and for a given domain size, the line indicates how changes in the significance thresholding impacts essentiality gold set recapitulation. The reference gold set of true positive and true negative genes were obtained as mentioned. **C**) Analysis on how library saturation impacts the number of essential genes when using the naïve approach, and the conservative TnSeeker method. **D**) ROC plot with the optimal domain size/significance threshold point indicated. Naïve Vs. TnSeeker method comparison are indicated for all analyzed transposon libraries in this study (including a *Mycobacterium tuberculosis* set (Carey *et al.*, 2018), a *S. pneumoniae* set (Liu et al MolSystBiol 2017) and a *Pseudomonas aeruginosa* set (Poulsen *et al.*, 2019)). ROC was also determined for different published datasets of BW25113. **E**) Venn diagram comparing the

essential genes obtained using different datasets (Emily C. A. Goodall *et al.*, 2018; Koo *et al.*, 2017; Price *et al.*, 2018; Rousset *et al.*, 2021), with TnSeeker. Rousset F. *et al*. used a CRISPRI library instead of a transposon. **F**) Venn diagram comparing the essential genes returned by TnSeeker, TRANSIT (DeJesus *et al.*, 2015) (Tn5Gaps), and ANUBIS (Miravet-Verde *et al.*, 2020) (Poisson) for the built BW25113 library, and an independent dataset by Koo BM. *et al.*. * indicates the number of overlapping genes when including genes deemed too small to be assayed by TnSeeker in the comparison. When possible, all comparisons were performed using standardized gene annotation names. Genes without such names were discarded.

TnSeeker was further benchmarked with 3 independent transposon libraries originating from *Streptococus pneumoniae* (Jan Willem Veening lab, unpublished), *Pseudomonas aeruginosa* (Poulsen et al., 2019), and *Mycobacterium tuberculosis* (Carey et al., 2018). Regarding the two latter, essentials returned by TnSeeker were compared to the essentials from TRANSIT, and the consensus from the OGEE database (Gurumayum *et al.*, 2021) (supplementary figure 5A and B). In both cases, TnSeeker optimized the accuracy return rate of true essentials above the naïve method (figure 3B and D). In *M. tuberculosis*, only 2 genes were not in common with the remaining datasets, contrasting once again with the TRANSIT method (506 differentially essential genes). For *P. aeruginosa*, a larger heterogeneity across all datasets was observed, possibly indicative of the organism larger genome and the higher saturation library (> 200.000 insertions), where the essentiality of more genes can be assayed. In fact, only 21 out of 5771 genes were deemed too small to be assayed, whereas in BW25113 the number was 202 out of 4313.

The dataset for *S. pneumoniae* was directly compared with previously published datasets from both Tn-seq and CRISPRi experiments (supplementary figure 5C). As this latter examines essentiality at the operon level, we report essentiality at both the operon and gene level.

TnSeeker infers domain level transposon orientation biases

Besides inferring essentiality, TnSeeker also examines biases in transposon insertion direction relative to gene orientation. For domains/genes with large amounts of insertions, such feature indicates if there is a negative bias towards having transposons oriented in a certain direction. Therefore, similar to how the significant lack of transposon insertions indicates a non-random event (such as an essential domain), so does the non-random distribution of transposon orientations over a certain domain indicate some kind of transposon orientation dependent significant bias.

By examining strain BW25113, around 1000 non-essential gene domains were significantly skewed in regards of transposon orientation (FDR corrected p-value <= 0.01). 88% had a transposon orientation bias towards the same orientation of the gene, indicating that in the majority of the cases an opposite gene strand transposon insertion was either not occurring due to local genomic context factors, or resulting in cell death.

Open reading frame prediction (ExPASy) of the 2 most heavily biased transposon genes in BW25113 (table 3 [*ydiY* and *ybhI*]) revealed alternative coding sequences (a 68 amino acid (a.a.) long sequence for *ydiY*, and a 25 a.a. for *ybhI*), albeit without canonical *E. coli* shine-Dalgarno sequences. *ydiY* was also the most significantly transposon orientation-biased gene for the UTI89 strain, whilst being in the top hits of Nissle 1917. Curiously, *ydiY* is directly upstream (and in the opposite strand) of *pfkB* II, a gene encoding an alternative enzyme in the glycolysis pathway. This orientation bias could thus be related to possible lethal disturbances on this gene. Another possibility could be the existence of a non-annotated essential gene between these two genes.

Despite reporting a large number of biases, relative insertion position analysis revealed, for most cases, an evenly spaced distribution of opposite insertions across the genes, with the biases originating from domains with highly concentrated same orientation insertions (figure 4), and thus unlikely to be of biological significance.

Table 3 | Top 8 most significant genes with a bias in transposon orientation in the BW25113 Tn5 library strain.

Total Tn insertions	Gene Description	Orientation ratio (+/total)	Orientation p-value	Gene orientation	Gene name
111	acid-inducible putative outer membrane protein	0.0089	6.76×10^{-31}	-	ydiY
138	putative outer membrane protein	0.0780	1.90×10^{-25}	-	yiaT
255	putative mannitol-specific PTS enzymes: IIB component/IIC component	0.1839	8.28×10^{-24}	-	NT12004_22_02874
395	putative sigma-54-interacting transcriptional activator	0.2523	2.91×10^{-22}	-	ygeV
117	putative LuxR family transcriptional regulator	0.8916	5.79×10^{-21}	+	yqeH
83	dTDP-glucose 4%2C6 dehydratase%2C NAD(P)-binding	0.0595	1.88×10^{-17}	-	rfbB
363	putative Zn-binding dehydrogenase	0.2727	3.13×10^{-17}	-	yggP
72	putative transporter	0.9324	1.49×10^{-16}	+	ybhI

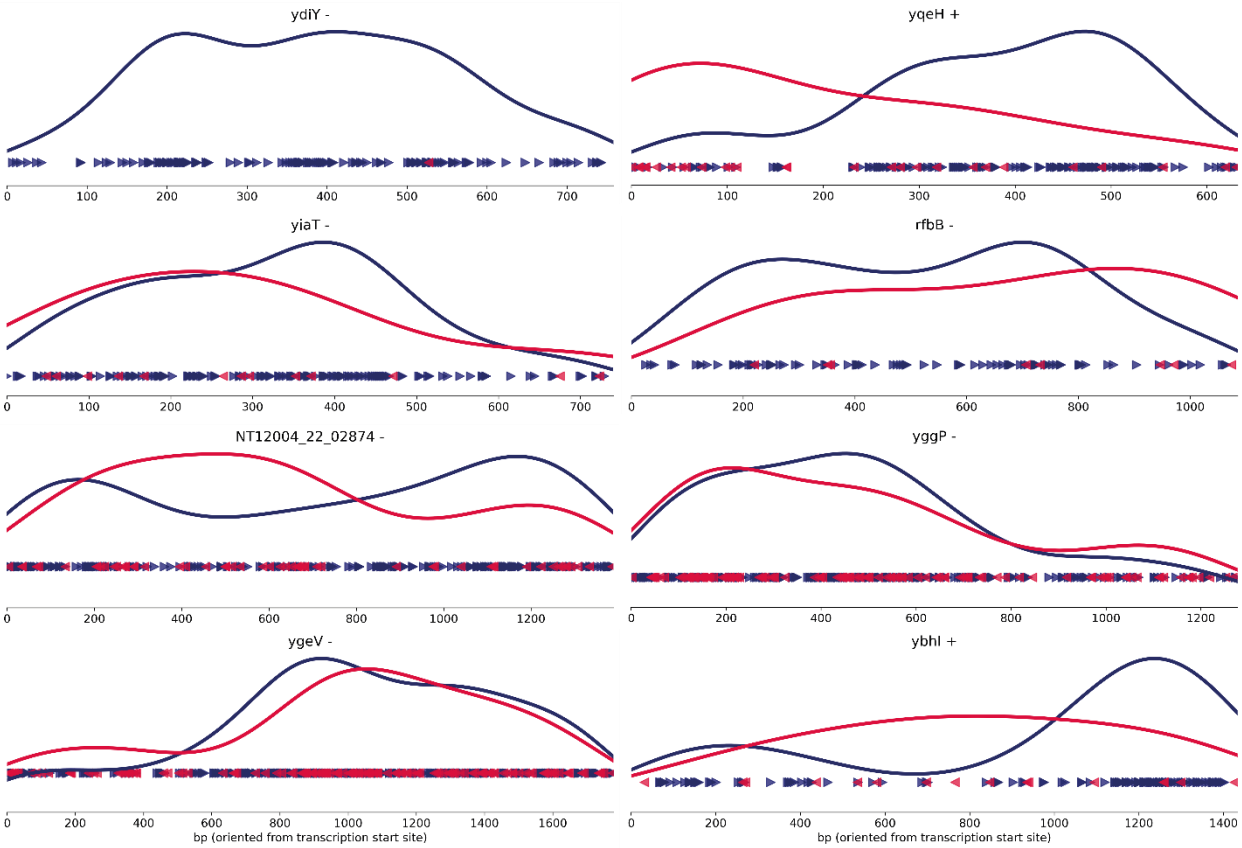


Figure 4 | Distribution of transposon insertion orientation for the top 8 most orientation biased genes in strain BW25113.

For each gene, the orientation and relative location of every transposon is indicated. Arrow direction indicates the transposon orientation relative to the gene (blue arrow: same direction; red arrow: opposite direction). Blue and red curves indicate the relative density distribution of the same colour parameters, respectively. Gene names and relative genome orientation are indicated in the titles.

Tn-seq reveals a broad *E. coli* pan-essentialome

Essentialome analysis of all the built Tn5 libraries revealed a core-essentialome (genes considered to be essential in all strains) of 70 genes (figure 5A), and a pan-essentialome (genes considered to be essential in at least 1 strain) of 664 genes (12.5% of which are hypothetical proteins) (figure5B). No correlation between the pan-essentialome and the phylogenetic distance was observed at the strain level, unlike with the pan-genome, which clustered based on phylogeny (supplementary figure 6).

We did not remove duplicated genes from this analysis due to the risk of ignoring possible essential genes, thus incurring the risk of over-reporting rather than under-reporting. Nonetheless, we determined that most of such genes are associated with transposons and bacteriophages, corresponding to 6.7% of the pan-genome,

6.3% of the pan-essentialome, and 0% of the core-essentialome. The most frequent reasoning for ignoring such genes is the possible biases arising from imprecise read mapping due to similar chromosome sequences, the presence of transposon coldspots (areas recalcitrant towards accepting transposon insertion events), and/or their mobile nature (not being in the genome anymore, and thus creating a false transposon free region). Indeed, several of these are unique essentials in the strain BW25113, having the highest number of unique essential genes, higher than the similarly saturated library UTI89 (table 2 and supplementary table 2) (comparison with lower saturating libraries might induce biases due to a higher number of limbo genes, so direct comparison was not attempted). Closer inspection of such genes revealed these to be mostly related with transposases and other mobile sequences, with only 15 not being related with these elements. 6 others were possibly duplicated genes as they shared the same putative function (Rhs family toxin). These were thus likely false essentials, bringing the total number of unique BW25113 essential genes to 9.

When removing all the prophages and other mobile elements from all the unique essentials across all strains, the total number of strain unique essential genes decreased from 198 to 159. We nonetheless report all these instances as the causes for these genes' essentiality (or lack of) might require consideration on a strain-by-strain basis, and consequently should not be easily dismissed.

Both the pan-genome and pan-essentialome size increased following a rarefaction curve model. Conversely, the core-essentialome started to plateau even after the combination of any given 3 strains (figure 5C1 and 5C2). A curve was fitted for all 3 different datasets and the theoretical asymptotic maximum determined to be 13 strains (~700 genes) for the pan-essentialome, 15 for the pan-genome (~10.000 genes), and 6 for the core-essentialome (~70 genes).

Clusters of Orthologous Genes (COG) enrichment analysis of the core-essentialome only revealed significant enrichment for genes involved in translation, such as ribosome biogenesis (Figure 5D1, category J). Moreover, and surprisingly, in the pan-essentialome genes related to transcription were significantly depleted (figure 5D1, category K), with mostly central metabolism and cell replication pathways being enriched. In both datasets, genes of unknown function, or yet still unlabeled, were also significantly depleted. At the strain level, similar patterns were observed for the essentialomes of the 8 analyzed strains (supplementary figure 4).

No significantly differential COGs were detected when considering only the strain specific essential genes. However, a closer examination revealed that 28% of these were labeled as hypothetical proteins, and 20% were annotated as part of a mobile genetic element, such as transposons or prophages. Interestingly, some iron-transport related genes were found, such as the genes encoding the iron enterobactin transporter FepA and FepG on strain IAI33, the ferric iron-catecholate transporter on Nissle 1917, and the ferrous ion transporter EfeO on IAI 16. Other lipid-related transporters were also detected. Moreover, several virulence genes related to adhesion, protease, and toxin production were also found. However, due to the low saturation of some of these libraries, their incompletely sequenced genome (only BW25113 and UTI89 have a complete assembled genome), or even the possible existence of plasmids that can vanish from the population, caution is required when considering all strain dependent essentialomes.

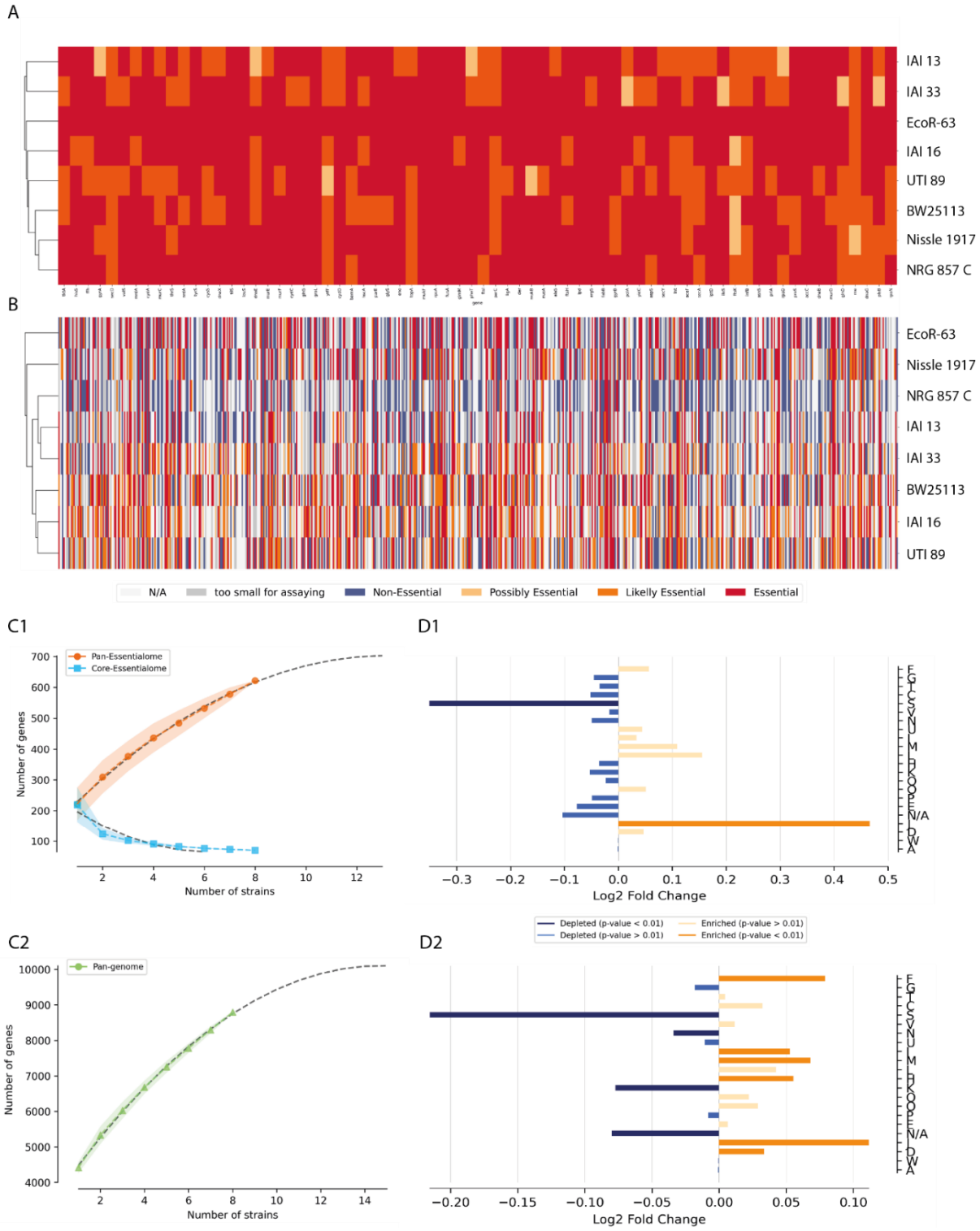


Figure 5 | *E. coli* Pan & Core essentialome.

A) Heatmap plot of the core essentialome for all the built transposon libraries. A gene was deemed 'core' if it was classified as essential by TnSeeker (getting a classification of esssential', 'likelly essential', or 'probably essential') in all the 8 analyzed transposon libraries. **B**) Heatmap plot of the pan-essentialome for all the built transposon libraries. A gene was considered to be part of the pan-essentialome if it was considered essential in, at least, 1 strain. **C**) Effect on the essentialomes of progressively considering the pooled essential genes of an increasing number of strains. The data was obtained by performing 100 independent

random sampling events of the essential genes of all 8 library combinations. Data fitting and extrapolation was performed using the Python library scipy curve_fit function to adjust the data to a polynomial function. **C1**) Core and pan-essentialomes. **C2**) Pan-genome. **D**) COG enrichment analysis for the built 8 Tn5 *E. coli* libraries (see methods for nomenclature description). **D1**) core-essentialome. **D2**) pan-essentialome.

TnSeeker data exploration reveals species and strain specific biases at the gene essentiality level

Gene orientation analysis revealed no significant strand bias for essential genes in both the BW25113 and UTI89 strains. No relation between gene essentiality and GC content was also observed, and the correlation between GC content and transposon insertion location has already been previously addressed (figure 1B).

Essential genes displayed a slight preference for regions closer to the origin of replication (ORI). In the BW25113 strain, essentials were 20% (median of relative gene location) closer to the ORI than non-essentials. A similar result was observed for UTI89 (16% closer), *P. aeruginosa* (13%), and *M. tuberculosis* H37Rv (18% closer). Regarding *S. pneumoniae* D39V, essentials were actually 12% further from the ORI when compared with the expected location of all genes (49% when using CRISPRi data from Xue Liu *et al.* (Xue Liu *et al.*, 2021)). Notably, D39V has the smallest genome (~2 Mb).

A positive bias in both read coverage and transposon insertion density was also observed towards the chromosome origin of replication, with negative bias at the terminus side (figure 6A, B, C, and D). In the case of BW25113, a sharp increase is seen at around position 2,750,000 of the chromosome. No known plasmids or significant mobiles elements are known to match this region in the used strain, and such result is possibly indicative of an unknown experimental artifact.

Besides chromosomal DNA, Tn-seq is also able to evaluate the essentiality of exogenous elements such as plasmids. Indeed, the UTI89 strain harbors a plasmid with 130 genes of which one, albeit a transposase, was deemed essential by TnSeeker.

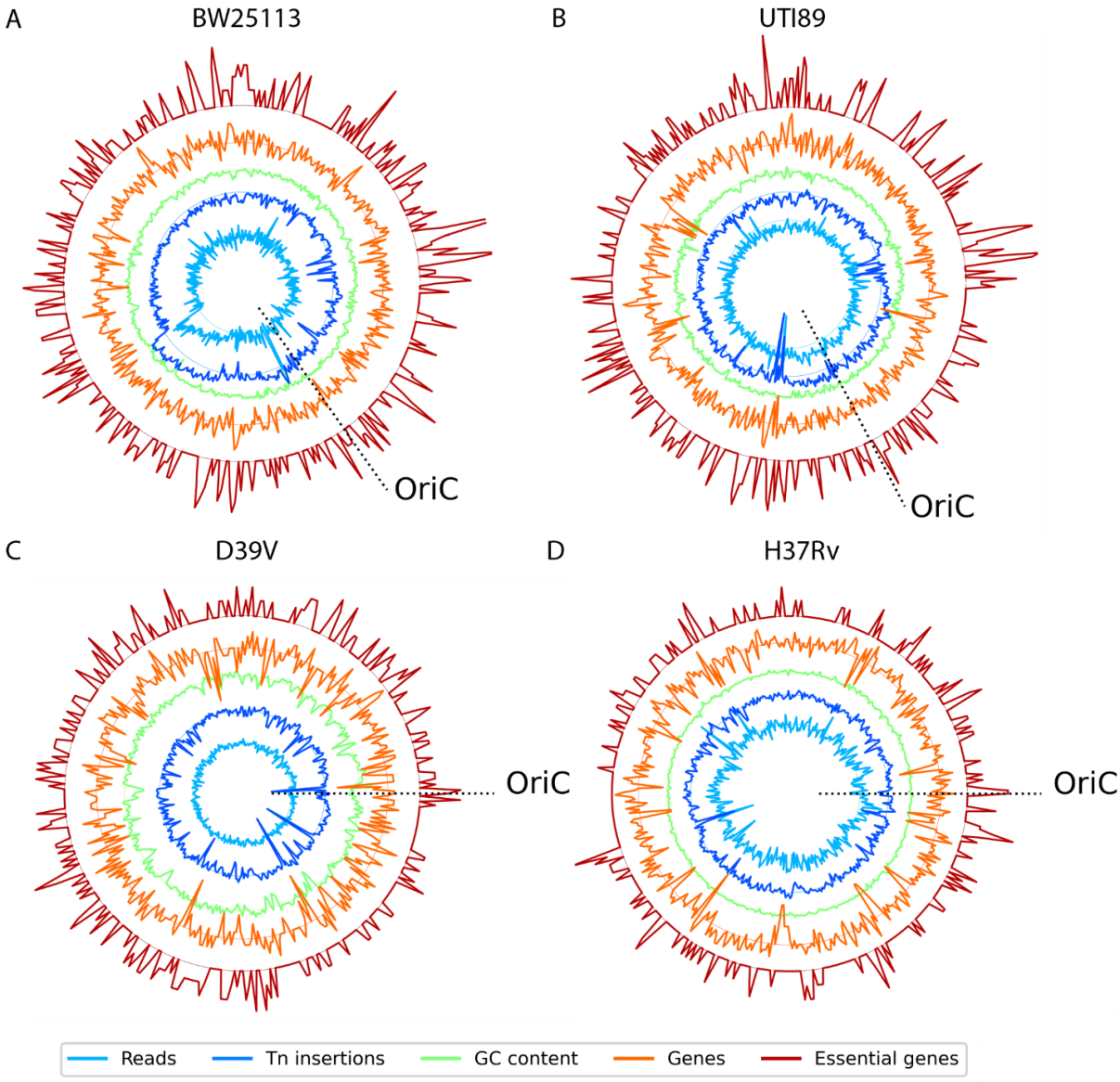


Figure 6 | Strain signature plots.

Genome mapped diagram of the library differences in regards to transposon insertions and reads, strain GC content, gene distribution, and essential gene abundance. Each parameter was discreetly calculated by subdividing the genome into bins of 10Kbp and by computing the local median. The genome medians were calculated for comparison and are indicated as concentric circles represented by the same colour as the parameter they refer to. **A**) this study BW25113. **B**) This study UTI89; **C**) *Streptococcus pneumoniae* (Jan Willem Veening lab, unpublished). **D**) *Mycobacterium tuberculosis* (Carey *et al.*, 2018)

Discussion

The advent of NGS and other high-throughput techniques such as Tn-seq have moved gene essentiality genetic studies beyond model organisms and optimal laboratory conditions, into the pan-genome era. We are now able to better explore the apparent paradox of reduced core-essentialomes, and the variability of the pan-essentialomes. Indeed, each strain carries its own signature essentialome, smaller than the sum of all, and bigger than the common overlap.

In this work we built transposon libraries for 8 *E. coli* strains and developed a novel Tn-seq analysis program, termed TnSeeker. TnSeeker uses a self-optimizing stringent approach to infer gene and domain essentiality, better recapitulating true positives while minimizing possible false positives, and avoiding the classification of large number of genes as essential. This latter was observed to be specifically true in lower saturating libraries when compared with other methods. TnSeeker performs such task by implementing 'limbo genes': genes whose essentiality cannot be determined for a given transposon library saturation. Moreover, essentiality inference is optimized based on a list of known essential and non-essential genes that should/not be present, to some degree, in the assayed strain (gold set genes). The outcome of TnSeeker will therefore vary depending on the genes used for validation, and on their conserved annotation across strains. To limit such biases, we used a list of representative well characterized essential gene names from several bacterial species (inferred core-genome from OGEE). Future versions of TnSeeker would benefit from homology-based comparison for such task, mitigating miss-annotation risks.

Due to the use of a large number of known non-essential genes in the validation set, and considering reported strain-to-strain variations in essentiality (including this study), the true negative rate is perhaps not as important as the true positive recall rate. An example of these variations is exemplified when comparing the reported essential genes from Goodall *et al.* and Koo BM. *et al.* with our picked gene gold set. A large portion of non-essentials were labelled as essential (variation is expected on this axis due to the less stringent picking of non-essential genes in the gold set), however less than half the expected true positive genes were deemed essential. Such variations might be explained by the choice of using the OGEE database for the gold set, which relies mostly on published Tn-seq data, where various analysis

methodologies are used. Despite these obvious biases, TnSeeker nonetheless stringently recapitulated known essentiality not only for *E. coli* BW25113, but also for other species.

Besides essentiality calling, TnSeeker also examines transposon insertion orientation biases. However, the biological significance of such might not be of strait forward interpretation. For example, nucleoid proteins have been reported to locally influence transposon orientation by mediating strand accessibility and transpososome stability during the transposition process (Garsin *et al.*, 2004; Swingle *et al.*, 2004). Biases would then reflect, in this case, the influence of these proteins on the transposon. Another hypothesis is related to the transposon cassette itself. The fact that in most of the cases the orientation bias is in favor of the gene orientation could originate from some toxic effect related with opposite strand gene expression continuing from the transposon resistance gene. The orientation bias would then be related with orientation lethality, and not with a specific transpososome effect. Such could relate with the action of either an antisense RNA, or through negative peptides.

So named due to their phenotype prevailing over the wild-type (WT) form, inactivating it, dominant negative peptides (DNP) can disrupt the function of WT proteins by either the creation of WT-DNP non-functional complexes, or by competition through substrate titration (Dorrity *et al.*, 2019; Herskowitz, 1987). Gene domains with significant transposon orientation biases may thus indicate the existence of DNP whose expression causes the loss of cellular viability. Such could be the case of *ydiY*, which we reported to have transposon orientation bias in various strains and a conserved local genomic context, pointing towards some local level conserved mechanism, or the existence of a yet uncharacterized gene.

TnSeeker is also capable of calculating essentiality across unassembled genomes in different contigs, or plasmids, and merging the results into a single file. Indeed, such was performed for most of the strains in this study, with essentiality being calculated on a contig basis (supplementary figure 2). However, despite the existence of plasmids in at least one strain (UTI89), no further comparisons in this regard were possible due to the remaining strains existing only in partially assembled contigs, thus making distinguishing between chromosomal DNA and plasmid difficult.

In 2021, Rousset F. *et al.* published a CRISPRi-seq screen analysis of 18 *E. coli* strains across multiple conditions, although only focusing on 3,400 common genes. Despite a major overlap in essentials for BW25113 in the same media, an

overestimation of gene essentiality by Rousset F. *et al.* was observed, even when compared with other methods (figure 3E). Such can possibly be attributed to operon effects arising from the CRISPRi method. Indeed, comparative analysis performed with the *S. pneumoniae* D39V dataset revealed a tendency of CRISPRi to classify all operon genes as essential if at least one is essential (supplementary figure 5C).

Despite such discrepancies, similar conclusions are drawn regarding the lack of essentialome correlation with phylogenic distance, with correlation only emerging when considering all the genes in the strain. This latter is expected as such analysis is closely related with how phylogeny is determined. We also reported larger core and pan-essentialome sizes, although probably related to the Tn-seq ability to examine all possible genetic locations, and thus going beyond the 3,400 features used by Rousset F. *et al*.. We also didn't remove mobile elements from our analysis, which correspond to at least 6% of the pan-essentialome. Such elements, despite the difficulties in assaying their true essentiality, might play a role not only in conditional essentiality, but in driving gene essentiality evolution. Indeed, an increasing pan-genome might be explained in part by non-orthologous gene displacement via horizontal gene transfer (NOD-HGT), and consequent changes in gene synteny by gene gain/loss (Forterre, 1999; Martinez-Carranza *et al.*, 2018; Rousset *et al.*, 2021). The reduced size of the core-essentialome, smaller than the smallest strain essentialome, and thus likely missing essentials required to support life, is also probably related to such phenomenon. In these cases, unrelated proteins/genes perform the same essential function in different organisms, albeit using a different sequence, and thus belonging only to the pan-essentialome whilst being outside the highly conserved core-essentialome. Such is supported by COG analysis of the core and pan-essentialome (figure 5D1 and 5D2), where metabolic pathways required to support life are seen at the pan-essentialome level, but not on the core-essentialome. Curiously, in this latter, only ribosome biogenesis related genes are enriched, hinting at these as being the only genes that are ultra-conserved, of difficult functional replacement, or least submitted to NOD-HGT.

Tn-seq data analysis is complex, and sensitive to methodological variations at all levels. Throughout this study we focused on methodology benchmarking, both by building *de novo E. coli* transposon libraries, and using published datasets. Straightforward comparisons, however, despite being required as a validation

necessity, are prone to never fully agree between different studies. Caution is thus always advised when drawing any hard conclusions from these cases, perhaps being more interesting in the future to understand the common overlaps across conditions and analysis methods.

Similarly to recent works (Carey *et al.*, 2018; Coe *et al.*, 2019; Poulsen *et al.*, 2019; Rousset *et al.*, 2021), in here we highlighted the insight that using different strains to assess species level essentiality might have on both finding conserved essential pathways, and strain (or species) specific essential targets.

Methods

Strains

The APA766 strain (harboring the Tn5 library plasmid pKMW7) was a gift from Adam Deutschbauer (Wetmore *et al.*, 2015). For culturing APA766, LB was supplemented with diaminopimelic acid (DAP) to a final concentration of 300µM, and kanamycin (50µg/ml). The *E. coli* strains Nissle 1917, IAI16, IAI13, IAI33, EcoR-63, NRG857C, BW25113, and UTI89 are available in the Ecoref panel (Galardini *et al.*, 2017) and were routinely cultured in LB media at 37ºC, unless stated otherwise.

Data availability

All the sequence, proteome, and annotation data corresponding to the *E. coli* strains Nissle 1917, IAI16, IAI13, IAI33, EcoR-63, NRG857C, BW25113, and UTI89 was used as is from Galardini *et al. P. aeruginosa* sequencing dataset was compiled from SRR8907313_1 and SRR8907318_2. *M. tuberculosis* sequencing dataset was compiled from SRR12234126, SRR12465951, and SRR12473643. S*. pneumoniae* sequencing dataset was compiled from .sam files available as unpublished Jan-Willem Veening Lab data. Any remaining datasets were downloaded from works referenced as required.

Transposon Library Building

An overnight culture of the strain of interest was used for either electroporation with pKMW7 or conjugation with the APA766 strain as described by Wetmore *et al.* (Wetmore *et al.*, 2015). Briefly, for the conjugation process, an overnight culture of APA766 was diluted to OD_{578}=1.0 and 200µl plated on LB Agar supplemented with DAP. After 1h at 37ºC, 200µl of a diluted culture (OD_{578}=1.0) of the strain of interest was added to the top of the APA766 culture. Conjugation was performed for 4h at 37ºC, at which point LB supplemented with Kanamycin (30µl/ml) was added to the plate and stirred with a spatula. Serial dilutions were performed at this point for total library CFU estimation (to infer the expected library saturation). The mix of either conjugated or electroporated cells was diluted to a starting OD_{578} ~ 0.3-0.01 in LB supplemented with Kanamycin (30µl/ml) and grown for 30 generations in continuous

exponential phase, at which point the transposon library was frozen at -80C and total DNA extracted.

Transposon Library Sequencing

Total DNA extracted from the built transposon libraries was used as template for the nested PCR required to extract the Tn5-chromosome borders, and attach the Illumina sequencing adaptors, as described by Anzai *et al.* (Anzai *et al.*, 2017). Briefly, a PCR was performed with the primers Seq_1.transposon and Seq_1.Chrom.1 (Das *et al.*, 2005). This latter binds the transposon sequence near its end junction, and the first carries a random base pair sequence for randomly annealing to the bacterial DNA downstream of the transposon junction. A second parallel PCR was performed using Seq_1.Chrom.4, which has a different GC content from Seq_1.Chrom.1.

Table 4 | 1st nested PCR reaction for assembling the Tn5 Illumina library.

Reagent	*Amount*
Q5 reaction buffer	5 µl
Q5 Hot Start Polymerase	0.25 µl
dNTP's (5µM)	2 µl
Seq.1.Transposon (10uM)	0.5 µl
Seq.1 Chrome 1 (run another PCR with Seq.1 Chrome 4 (10uM)	0.5 µl
DNA (150ng)	X
H_2O	For 25 µl

Table 5 | 1st nested PCR reaction cycling protocol.

Temperature (ºC)	*Time*	*Cycles*
98ºC	5 min	1x
98ºC	30 s	6X
30ºC	30 s	
72ºC	1.5 min	
98ºC	30 s	25X
45ºC	30 s	
72ºC	2 min	
72ºC	5 min	1x

The products of both PCRs, comprised of all transposon borders, were merged and cleaned using a PCR cleaning kit (QIAquick PCR purification kit). A final 3rd PCR was then performed, indexing each sample as appropriate with an Illumina sequencing specific nucleotide sequence present in the 'Seq_2.Chrom.X' primers. A mixture of primers with increasing number of "N" random base pairs was used. This offsets the

PCR sequence thus optimizing sequencing cluster formation by avoiding saturating the same reading channel across all sequences.

Table 6 | 2st nested PCR reaction for indexing and assembling the Tn5 Illumina library.

Reagent	*Amount*
Q5 reaction buffer	10 µl
Q5 Hot Start Polymerase	0. 5 µl
dNTP's (5µM)	4 µl
Mix of Seq.2.Trans_4N/5N/6N/7N (1uM)	0.5 µl
Seq.2 Chrome X (index primer) (1uM)	0.5 µl
Purified PCR	5 µl
H_2O	For 50 µl

Table 7 | 2st nested PCR reaction cycling protocol.

Temperature (ºC)	*Time*	*Cycles*
98ºC	3 min	1x
98ºC	30 s	25X
52ºC	30 s	
72ºC	30 s	
72ºC	5 min	1x

The resulting Illumina library was cleaned using SPRIselect magnetic beads (Beckman Coulter) (0.8X size restriction) and submitted to multiplexed either single-ended or pair-ended sequencing (150bp) in a MiSeq or HiSeq 4000 apparatus at EMBL GeneCore.

Transposon Library Sequencing Analysis

The TnSeeker pipeline processed the data corresponding to all analyzed libraries, unless indicated otherwise. Essentially, FASTQ files were trimmed based on the existence of a Tn5 border (AGATGTGTATAAGAGACAG) junction and of a quality passing *E. coli* chromosome sequence (Phred-score >= 10 across the length of the chromosome DNA sequence). The trimmed reads were then mapped to their respective strain FASTA file using Bowtie2 (--very-sensitive-local alignment profile). TnSeeker then extracted the location of all the valid transposon-chromosome border alignments (filtering based on MAPQ>=40, and the expected FLAG values), as well as the 10bp following a given insertion (for later transposon insertion bias frequency calculation). When appropriate, the transposon locations of several independent libraries were pooled together to increase the resolution of the essentiality calculation.

Essentiality Analysis

Essential gene analysis was performed by TnSeeker, unless stated otherwise. Essentiality is determined by comparing the obtained insertion frequency of a given transposon across all combinations of 2-mers (dinucleotides) in the genome (resulting in different insertion probabilities for each pair, depending on the used transposon insertion bias), with the ones observed for any given domain. Essentiality then is, in this case, defined by a domain that has less transposon insertions than the one expected by chance over the entire genome.

To conduct these comparisons, we applied a Poisson-Binomial distribution model based on the Python module Poibin, from Mika J. Straka. With these considerations, any transposon insertion bias, and both the domain's individual length and nucleotide composition are taken into consideration when calculating essentiality. The resulting probability values were FDR corrected using the Benjamini-Hochberg procedure. This process was iteratively repeated starting from a domain size corresponding to the library saturation (i.e., a domain size corresponding to the average insertions per length of genome [for example, if one insertion is expected every 10bp, the iteration starts with 10bp size domains]), increasing until the domain size was larger than the largest gene. Thresholding was performed, for each domain size iteration, by progressively decreasing the significance threshold at which a domain is deemed essential, and determining the true-positive and false-negative discovery rate from a gold set. The gold set was obtained by combining the common overlap of all the genes deemed essential (or not essential for the true negatives set) from studies available in the OGEE database for *Acinetobacter baumannii, Bacillus subtilis, Escherichia coli, Haemophilus influenza, Mycobacterium tuberculosis, Pseudomonas aeruginosa, Staphylococcus aureus, Streptococcus pneumoniae*, and *Vibrio cholerae*. 71 genes constitute the true-positive dataset, and 6229 the true-negatives.

COG enrichment analysis

The COG terms for the proteome of the analyzed transposon libraries strains were obtained using the online version of the eggNOG v5.0 program (Huerta-Cepas *et al.*, 2019). A custom-made Python script was used to perform enrichment analysis

on the various datasets using a two-sided fisher exact test, with FDR correction. One-letter abbreviations for the functional categories:

- J - translation, including ribosome structure and biogenesis;
- N/A - non-assigned;
- C - energy production and conversion;
- O - molecular chaperones and related functions;
- G - carbohydrate metabolism and transport;
- E - amino acid metabolism and transport;
- M - cell wall structure and biogenesis and outer membrane;
- P - inorganic ion transport and metabolism;
- K - transcription;
- R - general functional prediction;
- H - coenzyme metabolism;
- N - secretion, motility and chemotaxis;
- F – Nucleotide transport and metabolism;
- L - replication, recombination and repair;
- S - no functional prediction;
- D - cell division and chromosome partitioning;
- U – Intracellular trafficking, secretion, and vesicular transport;
- T – Signal transduction mechanisms;
- I – Lipid transport and metabolism
- Q – Secondary metabolites biosynthesis, transport, and catabolism
- V – Defense mechanisms

Table 3 | Primers used in this study.

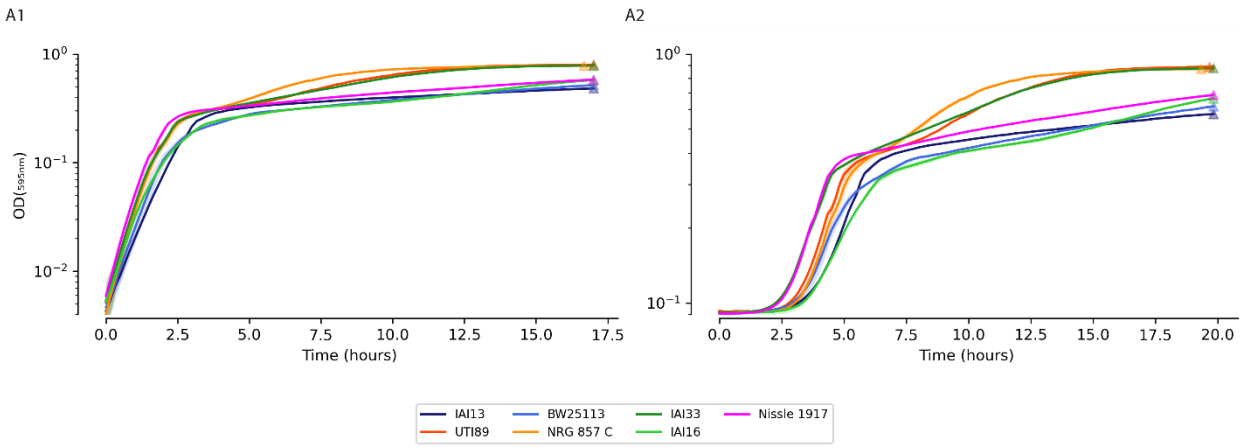
Primer Name	*Primer Sequence*	*Used Workflow*
Tn5_Transp_Fw	GGTAGTAAAGCCGCCCAGGAAG	Quality control during Tn5 optimization
Tn5_Transp_Rv	CCTGCGCCATCAGATCCTTG	Quality control during Tn5 optimization
Kan_test_Fw	ATGAGCCATATTCAACGGGAAACG	Quality control during Tn5 optimization
Kan_test_Rv	CGACTCGTCCAACATCAATACAACC	Quality control during Tn5 optimization
Transp_LOCALCHECK	CCAATTAACCAATTCTGATTAGAAAAACTCATCG	Quality control during Tn5 optimization
Nested_Transp_LOCHEK	AGAGACCTCGTGGACATCCC	Quality control during Tn5 optimization
Nested_Rand_Chromo	GGCCACGCGTCGACTAGTCA	Quality control during Tn5 optimization
Seq_1.transposon	CGATGAGTTTTTCTAATCAGAATTGGTTAATTGG	Illumina Sequencing
Seq_1.Chrom.1	GAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACGC	Illumina Sequencing
Seq_1.Chrom.2	GAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATAT	Illumina Sequencing
Seq_1.Chrom.3	GAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTCG	Illumina Sequencing
Seq_1.Chrom.4	GAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACTC	Illumina Sequencing
Seq_2.Transp_4N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCTGCAGGGATGTCCACGAGG	Illumina Sequencing
Seq_2.Transp_5N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNCTGCAGGGATGTCCACGAGG	Illumina Sequencing
Seq_2.Transp_6N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCTGCAGGGATGTCCACGAGG	Illumina Sequencing
Seq_2.Transp_7N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNCTGCAGGGATGTCCACGAGG	Illumina Sequencing
Seq_2.Chrom.1	CAAGCAGAAGACGGCATACGAGATAACGATGGGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.2	CAAGCAGAAGACGGCATACGAGATAAGCGCAAGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.3	CAAGCAGAAGACGGCATACGAGATAATTGCCGGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.4	CAAGCAGAAGACGGCATACGAGATACAGCTCAGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.5	CAAGCAGAAGACGGCATACGAGATACCTTGGAGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.6	CAAGCAGAAGACGGCATACGAGATACGAAGGTGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing

Seq_2.Chrom.7	CAAGCAGAAGACGGCATACGAGATACGAGATGGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.8	CAAGCAGAAGACGGCATACGAGATACGTCAACGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.9	CAAGCAGAAGACGGCATACGAGATACGTTCCTGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_2.Chrom.10	CAAGCAGAAGACGGCATACGAGATACTCGAGTGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
Seq_1.transp_alt1	ACGCTGCAGGTCGAC	Illumina Sequencing
Seq_2.Transp_4N_alt1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCGGTTGAGATGTGTATAAGAGAC	Illumina Sequencing
Transp_Sanger_1	AGACCGATACCAGGATCTTGC	Illumina Sequencing
Transp_Sanger_2	GAACTGCCTCGGTGAG	Illumina Sequencing
ilu.Seq_2.Chrom.1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.2	CAAGCAGAAGACGGCATACGAGATACACCGGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.3	CAAGCAGAAGACGGCATACGAGATGCCTGAGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.4	CAAGCAGAAGACGGCATACGAGATTGGATAGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.7	CAAGCAGAAGACGGCATACGAGATGGACTGGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing
ilu.Seq_2.Chrom.8	CAAGCAGAAGACGGCATACGAGATCAGTACGTGACTGGAGTTCAGACGTGTGCTCTT	Illumina Sequencing

Acknowledgements

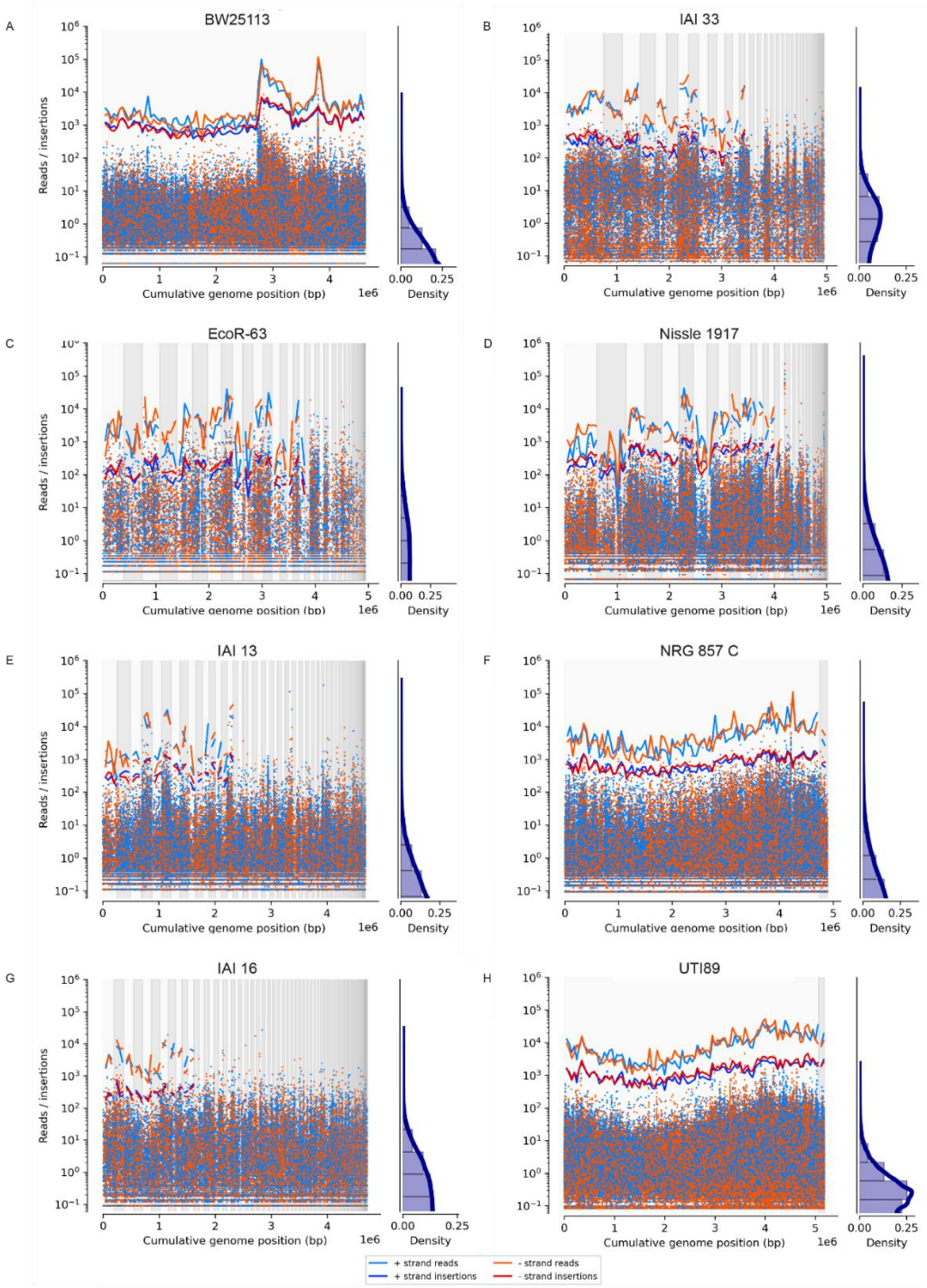
The author Afonso would like to thank Alexandra Koumoutsi for all her support and unvaluable help. Another big thank you goes to Vincent de Bakker, for both good comaraderie, and providing guidance through the statistics required to conduct this work.

Supplementary



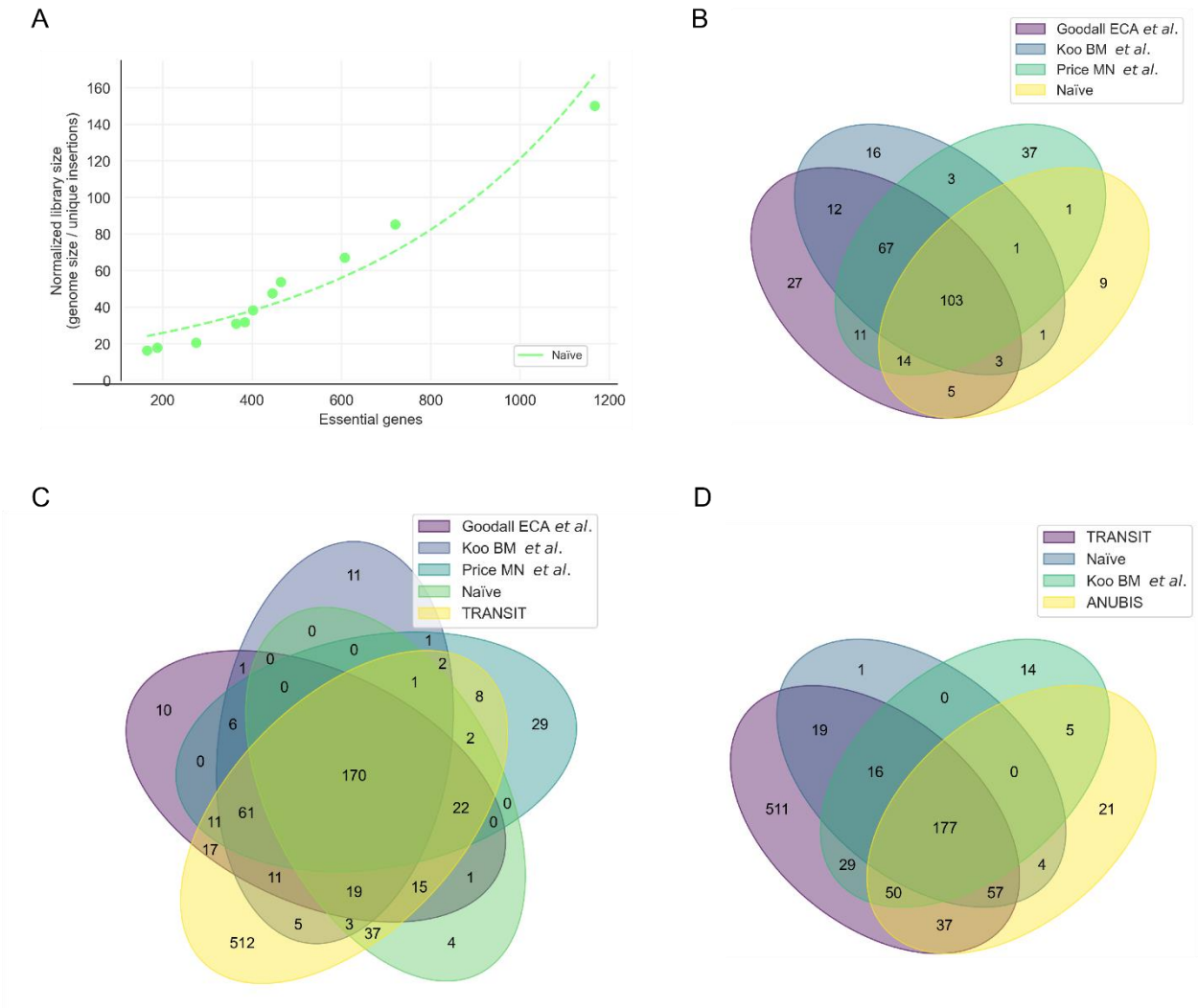
Supplementary figure 1 | Growth curves for the used Tn5 library strains.

A1) Normalized growth curves. The OD corresponding to the empty control was subtracted at each time point, for each strain. All curves were adjusted to start at their respective OD when the strain with the lowest OD crosses the value of 0.004 (lower readable value in the used instrument). **A2**) Un-normalized growth curves.



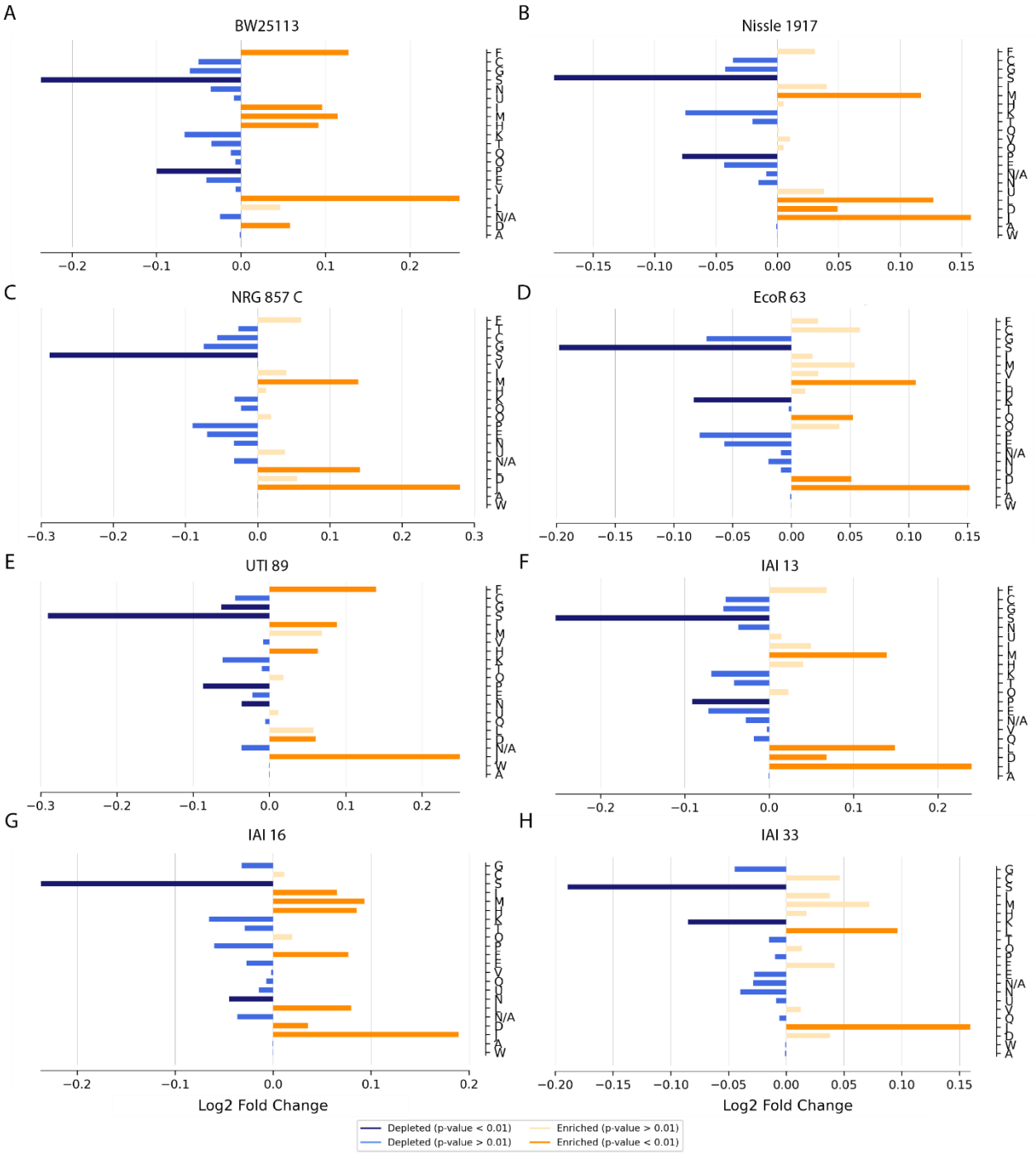
Supplementary figure 2 | Tn5 library read/insertion bias analysis.

Read distribution (reads per million [RPM]) across the chromosome (different contigs are indicated by different background shades, ordered from largest to smallest) for BW25113 (A), IAI33 (B), EcoR-63 (C), Nissle 1917 (D), IAI13 (E), NRG 857 C (F), IAI16 (G), and UTI89 (H). Number of reads and insertions in the positive and negative DNA strand, per bin of 50,000 bp, are indicated by the red/blue lines. Density plot represents the read quantity per insertion.



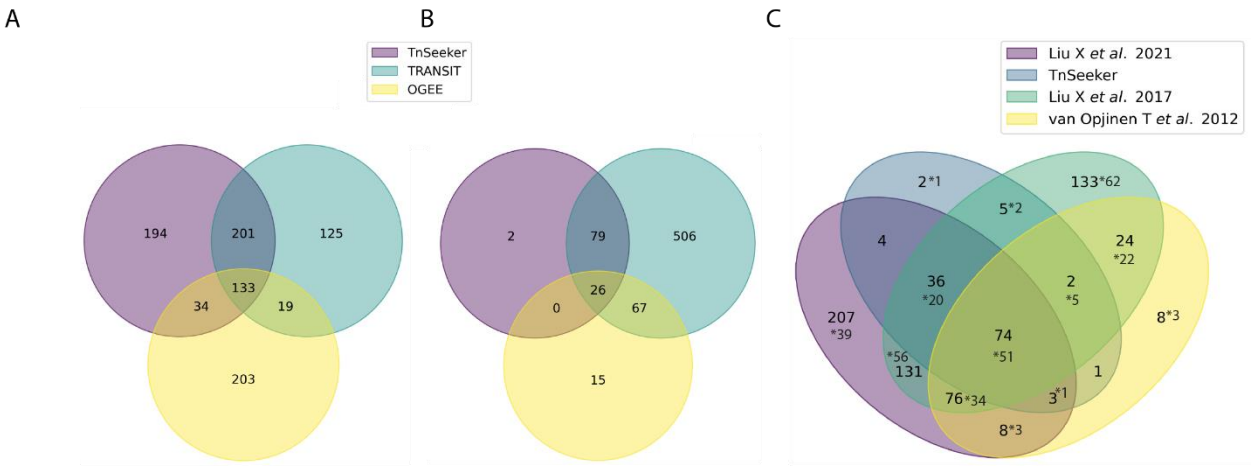
Supplementary figure 3 | Naïve approach to gene essentiality inference.

A) Distribution of essentialome size determined using a naïve approach according to library size for all Tn5 libraries in this study. The y-axis represents the average number of bp between contiguous transposon insertions, and thus the degree of library saturation. **B**, **C**) Comparison of essential genes inferred from the BW25113 Tn5 library using a naïve approach with those returned from other methods or studies. **B**) Essentials comparison with Goodall *et al.*, Koo BM. *et al.*, Price MN *et al*. (Emily C. A. Goodall *et al.*, 2018; Koo *et al.*, 2017; Price *et al.*, 2018). **C**) Same as B) but with the essentials returned when using the TRANSIT Tn5Gaps method (DeJesus *et al.*, 2015). **D**) Comparison of the essential genes returned with the naïve method, TRANSIT, the Koo BM *et al.* dataset, and ANUBIS when using the Poisson estimate with the same gold set as the ones used by TnSeeker (Miravet-Verde *et al.*, 2020). The ANUBIS dataset consists of 605 essential genes, but only 350 are shown due to incompatibilities in gene annotations.



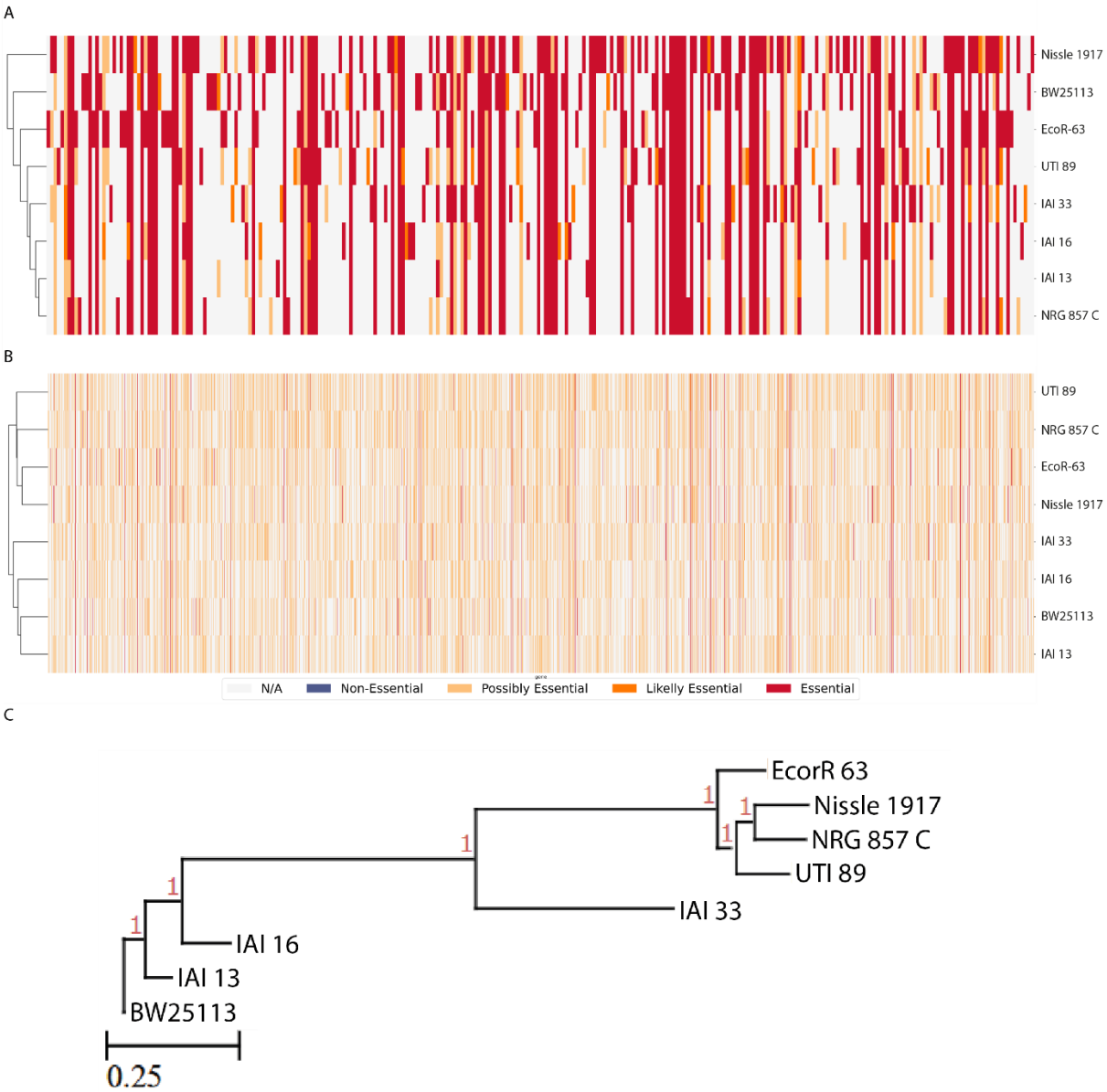
Supplementary figure 4 | COG enrichment analysis of the essentialome of all built Tn5 libraries.

See methods for COG terms categories description.



Supplementary figure 5 | Venn diagram overlapping the essentials inferred using TnSeeker, TRANSIT, and the OGEE database (Gurumayum *et al.*, 2021).

A) *Pseudomonas aeruginosa* (Poulsen *et al.*, 2019); **B**) *Mycobacterium tuberculosis* (Carey *et al.*, 2018) **C**) Venn diagram comparing the essentials inferred from an unpublished *Streptococcus pneumoniae* transposon library (Jan Willem Veening lab), with those from Opijnen vT. (van Opijnen & Camilli, 2012), Liu X. *et al.* 2017 (X. Liu *et al.*, 2017), and Liu X. *et al.* 2021 (Xue Liu *et al.*, 2021). These 2 latter refer to an operon level CRISPRI library. Analysis was thus also performed at the operon level for the transposon libraries, with its respective overlap being indicated by an *.



Supplementary figure 6 | Pan-essentialome phylogeny comparison.

Heatmap of the **A**) pan-essentialome **B**) pan-Genome when discarding genes “too small to be assayed” from all strains. Strains were vertically organized based on Pearson correlation of shared essentials. **B**) Phylogeny tree between all 8 analyzed strains in this study. parsnp was used for tree construction using default settings (Treangen *et al.*, 2014).

Supplementary table 1 | Tn5 mutagenesis efficiency of the strains used to build transposon libraries.

Strains in bold were downstream selected for transposon library building. Efficiency was calculated based on an *E. coli* cell number of $2.66x10^9$ cells per ml at OD_{600}=1. The number of CFUs after overnight selection was divided by the initial amount of CFUs. '-' means that no transformants were obtained. 'N/A' means that the experiment was not performed. 300ng of plasmid DNA were used for electroporation.

Strain	Tn5 mutagenesis efficiency by electroporation	Tn5 mutagenesis efficiency by conjugation
BW25113	**$1.4x10^{-3}$**	**$1.1x10^{-4}$**
UTI 89	**-**	**$3.7x10^{-4}$**
IAI 33	**$1.9x10^{-5}$**	**$1.1x10^{-4}$**
Nissle 1917	**$3.7x10^{-6}$**	**$3.7x10^{-6}$**
IAI 16	**$1.4x10^{-5}$**	**$2.8x10^{-4}$**
IAI 13	**$1.9x10^{-5}$**	**$4.7x10^{-4}$**
NRG 857 C	**$9.9x10^{-6}$**	**$1.5x10^{-3}$**
EcoR-63	**$3.3x10^{-4}$**	**$1.0x10^{-4}$**
EcoR-39	$1.1x10^{-7}$	$1.2x10^{-7}$
NILS 82	$1.6x10^{-6}$	N/A
SEPT362	$1.7x10^{-6}$	N/A
HM-345	$2.1x10^{-7}$	$2.2x10^{-5}$
H10407	$6.9x10^{-5}$	N/A
IAI63	$3.8x10^{-8}$	$4.6x10^{-5}$
EcoR-42	N/A	$5.9x10^{-6}$
HM-346	N/A	$1.6x10^{-7}$
HM-341	N/A	$3.8x10^{-7}$
EcoR-28	$9.7x10^{-9}$	N/A
E2348/69	$6.5x10^{-7}$	N/A
DE-COMM-2705	-	N/A
EcoR-70	$1.5x10^{-7}$	N/A
NILS 30	$1.4x10^{-6}$	N/A
Q42	-	$2.9x10^{-6}$
IAI36	$1.4x10^{-5}$	N/A
NILS 18	$5.2x10^{-5}$	N/A
EcoR-12	$1.2x10^{-4}$	N/A
HM-50	$1.4x10^{-4}$	N/A
UTI 83972	$1.6x10^{-4}$	N/A
HM605	$2.5x10^{-8}$	N/A
IAI 80	$7.5x10^{-9}$	N/A
NILS 79	$7.7x10^{-6}$	N/A
ZG-22.1	$2x10^{-5}$	N/A
EC958	-	-
NILS 49	-	-
NILS 34	-	-

Supplementary table 2 | Unique essential genes per strain.

Unique essential genes correspond to genes that are uniquely essential in only 1 of the libraries, having explicitly been deemed non-essential in the remaining 7 libraries, or not existing (the gene did not exhibit sufficient homology in the other strains to be considered equal). Uniqueness indicates whether a gene also exists in another strain or not, in this case being non-essential in the other strain(s).

Strain	Gene Name/ID	Functional Description	Uniqueness
BW25113	NT12004_22_01105	e14 prophage%3B isocitrate dehydrogenase%2C specific for NADP+	Unique
BW25113	NT12004_22_02049	IS3 transposase B	Unique
BW25113	NT12004_22_02920	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_01957	IS2 transposase TnpB	Unique
BW25113	NT12004_22_03568	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
BW25113	NT12004_22_00526	IS3 transposase B	Unique
BW25113	NT12004_22_00339	IS30 transposase	Unique
BW25113	NT12004_22_00646	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_01434	H repeat-associated putative transposase	Unique
BW25113	minD	membrane ATPase of the MinC-MinD-MinE system	Non Unique
BW25113	NT12004_22_01955	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_03885	translation elongation factor EF-Tu 2	Unique
BW25113	NT12004_22_03429	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_03262	translation elongation factor EF-Tu 1	Unique
BW25113	NT12004_22_02805	IS2 transposase TnpB	Unique
BW25113	NT12004_22_00682	Rhs family protein%2C putative polymorphic toxin%3B putative polysaccharide synthesis/export protein%3B putative neighboring cell growth inhibitor	Unique
BW25113	NT12004_22_01376	IS2 transposase TnpB	Unique
BW25113	NT12004_22_00014	IS186 transposase	Unique
BW25113	NT12004_22_01468	glutamate decarboxylase B%2C PLP-dependent	Unique
BW25113	NT12004_22_02152	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_00485	Rhs family putative polymorphic toxin	Unique
BW25113	NT12004_22_00680	Rhs family putative polymorphic toxin	Unique
BW25113	NT12004_22_01378	IS30 transposase	Unique
BW25113	NT12004_22_03521	Rhs family protein%2C putative polymorphic toxin%3B putative polysaccharide synthesis/export protein%3B putative neighboring cell growth inhibitor	Unique
BW25113	NT12004_22_03404	Rhs family putative polymorphic toxin%2C putative neighboring cell growth inhibitor	Unique
BW25113	NT12004_22_00569	IS186 transposase	Unique
BW25113	NT12004_22_03145	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_02980	IS2 transposase TnpB	Unique
BW25113	NT12004_22_01306	IS5 transposase and trans-activator	Unique

BW25113	NT12004_22_00352	IS2 transposase TnpB	Unique
BW25113	NT12004_22_00291	IS3 transposase B	Unique
BW25113	NT12004_22_01137	adhesin	Unique
BW25113	NT12004_22_00361	IS3 transposase B	Unique
BW25113	NT12004_22_01476	putative oxidoreductase	Unique
BW25113	NT12004_22_03879	glutamate racemase	Unique
BW25113	NT12004_22_00248	IS5 transposase and trans-activator	Unique
BW25113	NT12004_22_01991	IS5 transposase and trans-activator	Unique
BW25113	rhsE	Rhs family putative polymorphic toxin	Non Unique
BW25113	NT12004_22_04187	IS30 transposase	Unique
BW25113	NT12004_22_01000	IS3 transposase B	Unique
BW25113	NT12004_22_02356	IS186 transposase	Unique
BW25113	NT12004_22_00685	putative transposase	Unique
BW25113	NT12004_22_04178	IS2 transposase TnpB	Unique
BW25113	NT12004_22_00538	IS5 transposase and trans-activator	Unique
BW25113	paaH	3-hydroxyadipyl-CoA dehydrogenase%2C NAD+-dependent	Non Unique
ECOR-63	group_1690	hypothetical protein	Non Unique
ECOR-63	sacA	Sucrose-6-phosphate hydrolase	Unique
ECOR-63	NT12079_253_00003	hypothetical protein	Unique
ECOR-63	group_790	Chromosome partition protein Smc	Non Unique
ECOR-63	group_2320	hypothetical protein	Unique
ECOR-63	group_5869	CP4-57 prophage%3B integrase	Non Unique
ECOR-63	group_5922	Sel1 family TPR-like repeat protein	Unique
ECOR-63	group_6381	fused sensory histidine kinase in two-component regulatory system with KdpE: signal sensing protein	Non Unique
ECOR-63	toxA	Dermonecrotic toxin	Non Unique
ECOR-63	NT12079_253_04222	putative flagellar export pore protein	Unique
ECOR-63	caeA	Carboxylesterase A precursor	Unique
ECOR-63	group_1	hypothetical protein	Non Unique
ECOR-63	group_16546	DNA adenine methyltransferase%2C SAM-dependent	Unique
ECOR-63	group_6288	hypothetical protein	Unique
ECOR-63	group_3322	hypothetical protein	Unique
ECOR-63	group_627	hypothetical protein	Non Unique
ECOR-63	lgrD	3-oxoacyl-[acyl-carrier-protein] synthase I	Non Unique
ECOR-63	NT12079_253_02089	putative ABC superfamily sugar transporter periplasmic-binding protein	Unique
ECOR-63	group_2237	fused lipid transporter subunits of ABC superfamily: membrane component/ATP-binding component	Non Unique
ECOR-63	NT12079_253_04356	fimbrial usher outer membrane porin protein%3B FimCD chaperone-usher	Unique
ECOR-63	group_3167	putative multdrug exporter%2C MATE family	Non Unique
ECOR-63	group_6290	hypothetical protein	Unique
ECOR-63	group_16549	RNA polymerase remodeling/recycling factor ATPase%3B RNA polymerase-associated%2C ATP-dependent RNA translocase	Unique

ECOR-63	NT12079_253_00216	hypothetical protein	Unique
ECOR-63	cas3	CRISPR-associated nuclease/helicase Cas3 subtype I-F/YPEST	Non Unique
ECOR-63	NT12079_253_03668	hypothetical protein	Unique
ECOR-63	NT12079_253_04360	putative oxidoreductase	Unique
ECOR-63	group_16547	Serine protease AprX	Unique
ECOR-63	group_392	hypothetical protein	Non Unique
IAI13	NT12110_97_02204	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
IAI13	NT12110_97_03898	e14 prophage%3B isocitrate dehydrogenase%2C specific for NADP+	Unique
IAI16	eutA	reactivating factor for ethanolamine ammonia lyase	Non Unique
IAI16	NT12113_98_03875	glutamate racemase	Unique
IAI16	NT12113_98_01396	putative selenate reductase%2C periplasmic	Unique
IAI16	group_18548	hypothetical protein	Unique
IAI16	mhpC	2-hydroxy-6-ketonona-2%2C4-dienedioic acid hydrolase	Non Unique
IAI16	group_1105	protease%2C ATP-dependent zinc-metallo	Unique
IAI16	group_8515	protein disaggregation chaperone	Non Unique
IAI16	group_18536	hypothetical protein	Unique
IAI16	NT12113_98_02846	delta(2)-isopentenylpyrophosphate tRNA-adenosine transferase	Unique
IAI16	NT12113_98_02062	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
IAI16	group_12787	DNA cytosine methyltransferase	Unique
IAI16	efeO	inactive ferrous ion transporter EfeUOB	Non Unique
IAI16	frvB	putative PTS enzyme%2C IIB component/IIC component	Non Unique
IAI16	NT12113_98_02782	pyrimidine oxygenase%2C FMN-dependent	Unique
IAI33	group_2005	putative glycosyl transferase	Unique
IAI33	fepA	iron-enterobactin outer membrane transporter	Non Unique
IAI33	group_16670	hypothetical protein	Unique
IAI33	group_9432	Mobilization protein A	Unique
IAI33	sucC	succinyl-CoA synthetase%2C beta subunit	Non Unique
IAI33	group_7822	MreB assembly cytoskeletal protein	Non Unique
IAI33	NT12130_106_03654	putative flagellar export pore protein	Unique
IAI33	NT12130_106_03136	putative adhesin	Unique
IAI33	mutE	Methylaspartate mutase E chain	Unique
IAI33	group_774	hypothetical protein	Unique
IAI33	NT12130_106_04338	succinylarginine dihydrolase	Unique
IAI33	group_16596	hypothetical protein	Unique
IAI33	NT12130_106_03645	methyl-accepting chemotaxis protein II	Unique
IAI33	ydfI	putative NAD-dependent D-mannonate oxidoreductase	Non Unique
IAI33	NT12130_106_03739	Rac prophage%3B putative tail fiber protein	Unique
IAI33	group_19383	hypothetical protein	Unique

IAI33	NT12130_106_00716	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
IAI33	group_19381	hypothetical protein	Unique
IAI33	group_3872	fused lipid transporter subunits of ABC superfamily: membrane component/ATP-binding component	Unique
IAI33	yehP	VMA domain putative YehL ATPase stimulator	Non Unique
IAI33	hutU	Urocanate hydratase	Unique
IAI33	chbC	N%2CN'-diacetylchitobiose-specific enzyme IIC component of PTS	Non Unique
IAI33	group_19416	hypothetical protein	Unique
IAI33	fepG	iron-enterobactin transporter subunit	Non Unique
IAI33	group_3364	hypothetical protein	Unique
IAI33	NT12130_106_04562	hypothetical protein	Unique
IAI33	ipaH3	E3 ubiquitin-protein ligase SlrP	Unique
Nissle 1917	NT12010_146_04377	long-chain fatty acid outer membrane transporter	Unique
Nissle 1917	group_4672	hypothetical protein	Unique
Nissle 1917	group_7966	hypothetical protein	Non Unique
Nissle 1917	group_12072	hypothetical protein	Unique
Nissle 1917	group_7943	hypothetical protein	Unique
Nissle 1917	mtnK	Methylthioribose kinase	Unique
Nissle 1917	group_5858	Bifunctional protein Aas	Unique
Nissle 1917	group_678	Type-1 restriction enzyme R protein	Non Unique
Nissle 1917	group_7898	hypothetical protein	Unique
Nissle 1917	group_5785	hypothetical protein	Non Unique
Nissle 1917	group_2239	Undecaprenyl-phosphate mannosyltransferase	Unique
Nissle 1917	iucB	N(6)-hydroxylysine O-acetyltransferase	Non Unique
Nissle 1917	group_4657	hypothetical protein	Non Unique
Nissle 1917	group_1905	colicin IA outer membrane receptor and translocator%3B ferric iron-catecholate transporter	Unique
Nissle 1917	group_424	putative adhesin	Unique
Nissle 1917	group_2188	3-oxoacyl-[acyl-carrier-protein] synthase II	Unique
Nissle 1917	wcaJ	colanic biosynthesis UDP-glucose lipid carrier transferase	Non Unique
Nissle 1917	NT12010_146_00364	hypothetical protein	Unique
Nissle 1917	group_12109	hypothetical protein	Unique

Nissle 1917	NT12010_146_00726	putative oxidoreductase	Unique
Nissle 1917	group_7945	hypothetical protein	Unique
Nissle 1917	group_12003	CP4-57 prophage%3B integrase	Unique
Nissle 1917	sgcX	putative endoglucanase with Zn-dependent exopeptidase domain protein	Non Unique
Nissle 1917	group_1397	hypothetical protein	Unique
Nissle 1917	group_950	hypothetical protein	Non Unique
Nissle 1917	NT12010_146_02098	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
Nissle 1917	fabF_1	3-oxoacyl-[acyl-carrier-protein] synthase II	Unique
Nissle 1917	group_675	Sel1 family TPR-like repeat protein	Non Unique
Nissle 1917	group_12098	mannose-1-phosphate guanyltransferase	Unique
Nissle 1917	group_3177	hypothetical protein	Non Unique
Nissle 1917	group_12144	Rac prophage%3B integrase	Non Unique
Nissle 1917	group_2645	hypothetical protein	Unique
Nissle 1917	group_7881	sialic acid transporter	Unique
Nissle 1917	group_2205	hypothetical protein	Unique
Nissle 1917	NT12010_146_00353	CP4-44 prophage%3B antigen 43 (Ag43) phase-variable biofilm formation autotransporter	Unique
Nissle 1917	group_3126	hypothetical protein	Unique
Nissle 1917	group_1652	DNA recombination-mediator A family protein	Non Unique
Nissle 1917	group_12131	hypothetical protein	Unique
Nissle 1917	NT12010_146_01590	hypothetical protein	Unique
Nissle 1917	NT12010_146_00718	CP4-44 prophage%3B antigen 43 (Ag43) phase-variable biofilm formation autotransporter	Unique
Nissle 1917	group_775	adhesin	Unique
Nissle 1917	group_2244	Hsp70 family chaperone Hsc62%2C binds to RpoD and inhibits transcription	Non Unique
Nissle 1917	group_12167	hypothetical protein	Unique
Nissle 1917	yjhG	putative dehydratase	Non Unique
Nissle 1917	NT12010_146_00730	fimbrial usher outer membrane porin protein%3B FimCD chaperone-usher	Unique
Nissle 1917	pelX	Pectate disaccharide-lyase precursor	Non Unique
Nissle 1917	group_7916	hypothetical protein	Unique

Nissle 1917	group_12184	hypothetical protein	Non Unique
Nissle 1917	group_1656	Multifunctional CCA protein	Non Unique
Nissle 1917	group_1220	N%2CN'-diacetylchitobiose-specific enzyme IIC component of PTS	Non Unique
Nissle 1917	group_419	putative fimbrial-like adhesin protein	Non Unique
Nissle 1917	group_1917	hypothetical protein	Unique
Nissle 1917	group_2250	hypothetical protein	Unique
Nissle 1917	fhaB	Filamentous hemagglutinin	Unique
Nissle 1917	group_1916	Mobilization protein A	Non Unique
NRG 857C	NT12016_152_01504	hypothetical protein	Unique
NRG 857C	NT12016_152_03918	translation elongation factor EF-Tu 2	Unique
NRG 857C	NT12016_152_01079	e14 prophage%3B isocitrate dehydrogenase%2C specific for NADP+	Unique
NRG 857C	NT12016_152_01153	hypothetical protein	Unique
NRG 857C	NT12016_152_03569	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
UTI89	NT12097_202_02172	Rac prophage%3B putative DNA replication protein	Unique
UTI89	NT12097_202_04460	delta(2)-isopentenylpyrophosphate tRNA-adenosine transferase	Unique
UTI89	NT12097_202_03925	fused 4'-phosphopantothenoylcysteine decarboxylase/phosphopantothenoylcysteine synthetase%2C FMN-binding	Unique
UTI89	NT12097_202_03610	translation elongation factor EF-Tu 2	Unique
UTI89	NT12097_202_01469	hypothetical protein	Unique
UTI89	NT12097_202_01151	colicin IA outer membrane receptor and translocator%3B ferric iron-catecholate transporter	Unique
UTI89	pncB	nicotinate phosphoribosyltransferase	Non Unique
UTI89	NT12097_202_01486	hypothetical protein	Unique
UTI89	NT12097_202_00359	hypothetical protein	Unique
UTI89	NT12097_202_01467	hypothetical protein	Unique
UTI89	NT12097_202_01635	hypothetical protein	Unique
UTI89	NT12097_202_02158	hypothetical protein	Unique
UTI89	NT12097_202_01329	hypothetical protein	Unique
UTI89	NT12097_202_00041	hypothetical protein	Unique
UTI89	rdgC	nucleoid-associated ssDNA and dsDNA binding protein%3B competitive inhibitor of RecA function	Non Unique
UTI89	NT12097_202_03577	translation elongation factor EF-Tu 1	Unique
UTI89	NT12097_202_02173	hypothetical protein	Unique
UTI89	NT12097_202_01313	hypothetical protein	Unique
UTI89	NT12097_202_03806	glutamate decarboxylase A%2C PLP-dependent	Unique

UTI89	NT12097_202_01663	glutamate decarboxylase B%2C PLP-dependent	Unique
UTI89	NT12097_202_04595	hypothetical protein	Unique

Supplementary table 3 | Unique essential genes returned by TnSeeker for the BW25113 Tn5 library.

Genes not in common with any of the analyzed datasets (figure 3E)

Gene	Function
ksgA	Specifically, dimethylates two adjacent adenosines (A1518 and A1519) in the loop of a conserved hairpin near the 3'-end of 16S rRNA in the 30S particle. May play a critical role in biogenesis of 30S subunits
aceE	Component of the pyruvate dehydrogenase (PDH) complex, that catalyzes the overall conversion of pyruvate to acetyl-CoA and CO_2
glnD	in response to the nitrogen status of the cell that GlnD senses through the glutamine level. Under low glutamine levels, catalyzes the conversion of the PII proteins and UTP to PII-UMP and PPi, while under higher glutamine levels, GlnD hydrolyzes PII-UMP to PII and UMP (deuridylylation). Thus, controls uridylylation state and activity of the PII proteins, and plays an important role in the regulation of nitrogen
ubiF	2-octoprenyl-3-methyl-6-methoxy-1,4-benzoquinone hydroxylase activity
ychF	ATPase that binds to both the 70S ribosome and the 50S ribosomal subunit in a nucleotide-independent manner
fnr	Global transcription factor that controls the expression of over 100 target genes in response to anoxia. It facilitates the adaptation to anaerobic growth conditions by regulating the expression of gene products that are involved in anaerobic energy metabolism. When the terminal electron acceptor, O(2), is no longer available, it represses the synthesis of enzymes involved in aerobic respiration and increases the synthesis of enzymes required for anaerobic respiration
ydgN	Required to maintain the reduced state of SoxR. Probably transfers electron from NAD(P)H to SoxR
pnp	Involved in mRNA degradation. Catalyzes the phosphorolysis of single-stranded polyribonucleotides processively in the 3'- to 5'-direction
rpoN	Sigma factors are initiation factors that promote the attachment of RNA polymerase to specific initiation sites and are then released
mnmE	Exhibits a very high intrinsic GTPase hydrolysis rate. Involved in the addition of a carboxymethylaminomethyl (cmnm) group at the wobble position (U34) of certain tRNAs, forming tRNA- cmnm(5)s(2)U34
rep	it can initiate unwinding at a nick in the DNA. It binds to the single-stranded DNA and acts in a progressive fashion along the DNA in the 3' to 5' direction
glnG	Member of the two-component regulatory system NtrB NtrC, which controls expression of the nitrogen-regulated (ntr) genes in response to nitrogen limitation. Phosphorylated NtrC binds directly to DNA and stimulates the formation of open promoter- sigma54-RNA polymerase complexes

References

Anzai, I. A., Shaket, L., Adesina, O., Baym, M., & Barstow, B. (2017). Rapid curation of gene disruption collections using Knockout Sudoku. *Nat Protoc, 12*(10), 2110-2137. doi:10.1038/nprot.2017.073

Barquist, L., Mayho, M., Cummins, C., Cain, A. K., Boinett, C. J., Page, A. J., Langridge, G. C., Quail, M. A., Keane, J. A., & Parkhill, J. (2016). The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics, 32*(7), 1109-1111. doi:10.1093/bioinformatics/btw022

Berg, D. E., Davies, J., Allet, B., & Rochaix, J.-D. (1975). Transposition of R factor genes to bacteriophage λ. *Proc. Nat. Acad. Sci., 72*(9), 3628-3632.

Bouhenni, R., Gehrke, A., & Saffarini, D. (2005). Identification of genes involved in cytochrome c biogenesis in Shewanella oneidensis, using a modified mariner transposon. *Appl Environ Microbiol, 71*(8), 4935-4937. doi:10.1128/AEM.71.8.4935-4937.2005

Brian Green, Christiane Bouchier, Cécile Fairhead, Craig, N. L., & Cormack, B. P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mobile DNA, 3*.

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & Opijnen, T. v. (2020). A decade of advances in transposon-insertion sequencing. *Nat Rev Genet, 21*(9). doi:10.1038/s41576-020-0244-

Carey, A. F., Rock, J. M., Krieger, I. V., Chase, M. R., Fernandez-Suarez, M., Gagneux, S., Sacchettini, J. C., Ioerger, T. R., & Fortune, S. M. (2018). TnSeq of Mycobacterium tuberculosis clinical isolates reveals strain-specific antibiotic liabilities. *PLoS Pathog, 14*(3), e1006939. doi:10.1371/journal.ppat.1006939

Chao, M. C., Abel, S., Davis, B. M., & Waldor, M. K. (2016). The design and analysis of transposon insertion sequencing experiments. *Nat Rev Microbiol, 14*(2), 119-128. doi:10.1038/nrmicro.2015.7

Chao, M. C., Pritchard, J. R., Zhang, Y. J., Rubin, E. J., Livny, J., Davis, B. M., & Waldor, M. K. (2013). High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. *Nucleic Acids Res, 41*(19), 9033-9048. doi:10.1093/nar/gkt654

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., Fero, M. J., McAdams, H. H., & Shapiro, L. (2011). The essential genome of a bacterium. *Mol Syst Biol, 7*, 528. doi:10.1038/msb.2011.58

Coe, K. A., Lee, W., Stone, M. C., Komazin-Meredith, G., Meredith, T. C., Grad, Y. H., & Walker, S. (2019). Multi-strain Tn-Seq reveals common daptomycin resistance determinants in Staphylococcus aureus. *PLoS Pathog, 15*(11), e1007862. doi:10.1371/journal.ppat.1007862

Das, S., Noe, J. C., Paik, S., & Kitten, T. (2005). An improved arbitrary primed PCR method for rapid characterization of transposon insertion sites. *J Microbiol Methods, 63*(1), 89-94. doi:10.1016/j.mimet.2005.02.011

DeJesus, M. A., Ambadipudi, C., Baker, R., Sassetti, C., & Ioerger, T. R. (2015). TRANSIT--A Software Tool for Himar1 TnSeq Analysis. *PLoS Comput Biol, 11*(10), e1004401. doi:10.1371/journal.pcbi.1004401

DeJesus, M. A., & Ioerger, T. R. (2013). *Improving discrimination of essential genes by modeling local insertion frequencies in transposon mutagenesis data*. Paper presented at the Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics.

DeJesus, M. A., & Ioerger, T. R. (2016). Normalization of transposon-mutant

library sequencing datasets to improve identification of conditionally essential genes. *J Bioinform Comput Biol, 14*(3), 1642004. doi:10.1142/S021972001642004X

Dorrity, M. W., Queitsch, C., & Fields, S. (2019). High-throughput identification of dominant negative peptides in yeast. *nature methods, 16*, 413-416.

Emily C. A. Goodall, Ashley Robinson, Iain G. Johnston, Sara Jabbari, Keith A. Turner, Adam F. Cunningham, Peter A. Lund, Jeffrey A. Cole, & Hendersona, I. R. (2018). The Essential Genome of Escherichia coli K-12. *MBio, 9*(1). doi:10.1128/mBio.02096

Forterre, P. (1999). Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Micro, 33*(3), 457-465.

Galardini, M., Koumoutsi, A., Herrera-Dominguez, L., Cordero Varela, J. A., Telzerow, A., Wagih, O., Wartel, M., Clermont, O., Denamur, E., Typas, A., & Beltrao, P. (2017). Phenotype inference in an Escherichia coli strain panel. *Elife, 6*. doi:10.7554/eLife.31035

Gallagher, L. A., Bailey, J., & Manoil, C. (2020). Ranking essential bacterial processes by speed of mutant death. *Proc Natl Acad Sci U S A, 117*(30), 18010-18017. doi:10.1073/pnas.2001507117

Garsin, D. A., Urbach, J., Huguet-Tapia, J. C., Peters, J. E., & Ausubel, F. M. (2004). Construction of an Enterococcus faecalis Tn917-mediated-gene-disruption library offers insight into Tn917 insertion patterns. *J Bacteriol, 186*(21), 7280-7289. doi:10.1128/JB.186.21.7280-7289.2004

Gurumayum, S., Jiang, P., Hao, X., Campos, T. L., Young, N. D., Korhonen, P. K., Gasser, R. B., Bork, P., Zhao, X.-M., He, L.-j., & Chen, W.-H. (2021). OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. *Nucleic Acids Research, 49*.

Helmann, T. C., Deutschbauer, A. M., & Lindow, S. E. (2019). Genome-wide identification of Pseudomonas syringae genes required for fitness during colonization of the leaf surface and apoplast. *Proc Natl Acad Sci U S A, 116*(38), 18900-18910. doi:10.1073/pnas.1908858116

Herskowitz, I. (1987). Functional inactivation of genes by dominant negative mutations. *Nature, 329*, 219-222.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res, 47*(D1), D309-D314. doi:10.1093/nar/gky1085

Hutchison, C. A., Merrymana, C., Suna, L., Assad-Garciab, N., Richtera, R. A., Smitha, H. O., & Glassa, J. I. (2019). Polar Effects of Transposon Insertion into a Minimal Bacterial Genome. *American Society for Microbiology, 201*(19). doi:10.1128/JB

Koo, B. M., Kritikos, G., Farelli, J. D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J. M., Hachmann, A. B., Rudner, D. Z., Allen, K. N., Typas, A., & Gross, C. A. (2017). Construction and Analysis of Two Genome-Scale Deletion Libraries for Bacillus subtilis. *Cell Syst, 4*(3), 291-305 e297. doi:10.1016/j.cels.2016.12.013

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform, 17*(1), 154-179. doi:10.1093/bib/bbv029

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with

Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Langridge, G. C., Phan, M. D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., & Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res, 19*(12), 2308-2316. doi:10.1101/gr.097097.109

Larivière, D., Wickham, L., Keiler, K. C., & Nekrutenko, A. (2021). Reproducible and accessible analysis of transposon insertion data at scale. *BMC Microbiology, 21*(168). doi:10.1101/2020.05.19.105429

Liu, X., Gallay, C., Kjos, M., Domenech, A., Slager, J., van Kessel, S. P., Knoops, K., Sorg, R. A., Zhang, J. R., & Veening, J. W. (2017). High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. *Mol Syst Biol, 13*(5), 931. doi:10.15252/msb.20167449

Liu, X., Kimmey, J. M., Matarazzo, L., Bakker, V. d., Maele, L. V., Sirard, J.-C., Nizet, V., & Veening, J.-W. (2021). Exploration of Bacterial Bottlenecks and Streptococcus pneumoniae Pathogenesis by CRISPRi-Seq *Cell Host & Microbe, 29*, 107-120.

Lluch-Senar, M., Delgado, J., Chen, W. H., Llorens-Rico, V., O'Reilly, F. J., Wodke, J. A., Unal, E. B., Yus, E., Martinez, S., Nichols, R. J., Ferrar, T., Vivancos, A., Schmeisky, A., Stulke, J., van Noort, V., Gavin, A. C., Bork, P., & Serrano, L. (2015). Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol Syst Biol, 11*(1), 780. doi:10.15252/msb.20145558

Martinez-Carranza, E., Barajas, H., Alcaraz, L. D., Servin-Gonzalez, L., Ponce-Soto, G. Y., & Soberon-Chavez, G. (2018). Variability of Bacterial Essential Genes Among Closely Related Bacteria: The Case of Escherichia coli. *Front Microbiol, 9*, 1059. doi:10.3389/fmicb.2018.01059

Miravet-Verde, S., Burgos, R., Delgado, J., Lluch-Senar, M., & Serrano, L. (2020). FASTQINS and ANUBIS: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies. *Nucleic Acids Res, 48*(17), e102. doi:10.1093/nar/gkaa679

Nicholas A Bokulich, Sathish Subramanian, Jeremiah J Faith, Dirk Gevers, Jeffrey I Gordon, Rob Knight, David A Mills, & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *nature methods, 10*(1), 57-60.

Nlebedim, V. U., Chaudhuri, R. R., & Walters, K. (2021). Probabilistic Identification of Bacterial Essential Genes via insertion density using TraDIS Data with Tn5 libraries. *Bioinformatics*. doi:10.1093/bioinformatics/btab508

Poulsen, B. E., Yang, R., Clatworthy, A. E., White, T., Osmulski, S. J., Li, L., Penaranda, C., Lander, E. S., Shoresh, N., & Hung, D. T. (2019). Defining the core essential genome of Pseudomonas aeruginosa. *Proc Natl Acad Sci U S A, 116*(20), 10072-10080. doi:10.1073/pnas.1900570116

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., Blow, M. J., Arkin, A. P., & Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature, 557*(7706), 503-509. doi:10.1038/s41586-018-0124-0

Pritchard, J. R., Chao, M. C., Abel, S., Davis, B. M., Baranowski, C., Zhang, Y. J., Rubin, E. J., & Waldor, M. K. (2014). ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing.

PLoS Genet, 10(11), e1004782. doi:10.1371/journal.pgen.1004782

Rahman, A., Timmerman, L., Gallardo, F., & Cardona, S. T. (2022). Identification of putative essential protein domains from high-density transposon insertion sequencing. *Sci Rep, 12*(1), 962. doi:10.1038/s41598-022-05028-x

Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernandez-Rodriguez, J., Clermont, O., Denamur, E., Rocha, E. P. C., & Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the Escherichia coli species. *Nat Microbiol, 6*(3), 301-312. doi:10.1038/s41564-020-00839-y

Rubin, E. J., Akerley, B. J., Novik, V. N., Lampe, D. J., Husson, R. N., & Mekalanos, A. J. (1999). In vivo transposition of mariner-based elements in entericbacteria and mycobacteria. *Proc. Natl. Acad. Sci., 96*, 1645-1650.

Solaimanpour, S., Sarmiento, F., & Mrazek, J. (2015). Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. *PLoS One, 10*(5), e0126070. doi:10.1371/journal.pone.0126070

Swingle, B., O'Carroll, M., Haniford, D., & Derbyshire, K. M. (2004). The effect of host-encoded nucleoid proteins on transposition: H-NS influences targeting of both IS903 and Tn10. *Mol Microbiol, 52*(4), 1055-1067. doi:10.1111/j.1365-2958.2004.04051.x

Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology, 15*(524).

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods, 6*(10), 767-772. doi:10.1038/nmeth.1377

van Opijnen, T., & Camilli, A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res, 22*(12), 2541-2551. doi:10.1101/gr.137430.112

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio, 6*(3), e00306-00315. doi:10.1128/mBio.00306-15

Zomer, A., Burghout, P., Bootsma, H. J., Hermans, P. W., & van Hijum, S. A. (2012). ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One, 7*(8), e43012. doi:10.1371/journal.pone.0043012

This page is intentionally left blank.

Chapter 4

On new methodologies of mutant libraries usage

Afonso M. Bravo[1*], Alexandra Koumoutsi[2*], Nazgul Sakenova[2*], Tümay Klemens[2*], Jan-Willem Veening[1], Athanasios Typas[2]

[1]Department of Fundamental Microbiology, Faculty of Biology and Medicine, University of Lausanne, Biophore Building, Lausanne 1015, Switzerland.
[2]Genome Biology Unit, EMBL, Heidelberg, Germany

*The author Afonso M. Bravo would like it to be known that he considers all * indicated authors should be considered as joint first authors due to their joint equal struggles in assembling the data presented in this chapter.

Afonso M. Bravo conceived and designed experiments, performed experiments, analyzed data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft. Alexandra Koumoutsi conceived and designed the experiments, performed experiments, authored or reviewed drafts of the chapter, and approved the final draft. Nazgul Sakenova performed bioinformatics experiments and analyzed data. Tümay Klemens performed statistical and computational analysis experiments, and analyzed data. Jan-Willem Veening conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft. Athanasios Typas conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Abstract

Decades after its inception, transposon mutagenesis continues to enable high-throughput genetic analysis of model and non-model organisms alike. Classic protocols, however, are typically laborious when applied to more than a few tens of conditions, or when the isolation of single mutants is required.

In here, we implemented random barcode Tn-seq (RB-Tnseq) in 7 different *Escherichia coli* strains. We used a randomly barcoded transposon vector to generate saturated barcoded transposon libraries, and cataloged all respective barcode–to-genomic locations. We found that the majority of the retrieved barcodes were not unique, with any given non-unique barcode occurring on average on 15 different insertions. Nonetheless, 20 to 50% of all insertions had at least one unique barcode, with most genes having at least one uniquely barcoded insertion. Across all random-barcodes, populations of both low abundance and low nucleotide diversity barcodes were found. However, such phenomenon was found to be mitigated by the use of read cutoffs across different technical replicates. This suggests random barcodes are only optimally recapitulated at certain read abundances, with low diversity forms potentially arising from sequencing artifacts.

Following the creation of the randomly barcoded transposon libraries, we optimized and applied the DNA SUDOKU arraying protocol to the UTI89 library. By using a semi-robotized approach, such method allows the deconvolution of mutants from a pooled population mixture into an ordered curated library format. Ultimately, from a starting population of 50,688 initial colonies, we were able to obtain isolated transposon mutants for 3,614 out of a total of 4,998 genes.

We intend on repooling the obtained condensed library for multiple condition testing using high coverage RB-Tnseq studies, which lacked significance in our current study. Moreover, we plan on arraying and curating not only other *Escherichia coli* transposon libraries, but also expand to CRISPRi libraries of other bacteria. It is our belief that such tools will be able to be used by the larger community, and thus facilitate research into novel genes and organisms.

Introduction

Barcodes and Transposons

Transposon mutagenesis coupled to next-generation sequencing (Tn-seq) has revolutionized experimental biology by allowing the systematic evaluation of gene functions in both model and non-model organisms (Deutschbauer A *et al.*, 2011; Deutschbauer *et al.*, 2014; van Opijnen *et al.*, 2009; van Opijnen & Camilli, 2013). In previous chapters we have explored how such method can be leveraged for essential gene inference using what we have defined as 'snapshot' Tn-seq. In here, we further explore Tn-seq applications by using random barcode Tn-seq (RB-Tnseq) (Wetmore *et al.*, 2015). RB-Tnseq was firstly implemented in yeast as DNA Bar-Seq, and only later translated into bacteria (Oh *et al.*, 2010; Smith AM *et al.*, 2009; Wetmore *et al.*, 2015). By introducing random barcodes into each transposon insertion (figure 1A), and measuring their relative abundance, RB-Tnseq can determine the fitness of each labelled insertion across any type of condition. In practice this implies that the sequencing of the barcode-transposon-chromosome region is only required once, to associate the found barcodes to the DNA locations. Any later studies using the same library need only to aligning and count the recovered barcodes using as reference the original library.

RB-Tnseq therefore diverges from other Tn-seq approaches in the way that it requires remarkably less laborious laboratory work and sequencing capacity. Whereas the first requires a multi-step process of assembling Illumina libraries following noisy transposon-chromosome PCR procedures every time an assay is performed; the latter requires only a single-step PCR to amplify the random barcodes for sequencing. This increases efficiency in capturing and accurately sequencing barcodes, thereby reducing the sequencing throughput needed to obtain significant coverage of all barcoded insertions (Helmann *et al.*, 2019; Wetmore *et al.*, 2015).

Brave new Sudoku

When handling large mutant libraries, it might be advantageous to create arrayed sub libraries containing only a specific selection of mutants. For example, by selecting only mutants with transposon insertions at the beginning of a gene, it is possible to reduce the total amount of mutants from the several hundreds of thousands normally obtained from a saturating transposon library, to only ~4,000 (in the case of *E. coli*). Such mutants can then be used as any other arrayed mutant library: as a background for systematic phenotypic assays on agar surfaces, chemogenomics, synthetic lethality, or any other kind of genomic studies. Moreover, these condensed libraries can be pooled again and used for conditional gene essentiality inference in any specific condition. This would achieve much higher read coverage when compared with their genome wide transposon libraries, overcoming the typical bottleneck effects arising from the use of large mutant libraries (Charlesworth, 2009), and increasing both analysis efficiency and sequencing throughput. Arraying a library collection is also advantageous due to the creation of a catalog. Usually, when performing transposon based genetic screens, a few thousands of transposon mutants are isolated based on a given relevant characteristic, with the majority later being discarded. Such creates the need to randomly re-isolate mutants every time a library is used. Moreover, due to the large nature of transposon libraries, either the chosen work size is often not representative of all genes, or large amounts of resources are wasted in ensuring that all non-essential genes are screened (Baym *et al.*, 2016; Gallagher *et al.*, 2013). Considering the impact that popular cataloged mutant libraries such as the KEIO collection or the Yeast Knockout Collection had on biology, the utility and need for curated small mutant libraries in both model and non-model organisms is beyond description.

Common approaches for arraying mutants from pooled libraries have normally relied on manual laborious work for the individual characterization of mutants. Indeed, mutants had to be randomly picked, arrayed, and individually sequenced. When considering the large libraries currently obtained by Tn-seq, such an approach is unfeasible. A new arraying method, termed DNA SUDOKU due to how it resembles the solving of a Sudoku puzzle, emerged as an alternative. Instead of individually sequencing all the picked mutants, the mutants can be pooled together, with their identity being spatially encoded by use of a 4-dimensional pooling scheme (figure 1B).

As only the pools need to be indexed for sequencing, the number of required primers, indexes, and overall labor is greatly reduced. When using this method, or similar spatial pooling coordinate-based approaches, hundreds of thousands of mutants can be sequenced and located using only a few barcodes (Anzai *et al.*, 2017; Baym *et al.*, 2016; Erlich *et al.*, 2009; Vandewalle *et al.*, 2015).

DNA SUDOKU can be further streamlined by robotics. Indeed, the use of automatic colony pickers and liquid handlers, in conjunction with high-throughput DNA extraction and PCR protocols can result in a further 30- to 100-fold improvement in both speed and cost. The deconvoluting DNA SUDOKU analysis method can then be used for inferring the location of all arrayed mutants (figure 1B, 1C, 1D, and 1E) (Anzai *et al.*, 2017; Baym *et al.*, 2016; Schmitz *et al.*, 2021). To this combination of library arraying and sequencing methods, we shall henceforth simply refer to as SUDOKU.

Following arraying and location inference from a pooled collection of hundreds of thousands of mutants, the size of the initially arrayed library can then be further reduced according to the required needs. For example, a full non-essential gene transposon disruption library, with one transposon insertion per gene, would only require 10/11 384-well plates for a typical *E. coli* lab strain: 1 well per gene. With small adjustments, SUDOKU can be adapted to function with any kind of genotypic strain differences, with applications ranging from transposon libraries, to CRISPRi, to any other kind of library with a plethora of distinguishable characteristics.

In this work, we successfully implemented and applied SUDOKU to the UTI89 transposon library and explored how such method could be improved. We also further built on the RB-Tnseq technique and developed a pilot study using the same *E. coli* strains previously mentioned in chapter 3. Using these, we performed both an in-depth analysis of RB-Tnseq, and the required optimizations required for a larger scale implementation. Ultimately, we combined both RB-Tnseq with SUDOKU to create a curated barcoded transposon gene wide arrayed library, opening the doors to further advanced massive screening works and genetics not only in UTI89, but also in other organisms.

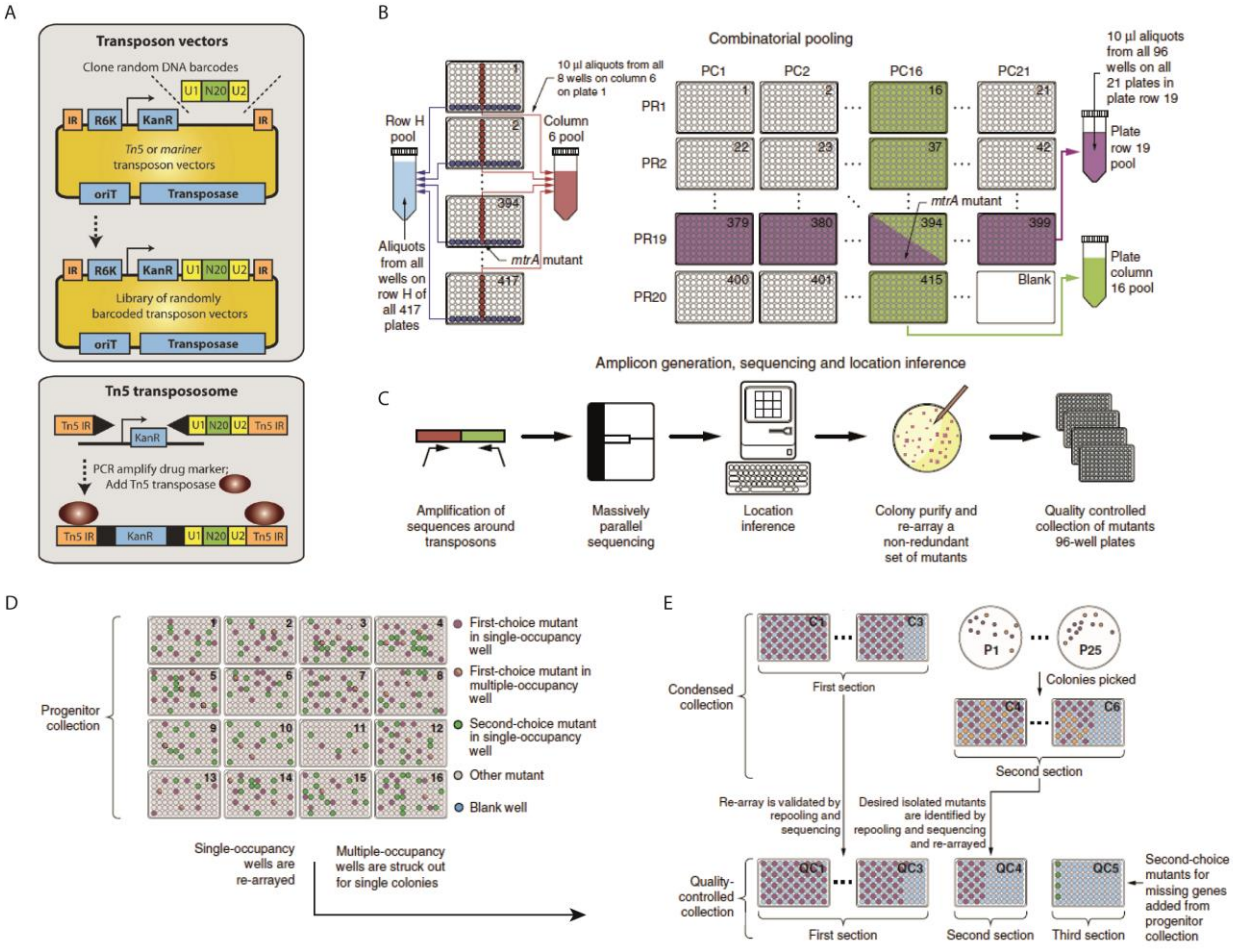


Figure 1 | Overview of DNA SUDOKU and the Tn5 mutagenesis vector

A) Organization of the used randomly barcoded Tn5 transposon cassette. **B**) Following the creation and arraying of a mutant library, SUDOKU can be carried out following a combinatorial pooling method where each pool, composed by different combinations of wells, is created, and encoded by PCR with its own index. By following the pooling scheme, any given well is only present in any 4 pools, different from well to well, with their respective location within the matrix being the only possible intersection of said pools (all the 4 pools where sequencing reads are detected). **C**) Overall schematic of the mutant inference and purification step. **D**) After sequencing and mutant location inference, individual mutants of interest can be picked and arrayed into a smaller condensed curated library. **E**) Re-arraying and purification of the mutants of interest can be done several times until a complete curated library is achieved. Figures were adapted from Wetmore *et al.* (**A**) and Baym *et al.* (**B**, **C**, **D**, and **E**) (Baym *et al.*, 2016; Wetmore *et al.*, 2015)

Results

The pKMW7 Tn5 vector creates randomly barcoded transposon libraries in *E. coli*.

Besides carrying the Tn5 transposon system used to build the transposon libraries previously reported in chapter 3, the pKMW7 vector is also randomly barcoded: Each individual vector carries a transposon cassette with a random 20bp long barcode downstream of the resistance gene, before the Tn5 transposition recognition site. All Tn5 libraries created using this system thus carry a randomly barcoded transposon insertion (figure 1C).

For the purpose of performing RB-Tnseq, *de novo* Tn5 random transposon libraries were built for 7 out of the 8 previously used *E. coli* strains, and the respective obtained data analyzed using TnSeeker (chapter 3). TnSeeker further associated all transposon insertions with their respective found barcodes, besides also reporting the absolute read abundance of each. Using a custom-made Python script, we characterized all barcoded insertion events. Genome insertion distribution analysis revealed that barcodes were found to be equally distributed across the entire genome, revealing a slight bias towards the origin of replication, similarly to what was previously shown in chapter 3 (supplementary figure 1). We also report that despite the large abundance of different barcodes, apparently more than the total amount of insertions, the majority of these are non-unique (each non-unique barcode was on average observed in 15 randomly different insertion locations) (table 1 and figure 2).

Table 1 | Summary of the transposon libraries used for RB-Tnseq.

Only insertions with barcodes with at least 3 reads within the first 10% and last 90% of a gene were considered.

Strain	Total barcodes	Unique barcodes	Insertions with unique barcodes	Unique insertions (MAPQ>=40, Phred>=10)	Library size (CFUs)
BW25113	23,195	13,087	5,118	18,301	300,000
UTI 89	129,678	28,083	19,578	110,162	300,000
IAI 33	157,413	37,642	18,840	95,254	181,000
Nissle 1917	119,803	31,455	9,811	50,927	800,000
IAI 16	120,716	44,690	18,505	90,468	320,000
IAI 13	157,195	28,221	9,238	58,713	300,000
NRG 857 C	949	617	226	6,077	290,000

For strain UTI89, out of 129,678 valid total barcodes, 28,083 uniquely occurred once (barcodes uniquely associated with any given single location) (Table 1. See methods). Such discrepancy in absolute values is also seen when considering unfiltered barcodes (barcodes with at least 1 read, existing in insertions located within the first 10% and last 90% of a gene) (supplementary figure 2A, 2B, 2C, and 2D). Indeed, in some strains the sum of all unfiltered different barcodes rises to above 1M, with some insertions having more than 10,000 different barcodes, and capturing close to 1M reads altogether. Such extreme values are also present on the lower side, with 50% of all insertions having two or less unique barcodes (median of 2 unique barcodes per insertion). When considering each insertion, and both unique and non-unique barcodes, we can observe 2 independent prevalent barcode forms (nucleotide sequences): barcodes that differ following a random distribution, with each position having a 3/4 of chance of being the same as any other barcode and thus having a Hamming distance of 15 out of 20 (bp); barcodes that marginally differ from these latter by only 1 or 2 bp, with a Hamming distance of 1 or 2 out of 20 (figure 2F and 2G). Despite such biases at the insertion level, with the existence of multiple barcode forms, the unique barcodes across all insertions mostly follow a random distribution in sequence, and thus allow for a confident differentiation of barcode-insertion pairs (figure 2H). These low abundance and low diversity barcodes nonetheless persist even when considering progressively increasing read-cutoffs, with similar outputs being shown for read cutoffs of 1, 3, 10, and even 30. From these, the latter returned the least amount of low differing barcodes per insertion, albeit at the cost of a reduced

number of total barcodes. Similarly, when sorting all the library barcodes by relative abundance per insertion (i.e.: the barcode with the most amount of reads in any given insertion would be ranked 1 for that insertion), and then plotting the read distribution of all these, and then redoing the same for all barcodes ranked 2, and 3, we observe the quickest decline in barcode read distribution is between rank 1, and all the others, when all barcodes are considered (at least 1 read) (figure1E). Such effect is mitigated as the minimum read cutoff for a barcode to be considered in this ranked distribution increases. Thus, when considering barcodes with at least 1 read (all), most of the insertions with more than 3 unique barcodes only seem to have a single highly prevalent form, followed by much less abundant barcodes (figure1E). Increasing the read cutoff reduces this discrepancy, possibly indicative of the filtering out of low abundance barcodes, the majority, and thus increasing the barcode read distribution. As a compromise we decided to only consider barcodes with more than 3 identical reads as ‘true’, thus allowing for some maneuverability in the filtering of “false positive” barcodes, which should be validated in subsequent experiments. For strain UTI89, 28,083 barcodes matched these criteria, with on average each gene having more than 10 unique barcodes (figure 2A, 2B, 2C, and 2D). More stringent read criteria were also attempted (using 18 reads [0.3 RPM]), but only 748 barcoded insertions matched these criteria.

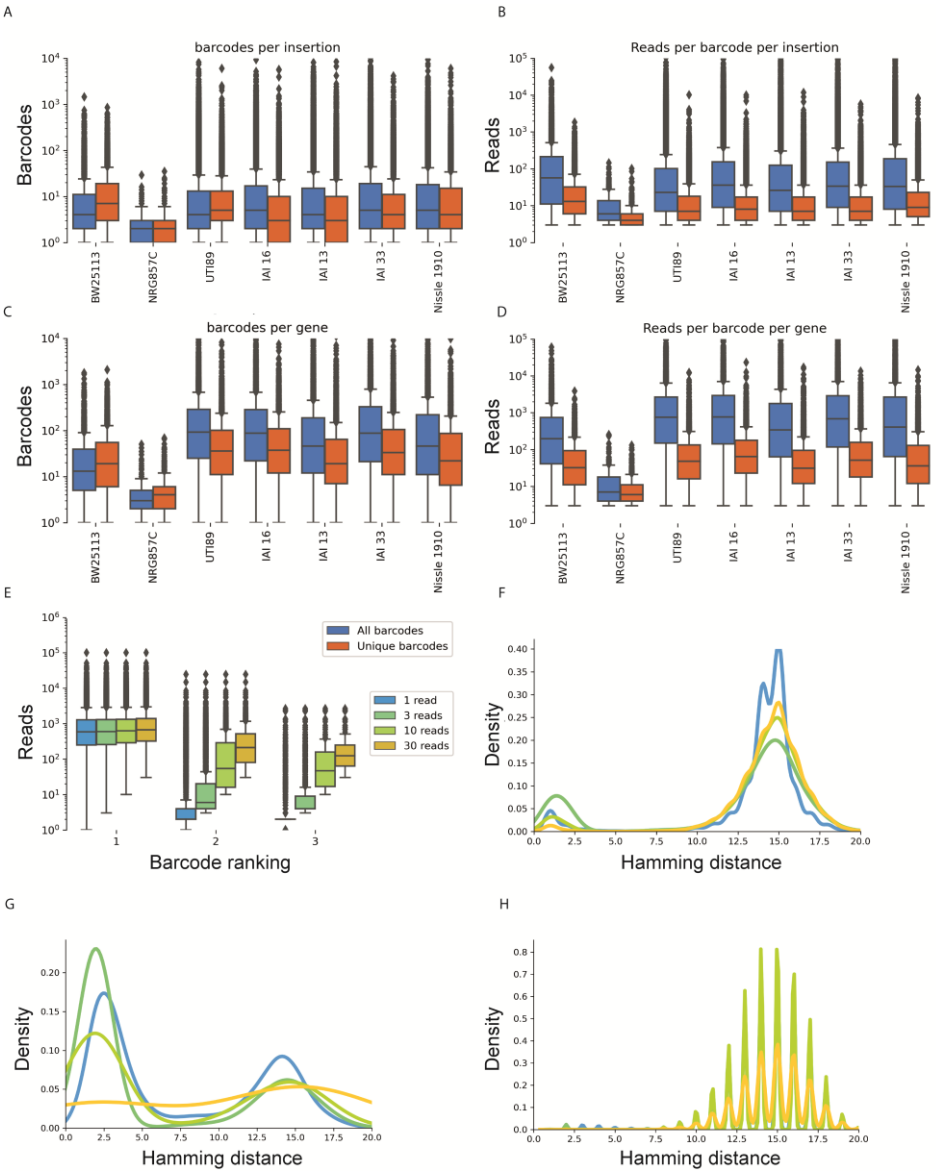


Figure 2 | Differently filtered barcode numbers and read distribution across all libraries used for RB-Tnseq.

For each strain, the distribution of the absolute number of all barcodes and unique barcodes is plotted for each insertion (**A**) or gene (**C**). The same distribution for reads is also shown for insertions (**B**) or genes (**D**). **E**) For each insertion of the UTI89 strain, the unique barcodes were filtered based on absolute read numbers using different read cutoffs. For each position, the most abundant barcode was ranked 1, and so forth. All barcodes corresponding to any given rank were grouped, and their read abundance plotted as a boxplot. Each rank position thus represents the distribution of the entire barcoded library, when barcodes with at least 1,3,10, or 30 reads are considered. **F**, **G**) and **H**) Density distribution of the effect that different read cutoffs have on recalling different barcodes as measured by determining the hamming distance between any combination of barcodes per insertion. A higher Hamming distance signifies a higher barcode diversity, and a small value implies the existence of highly similar barcodes. **F**) Distribution of all unique and non-unique barcodes. **G**) Distribution of only unique barcodes, per insertion. **H**) Distribution of the comparison of all unique barcodes across all insertions.

More than meets the eye: barcode sequencing requires ultra-deep sequencing for barcode-to-location associations

Using the previously uniquely associated barcode-insertion lists for each strain, we performed RB-Tnseq with all the built libraries. The libraries were submitted to 7 different conditions for 8 generations, and sequenced (see methods). Using 2FAST2Q (chapter 2) (Bravo *et al.*, 2022), barcode sequences were extracted and counted by back aligning to the master barcode lists for location inference. For strain UTI89, the largest library by unique insertions, barcodes associated with 3,388 different genes were recovered (18±55 unique barcodes per gene). For obtaining differential barcode fitness for each insertion, and thus gene, the barcode abundance of each insertion for each condition was compared with the basal condition using MAGeCK (Li *et al.*, 2014). Despite observing overall differences in read distribution across some of the tested conditions, namely between the basal and the exposure to sub inhibitory concentrations of doxycycline condition, no significant differential fitness was observed for any gene, with the same result being observed when using the more stringent barcoded list of more than 18 reads per barcode. This was probably due to large differences in the individual fitness of each barcode within each gene, and insertion. Such could have been exacerbated by not having replicates as an unambiguity factor, and the low read coverage per barcode. Moreover, comparison between the aligned barcodes and the total barcodes that can be extracted from each sample without alignment, revealed a 10-fold discrepancy in total amount of reads (figure 3B1 and 3C1). Such gap was still observed when not performing read filtering for any of the master list barcodes (figure 3A1). Thus suggests an incomplete sequencing of all the barcodes present in the original library, as new barcodes were now found. Such barcodes displayed a median Hamming distance of 15 in all the samples, indicating that no low differing barcodes “forms” were found (Hamming distance of 15 in a 20bp long DNA segment implies ¾ of the sequence is different, as expected by random chance). This might indicate that the low abundance alternative barcode ‘forms’ initially found in the original library are possibly an artifact of the original sequencing building protocol, as no such barcodes were found in these sequencing samples.

Considering all these factors, the RB-Tnseq performed in this study was inconclusive regarding the differential fitness of any barcoded gene.

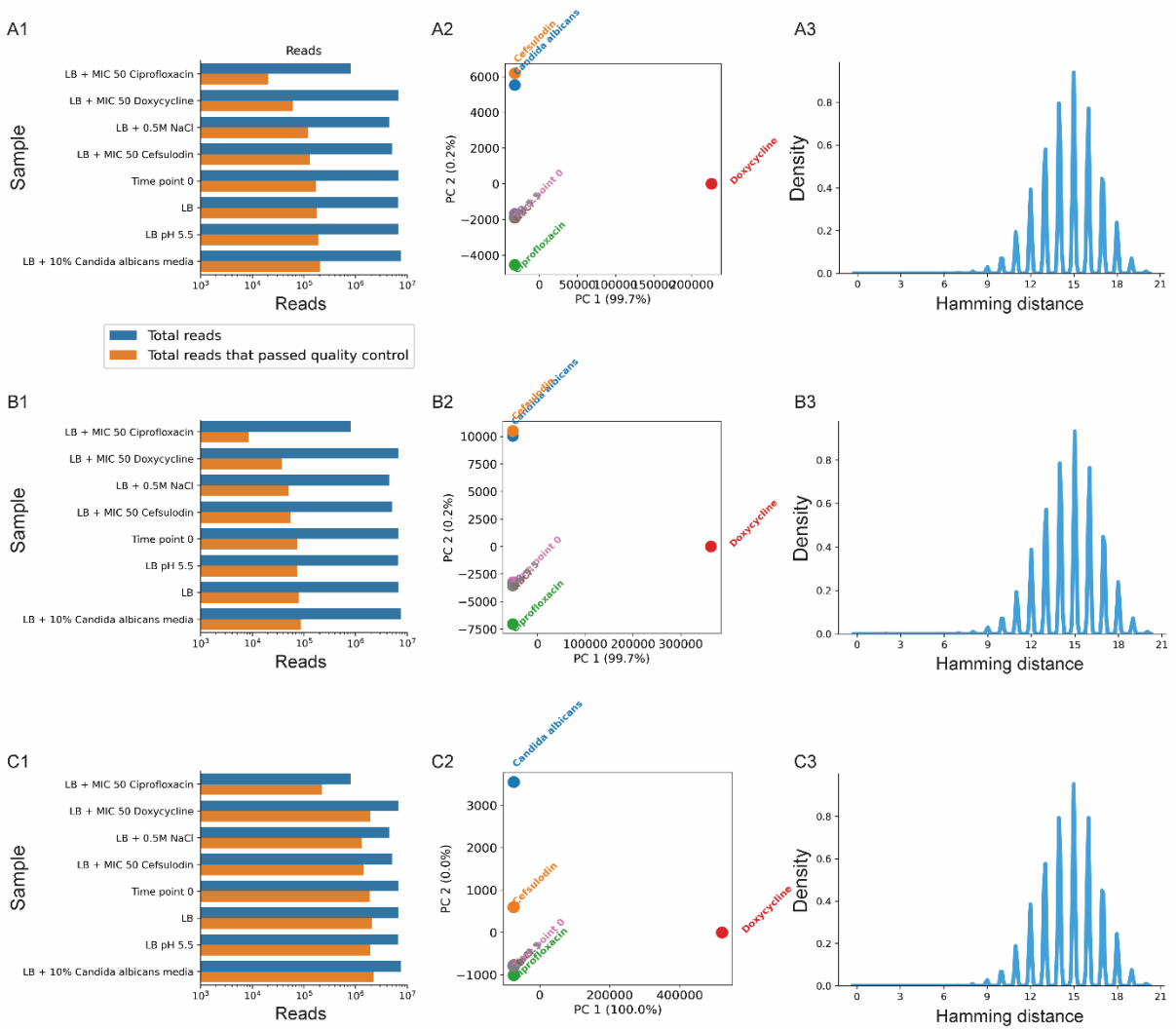


Figure 3 | Recovered reads from the RB-Tnseq for the UTI89 transposon library.

A1) Absolute number of recovered and aligned reads ('total reads that passed quality control') for the UTI89 strain when considering all unique barcodes (reads >=1 in the original library), across all conditions, **A2**) PCA of the normalized reads per million (RPM) per condition. A3) Hamming distance density plot of all the found barcodes. **B1**) Same as A1 but considering filtered barcodes (reads >=3 in the original library). **B2** and **B3**) Same as A2 and A3 but using the barcodes of B1. **C1**) Absolute number of all recovered barcodes found in the samples, independently of alignments with the original library sample. **C2** and **C3**) same as before but with the barcodes of C1.

Library biases influence the required number of mutants to achieve an all-encompassing arrayed transposon library

Considering the previous success at building saturated randomly barcoded pooled Tn5 transposon libraries (table 1; chapter 3: table 2), we proceeded into arraying and deconvoluting these into curated collections of single gene transposon knock-outs using the SUDOKU method. To this end, as a pilot experiment, we used the already built UTI89 strain transposon library described in chapter 3 (chapter 3: table 2).

As the initial process of SUDOKU involves the plating and the random picking of library mutants from agar plates, we firstly determined how many colonies would need to be isolated from the pooled transposon library to obtain a representative panel with a disruption mutant for all theoretically available genes. We first performed a simulation assuming an absolute random distribution of transposon insertions across the UTI89 genome, with each insertion having an equal chance of being picked for arraying. Considering our preference for having mutants with a disruption between the first 20% and last 80% of any gene, we determined the final library size should consist of 50,000 different colonies to obtain ~75% coverage of all UTI89 genes. Such scenario, however, differed from what was later observed upon more in-depth sequencing of the used UTI89 transposon library. Indeed, some transposon insertions were overrepresented, with a bias towards GC rich content regions also being shown (chapter 3). A new simulation was thus performed using the skewness seen in the original library as a baseline parameter. Under these circumstances, the expected number of mutants for the final picked library size of 50,000 was determined to only encompass around 50% of all genes. As the mutants picked vs. suitably disrupted genes curve started to saturate at a library size 50,000 mutants, we proceeded with the SUDOKU arraying method using the 50,000 (total of 50,688) as the total mutant target for the arrayed library (figure 4).

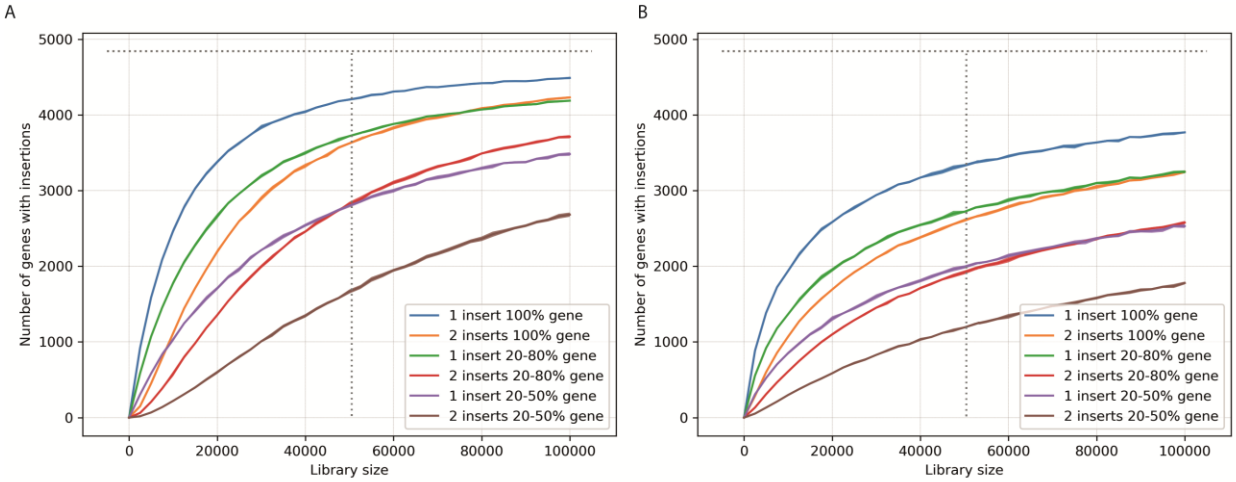


Figure 4 | Colony picking simulation for the used UTI89 transposon library.

A) Colony picking simulation for UTI89 assuming random transposon insertions with equal chance of being picked. **B**) Simulation using as assumption the same insertion distribution present in the UTI89 transposon library used for the SUDOKU arraying method. The vertical dashed grey line indicates the chosen library size of 50,000 mutants. The horizontal line indicates to total amount of genes in the UTI89 strain.

The correct inference of transposon mutants is improved by technical replicates cross validation

We next performed library arraying on the pooled UTI89 transposon library using the published SUDOKU method (Anzai *et al.*, 2017; Erlich *et al.*, 2009). Heavy technical optimizations were required to establish the entirety of this protocol in the laboratory (see methods for details). In here, however, we will focus on the post-initial arraying step of SUDOKU: the data analysis, and the current outcome.

Ultimately, the arrayed SUDOKU library was independently sequenced 3 times, with initial quality control being performed by both the TnSeeker pipeline (chapter 3), and other custom-made Python scripts. When considering all the combined replicates and sequencing pools, we observed insertions across the entire length of the genome in the same pattern as the original library, thus indicating no specific biases for certain mutants upon colony picking, or in the overall process. When considering read abundance per insertion, 2 distinct populations are seen, the first contains insertions with low abundance reads, and the second insertions with around 1000 reads (figure 5 B). To evaluate the influence that read abundance might have on unique insertion determination, we calculated the total number of insertions when assuming distinct minimal read-thresholds. Curiously, despite stringent read pre-processing parameters (see methods), when considering all insertions with at least 1 read, 76,834 unique insertions were reported across all pools in all replicates combined (figure 5D1). Such is higher than the absolute maximum number of picked mutants (50,688), indicating either the occurrence of more than one insertion per mutant, or the lack of stringency upon unique insertion determination. The first requires the occurrence of several highly unlikely events (a large population of cells with at least 2 transpositions events), and has been discarded as significantly occurring in both the current, and previous works (Wetmore *et al.*, 2015) (evaluated by well-specific sanger sequencing). We therefore attributed such overestimation to miss alignments arising from artifacts in the Illumina library building process. To further explore this hypothesis, we checked how iteratively increasing the total read cutoff influenced the intersection of common insertions across all 3 sequencing replicate datasets. An optimal cutoff value of 15 reads, corresponding to 33,612 uniquely found insertions, and an overlap of 68% (insertion wise) across all replicates was deemed optimal (figure 5C and 5D2). 3,614 different genes had a transposon mutant with these settings, similar to what was predicted in the simulation

when considering library biases (figure 4B). Curiously, 4,200 genes are found when no cutoffs are used, the same as expected from the random picking simulation (figure 4A). Considering a high stringency in mutant location inference is desired, and that such numbers fall within the expected total number of mutants, a cutoff of 15 reads (0.1 RPM) is perhaps the most advisable for any downstream SUDOKU pool deconvolution, especially when using all the available data for pool deconvolution and mutant location inference.

We next determined the total amount of unique barcodes, and how imposing read cutoffs influences unique barcode distribution. Similarly to before, the cutoff determination can be based on total percentage overlap between the 3 independent sequencing samples. Interestingly, 40 reads (0.3 RPM) per barcode were required to maximize barcode overlap (figure 5C and 5E2). As such values are largely dependent on dataset sequencing depth, and on the sequencing of individual SUDOKU pools, the usage of RPM might be more suitable for any downstream applications.

Overall unique barcode diversity was also maximized when considering a cutoff of 40 reads, with fewer low diversity barcodes being observed at such values than other values (figure 5A1). Such situation was observed for intra-insertion unique, and all (inter-unique and inter-non-unique) barcodes (figure 5A2 and 5A3).

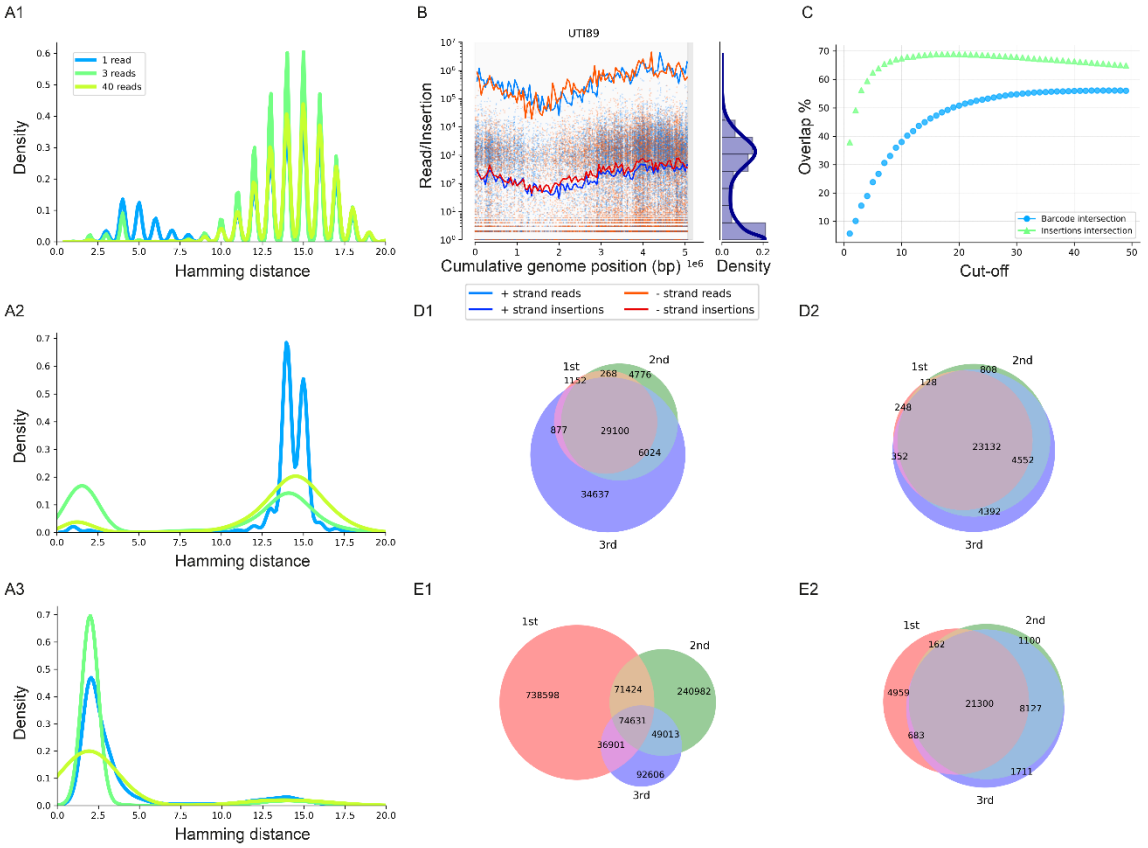


Figure 5 | SUDOKU arrayed UTI89 transposon library statistics.

The transposon insertions corresponding to all 3 sequencing replicates were combined and analyzed together. **A**) Density distribution of the effect that different read cutoffs have on recalling different barcodes as measured by determining the Hamming distance between any combination of barcodes per insertion. A higher Hamming distance signifies a higher barcode diversity, and a small value implies the existence of highly similar barcodes. **A1**) Diversity of all unique barcodes across all insertions. **A2**) Diversity across all barcodes in all samples. **A3**) Diversity of intra-insertion barcodes. **B**) Distribution of reads and insertions across the UTI89 genome. **C**) Rarefaction curve for the influence of different minimal read cutoffs on the total % overlap of either unique barcodes or transposon insertions across the 3 replicates. **D** and **E**) Venn diagram with the overlap of **D**) insertions and **E**) unique barcodes, using a read cutoff of either 1 read (**D1** and **E1**), 40 reads (**E2**), or 15 reads (**D2**).

Solving the SUDOKU: arraying recapitulates mutants from pooled libraries

Following SUDOKU pool sequencing and quality control, we applied the published SUDOKU deconvolution Python pipeline for inferring the location of all sequenced transposon mutants within the physical plate matrix (Anzai *et al.*, 2017; Erlich *et al.*, 2009). Despite successful attempts with the provided test data (1000 reads, corresponding to 181KB of computer space), we were unable to correctly use the program with our dataset. Initial problems persisted at the level of total running time and RAM usage. Indeed, we estimated several years would be required to process the entirety of our sequencing data, whilst requiring several TB of RAM to run in the EMBL HPC cluster. We optimized the program to perform operations in parallel and decrease RAM usage, however feasible total running times were still only possible when heavily using the cluster. Moreover, several code rewrites and bug fixes were required to correctly process our transposon dataset and infer the mutants' correct location. Ultimately, we recreated the entire program using a simpler, more accurate, mutant location inference statistical method, and by partially adapting Python code snippets from both TnSeeker and 2FAST2Q. Such improvement made the SUDOKU pipeline capable of running on its entirety on a common laptop within a few hours. Furthermore, the program is now capable of not only deconvoluting libraries of transposon mutants, but also CRISPRi.

Despite successfully having implemented the SUDOKU arraying pipeline, we found the location of 50% of all mutants to be totally unpredictable. Such a problem could not be fully mitigated by adjusting read-thresholds, indicating some other methodology related issue. We were nonetheless able to infer the location of mutants targeting 3,503 different genes (~70% of all genes), out of a total of 3,850 detected genes. Out of these, we were able to detect 3,067 genes with a transposon insertion within 10%-80% of the gene, with 2,737 genes likely existing as a pure mutant population (wells with a single unique mutant). The remaining mutants potentially exist as single well mixed populations in need of further purification. The inferred total gene mutants obtained correspond to the predicted limit for the library size (figure 4B), indicating the robustness of the technique in recapitulating mutants from a pooled to an arrayed library. Works are ongoing for further condensing the library into single occupancy wells for each curated gene transposon mutant.

Discussion

Similarly to the previous chapter, here we continued exploring next-generation sequencing applications in transposon libraries. We used a randomly barcoded Tn5 transposon vector to generate mutant libraries and associate any found barcodes to insertion locations. We found that, in most cases, most of the barcodes are repeated across more than 1 insertion. This issue does not arise from a lack of barcode variability, as when comparing the hamming distance of all found barcodes, a random nucleotide distribution is consistently observed. We found that the most prevalent barcode forms have a Hamming distance of 15, indicating that they differ, position wise, on ¾ of their nucleotide sequence from any other observed barcode. Barcode repeatability could result from a bottleneck in the used method for transposon building as all transposon libraries were done by conjugation with the pKMW7 vector carrying strain on an agar surface. It is thus possible that a local overrepresentation of certain types of barcode carrying strains among the conjugation matrix might have resulted in an enrichment for certain barcodes among the receiving strains. This would result in different transposon insertion events carrying the same barcode.

When looking at the Hamming distance of unique barcodes per insertion, we consistently observed the occurrence of low Hamming distance forms, indicating these to be low abundance derivatives of a single unique barcode. These seemed to mostly exist as lower read abundance forms, being able to be filtered out of the population by using highly stringent library dependent cutoffs, while not decreasing non-unique barcodes. Total unique barcodes thus do not provide a good representation of the actual number of the total uniquely barcoded population, or unique insertions, as a portion of these arise from artifacts possibly introduced during the sequencing library creation procedure. The importance of such an issue, however, can be mitigated by the type of desired downstream applications. For example, when comparing barcodes from the same library across multiple conditions, the relevance is on assuring that the barcodes are only associated to a certain unique location, not necessarily to which barcode of that same location. In this case, it is only necessary to guarantee that there aren’t similar barcodes across different locations, and such is indeed the case (figure 2H and 4A1).

Using the associated list of linked unique-barcodes to insertions as a barcode guide, we performed RB-Tnseq across multiple conditions for all the 7 libraries. Such

would allow for the high-throughput differential fitness determination of all insertions and genes in all these scenarios. Upon sequencing, however, hardly any reads matched the original barcoded population. Indeed, only 1 in 10 of all sequenced barcodes aligned to the reference (figure 3B1 and 3C1). Such result possibly indicated an issue related with the sequencing depth of the original libraries, where barcodes might not be found due to their relative low abundance, or even be considered unique while in reality existing across multiple insertions. This phenomenon is clearly present in the UTI89 arrayed library, where at most only 55% of all unique barcodes were shared among the 3 sequencing replicates of the same initial population, despite stringent read-cutoffs (figure 5C). RB-Tnseq thus requires deep sequencing of the initial libraries to be accurate, a possibly costly process when using transposon libraries carrying hundreds of thousands of different barcoded insertions. The lack of significant differential fitness observed for all genes in the performed RB-Tnseq assays can thus be explained by these issues, and the low coverage of the found barcodes.

In this work we also applied the SUDOKU arraying technique to a randomly picked subset of 50,688 different colonies from the UTI89 transposon library. Extensive recoding and fixing of the originally SUDOKU analysis pipeline was required and is still ongoing, however we have reduced the total running time and hardware requirements, with the deconvolution program now being able to run on a laptop.

We have inferred the location of pure transposon mutants for 70% of all UTI89 genes. When considering that some essential genes do not accept transposon insertions, the actual percentage of recovered mutants is higher. Despite such numbers, the locations of a vast portion of all insertions remained undetermined. Due to the nature of SUDOKU's combinatorial pooling, it's possible that confounding effects arising from well or sequencing pool cross contamination from other pools/wells creates an unsolvable puzzle for the location inference of various transposon insertions. This effect would render any given mutant as being putatively present in a plethora of pools and wells, and thus create difficulties in true location inference. We observed that imposing read cutoffs limits such effects, however, this mostly only buffers false positives in the mutant insertion alignment pipeline. This is exemplified in figure 5D2, where the biggest overlap in recapitulated mutants among all 3 sequencing replicates is observed when a threshold of 15 reads is used. Further methodology optimizations are perhaps then required upon the laboratorial pool sequencing creation and colony arraying process to avoid biological derived errors. Another

possibility would be to subdivide the current SUDOKU layout into smaller ones (for example, arraying 3 times 17K mutants). Such would limit any potential mutant overlap within the sequencing pools.

Despite these setbacks, we are currently improving the location inference algorithm to calculate the most likely location of ambiguously located insertions. This will allow to confidently extract more mutants from the currently arrayed library. The presence and absence of unique barcodes, after cutoff filtering, can also be further leveraged for correct location inference. In this case, both barcodes and insertions would need to be present in a certain relative amount for the mutant location to be considered as valid, and not as noise from methodological errors.

It would be interesting in the future to retry RB-Tnseq using the barcodes found in not only this arrayed library, but also the newer more saturating libraries characterized in chapter 3. To maximize correct unique barcode retrieval, it might also be interesting to sequence each library with more than one technical replicate and perform overlap read cutoff optimization, as described for the SODOKU library. Indeed, some replicates already exist for some of the libraries, opening the door for further trials. Ultimately, it is our goal to be able to combine the curated barcoded SUDOKU arrayed library into unique combinations of smaller-scale libraries. Such could, for example, allow RB-Tnseq to be performed with a high degree of coverage on a subset of picked cataloged mutants.

Considering the relative success in retrieving arrayed mutants from a pooled library, we intend on applying the SUDOKU arraying method to both other transposon libraries and CRISPRi libraries. Due to their reduced size and more homogenous distribution of mutants, these latter would require smaller numbers of picked mutants, thus reducing cost, time, and simplifying the creation of the sequencing pools. Ultimately, it is our belief these approaches will contribute to accelerated research into non-model organisms and strains, expediting so far unknown biology.

Methods

Strains

The APA766 strain (harboring the randomly barcoded Tn5 library plasmid pKMW7) was a gift from Adam Deutschbauer. For culturing APA766, LB was supplemented with diaminopimelic acid (DAP) to a final concentration of 300µM, and Kanamycin (50µg/ml). The *E. coli* strains Nissle 1917, IAI16, IAI13, IAI33, NRG857C, BW25113, and UTI89 are available in the Ecoref panel (Galardini *et al.*, 2017) and were routinely cultured in LB media at 37ºC, unless stated otherwise.

All Tn5 library strains were built and maintained as described in chapter 3.

Fitness Assay

An aliquot of the appropriate library strain was defrosted into 20ml of LB and grown until OD_{578} = 1.0 at 37ºC, 200rpm. At this point, 5ml were used for total DNA extraction (basal point, time 0), with the remaining culture being diluted to OD_{578} = 0.005 in 25ml of LB supplemented with one of the various tested compounds: LB + 10% Candida albicans spent media; LB at pH 5.5 (MES buffer); LB (control), LB with 0.5M of NaCl; LB + MIC50 of Ciprofloxacin; LB + MIC10 of Doxycycline, LB + MIC50 of Cefsulodin (table 2). After 8 generations at 37ºC and 200rpm, when OD_{578} ~ 1.1, 10ml were used for total DNA extraction (end point).

Table 2 | MIC per antibiotic per *Escherichia coli* strain

MIC were determined by performing an e-test with the indicated antibiotics and strains of interest.

Strain	*Ciprofloxacin (µg/ml)*	*Doxycycline (µg/ml)*	*Cefsulodin (µg/ml)*
IAI 33	0.004	1	0.12
NRG 857 C	0.008	16	0.25
Nissle 1917	0.008	8	0.5
UTI 89	0.008	6	0.25
IAI 16	0.004	0.75	0.12
IAI 13	0.004	1	0.25

Fitness Assay Sequencing

Random-Barcode transposon sequencing was performed as described by Wetmore *et al* (Wetmore *et al.*, 2015). Briefly, a single step Illumina library building PCR was performed with primers with Illumina adaptor sequences, and delimiting the barcode sequence (table 3 and table 4). The resulting product was purified using SPRIselect magnetic beads and sequenced on a HiSeq 2500 apparatus using single ended 50bp reads.

Table 3 | PCR reaction for generating the barcode Illumina library for the barcode fitness assay.

Reagent	*Amount*
Q5 reaction buffer	10 µl
Q5 Hot Start Polymerase	0.5 µl
dNTP's (5µM)	4 µl
Mix of Seq.2.Trans_4N/5N/6N/7N (10uM)	2.5 µl
RB_Seq.X (X=index primer) (10uM)	2.5 µl
DNA (200ng)	X
H_2O	For 50 µl

Table 4 | PCR reaction cycling protocol.

Temperature (ºC)	*Time*	*Cycles*
98ºC	4 min	1x
98ºC	30 s	25X
30ºC	30 s	
72ºC	30 s	
72ºC	5 min	1x

Tn5 associated Random-Barcode Extraction

The random barcodes corresponding to all insertions present in the original libraries were extracted using TnSeeker (chapter 3). Filtering was performed to obtain all 20bp long barcodes that were associated with only a single chromosome location, and that had at least 3 reads. Finally, only barcodes in insertions occurring after the first 10%, and before the last 90% of a gene were considered.

Fitness Assay Sequencing Analysis

Random-barcodes were extracted from the fitness assay .fastq files using 2FAST2Q (chapter 2). All reads were searched for the barcode delimiting conserved

upstream sequence, CTGCAGGGATGTCCACGAGGTCTCT, allowing for 2 mismatches. When present, the following 20bp were aligned and counted against the reference barcodes found in the originally transposon library. Alignment was performed allowing for 0 mismatches on the barcode. The resulting read count table was filtered based on the presence of transposon insertion sites with more than 3 reads in the initial condition (at time 0), and the test condition. An extra filtering was performed to only consider genes with more than 30 reads across all barcoded insertions within any given gene. Differential barcode analysis was performed using MAGeCK (Li *et al.*, 2014).

SUDOKU arraying

The UTI89 transposon library (chapter 3) was plated at OD_{578} = 1 in rectangular LB+Kan (30μl/ml) containing plates at an enough density to obtain isolated colonies. Following overnight incubation at 37ºC, all plates were loaded into the RapidPick Lite Colony Picker system from Hudson Robotics for individual colony picking (robot picking settings: dither=0; dwell=0; and place value=1). Automatic arraying of the colonies was performed into 384 well plates with 100μl of LB+Kan (30μl/ml) media/well. Due to the relative high number of colonies to be picked, this process was repeated every day for a few days. At the end of each day, all arrayed colony containing 384 well plates were incubated overnight at 37ºC, 900rpm.

The following day, all overnight 384 well plates were submitted to robotic arraying into new 384 plates, the 'master collection' plates, for freezing at -80ºC by following the schematic on table 6. Simultaneously, all wells were respectively pipetted into 4 different pools depending on the sequencing pool they belonged to (1 pool contained all the same rows, 1 pool all same columns. The 2 other pools were organized as described in table 6,7, and 8) (figure 1). The pool decision depended on the well/plate spatial organization within the virtual plate matrix (table 8). For the creation of the sequencing pools, all 384well plates were virtually demultiplexed into 96well plates as demonstrated in table 6 and 7. A liquid handling robot (BioMek, Beckman Coulter) was used for all these operations. All 'master collection' plates were frozen as soon as possible, with sequencing pools being centrifuged, and all pellets being stored and frozen in 50ml tubes for later use. This process was performed in

parallel with the robotic colony picking, and repeated until ~50,000 colonies were arrayed (3-4 days).

Following SUDOKU arraying, DNA was extracted from all created sequencing pools. Illumina libraries were assembled and sequenced as described (chapter 3). The used sequencing indexes for multiplexing are indicated on table 8.

Table 5 | Primers used in this study

Primer Name	*Primer Sequence*	*Used Workflow*
Seq_2.Transp_4N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCTGCAGGGATGTCCACGAGG	Fitness Assay Sequencing
Seq_2.Transp_5N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNCTGCAGGGATGTCCACGAGG	Fitness Assay Sequencing
Seq_2.Transp_6N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCTGCAGGGATGTCCACGAGG	Fitness Assay Sequencing
Seq_2.Transp_7N	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNCTGCAGGGATGTCCACGAGG	Fitness Assay Sequencing
RB_Seq.1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.3	CAAGCAGAAGACGGCATACGAGATGCCTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing

RB_Seq.10	CAAGCAGAAGACGGCATACGAGATAAGCCAG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.11	CAAGCAGAAGACGGCATACGAGATGTAGCCG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.12	CAAGCAGAAGACGGCATACGAGATTACAAGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.13	CAAGCAGAAGACGGCATACGAGATTTGACTG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.14	CAAGCAGAAGACGGCATACGAGATGGAACTG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.15	CAAGCAGAAGACGGCATACGAGATTGACATG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.16	CAAGCAGAAGACGGCATACGAGATGGACGG GTGACTGGAGTTCAGACGTGTGCTCTTCCGA TCTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.17	CAAGCAGAAGACGGCATACGAGATCACTACG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.18	CAAGCAGAAGACGGCATACGAGATGCGGACG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.19	CAAGCAGAAGACGGCATACGAGATTATCGCG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.20	CAAGCAGAAGACGGCATACGAGATGGCCACG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.21	CAAGCAGAAGACGGCATACGAGATCGAAACG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.22	CAAGCAGAAGACGGCATACGAGATCGTACGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.23	CAAGCAGAAGACGGCATACGAGATCCACGCG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.24	CAAGCAGAAGACGGCATACGAGATGCTACCG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.25	CAAGCAGAAGACGGCATACGAGATATCAGTG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.26	CAAGCAGAAGACGGCATACGAGATGCTCATG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.27	CAAGCAGAAGACGGCATACGAGATAGGAATG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing

RB_Seq.28	CAAGCAGAAGACGGCATACGAGATCATTTGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.29	CAAGCAGAAGACGGCATACGAGATTAGTTGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.30	CAAGCAGAAGACGGCATACGAGATCCGGTGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.31	CAAGCAGAAGACGGCATACGAGATATCGTGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.32	CAAGCAGAAGACGGCATACGAGATTGAGTGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.33	CAAGCAGAAGACGGCATACGAGATCGCCTGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.34	CAAGCAGAAGACGGCATACGAGATGCCATGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.35	CAAGCAGAAGACGGCATACGAGATAAAACGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.36	CAAGCAGAAGACGGCATACGAGATTGTTGGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.37	CAAGCAGAAGACGGCATACGAGATATTCCGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing
RB_Seq.38	CAAGCAGAAGACGGCATACGAGATAGCTAGG TGACTGGAGTTCAGACGTGTGCTCTTCCGAT CTGTCGACCTGCAGCGTACG	Fitness Assay Sequencing

Table 6 | SUDOKU arraying matrix layout

All the entries in each row and column correspond to 384 well plates, in total capable of holding 50,688 individual mutants. All mutants were directly arrayed from the overnight incubation 384 well plates into fresh 384-plates, upon which the sequencing pools were created following a demultiplexing procedure in the following configuration (see table 7), and frozen at -80ºC.

384-plates	A	B	C	D	E	F
I	1	2	3	4	5	6
II	7	8	9	10	11	12
III	13	14	15	16	17	18
IV	19	20	21	22	23	24
V	25	26	27	28	29	30
VI	31	32	33	34	35	36
VII	37	38	39	40	41	42
VIII	43	44	45	46	47	48
IX	49	50	51	52	53	54
X	55	56	57	58	59	60
XI	61	62	63	64	65	66
XII	67	68	69	70	71	72
XIII	73	74	75	76	77	78
XIV	79	80	81	82	83	84
XV	85	86	87	88	89	90
XVI	91	92	93	94	95	96
XVII	97	98	99	100	101	102
XVIII	103	104	105	106	107	108
XIX	109	110	111	112	113	114
XX	115	116	117	118	119	120
XXI	121	122	123	124	125	126
XXII	127	128	129	130	131	132

Table 7 | Demultiplexing schematic from the 384 well plate matrix to the pool row and column sequencing matrix

Six different 384 well plates are indicated in the schematic (outer table, grey). To reduce the amount of mutants per pool, each 384-plate was demultiplexed into four different pools (inner table). The pools depended on the spatial arrangement of the 384-well plate matrix, and the spatial organization of the sequencing pools is indicated on table 8.

384-plates	I	II	III	IV
A	PR1:PC1	PR1:PC2	PR1:PC3	PR1:PC4
	PR2:PC1	PR2:PC2	PR2:PC3	PR2:PC4
	PR3:PC1	PR3:PC2	PR3:PC3	PR3:PC4
	PR4:PC1	PR4:PC2	PR4:PC3	PR4:PC4
B	PR5:PC1	PR5:PC2	PR5:PC3	PR5:PC4
	PR6:PC1	PR6:PC2	PR6:PC3	PR6:PC4
	PR7:PC1	PR7:PC2	PR7:PC3	PR7:PC4
	PR8:PC1	PR8:PC2	PR8:PC3	PR8:PC4

Table 8 | SUDOKU sequencing pools matrix layout.

All the entries in each row and column represent the virtual demultiplexed pool well plates present in the same sequencing pools (24 pool row (PR) and 22 pool columns (PC) pools).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22
PR1	1	25	49	73	97	121	145	169	193	217	241	265	289	313	337	361	385	409	433	457	481	505
PR2	2	26	50	74	98	122	146	170	194	218	242	266	290	314	338	362	386	410	434	458	482	506
PR3	3	27	51	75	99	123	147	171	195	219	243	267	291	315	339	363	387	411	435	459	483	507
PR4	4	28	52	76	100	124	148	172	196	220	244	268	292	316	340	364	388	412	436	460	484	508
PR5	5	29	53	77	101	125	149	173	197	221	245	269	293	317	341	365	389	413	437	461	485	509
PR6	6	30	54	78	102	126	150	174	198	222	246	270	294	318	342	366	390	414	438	462	486	510
PR7	7	31	55	79	103	127	151	175	199	223	247	271	295	319	343	367	391	415	439	463	487	511
PR8	8	32	56	80	104	128	152	176	200	224	248	272	296	320	344	368	392	416	440	464	488	512
PR9	9	33	57	81	105	129	153	177	201	225	249	273	297	321	345	369	393	417	441	465	489	513
PR10	10	34	58	82	106	130	154	178	202	226	250	274	298	322	346	370	394	418	442	466	490	514
PR11	11	35	59	83	107	131	155	179	203	227	251	275	299	323	347	371	395	419	443	467	491	515
PR12	12	36	60	84	108	132	156	180	204	228	252	276	300	324	348	372	396	420	444	468	492	516
PR13	13	37	61	85	109	133	157	181	205	229	253	277	301	325	349	373	397	421	445	469	493	517
PR14	14	38	62	86	110	134	158	182	206	230	254	278	302	326	350	374	398	422	446	470	494	518
PR15	15	39	63	87	111	135	159	183	207	231	255	279	303	327	351	375	399	423	447	471	495	519
PR16	16	40	64	88	112	136	160	184	208	232	256	280	304	328	352	376	400	424	448	472	496	520
PR17	17	41	65	89	113	137	161	185	209	233	257	281	305	329	353	377	401	425	449	473	497	521
PR18	18	42	66	90	114	138	162	186	210	234	258	282	306	330	354	378	402	426	450	474	498	522
PR19	19	43	67	91	115	139	163	187	211	235	259	283	307	331	355	379	403	427	451	475	499	523
PR20	20	44	68	92	116	140	164	188	212	236	260	284	308	332	356	380	404	428	452	476	500	524
PR21	21	45	69	93	117	141	165	189	213	237	261	285	309	333	357	381	405	429	453	477	501	525
PR22	22	46	70	94	118	142	166	190	214	238	262	286	310	334	358	382	406	430	454	478	502	526
PR23	23	47	71	95	119	143	167	191	215	239	263	287	311	335	359	383	407	431	455	479	503	527
PR24	24	48	72	96	120	144	168	192	216	240	264	288	312	336	360	384	408	432	456	480	504	528

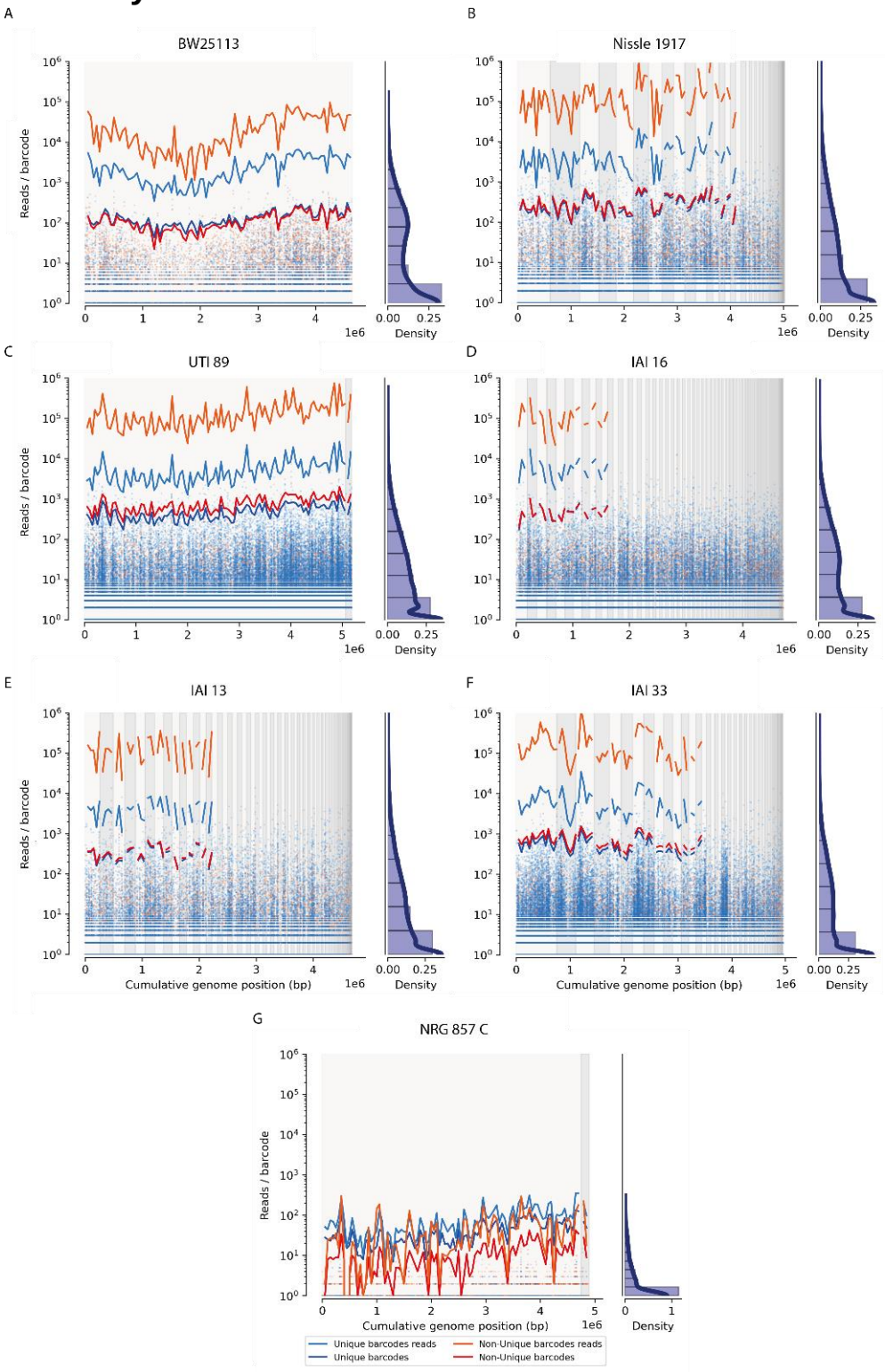
Table 9 | Used sequencing index per SUDOKU pool

The following Illumina indexes were used to make the primers ‘Seq_2.Chrom.X’ prior to the last PCR in the library building process.

Pool	*Index*
A	GGTTCA
B	TTGGAC
C	TCCAGG
D	CAGCGT
E	CAAGTG
F	AAGCTG
G	CAAGGT
H	CAGCTG
1	AAGCGT
2	ACAGTG
3	AACGTG
4	CCAGTG
5	ACAGGT
6	ATACGG
7	ACGCTG
8	CACGTG
9	AACGGT
10	CCAGGT
11	CTACGG
12	CAGATG
PR01	ACGCGT
PR02	CACGGT
PR03	CAGAGT
PR04	ACCGTG
PR05	ACGATG
PR06	ACCGGT
PR07	ATCCGG
PR08	CTAAGG
PR09	CCGATG
PR10	ACGAGT
PR11	TAACGG
PR12	CCGAGT
PR13	ATCAGG
PR14	CTCAGG
PR15	TCACGG
PR16	TACCGG
PR17	TCAAGG
PR18	TACAGG
PR19	CAGGTT
PR20	CTAGTG
PR21	ACGGTT

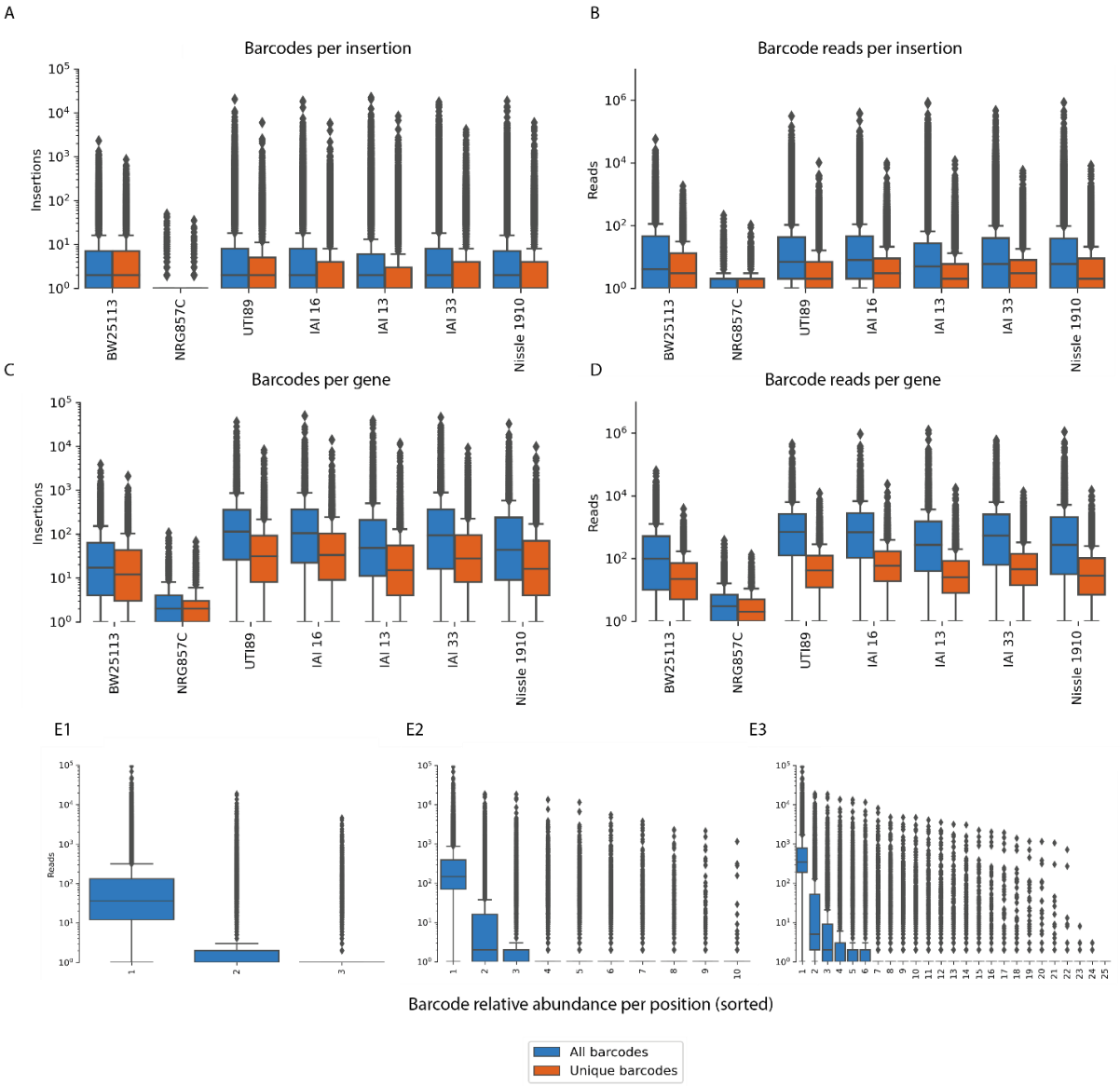
PR22	ATGCTG
PR23	CTAGGT
PR24	ATGCAG
PC01	ATGCGT
PC02	ATCGTG
PC03	ATCGGT
PC04	CTGATG
PC05	TAGCTG
PC06	CTGAGT
PC07	TAGCGT
PC08	TCAGTG
PC09	TACGTG
PC10	TCAGGT
PC11	TTACGG
PC12	TACGGT
PC13	TCGATG
PC14	TCGAGT
PC15	TTCAGG
PC16	AAGGTC
PC17	CAGGTC
PC18	CAGGAT
PC19	ATAGGC
PC20	ACGGTC
PC21	CTAGAG
PC22	ACGGAT

Supplementary



Supplementary figure 1 | Distribution of barcodes across the genome of the used strains used for RB-Tnseq.

Read distribution across the chromosome (different contigs are indicated by different background shades, ordered from largest to smallest) for BW25113 (**A**), Nissle 1917 (**B**), UTI 89 (**C**), IAI 16 (**D**), IAI 13 (**E**), IAI 33 (**F**), and NRG 857 C (**G**). Number of reads and barcodes per unique, or across all barcodes, per bin of 50,000 bp, are indicated by the red/blue lines. Density plot represents the read quantity per insertion.



Supplementary figure 2 | Unfiltered barcode numbers and reads distribution across all libraries used for RB-Tnseq.

For each strain, the distribution of the absolute number of all barcodes and unique barcodes is plotted for each insertion (**A**) or gene (**C**). The same distribution for reads is also shown for insertions (**B**) or genes (**D**). **E**) For each insertion of the UTI89 strain, the most abundant barcode was ranked 1, and so forth. All barcodes corresponding to any given rank were grouped, and their read abundance plotted as a boxplot. Each rank position thus represents the distribution of the entire barcoded library, when barcodes with at least **E1**) 3 barcodes; **E2**) 10 barcodes; or **E3**) 25 barcodes are considered.

References

Anzai, I. A., Shaket, L., Adesina, O., Baym, M., & Barstow, B. (2017). Rapid curation of gene disruption collections using Knockout Sudoku. *Nat Protoc, 12*(10), 2110-2137. doi:10.1038/nprot.2017.073

Baym, M., Shaket, L., Anzai, I. A., Adesina, O., & Barstow, B. (2016). Rapid construction of a whole-genome transposon insertion collection for Shewanella oneidensis by Knockout Sudoku. *Nat Commun, 7*, 13270. doi:10.1038/ncomms13270

Bravo, A. M., Typas, A., & Veening, J.-W. (2022). 2FAST2Q: a general-purpose sequence search and counting program for FASTQ files. *PeerJ, 10*. doi:10.7717/peerj.14041

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet, 10*(3), 195-205. doi:10.1038/nrg2526

Deutschbauer A, Price MN, Wetmore KM, Shao W, B. J., Xu Z, Nguyen M, Tamse R, Davis RW, & AP., A. (2011). Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genomewide fitness profiling across 121 conditions. *PLoS Genet, 7:e1002385*.

Deutschbauer, A., Price, M. N., Wetmore, K. M., Tarjan, D. R., Xu, Z., Shao, W., Leon, D., Arkin, A. P., & Skerker, J. M. (2014). Towards an informative mutant phenotype for every bacterial gene. *J Bacteriol, 196*(20), 3643-3655. doi:10.1128/JB.01836-14

Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M., & Hannon, G. J. (2009). DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res, 19*(7), 1243-1253. doi:10.1101/gr.092957.109

Galardini, M., Koumoutsi, A., Herrera-Dominguez, L., Cordero Varela, J. A., Telzerow, A., Wagih, O., Wartel, M., Clermont, O., Denamur, E., Typas, A., & Beltrao, P. (2017). Phenotype inference in an Escherichia coli strain panel. *Elife, 6*. doi:10.7554/eLife.31035

Gallagher, L. A., Ramage, E., Patrapuvich, R., Weiss, E., Brittnacher, M., & Manoil, C. (2013). Sequence-defined transposon mutant library of Burkholderia thailandensis. *MBio, 4*(6), e00604-00613. doi:10.1128/mBio.00604-13

Helmann, T. C., Deutschbauer, A. M., & Lindow, S. E. (2019). Genome-wide identification of Pseudomonas syringae genes required for fitness during colonization of the leaf surface and apoplast. *Proc Natl Acad Sci U S A, 116*(38), 18900-18910. doi:10.1073/pnas.1908858116

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology, 15*(554).

Oh, J., Fung, E., Schlecht, U., Davis, R. W., Giaever, G., St Onge, R. P., Deutschbauer, A., & Nislow, C. (2010). Gene annotation and drug target discovery in Candida albicans with a tagged transposon mutant collection. *PLoS Pathog, 6*(10), e1001140. doi:10.1371/journal.ppat.1001140

Schmitz, A. M., Pian, B., Medin, S., Reid, M. C., Wu, M., Gazel, E., & Barstow, B. (2021). Generation of a Gluconobacter oxydans knockout collection for improved extraction of rare earth elements. *Nat Commun, 12*(1), 6693. doi:10.1038/s41467-021-27047-4

Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, R. F., Giaever G, & Nislow C. (2009).

Quantitative phenotyping via deep bar code sequencing. *Genome Res, 19*, 1836–1842.

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods, 6*(10), 767-772. doi:10.1038/nmeth.1377

van Opijnen, T., & Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol, 11*(7), 435-442. doi:10.1038/nrmicro3033

Vandewalle, K., Festjens, N., Plets, E., Vuylsteke, M., Saeys, Y., & Callewaert, N. (2015). Characterization of genome-wide ordered sequence-tagged Mycobacterium mutant libraries by Cartesian Pooling-Coordinate Sequencing. *Nat Commun, 6*, 7106. doi:10.1038/ncomms8106

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio, 6*(3), e00306-00315. doi:10.1128/mBio.00306-15

Chapter 5

Streptococcus pneumoniae Vs. the World: High-throughput analysis of competition mechanisms

Afonso M. Bravo[1], Christopher Forbes-Jaeger[1], Jessica Burnier[1], Athanasios Typas[2], Jan-Willem Veening[1]

[1]Department of Fundamental Microbiology, Faculty of Biology and Medicine, University of Lausanne, Biophore Building, Lausanne 1015, Switzerland.

[2]Genome Biology Unit, EMBL, Heidelberg, Germany

Afonso M. Bravo conceived and designed the experiments, performed experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft. Christopher Forbes-Jaeger performed experiments and analyzed the data. Jessica Burnier performed experiments. Athanasios Typas conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft. Jan-Willem Veening conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Abstract

It is known that bacteria often exist as multi-species communities, yet little is known about how gene essentiality changes at this level. To study such interactions, we used a pooled IPTG-inducible CRISPRi library targeting all known genetic *loci* in *Streptococcus pneumoniae*, an opportunistic human pathogen. To better mimic the *S. pneumoniae* natural habitat, we allowed pneumococci to compete with other bacteria in a nasopharyngeal epithelial cell matrix. Upon co-culture with *Staphylococcus aureus*, and wild type (WT) *S. pneumoniae*, we identified pneumococcal genes that display a fitness defect only in the presence of the competitor. Moreover, we demonstrate how two of these genes, an efflux pump (*SPV 686/7/8*, here renamed as *arpABC*) and a serine protease (*prtA*), are ubiquitous to *S. pneumoniae* competition with different species, including itself. Intriguingly, we show that the efflux pump mutant doesn't grow when *S. aureus* is present, and the pH of the environment is acidic (pH ≤6), suggesting an active role of this protein complex in counteracting adverse *S. aureus*-induced changes in the environment.

We further explored *S. pneumoniae* and *S. aureus* interaction, and demonstrated by confocal microscopy that pneumococcal strain D39V mainly inhabits the upper area of a *S. aureus* – Detroit 562 matrix. Indeed, under these conditions, we show that core essential pathways related with cell wall and peptidoglycan production confer a fitness disadvantage to *S. pneumoniae*, thus highlighting how assumed essential metabolism can be compensated by using possible opportunistic mechanisms.

Together, this study is one of the first in its kind to explore how gene fitness is affected by competing species at a genome-wide level under different environments, and will provide a better understanding of genes important for pneumococcal ecology.

Introduction

The human nasopharynx is actively colonized by a unique niche of microorganisms. Different in abundance and taxa from that of the oral and oropharynx, the active maintenance of such a distinct ecological landscape has been associated with a decreased incidence of respiratory tract infections (Flynn & Dooley, 2021; Man *et al.*, 2019; Matthew S. Kelly *et al.*, 2017; Siegel & Weiser, 2015).

Streptococcus pneumoniae (Sp) is a Gram-positive bacterium that can asymptomatically colonize the nasopharynx of healthy individuals. Innocuous Sp infection, however, can progress to opportunistic disease. Such process normally initiates with Sp translocating to the sterile environment of the internal organs, where it can cause pneumonia, meningitis, otitis media, and/or sepsis (Loughran *et al.*, 2018; Subramanian *et al.*, 2019). Sp versatility as both a commensal and pathogen can be attributed to its repertoire of cell attachment and host-defense/evading components (Subramanian *et al.*, 2019; Weiser, 2010). For example, cellular proteases are typically used as part of normal homeostasis processes, but can also play a role in preventing the action of the host complement immune system (Marquart, 2021). At its core, it is the regulation of these dual-purpose elements that determine the shift between the commensal and pathogen states. The exact mechanisms that induce this switch are still underexplored, however, several reports have highlighted the role that inter- and intra-species competition might have on this phenomenon (Shak *et al.*, 2013; Weiser, 2010).

Upon infection, Sp must establish itself into a niche occupied not only by other species, but possibly also by other pneumococcal strains. Indeed, studies indicate that between 35 to 43% of all Sp hosts carry multiple colonizing serotypes (Tonkin-Hill *et al.*, 2022; Turner *et al.*, 2011). The relative recent introduction of pneumococcal conjugate vaccines (PCVs) has shifted the frequency and colonization dynamics of Sp strain types, allowing previously uncommon strains to become more prevalent in a process known as serotype replacement, and thus highlighting the importance of Sp co-colonization and interspecies competition studies (Lo *et al.*, 2019; Weiser, 2010).

As an hedge against competitors, Sp can deploy the action of toxic peptides known as bacteriocins, and thereby eliminate strains not expressing the respective immunity factors (Aggarwal *et al.*, 2020; Lehtinen *et al.*, 2022; Shak *et al.*, 2013; Weiser, 2010; Wholey *et al.*, 2019). Several of these bacteriocin systems have been

characterized, with the *blp* regulon perhaps being one of the most ubiquitous. *Blp* activation relies on a quorum-sensing pheromone, BlpC, and has been implicated to crosstalk with the also quorum related Sp genetic transformation competence system. Interestingly, both pathways are activated in a pH-dependent manner, possibly suggesting a link between intra-species competition and exogenous DNA uptake (Dawid *et al.*, 2007; Kjos *et al.*, 2016; Shak *et al.*, 2013; Wang *et al.*, 2018; Weiser, 2010). Indeed, BlpC can be exported by its cognate transporter BlpAB as well by the competence pheromone exporter ComAB (Kjos et al 2016, Wholey et al 2016). Despite not being a decisive factor in human co-colonization (Valente *et al.*, 2016), *blp* harboring strains have been demonstrated to have competitive advantages in a mouse colonization model (Dawid *et al.*, 2007). Such results indicate a possible involvement of not only a plethora of bacteriocin systems, but also of other yet unknown variables.

Besides interspecies competition, Sp must also compete with other bacterial species, viruses, and the host immune system for a place in the nasopharynx niche. Similarly to Sp, *Staphylococcus aureus* (Sa) is also a common human commensal and/or opportunistic pathogen, being asymptomatically carried by around 20% of the population, where it preferentially inhabits the nares (J Kluytmans *et al.*, 1997; van Belkum *et al.*, 2009). Sa is a leading cause of both hospital and community-acquired skin and soft tissue infections (Olaniyi *et al.*, 2017). Epidemiological studies have also demonstrated an inverse correlation between the carriage of Sp and Sa, with the presence of pneumococcus apparently inhibiting Sa colonization. Indeed, immunocompetent pneumococcal carriers are 50% less likely to carry Sa. These observations have led several studies to speculate about the possible connection between the introduction of PCV and an increase in Sa carriage. Such hypothesis, however, seem to be as of yet, unsubstantiated (Bogaert *et al.*, 2004; Lee *et al.*, 2009; Gili Regev-Yochay *et al.*, 2004; Reiss-Mandel & Regev-Yochay, 2016; Shak *et al.*, 2013).

Several explanatory mechanisms have been advanced as a reasoning for this negative interaction. For example, Sa co-colonization with Sp has been demonstrated to be inhibited via the effect of a cross-reactive antibody. Indeed, in this case antibodies against the Sp cell wall associated dehydrogenase SP 1119 have been shown to cross-react against P5CDH, a similar dehydrogenase in Sa. Such interaction induces a reduction in Sa carriage following Sp colonization in mice (Lijek *et al.*, 2012). Such an effect, however, doesn't fully explain how, in some cases, Sp and Sa might

co-colonize the same host, or how Sa carriage might follow Sp infection later in life. *In vitro* studies have also shown that H_2O_2 production is a major component of Sp bactericidal activity against Sa (Bryant *et al.*, 2016; G. Regev-Yochay *et al.*, 2006). Such effect, however, is conflicting when translating to *in vivo* models, with reports often claiming a neutral influence of H_2O_2 on co-colonization (Lijek & Weiser, 2012; E. Margolis, 2009; Park *et al.*, 2008; G. Regev-Yochay *et al.*, 2008; Reiss-Mandel & Regev-Yochay, 2016). Indeed, recent work has shown that Sp can eradicate Sa biofilms independently of H_2O_2 production. By using Transwell devices, it was demonstrated that this mechanism required direct physical contact (Khan *et al.*, 2016). Curiously, another study demonstrated that both Sp and Sa can form stable biofilm communities in the presence of epithelial cells (Reddinger *et al.*, 2018). Additional factors, beyond the presence of H_2O_2, are thus at play regarding Sp and Sa co-colonization. Such results highlight the differences between *in vitro* and *in vivo* studies, and strengthen the case for more comprehensive 'natural-like' testing setups.

Curiously, Sp carriage has also been associated with positive inter-bacterial interactions. Indeed, a study has found a positive correlation between Sp and the simultaneous co-colonization of both *Haemophilus influenzae* and *Moraxella catarrhalis* (Pettigrew *et al.*, 2008). Other works have further demonstrated that established Sp populations facilitate colonization of *H. influenzae* alone, an effect also seen in longitudinal studies, but not observed *in vitro* (Lijek & Weiser, 2012; E Margolis *et al.*, 2010; Neto *et al.*, 2003; Pericone *et al.*, 2000). *In vivo* co-colonization with *H. influenzae*, however, results in rapid Sp clearance from the nasopharynx, probably due to an increased inflammatory response, and demonstrating how one species can modulate the host to eliminate competitors (Ratner *et al.*, 2005).

Taken together, these studies demonstrate the importance that different experimental conditions might have on recapitulating the true state of the natural environment. So far co-colonization *in vitro* studies using epithelial cell substrates have often focused on the host response side, with the activity of the 'invader' seldom being evaluated on a genome wide scale, especially in regards to gene essentiality (Asmat *et al.*, 2011; S. Novick *et al.*, 2017; Ratner *et al.*, 2005; van Opijnen & Camilli, 2012; Weight *et al.*, 2019). In here, we leveraged CRISPRi-Seq for assaying Sp responses to the presence of competing species, such as Sa. We then further elaborated on this model and explored co-colonization mechanism on a human epithelial cell matrix format (figure 1). This is, to our knowledge, the first study where the fitness of both

essential and non-essential genes is determined in such settings. We demonstrate how SPV 686/7/8 (renamed ArpABC), a general purpose antimicrobial resistance related efflux pump, becomes a conditional essential gene when Sa is co-cultured with Sa, but only at pH 6.

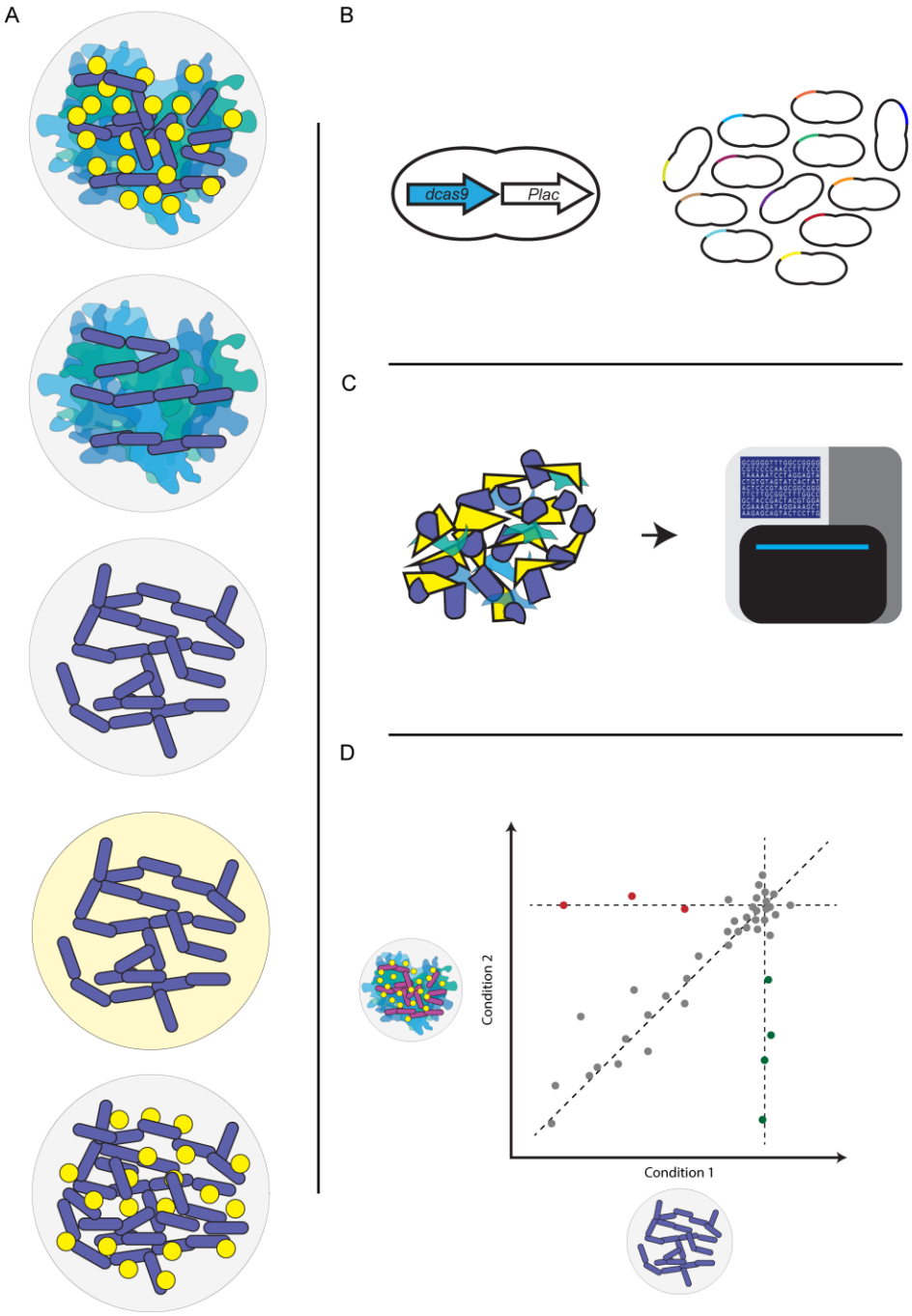


Figure 1 | Overview of the used CRISPRi-Seq methodology

The D39V IPTG-inducible CRISPRi library (see methods) was used in combination with different co-culture conditions: When relevant, Sp (purple rounded rectangles) was incubated with Sa (yellow rectangles) or Detroit 562 human cells (blue cells) in both liquid (supplemented RPMI, yellow) and solid (C+Y agar, white background) media (**A**). At the end of the co-culture experiment, cells were collected and the CRISPRi library sequenced (**B-C**). Differential analysis reveals genetic features relevant for each condition Vs. another condition (**D**).

Results

The ArpABC MacAB-like efflux pump is crucial for successful competition

To identify genes and pathways required for Sp survival in the presence of various competitors, we developed a solid-based (C+Y agar) CRISPRi-seq assay (Fig. 1; see methods). By plating Sp in the presence of competitors on a solid surface, we were able to screen for contact-dependent interactions within a static niche. To determine which genes are required for the survival of Sp on an established Sa environment, we inoculated a previously described Sp D39V CRISPRi library (Liu *et al.*, 2017; Liu *et al.*, 2021) onto an agar surface on which Sa was growing for 4h (see methods). We observed that only a limited number of Sp genes became conditionally essential in these settings (figure 2A, supplementary figure 2A). The top two hits were sgRNAs targeting *SPV_0686/7/8* (log_2 fold change = -5, p-value = $2.3x10^{-33}$), and *prtA* (log_2 fold change = -1.5, p-value = $1.8x10^{-3}$) (figure 2A). The *spv_0686/7/8* operon (also known as *sp0785/6/7* and spr0693/4/5) encodes an ATP-binding cassette (ABC)-type MacAB-like efflux pump and was previously associated with both antibiotic and LL-37 antimicrobial peptide resistance (Majchrzykiewicz *et al.*, 2010; Yang *et al.*, 2018). Due to its related activity with antimicrobial export, and now also with its discovered competition mechanism, we renamed *spv_0686/7/8* as *arpABC,* antimicrobial and competition related pump A/B/C.

PrtA (Pneumococcal Protease A) is a cell wall associated S8 serine protease belonging to a protein class that has been implicated in the cleavage of lantibiotic leader sequences. Lantibiotics are a sub group of bacteriocins, named after the non-proteinogenic amino acid lanthionine (Marquart, 2021).

To validate the observed fitness defects of CRISPRi knockdowns of both *prtA* and *arpABC* under the tested conditions, we generated knock-out (KO) mutants in a background that constitutively expresses firefly luciferase. By tracking luminescence over time, we can track the population's metabolic activity as a proxy for cell density (Sorg *et al.*, 2015). Luminescence of wild type (WT) and the Δ*prtA* and Δ*arpABC* mutants were measured in the presence and absence of Sa. We observed that when grown in the presence of an established Sa matrix, both mutants exhibited a relative fitness defect (fitness defect of ~0.3-0.4x compared with the WT), in line with our CRISPRi-seq screen (figure 2A, 2B, 2E). Interestingly, both mutants also competed less well against *Escherichia coli* and Sp (figure 2E, 2H). A strain in which the

Δ*arpABC* KO was complemented by expression of *arpABC* from an ectopic locus showed a WT-like phenotype when grown in the presence of SA (figure 2E, 2H, 2F).

Unlike what was observed for Sa, when Δ*arpABC* Sp had to compete with WT Sp, the luciferase-activity profile mimicked the respective single-culture conditions, albeit at a faster pace (activity is abolished before 10h). Conversely, Sa competition prolonged this state, with Sp metabolic activity “plateauing” for longer (figure 2E, 2H). This lowered plateaued metabolic state might result in a lower rate of division, and consequently in a lower total amount of elapsed generations per unit of time. Such phenomenon is in alignment with the obtained CRISPRi data for Sp co-cultured with Sa, where a concomitant relative decrease in total generations (measured by sgRNA abundance skewness from the diagonal) was observed when the relative initial concentration of Sa was increased (supplementary figure 3) (see methods for CRIPRi-Seq normalization procedures).

We next assessed whether these effects could be recapitulated at different initial relative Sp to Sa CFU ratios without any pre-established Sa matrix. Indeed, we only observed a fitness defect for Δ*arpABC* when the Sp CFU ratio was at least 100x lower than that of Sa. Similar cell density dependent results were also obtained when CRISPRi-seq was performed using different relative CFU ratios of either Sp WT or Sa as competitors. A generalized pronounced effect on differential gene fitness was only observed starting at a ratio of 1 Sp to 10 competitors, with the effect being more pronounced for Sp WT (figure 2D, 2G, supplementary figure 2B, supplementary figure 3).

Altogether, these results indicate that the *arpABC* operon is an important player in Sp adaptation to high bacterial density competition environments.

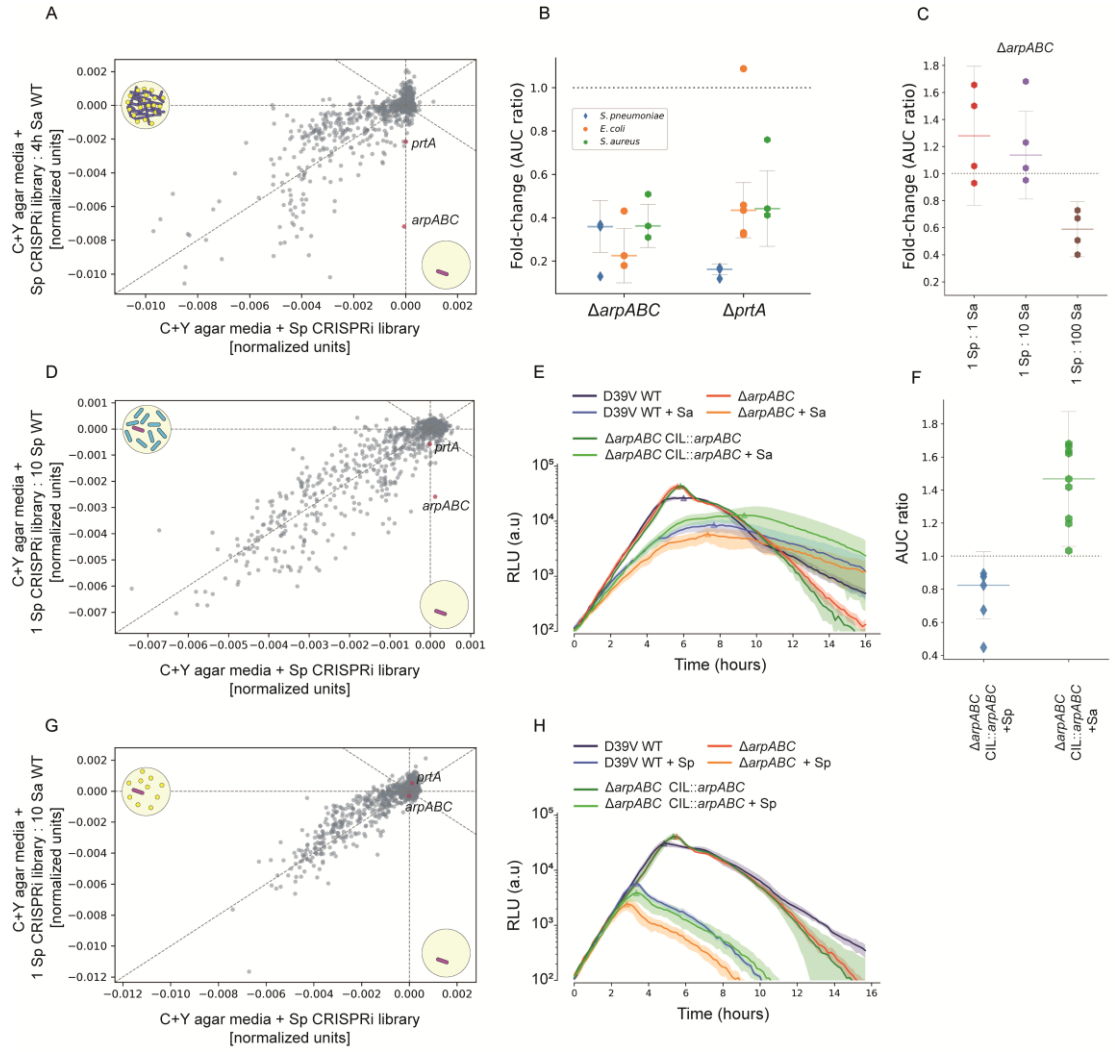


Figure 2 | PrtA and ArpABC negatively impact Sp competition fitness with other bacteria

A) D39V CRISPRi library normalized fitness comparison between the presence and absence of Sa incubated for 4h in C+Y agar. Normalized values correspond to the L1 Norm of the $\log_2$ fold-changes (see methods). A value of 0 indicates a neutral fitness effect. Both PrtA and ArpABC are indicated in red, and only display a negative differential fitness in the Sp + Sa condition. **B**) Relative luminesce area under the curve (AUC) fold-change difference between D39V WT (fold-change of 1) and (mutant / WT) when inoculated in the presence / absence (+competitor/-competitor) of a competitor incubated for 4h in agar C+Y prior to addition of Sp, respectively (see **E** and **H**). **C**) Same as **B**, but Sa was either simultaneously added or not to agar C+Y with either D39V WT or Δ*arpABC* at different starting relative CFU ratios (1:1,1:10, or 1:100). **D**) Same as **A**, but Sp D39V WT was simultaneously added to the Sp library in a 1:10 CFU ratio. **E)** Relative luminescence growth curves of D39V WT, Δ*arpABC*, or the complemented mutant Δ*arpABC* + CIL::*arpABC*. Sp was added at a concentration of 100,000 CFU/ml. **E**) In the presence of Sa (125,000 CFU/ml) incubated for 4h. **H**) In the presence of WT Sp (100 CFU/ml) incubated for 4h. **F)** Similar to B and C, but regarding the complementation strain Δ*arpABC* + CIL::*arpABC*. **G**) Same as D, but Sa WT was used as the competitor instead of Sp WT.

arpABC as a conditional essential gene: Acidic pH negatively impacts Sp survivability in the presence of Sa when _arpABC_ is absent

pH has been demonstrated to modulate the efflux of the Resistance Nodulation Division (RND) family of transporters (Martins et al., 2009). Considering *arpABC* belongs to this family of proteins, we evaluated the putative impact that environmental pH has on Sp fitness when competing with Sa.

We submitted both the WT D39V and the *ΔarpABC* to competition with Sa as previously indicated. As shown in Fig. 3, we observed that the fitness of the *ΔarpABC* mutant decreased with the decrease in the initial media pH, but only in the presence of Sa (figure 3A, 3B, 3C1, 3C2). Indeed, at pH 6.1, *ΔarpABC* growth was only 0.15 that of the WT under a similar condition (figure 3B). At a pH lower than 6, *ΔarpABC* was unable to grow in the presence of Sa, and thus no growth profile is available. In the absence of Sa, however, equivalent growth to that of D39V WT was observed under all pH conditions (figure 3A, 3C1). We also noted that in the presence of Sp, Sa naturally acidifies the growth media more than Sp by itself, possibly further exacerbating the pH impact on ArpABC mediated activity. Considering the normal nasopharynx pH ranges between 6.1 and 7.9 (Brunworth *et al.*, 2012), such observations hint at ArpABC being a key component in Sp adaptation within the nasopharynx microbiome.

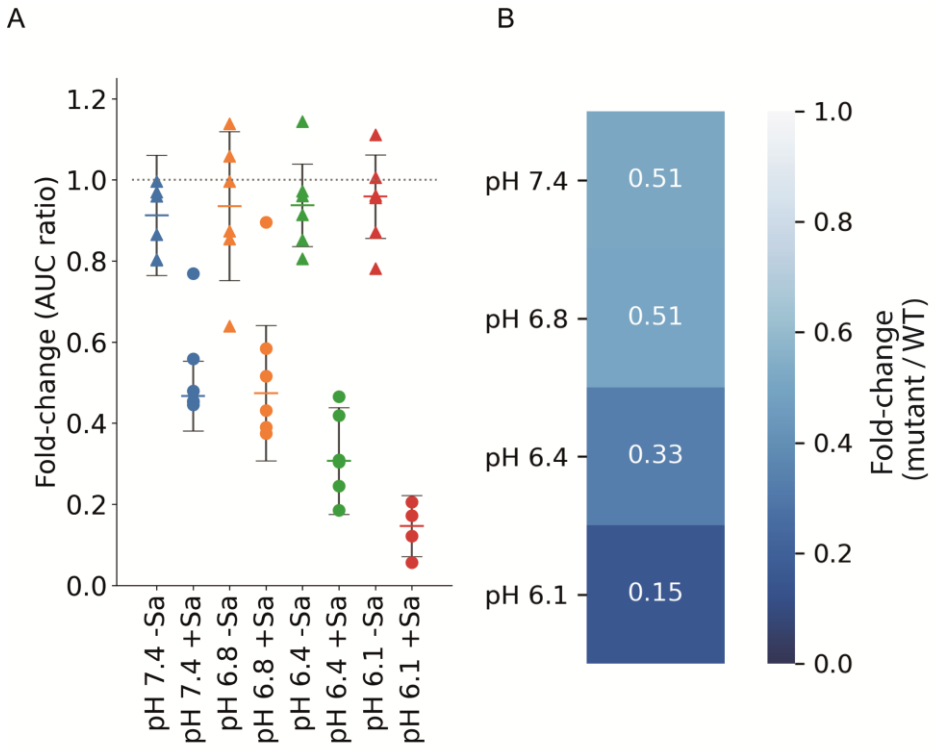


Figure 3 | ArpABC is required for competition with Sa in acidic environments

A) Growth fold-changes as measured by relative luminescence AUC between D39V WT and Δ*arpABC* in the presence/absence of Sa incubated for 4h in agar C+Y prior to addition of Sp, at different starting pH (mutant/WT AUC). **B**) Same as A, but relative fold-changes are normalized to the respective mutant/WT fold-change, for any given pH ((mutant/WT fold-change +Sa) / (mutant/WT fold-change -Sa)), and indicated as a heatmap where lower numbers indicate a worse fitness.

Sa dislodges Sp D39V from a Detroit 562 cell matrix

Several reports have highlighted the differences between *in vitro* (bacteria grown in laboratory conditions) conclusions and their applicability *in vivo* (bacteria during host infection), with the role of H_2O_2 on Sp colonization being one of such examples (see introduction). We therefore next sought to explore the impact Sa has on Sp co-colonization on a setting closer to the Sp natural environment, the human nasopharynx. To this end, we cultured pharynx epithelial cells (Detroit 562) to a confluent layer and infected them with either Sp, Sp and Sa, or Sa (see methods).

To maintain both Sp and Sa active growth, while minimizing damage to the Detroit 562 cells, we first determined the optimal multiplicity of infection (MOI) for both bacterial species. By tracking Sp metabolic activity via luminescence, and observing the spatial interaction, morphology, and distribution of Sp, Sa, and the Detroit 562 cells, we determined a multiplicity of infection (MOI) of 5 for Sp, with an infection time of 6h, would be the most suited. This would allow the time frames required to perform microscopy and genetic analyses on the different interactions (see methods) (figure 4) (supplementary figure 5 and 6).

When co-culturing Sp and Sa with Detroit 562, we observed that both species seem to co-exist within 2 exclusion zones: Vertically, with Sp mostly being dislodged to the top of the Sa matrix; and horizontally, with Sp concentrating within the gaps in the Sa matrix. Interestingly, Sa maintained a close association with the pharynx cells independently of the presence of Sp. This differed from Sp, where a looser association with the Detroit 562 cells was observed, even when in single-culture conditions (figure 4A, 4B, 4C). Such weaker connection to the pharynx cell matrix could readily dislodge Sp upon agitation. Indeed, upon performing scanning electron microscopy (SEM) on Sp, Sa, and Detroit 562 co-culture settings, we were unable to observe Sp, possibly due to the several washing stages required for SEM sample preparation (figure 4D, 4E).

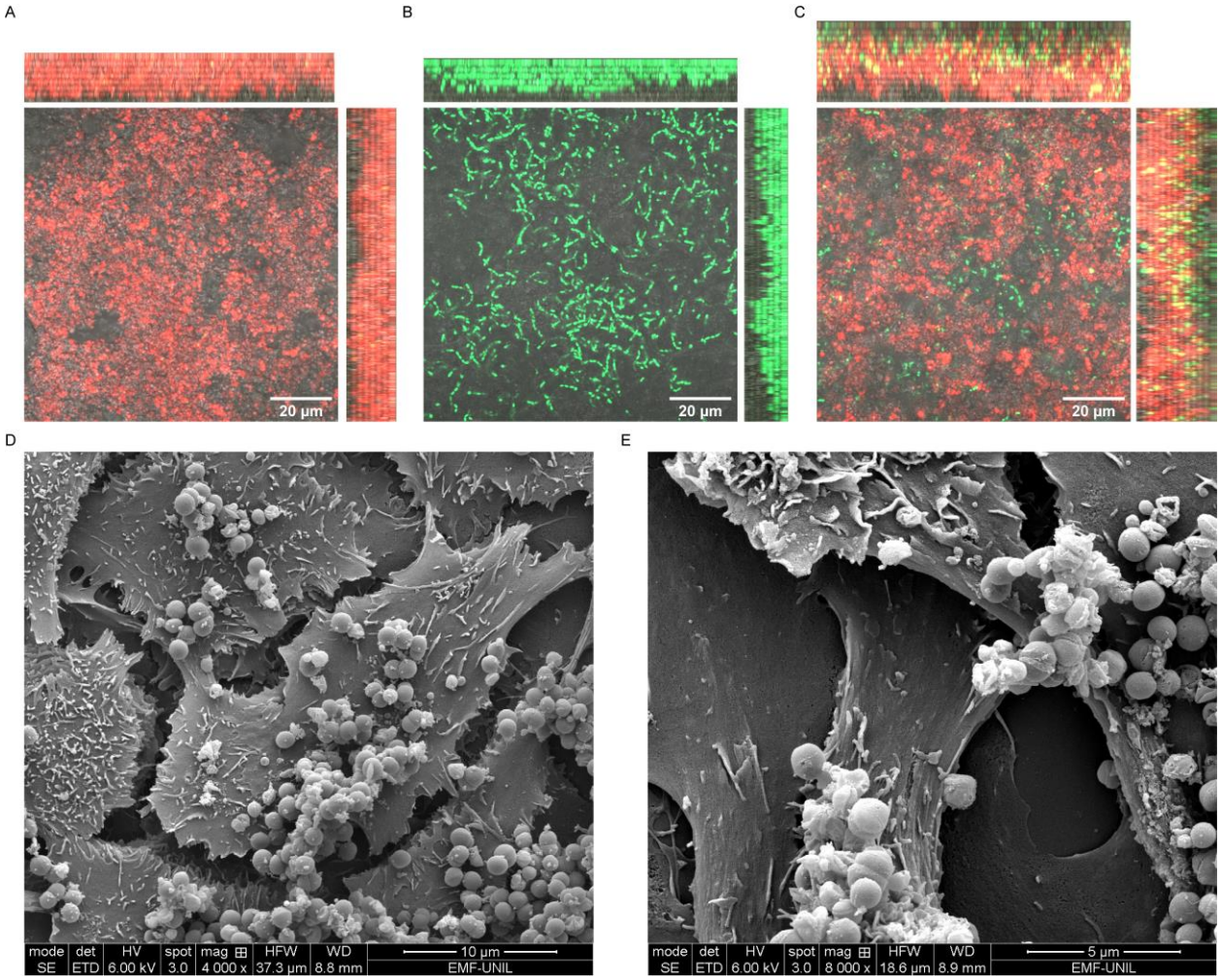


Figure 4 | Sp, Sa, or Sp and Sa on a human pharynx matrix (Detroit 562 cells)

Sp (green), Sa (red), or Sp and Sa were co-cultured with human pharynx Detroit 562 cells for 6h. The microscopy images were obtained by merging a Z-stack of bright field with both the red and green fluorescence fields (see methods). (**A**, **B**, **C**). **D** and **E**) SEM image of Detroit 562 cells inoculated with both Sp and Sa (**D**), or just Sp (**E**).

The Sp cell wall plays a key role on Sp competition with Sa in a human pharynx cell matrix

To further understand Sp interactions with both Detroit 562 and Sa, we next applied CRISPRi-seq under these settings. Considering that we previously demonstrated that the relative ratio of Sp Vs. competitor is a key factor in Sp competition, we submitted Detroit 562 cells to different inoculation times with Sa prior to adding the D39V Sp CRISPRi library. The Sp library was either simultaneously added with Sa, or following pre-incubation of the Detroit 562 cells for either 1h 45min, or 3h (see methods). These conditions were chosen based on their respective impact on the metabolic activity of Sp, as measured by relative luminescence (supplementary figure 5) .

Adding Sa at different time points to Detroit 562 cells resulted in different gene fitness requirements on Sp. We observed that the previously reported genes, *prtA* and *arpABC*, are not significantly observed in any of the tested conditions (figure 5C; supplementary figure 7). However, a borderline significant fitness defect for *prtA* was seen when Sp was simultaneously incubated with Sa on an epithelial cell matrix (log_2 fold change = -1.35, p-value = 0.08). In the same condition, genes involved in Sp cell wall/shape become more critical for survival, in line with the results obtained when Sp is cultured with just Detroit 562 cells, possibly indicating a reduced effect of Sa in the environment (*tar* operon, *rodZ*, *rodA*, *mreC*) (figure 5A, 5D). Conversely, when Sa is pre-incubated alone with Detroit 562 prior to Sp (for either 1h45min, or 3h), cell wall related genes associated with peptidoglycan biosynthesis, such as *pbp2x*, *ddl*, or the *potABCD* operon, gain a fitness defect, not being as required under these conditions as when Sp is by itself (x-axis of both figure 5B, supplementary figure 7G). This effect seems to be dependent on the presence of the Detroit 562 cells, as when these are absent, such pathways do not display a differential fitness defect (supplementary figure 7A, 7B, 7E). Also dependent on the presence of Detroit 562 cells, is the increased Sp requirement for the mevalonate pathway (*mva* gene family), with the presence of Sa decreasing the need for these metabolites under the same conditions (supplementary figure 7C, 7D). When no Detroit 562 cells are added, such effect is also seen, with the mevalonate pathway being less required when Sp is co-cultured with Sa than when grown by itself (supplementary figure 7). Considering that a depletion in the mevalonate pathway has been previously associated with a reduced

amount of peptidoglycan precursors (Dewachter *et al.*, 2022), the observed negative effects in the presence of Detroit 562 cells can be linked with the also negative gene fitness of several cell wall related genes, under the same conditions. The amelioration of the mevalonate pathway fitness in the presence of Sa could then be due to metabolic precursor sharing, and thus indicate some degree of metabolic synergy between both species when in RPMI media (supplementary figure 7C).

After 3h of Sa pre-incubation, most of the observed Sp gene fitness defects seem to originate from Sa itself, as the same overall gene fitness of all genes is observed in both the presence or absence of Detroit 562 cells (supplementary figure 7A). Such could indicate that after 3h, Sp mainly interacts with Sa, irrespective of the pharynx cell matrix. Indeed, we have demonstrated that Sp D39V mainly co-exists on top of the Sa-Detroit 562 matrix (figure 4C).

The *cps* operon encodes the Sp capsule, and, in D39V, the serotype 2 capsule. Interestingly, in all conditions where Sa was present, CRISPRi knockdown of the *cps2* operon consistently exhibited a positive fitness defect. This implies that this operon is detrimental for growth only when Sa is present, independently of the presence of Detroit 562 cells. Alternatively, reduced capsule could lead to increased adherence to the epithelial cells (Kjos *et al.*, 2015) thereby providing a competitive advantage in the presence of Sa. Exceptionally, *cps2* displays neutral fitness only when Sp and Sa co-exist in C+Y agar, indicating *cps2* to be detrimental for Sp competition in liquid RPMI media, but having no effect when cells are on a C+Y agar surface (figure 5C, supplementary figure 7F). Under these latter settings, cell wall related pathways once again play a key role in Sp adaptation to Sa. Although such effects could derive from the longer Sa incubation time in the C+Y agar condition than on RPMI, these results nonetheless contribute to show the prominent role of the cell wall on Sa competition. Curiously, no significant cell wall related genes were observed in the initial agar based CRISPRi screen (figure 2A), again highlighting the condition dependent role of gene essentiality/fitness.

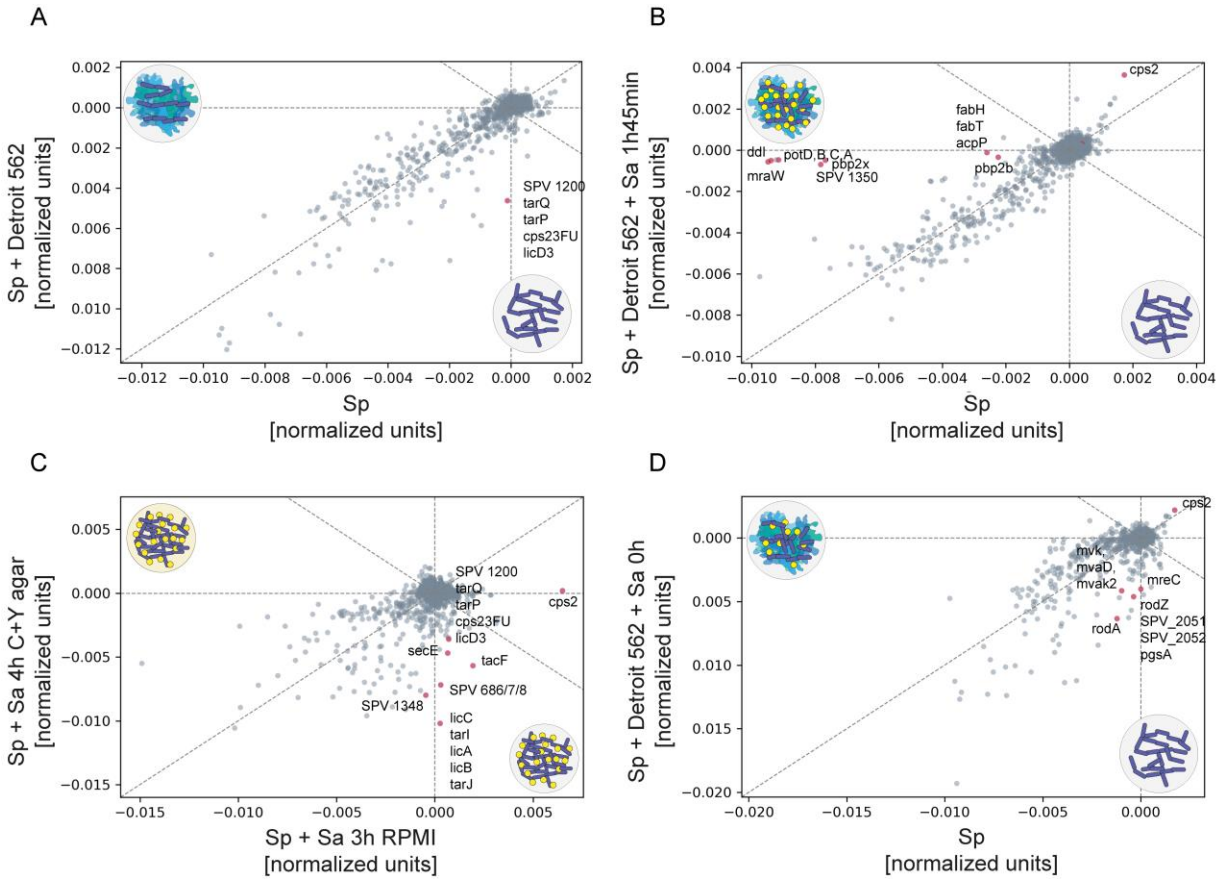


Figure 5 | Sp cell wall pathway fitness changes in regards to the presence of both Sa and human pharynx cells (Detroit 562)

D39V CRISPRi library normalized fitness comparison between the presence and absence of Sa, and/or Detroit 562 cells in RPMI media. Sp was either added simultaneously with Sa (**D**), or after 1h45min (**B**), or 3h (x-axis, **C**). Without Sa, but with Detroit 562 epithelial cells (**A**). In the case of **C** the conditions used (C+Y agar, y-axis) correspond to the same used in the assay indicated in figure 2A. The normalized values correspond to the L1 Norm of the $\log_2$ fold-changes (see methods). A value of 0 indicates a neutral fitness effect.

Discussion

Sp survival as either a human commensal or a pathogen is dependent on the crucial initial step of nasopharynx colonization. A plethora of studies have approached Sp requirements for adherence with human pharynx cells (Mlacha *et al.*, 2013; S. Novick *et al.*, 2017; Weight *et al.*, 2019), however only a few have dwelled into gene essentiality under such conditions (van Opijnen & Camilli, 2012). Indeed, transcriptional responses only describe how a gene expression changes, not how a gene changes the overall fitness of a cell when under a certain condition. The two have often been demonstrated to be uncorrelated, especially regarding core essential genes, where constant expression profiles are commonly seen across conditions (Jensen *et al.*, 2017). In here we have expanded on this knowledge and examined Sp gene essentiality using a three-part system, where a Sp CRISPRi library was grown in the presence of human Detroit 562 pharynx cells, and Sa. Unlike previous Tn-Seq studies, CRISPRi-Seq allows for the examination of the relative fitness of core essential genes (in this context defined as genes essential under all conditions), thus opening the door for new research avenues (de Bakker *et al.*, 2022; Liu *et al.*, 2021; van Opijnen *et al.*, 2009; van Opijnen & Camilli, 2012, 2013; van Opijnen *et al.*, 2014). We believe this is the first study where the relative essentiality of most Sp operons is examined under these settings.

Both ArpABC and PrtA were observed to be required by Sp to compete against other bacteria (itself, Sa, and *E. coli*), albeit only in close contact solid media conditions (figure 2A, figure 5, supplementary figure 7). PrtA, however, exhibited a borderline significant fitness defect when Sp was simultaneously inoculated with Sa on Detroit 562 cells ($\log_2$ fold change = -1.35, p-value = 0.08). PrtA has been previously demonstrated to worsen Sp systemic infection in mice, although not for all Sp strains (Bethe *et al.*, 2001; Mahdi *et al.*, 2015; Marquart, 2021). Considering PrtA is released into the extracellular milieu following maturation, it could be involved in the degradation of extracellular matrix components. Indeed, PrtA has been shown to interact with collagen (Frolet *et al.*, 2010), and S8 serine peptidases have been demonstrated to process lantibiotics into their active forms (Y. Zhang *et al.*, 2022). Besides its involvement in *in vivo* infection, PrtA could also have a role in either hindering the growth of competing bacteria via matrix degradation, or in self-defense by cleaving potentially harmful peptides secreted by neighboring cells. Considering *prtA* has been

demonstrated to be upregulated in the presence of various antimicrobial peptides, such as Bacitracin, LL-37, and Nisin (Majchrzykiewicz *et al.*, 2010), such results are in line with Sp population dynamics, which are often characterized by their aggressive fratricide strategies, with competent Sp inducing the lysis of non-competent cells (Claverys & Havarstein, 2007; Guiral *et al.*, 2005).

We have also demonstrated that the ABC-type MacAB-like efflux pump ArpABC (Yang *et al.*, 2018) increases Sp growth when in the presence of Sa, with its activity being essential for Sp survival with Sa in acidic pH (~6) (figure 3). Interestingly, a MacAB-like efflux pump has also been shown to protect *Salmonella enterica* from oxidative stress by the excretion of a soluble anti-H_2O_2 molecule (Bogomolnaya *et al.*, 2013). Similarly to *prtA*, *arpABC* has been demonstrated to be upregulated, and confer increased resistance in the presence of both bacitracin and LL-37 (Majchrzykiewicz *et al.*, 2010). Moreover, *arpABC* is also upregulated when competence stimulating peptide-1 (CSP) is added to the media, and in the initial steps of Sp A549 cells infection (1h) (Aprianto *et al.*, 2018). Another study also related this efflux pump with a decreased susceptibility to antimicrobials, such as erythromycin, fosfomycin, and fusidic acid (Marrer *et al.*, 2006). Genomic and transcriptome analysis also showed that *arpABC* has a complex regulation, with two distinct promoters, RpoD and GntR (Slager *et al.*, 2018). GntR, one of the largest bacterial transcription factor families, is also known to negatively regulate another ABC transporter involved in antimicrobial resistance, SPV 1525/6 (Majchrzykiewicz *et al.*, 2010). Altogether, these results hint at the multi-purpose function of this pump. In the future, it will be interesting to see the exact role of GntR on the here observed fitness. Furthermore, whether the entire ArpABC operon or just some of its subunits are required for the reported phenotypes still needs to be tested. ArpC, however, has been demonstrated to be essential for a fully functioning ArpABC, with ArpA cooperating with ArpBC to form a continuous efficient tunnel for subtract transportation from the extracellular membrane face to the peptidoglycan layer (figure 6) (Yang *et al.*, 2018).

Sa has recently been demonstrated to inhibit *Pseudomonas aeruginosa* in a glucose and pH dependent manner by production of acetoin, acetic acid, and other oligopeptides or cyclic peptides (Kvich *et al.*, 2022). The observed decrease in Δ*arpABC* survival when in competition with Sa at low pH could thus be related with ArpABC excreting Sa-produced low pH activated compounds. Such could also explain

why *arpABC* did not display any significant differential fitness when Sp was competed with Sa in liquid RPMI media. Unlike C+Y, RPMI is buffered at pH 7.4. The closer contact of Sp with its competitors when in an agar matrix can also be a factor due to the possible exacerbation of any possible effects due to higher local concentrations of various compounds. Further testing would be required to determine if these effects are also observed at a lower pH in the presence of pharynx cell, or when in competition with other bacteria. Such conditions would not be physiologically unrealistic, as nasopharynx pH can be as low as 5 in diseased individuals (Brunworth *et al.*, 2012). *arpABC* thus emerges as a conditional essential gene, being directly involved in the Sp general antimicrobial defense system.

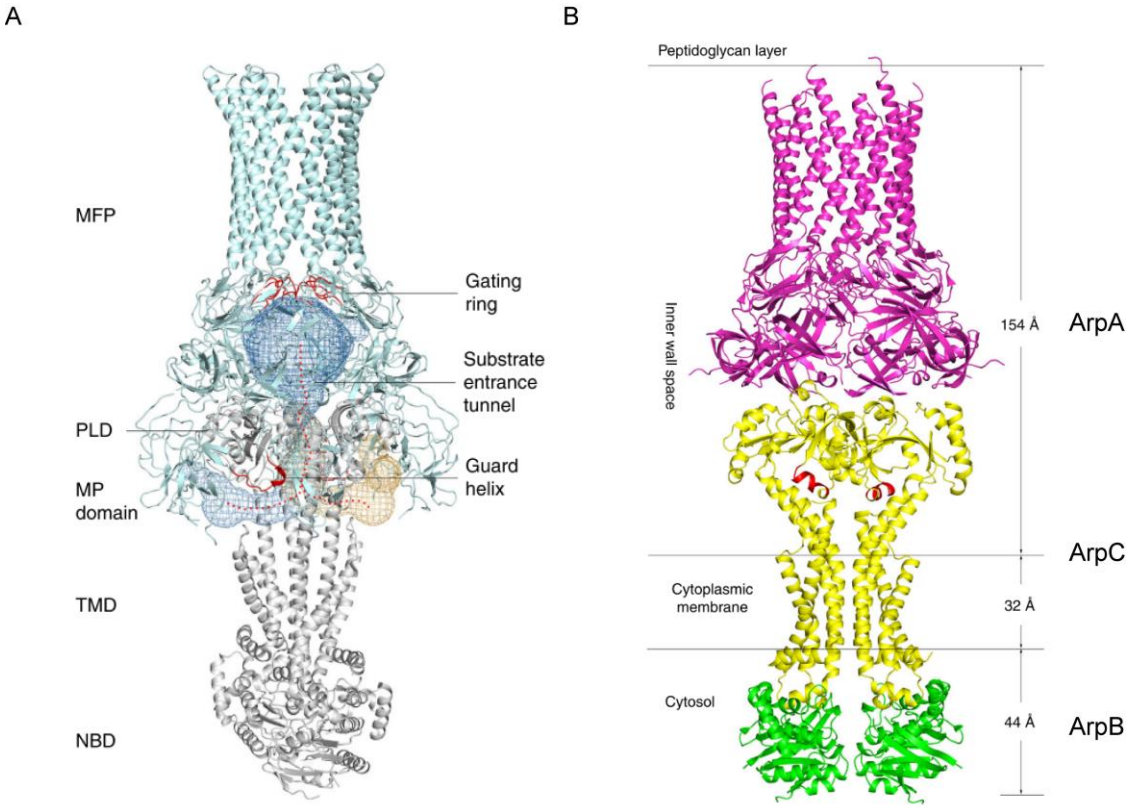


Figure 6 | ArpABC protein assembly

A) ArpABC substract-entrance tunnel as calculated by the program CAVEAR 3.0.1. MFP (membrane fusion protein), PLD (periplasmatic domain), MP (membrane proximal), TMD (transmembrane domain), NBD (nucleotide-binding domains) are indicated. ArpA augments ArpBC activity by two-fold. **B**) Simulated modle of ArpABC, built by manually superimposing the structures of AprA, and ArpBC. Adapted from Yang *et al*. (Yang *et al.*, 2018).

The *lic* operon is upregulated in the presence of pharynx cells (S. Novick *et al.*, 2017; Orihuela *et al.*, 2004). In here we report the same operon to have a fitness defect in both the presence of Detroit 562 cells and absence of Sa (log_2 fold change = -5, p-value = $1x10^{-9}$), when compared with all other conditions (presence/absence of Detroit

562 cells with Sa, and Sp by itself). Lic is involved in the metabolism of phosphorylcholine, which binds to the teichoic acids and lipoteichoic acids present in Sp surface. Considering phosphorylcholine is required for increased adherence to human cells (J. R. Zhang *et al.*, 1999), such results suggest that an increased adherence to the underlying cell matrix increases Sp fitness, but isn't as required when a competitor is present, or when a cell matrix is absent (neutral fitness). Such close contact has been associated with the first stage of invasive disease (Hammerschmidt *et al.*, 2005; S. Novick *et al.*, 2017). The limited effect observed when Sa is present could thus be caused by the protective effect one established species can have on modulating infection by another invading species. Indeed, we observed Sa tendency to occlude Sp D39V from direct contact with the pharynx matrix (figure 4). Similar results were obtained when Sp was co-cultured into a Sa agar matrix (figure 5C), and further highlight the overall net positive fitness effect on Sp when there is close contact with the surrounding matrix.

The requirement for essential cell wall related pathways (involved in the peptidoglycan pathway, figure 5B) was the most reduced when Sa was present, with similar mutants displaying a severe fitness defect when D39V is by itself. The largest effect was observed when Sa was co-cultured the longest time with Detroit 562 cells (figure 5B). Together, such results exemplify how assumed core essential functions, such as peptidoglycan synthesis, can be considered conditionally essential depending on the conditions. In this case, such reduced requirement might arise from Sp leveraging the existing Sa matrix as a structural support for its own cells, and therefore gaining a fitness advantage by bypassing an otherwise costly essential pathway. Such hypothesis also explains the greater decrease in capsule (*cps*) requirement observed when Sa is present (figure 5). Such effect, however, is not observed in solid agar media, where the fitness advantage arises mainly from the reduction in cell wall metabolism, as opposed from a concomitant reduction in cell wall and capsule production, as indicated by the neutral fitness of this latter (figure 5C). Further testing would be required to precisely ascertain the impact of capsule production on Sp interaction with Sa, however it is possible that the lack of a capsule would facilitate Sp dissemination within a Sa matrix.

Capsule shredding has been reported to increase Sp adherence, and as a first step in Sp transition into disease, with unencapsulated Sp strains significantly adhering better to human cells than encapsulated ones (Hammerschmidt *et al.*, 2005;

S. Novick *et al.*, 2017). Such results agree with our observations, where the current Sp D39V strain fails to be detected by SEM when using a cell matrix model, probably due to washing of the low adhered cells (figure 3D, 3E).

With this work, we have explored the dynamic role of gene essentiality under diverse complex environments, particularly how non-essential genes might transition to essential, and how previously known core essentials shift to non-essential. We have also demonstrated how changes in relative bacteria concentration, pH, and media type have on modulating Sp response to the same stimulus type: competition with other bacteria. Altogether, such results serve as a reminder of the infinitely intricate bacterial regulation system, and how *in vitro* relevance must always be considered in the context of all possible bacterial natural environments.

Methods

Strains and libraries

The Sp strain D39V and its derivatives were either cultured at 37°C without shaking in liquid C+Y (pH 6,8), or at 37°C with 5% CO_2 in C+Y (pH 6.8) 1% agar (Martin *et al.*, 1995), unless stated otherwise. See table 1 for the full list of used strains.

The used Sp D39V CRISPRi library consist of an IPTG-inducible CRISPRi system with 1498 pooled individual different guide RNAs (sgRNAs) targeting 2,111 genetic features (de Bakker *et al.*, 2022; Liu *et al.*, 2017). IPTG induction activates dCas9 expression, ultimately resulting in the transcriptional arrest of the sgRNAs´ binding site, usually located at the transcriptional start site of an operon.

Sa strain NCTC8325-4 was either grown in BHI media at 37ºC with shaking at 200rpm, or in LBA at 37ºC, unless stated otherwise.

Detroit 562 (ATCC® CCL-138™) cells were routinely grown in adherent cell culture flasks in Dulbecco's Modified Eagle Medium (DMEM) (ThermoFisher Scientific) supplemented with 10% FCS, 25mM HEPES, and 50U/ml of PS (Pen-strep) at 37ºC with 5% CO_2.

Electron Microscopy

Infection of Detroit 562 cells with Sa NCTC8325-4 (MOI of 5) and/or Sp D39V WT (MOI of 5) was performed as described (see ‘Sp Vs. Detroit 562 / Sp CRISPRi-Seq’ section). SEM imagining was performed in collaboration with the Electron Microscopy Facility at the University of Lausanne.

Confocal Microscopy

Briefly, 200µl of DMEM Detroit 562 cells at a concentration of $1X10^6$ cells/ml were inoculated into an 8 well ibidi microscopy chamber and incubated at 37ºC, 5%CO_2. 2 days after, at a concentration of ~$2X10^6$ cells/ml, the wells were washed with 200µl of PBS and inoculated with 200µl of RPMI media. When required, fluorescent Sp (VL1978) (MOI=5) and/or Sa (VL4874) (MOI=10) were added to the media. Imaging was performed after 6h of incubation at 37ºC, 5%CO_2, using a Zeiss LSM 900 confocal microscope. The resulting bright, and red/green fluorescence fields

were merged and processed using the Fiji ImageJ software (Schindelin *et al.*, 2012). For Z-stack creation, slices of 0.2μm were used.

Table 1 | Bacterial strains used in this study

See table 2 and 3 for further description on all built strains, and used primers.

Strain name	*Description*	*Reference*
Sa NCTC8325-4	MSSA strain, derivative of NCRC8325, cured of phages	(R. Novick, 1967)
Sa VL4874 (red)	NCTC8325-4, pSRFPS1(pKK30)-tmp^r	Lab collection
Sp D39V WT (VL1)	Serotype 2	(Slager *et al.*, 2018)
Sp VL1978 (green)	D39V, HlpA::HlpA-mNeongreen-cm^r	Lab collection
Sp D39V WT (luminescent) (VL551)	D39V, cep::spec^r-P3-Luciferase	Lab collection
Δprta (VL3893)	D39V, Δ*SPV 558*::ery^r, cep::spec^r-P3-luciferase,	This study
ΔarpABC (VL3933)	D39V, Δ*arpABC*::ery^r, cep::spec^r-P3-luciferase	This study
Complementation of ΔarpABC (VL5303)	D39V, Δ*arpABC*::ery^r, cep::spec^r-P3-luciferase, cil::*arpABC*-kan^r	This study

Strain building

For transformation with the appropriate DNA fragments, Sp was incubated in C+Y at 37°C to OD595 ~0.11. Competence was induced by addition of 100 ng/ml synthetic CSP-1, followed by 12 min of incubation at 37°C. 100µl of the resulting culture were sub-divided into as many tubes as transformation reactions (100µl of cells per transformation reaction). DNA was added and uptake occurred for 20 min at 30°C. cells were allowed to recover for up to 1.5h. Transformants were selected by mixing the cell culture with Columbia agar supplemented with 4% defibrinated sheep blood (CBA, Thermo Scientific). Appropriate antibiotics were added as needed. Incubated proceeded overnight at 37°C, 5% CO_2. Successful transformants were confirmed by PCR and Sanger sequencing (Microsynth). Successfully built strains were stocked at OD_{595} 0.3/0.4 in C+Y with 15% glycerol at -80°C.

Table 2 | Description for building all the strains used in this study

Strain name	*Description, and primers*
∆prta (VL3893)	The primers OVL5488, OVL5489, OVL5490, and OVL5491 were used to respectively PCR ~1000bp of the upstream and downstream regions of *SPV 558* (*prtA*). OVL1767 and OVL1768 were used to PCR the ery^r gene. BsaI was used to cut all fragments. Following ligation and PCR of the obtained correct fragment, transformation was carried as described.
∆arpABC (VL3933)	The primers OVL5438, OVL5439, OVL5440, and OVL5441 were used to respectively PCR ~1000bp of the upstream and downstream regions of *SPV 686-8* (*arpABC*). OVL1767 and OVL1768 were used to PCR the ery^r gene. BsaI was used to cut all fragments. Following ligation and PCR of the obtained correct fragment, transformation was carried as described.
Complementation of ∆arpABC (VL5303)	The primers OVL5457, OVL7804, OVL7807, and OVL828 were used to respectively PCR ~900bp of the upstream and downstream regions of the cil::kan^r locus. The primers OVL7805 and OVL7806 were used to PCR the native *arpABC* gene. BsmBI was used to cut all fragments. Following ligation and PCR of the obtained correct fragment, transformation was carried as described.

Table 3 | Primers used in this study

Primer Name	*Primer Sequence*	*Primer Description*
OVL5438	CTTCGAAATGAATGGTAATGC	ISU_Fw_oper278
OVL5439	CGAAGTGGTCTCGAGTAAATGAAACTCCTTTTCTTTTTTACA	ISU_Rv_oper278
OVL5440	CGAAGTGGTCTCGAAGCGATCAACAAGATGGACACTC	ISD_Fw_oper278
OVL5441	ATTTAACATCCAACATCATAAGAAGG	ISD_Rv_oper278
OVL5488	AGTGAAGATTGTGCAGAGA	ISU_FW_SPV558
OVL5489	CGAAGTGGTCTCGAGTATTTAATTCCTTACATATTTATTTAACTTCCA	ISU_RV_SPV558
OVL5490	CGAAGTGGTCTCGAAGCGACAAAAGCTATAGAAAAAAATGGT	ISD_FW_SPV558
OVL5491	GAGCCAGAATATTTGTTTGACT	ISD_RV_SPV558
OVL5457	CAATCCACATCGGCCAGATCGTTATTC	Kan_R
OVL7804	TTGTGGCGTCTCGGAGTGGTACCGGCTGCATGCATCG	1_CIL_kan_UP_R
OVL828	AATGATACGGCGACCACCGAGATCTACACAGAGTAGATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAACATAAAGAAAGGCCCGGCGC	1_CIL_kan_DW_R
OVL7805	TTGTGGCGTCTCGACTCTGCTACGCACAAAAAATTGC	2_SPV_686/7/8_F

OVL7806	TTGTGGCGTCTCGTCCAAAAACAAGATAGACGAGTGTCC	3_SPV_686/7/8_R
OVL7807	TTGTGGCGTCTCCTGGATCCCTCCAGTAACTCGTC	4_CIL_KAN_DW_F
OVL1767	CGAAGTGGTCTCGTCATGAACAAAAATATAAAATATTCTCAAAACT	Ery GD_F
OVL1768	CGAAGTGGTCTCGGCTTATTTCCTCCCGTTAAATAATAGAT	Ery GD_R
Read1-custom	CTTGACATTGCACTGTCCCCCTGGTATAATAACTATA	Illumina MiniSeq custom primer
N501	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATTCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N504	AATGATACGGCGACCACCGAGATCTACACAGAGTAGATCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGTCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N506	AATGATACGGCGACCACCGAGATCTACACACTGCATATCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTTCGTCGGCAGCGTCAGATGTGTATA	Illumina MiniSeq indexing primer
N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N708	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N709	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer
N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGTGTAT	Illumina MiniSeq indexing primer

Agar-based luminescence assay

200µl of C+Y 1% agar supplemented with 0.1 mg/ml of luciferin were added to a 96-well black polystyrene microplate plate. When required, the pH of the C+Y media was adjusted by adding HCl. The needed luminescent Sp strains were defrosted and used to inoculate 2ml of liquid C+Y at a 1/100 dilution. All strains were incubated at 37ºC until OD_{600} ~0.1, at which point another dilution was performed to a final OD_{600} of 0.0001 (100,000 CFU/ml). When appropriate, such dilution was mixed with non-luminescent competitors at the desired CFU ratios (1:1;1:10:1:100;1:1000). In the case of Sa, an appropriately diluted 16h overnight culture (~5M CFU/ml) of NCTC8325-4 was used for all assays. When pre-incubating the C+Y with competitors for 4h, an OD_{600} of 0.000001 (100 CFU/ml) and an OD_{600} of 0.001 (125,000 CFU/ml) were used for Sp D39V WT and Sa NCTC8325-4, respectively. Luminescence was measured every 10min for at least 16h using a plate reader (Infinite F200, Tecan). All resulting data analysis were performed using a custom made Python3 script. Prior to AUC determination, time points were trimmed in such a way that the first data point of each sample corresponded to the first point where the luminescence value was above 100 (a.u.). Such threshold was defined based on background noise level fluctuations, and allowed for correction of experimental artifacts by aligning all samples to time 0 at the start of exponential growth. AUC was then calculated between 0h and 8h, which normally corresponded to the onset of stationary phase, and the decrease in luminescence (metabolic activity). Such time frames also corresponded to the times used for CRISPRi-Seq. All relative fold-changes between samples were calculated by dividing the respective AUC, where appropriate.

Sp Vs. Detroit 562 / Sp CRISPRi-Seq

Following inoculation into 12-well flat bottom, sterile, tissue-culture treated plates at an initial concentration of $1X10^6$ cells/ml, the cells were incubated for 2/3 days at 37ºC, 5%CO_2 until confluent (~$2X10^6$ cells/ml).

Wells were washed with 200µl of PBS prior to the addition of 5ml RPMI 1640 (ThermoFisher Scientific) supplemented with 1% FCS, 10nM HEPES, and 0.5% yeast extract. 1mM of IPTG was added when required. When appropriate, RPMI was supplemented with the Sp D39V CRISPRi library. To this end, an aliquot of the library was defrosted and diluted 1/40 in C+Y, and at OD_{578} = 0.1, diluted again to match a

MOI of 5. For Sa, a 16h overnight culture (~5M CFU/ml) of NCTC8325-4 was used at a MOI of 5 (when Sa was either added simultaneously with Sp, or 1h45min before Sp), or with a MOI of 50 when Sa was pre-incubated for 3h prior to adding Sp. 4 replicates per condition were used, combining all possible arrangements: +/- Detroit 562 cells; +/- Sa. Sa was either added simultaneously with Sp, or Sa was pre-incubated for either 1h45min, or 3h, prior to Sp. All plates were inoculated for 6h at 37ºC, 5%CO2.

Total DNA was extracted by adding 3 ml of $(NH_4)_2SO_4$ solution for every 1 ml of media. All adherent cells were scrapped and mixed. The mix was incubated at room temperature for 5 min and the sample solutions were collected in 15 ml tubes. 8 ml of TE was added and the samples centrifuged at 10.000 rpm for 20 min for pellet recovery. Total DNA was extracted for each sample (FastPure Blood/cel/Tissue/Bacetria DNA isolation; Vazyme). Illumina libraries were built and indexed by PCR (table 3), and sequenced in an Illumina MiniSeq device, as described (de Bakker *et al.*, 2022).

Sp Vs. Sa C+Y agar competition CRISPRi Assay

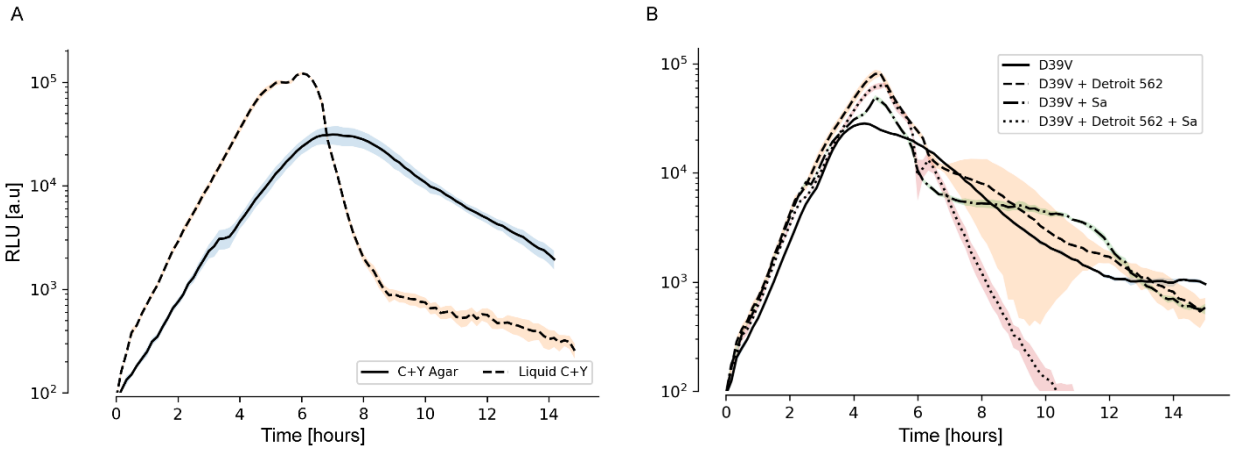
100µl of 16h overnight culture (~5M CFU/ml) of Sa NCTC8325-4 was inoculated into C+Y agar plates supplemented with and without 1mM IPTG, and grown for 4h at 37ºC, 5% CO_2. An aliquot of the Sp D39V CRISPRi library was defrosted and diluted 1/40 in C+Y. At OD_{578} = 0.1, and upon reaching 4h of Sa growth, 300µl of library were added to the control (C+Y agar with and without IPTG), and the previous Sa plates. After 8h of incubation at 37ºC, 5% CO_2, 2ml of C+Y were added to each plate, mixed, and the cell suspension collected. Total DNA was extracted for each sample (FastPure Blood/cel/Tissue/Bacetria DNA isolation; Vazyme). Illumina libraries were built and indexed by PCR (table 3), and sequenced in an Illumina MiniSeq device, as described (de Bakker *et al.*, 2022).

CRISPRi-Seq analysis

CRISPRi read counts were obtained from .fastq files using 2FAST2Q (chapter 2) (Bravo *et al.*, 2022). Fold change data analysis was performed using R DeSeq2 package (Love *et al.*, 2014). Essentially, the obtained reads per sgRNA were normalized, and differential analysis between the induced (+IPTG) and non-induced (-IPTG) samples were performed for each condition group (e.g.: Sp and Sa +IPTG Vs. Sp and Sa -IPTG). The obtained log_2 fold-changes indicated the relative fitness

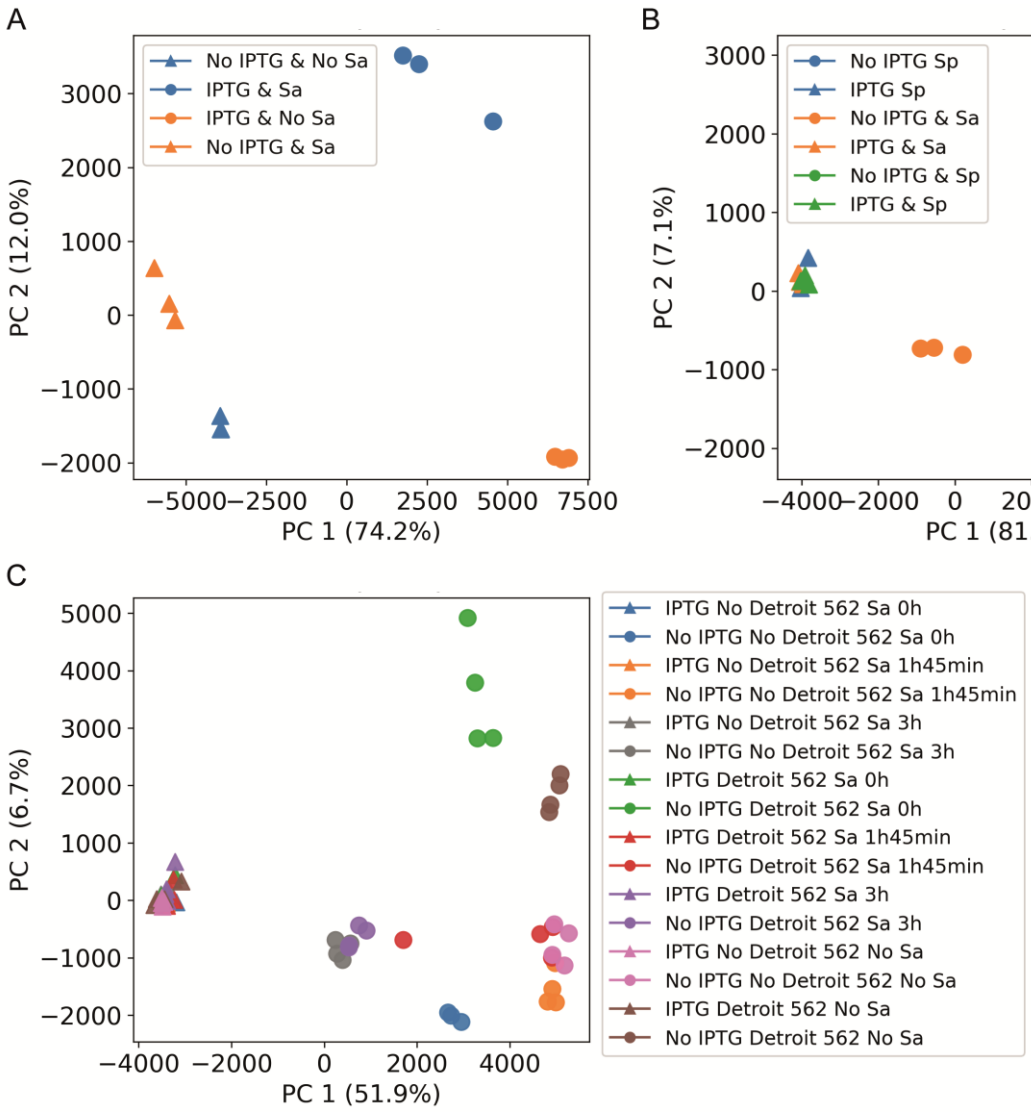
differences for each sgRNA under the tested condition. Comparison across different conditions was performed using a custom Python3 script. As log_2 fold-changes directly relate with a sample's total number of generations, different conditions might have different relative fold-changes for the same gene fitness, and thus incur in the risk of erroneously indicating a difference between conditions when different conditions are compared. To mitigate this possibility, for each pair of conditions (e.g.: Sp Vs. Sp and Sa), log_2 fold-changes were normalized based on the l1 Norm, calculated using the Scikit-learn Python module (Pedregosa *et al.*, 2011). Normalization corrected differences in the generation times between the 2 different conditions being compared, thus allowing the inference of sample specific differential fitness genes without log_2 fold-change generational skewness.

Supplementary



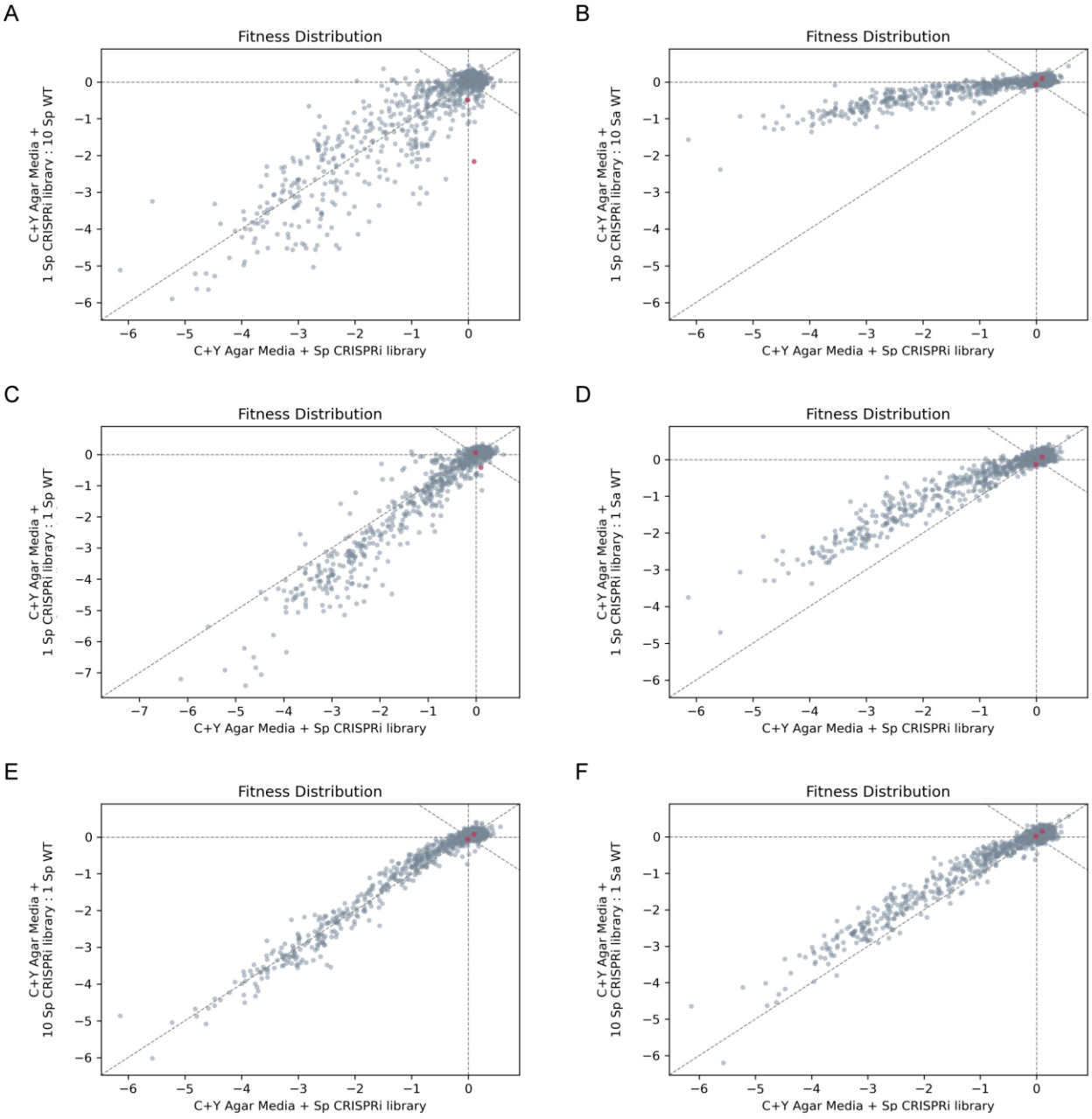
Supplementary figure 1 | Luminescence curves for *S. pneumoniae* D39V constitutively expressing the firefly luciferase gene under different co-culture conditions.

Luminescence was tracked as described and is represented as arbitrary units (a.u). D39V metabolic activity in: **A**) C+Y agar, and C+Y liquid. **B**) RPMI+Y, and RPMI+Y with Detroit 562 nasopharynx cells.



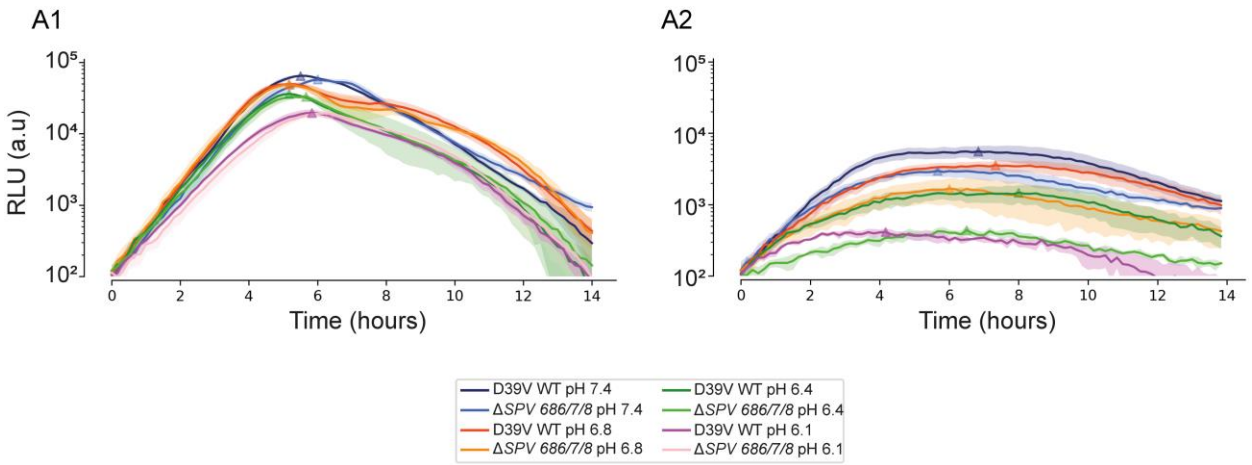
Supplementary figure 2 | PCA plot of the normalized reads per million (RPM) for the Sp CRISPRi library.

A) PCA distribution of the normalized reads of the 3 replicates per experimental condition (+/- 1mM IPTG; +/- Sa) from the experimental setup used to obtain the data shown in figure 2A. **B**) PCA distribution of the normalized reads of the 3 replicates per experimental condition (+/- 1mM IPTG; +/- Sa; +/- Sp WT) from the experimental setup used to obtain the data shown in figure 2F and 2G. **C**) PCA distribution of the normalized reads of the 4 replicates per experimental condition (+/- 1mM IPTG; +/- Sa; +/- Detroit 562) from the experimental setup used to obtain the data shown in figure 4.



Supplementary figure 3 | Un-normalized fitness comparisons at different starting ratios of Sp D39V CRISPRi library, and competitors

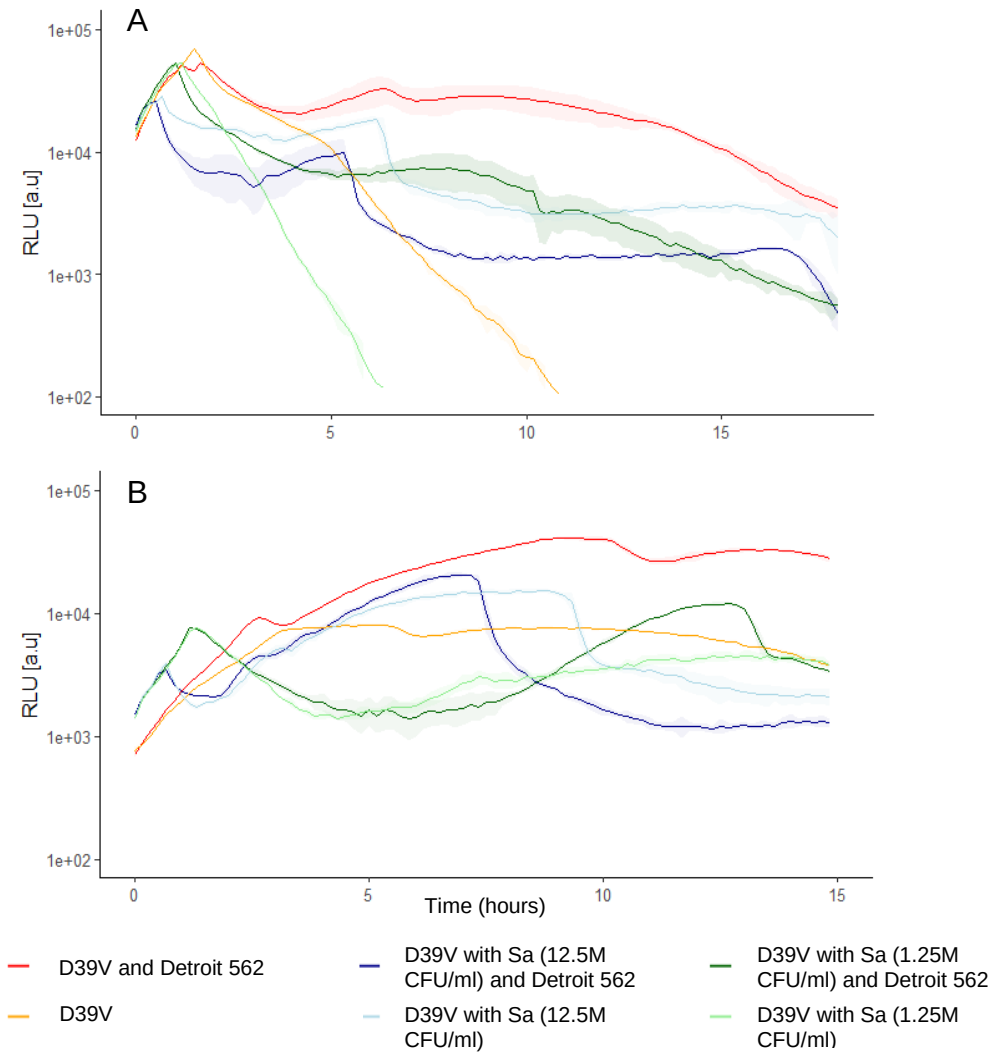
Fitness distribution of the Sp CRISPRi library co-cultured with either Sp WT or Sa when using a relative CFU ratio of either 10:1, 1:1, or 1:10 of Sp to competitor. Axis units correspond to $\log_2$ fold-changes.



[4]

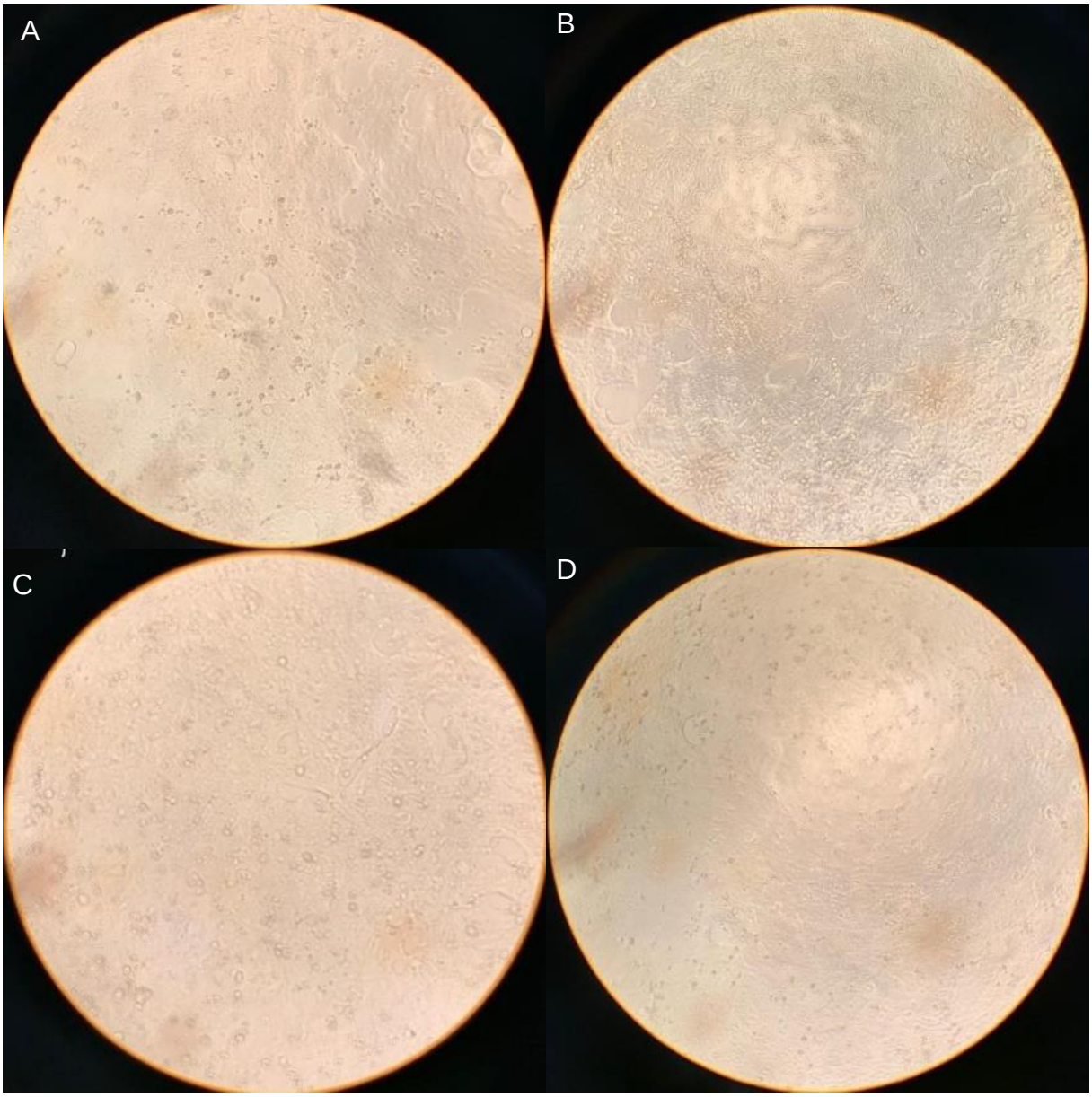
Supplementary figure 4 | Relative luminesce of D39V WT and Δ*arpABC* under different pH conditions

A1) Absence of Sa. **A2**) Presence of Sa.



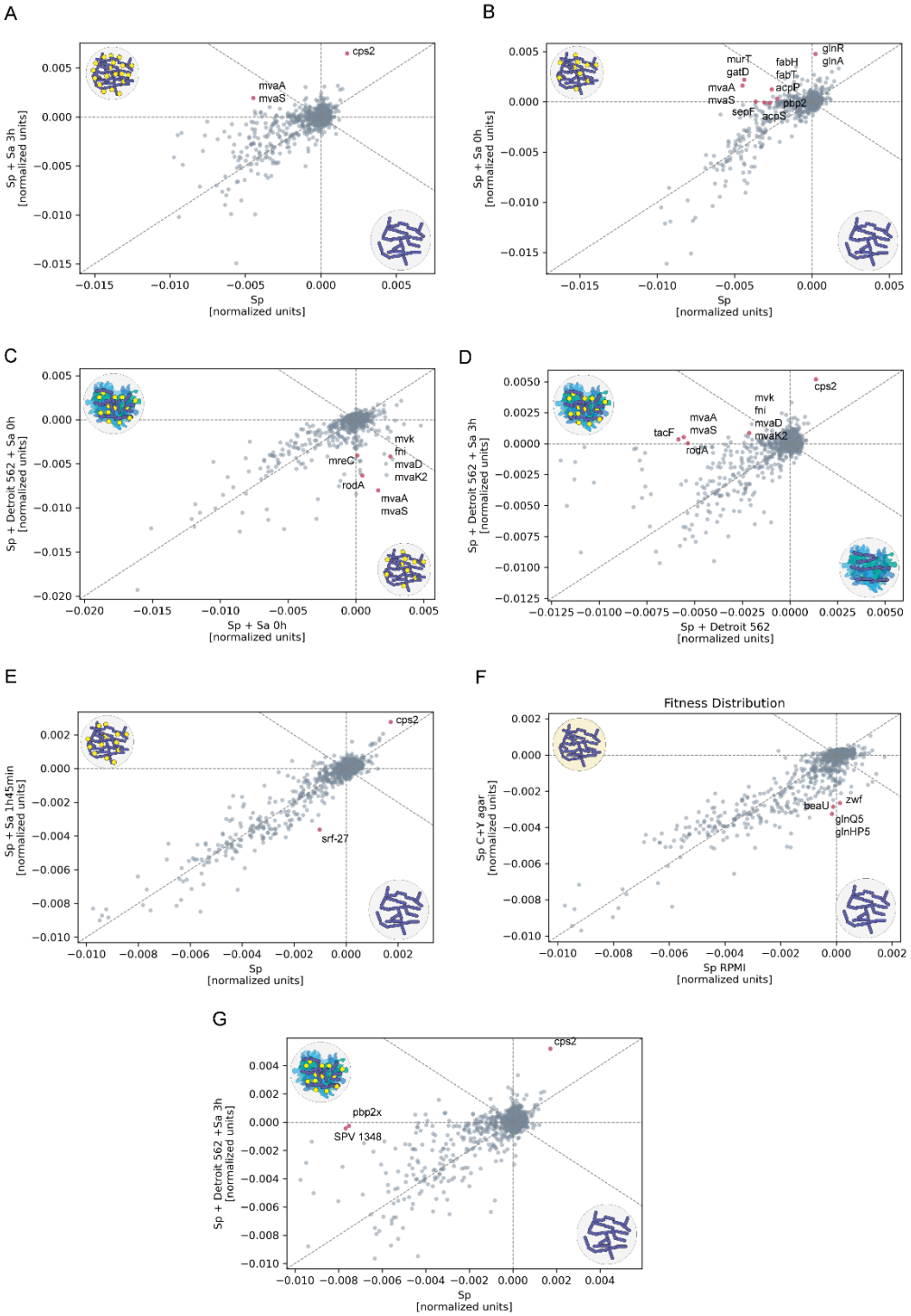
Supplementary figure 5 | Sp metabolic activity when in the presence of the tested CRISPRi-Seq conditions.

A) At a starting Sp concentration of 10M CFU/ml luminescence rapidly decreased following the start of the experiment, indicating either death or decreasing metabolic activity. This condition was thus deemed unsuitable for CRISPRi-Seq. **B)** Starting Sp concentration of 1M CFU/ml (MOI=5). This Sp concentration was chosen for the CRISPRi screen Sp concentration, with Sp reacting differently to the presence and absence of both epithelial cells, and Sa, whilst maintaining active metabolic activity.



Supplementary figure 6 | Bright field microscopy image of Detroit 562 when co-cultured with Sp (MOI=5) and Sa (MOI=6)

Bright field microscopy picture (amplification: 100x) of Detroit 562 cells with Sp (MOI=5) and Sa (MOI=6) at **A**) 1h post infection. **B**) 6h post infection. **C**) Detroit 562 cells in RPMI media 1h into the experiment. **D**) 6h into the experiment.



Supplementary figure 7 | The mevalonate pathway is less required when Sa is co-cultured with Sa

D39V CRISPRi library normalized fitness comparison with the absence/presence of Sa, pre incubated for either 1h45min, 3h, or simultaneously added to RPMI media. The normalized values correspond to the L1 Norm of the $\log_2$ fold-changes (see methods).

References

Aggarwal, S. D., Yesilkaya, H., Dawid, S., & Hiller, N. L. (2020). The pneumococcal social network. *PLoS Pathog, 16*(10), e1008931. doi:10.1371/journal.ppat.1008931

Aprianto, R., Slager, J., Holsappel, S., & Veening, J. W. (2018). High-resolution analysis of the pneumococcal transcriptome under a wide range of infection-relevant conditions. *Nucleic Acids Res, 46*(19), 9990-10006. doi:10.1093/nar/gky750

Asmat, T. M., Agarwal, V., Rath, S., Hildebrandt, J. P., & Hammerschmidt, S. (2011). Streptococcus pneumoniae infection of host epithelial cells via polymeric immunoglobulin receptor transiently induces calcium release from intracellular stores. *J Biol Chem, 286*(20), 17861-17869. doi:10.1074/jbc.M110.212225

Bethe, G., Nau, R., Wellmer, A., Hakenbeck, R., Reinert, R. R., Heinz, H.-P., & Zysk, G. (2001). The cell wall-associated serine protease PrtA: a highly conserved virulence factor of Streptococcus pneumoniae. *FEMS Microbiology Letters, 205*(1), 99–104.

Bogaert, D., van Belkum, A., Sluijter, M., Luijendijk, A., de Groot, R., Rümke, H. C., Verbrugh, H. A., & Hermans, P. W. M. (2004). Colonisation by Streptococcus pneumoniae and Staphylococcus aureus in healthy children. *The Lancet, 363*(9424), 1871-1872. doi:10.1016/s0140-6736(04)16357-5

Bogomolnaya, L. M., Andrews, K. D., Talamantes, M., Maple, A., Ragoza, Y., Vazquez-Torres, A., & Andrews-Polymenis, H. (2013). The ABC-type efflux pump MacAB protects Salmonella enterica serovar typhimurium from oxidative stress. *MBio, 4*(6), e00630-00613. doi:10.1128/mBio.00630-13

Bravo, A. M., Typas, A., & Veening, J.-W. (2022). 2FAST2Q: a general-purpose sequence search and counting program for FASTQ files. *PeerJ, 10*. doi:10.7717/peerj.14041

Brunworth, J. D., Garg, R., Mahboubi, H., Johnson, B., & Djalilian, H. R. (2012). Detecting nasopharyngeal reflux: a novel pH probe technique. *Ann Otol Rhinol Laryngol, 121*(7), 427-430.

Bryant, J. C., Dabbs, R. C., Oswalt, K. L., Brown, L. R., Rosch, J. W., Seo, K. S., Donaldson, J. R., McDaniel, L. S., & Thornton, J. A. (2016). Pyruvate oxidase of Streptococcus pneumoniae contributes to pneumolysin release. *BMC Microbiol, 16*(1), 271. doi:10.1186/s12866-016-0881-6

Claverys, J. P., & Havarstein, L. S. (2007). Cannibalism and fratricide: mechanisms and raisons d'etre. *Nat Rev Microbiol, 5*(3), 219-229. doi:10.1038/nrmicro1613

Dawid, S., Roche, A. M., & Weiser, J. N. (2007). The blp bacteriocins of Streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo. *Infect Immun, 75*(1), 443-451. doi:10.1128/IAI.01775-05

de Bakker, V., Liu, X., Bravo, A. M., & Veening, J.-W. (2022). CRISPRi-seq for genome-wide fitness quantification in bacteria. *Nature Protocols, 17*, 252–281

Dewachter, L., Dénéréaz, J., Liu, X., Bakker, V. d., Costa, C., Baldry, M., Sirard, J.-C., & Veening, J.-W. (2022). Amoxicillin-resistant Streptococcus pneumoniae can be resensitized by targeting the mevalonate pathway as indicated by sCRilecs-seq. *Elife, 11*(e75607).

Flynn, M., & Dooley, J. (2021). The microbiome of the nasopharynx. *J Med Microbiol, 70*(6). doi:10.1099/jmm.0.001368

Frolet, C., Beniazza, M., Roux, L., Gallet, B., Noirclerc-Savoye, M., Vernet, T., & Guilmi, A. M. D. (2010). New adhesin functions of surface-exposed pneumococcal proteins. *BMC Microbiology, 10*(190).

Guiral, S., Mitchell, T. J., Martin, B., & Claverys, J.-P. (2005). Competence-programmed predation of noncompetent cells in the human pathogen Streptococcus pneumoniae: Genetic requirements. *PNAS, 102*(24), 8710–8715

Hammerschmidt, S., Wolff, S., Hocke, A., Rosseau, S., Muller, E., & Rohde, M. (2005). Illustration of pneumococcal polysaccharide capsule during adherence and invasion of epithelial cells. *Infect Immun, 73*(8), 4653-4667. doi:10.1128/IAI.73.8.4653-4667.2005

J Kluytmans, Belkum, A. v., & Verbrugh, H. (1997). Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews, 10*(3), 505 - 520.

Jensen, P. A., Zhu, Z., & van Opijnen, T. (2017). Antibiotics Disrupt Coordination between Transcriptional and Phenotypic Stress Responses in Pathogenic Bacteria. *Cell Rep, 20*(7), 1705-1716. doi:10.1016/j.celrep.2017.07.062

Khan, F., Wu, X., Matzkin, G. L., Khan, M. A., Sakai, F., & Vidal, J. E. (2016). Streptococcus pneumoniae Eradicates Preformed Staphylococcus aureus Biofilms through a Mechanism Requiring Physical Contact. *Front Cell Infect Microbiol, 6*, 104. doi:10.3389/fcimb.2016.00104

Kjos, M., Aprianto, R., Fernandes, V. E., Andrew, P. W., van Strijp, J. A., Nijland, R., & Veening, J. W. (2015). Bright fluorescent Streptococcus pneumoniae for live-cell imaging of host-pathogen interactions. *J Bacteriol, 197*(5), 807-818. doi:10.1128/JB.02221-14

Kjos, M., Miller, E., Slager, J., Lake, F. B., Gericke, O., Roberts, I. S., Rozen, D. E., & Veening, J. W. (2016). Expression of Streptococcus pneumoniae Bacteriocins Is Induced by Antibiotics via Regulatory Interplay with the Competence System. *PLoS Pathog, 12*(2), e1005422. doi:10.1371/journal.ppat.1005422

Kvich, L., Crone, S., Christensen, M. H., Lima, R., Alhede, M., Alhede, M., Staerk, D., & Bjarnsholta, T. (2022). Investigation of the Mechanism and Chemistry Underlying Staphylococcus aureus' Ability to Inhibit Pseudomonas aeruginosa Growth In Vitro. *Journal of Bacteriology, 10.1128/jb.00174-22*.

Lee, G. M., Huang, S. S., Rifas-Shiman, S. L., Hinrichsen, V. L., Pelton, S. I., Kleinman, K., Hanage, W. P., Lipsitch, M., McAdam, A. J., & Finkelstein, J. A. (2009). Epidemiology and risk factors for Staphylococcus aureus colonization in children in the post-PCV7 era. *BMC Infect Dis, 9*, 110. doi:10.1186/1471-2334-9-110

Lehtinen, S., Croucher, N. J., Blanquart, F., & Fraser, C. (2022). Epidemiological dynamics of bacteriocin competition and antibiotic resistance. *Proc Biol Sci, 289*(1984), 20221197. doi:10.1098/rspb.2022.1197

Lijek, R. S., Luque, S. L., Liu, Q., Parker, D., Bae, T., & Weiser, J. N. (2012). Protection from the acquisition of Staphylococcus aureus nasal carriage by cross-reactive antibody to a pneumococcal dehydrogenase. *Proc Natl Acad Sci U S A, 109*(34), 13823-13828. doi:10.1073/pnas.1208075109

Lijek, R. S., & Weiser, J. N. (2012). Co-infection subverts mucosal immunity in the upper respiratory tract. *Curr Opin Immunol, 24*(4), 417-423. doi:10.1016/j.coi.2012.05.005

Liu, X., Gallay, C., Kjos, M., Domenech, A., Slager, J., Kessel, S. P. v., Knoops, K., Sorg, R. A., Zhang, J.-R., & Veening, J.-W. (2017). High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. *Mol Syst Biol, 13*(5), 931. doi:10.15252/msb.20167449

Liu, X., Kimmey, J. M., Matarazzo, L., Bakker, V. d., Maele, L. V., Sirard, J.-C., Nizet, V., & Veening, J.-W. (2021). Exploration of Bacterial Bottlenecks and Streptococcus pneumoniae Pathogenesis by CRISPRi-Seq *Cell Host & Microbe, 29*, 107-120.

Lo, S. W., Gladstone, R. A., van Tonder, A. J., Lees, J. A., du Plessis, M., Benisty, R., Givon-Lavi, N., Hawkins, P. A., Cornick, J. E., Kwambana-Adams, B., Law, P. Y., Ho, P. L., Antonio, M., Everett, D. B., Dagan, R., von Gottberg, A., Klugman, K. P., McGee, L., Breiman, R. F., Bentley, S. D., Brooks, A. W., Corso, A., Davydov, A., Maguire, A., Pollard, A., Kiran, A., Skoczynska, A., Moiane, B., Beall, B., Sigauque, B., Aanensen, D., Lehmann, D., Faccone, D., Foster-Nyarko, E., Bojang, E., Egorova, E., Voropaeva, E., Sampane-Donkor, E., Sadowy, E., Bigogo, G., Mucavele, H., Belabbès, H., Diawara, I., Moïsi, J., Verani, J., Keenan, J., Nair Thulasee Bhai, J. N., Ndlangisa, K. M., Zerouali, K., Ravikumar, K. L., Titov, L., De Gouveia, L., Alaerts, M., Ip, M., de Cunto Brandileone, M. C., Hasanuzzaman, M., Paragi, M., Nurse-Lucas, M., Ali, M., Elmdaghri, N., Croucher, N., Wolter, N., Porat, N., Eser, Ö. K., Akpaka, P. E., Turner, P., Gagetti, P., Tientcheu, P.-E., Carter, P. E., Mostowy, R., Kandasamy, R., Ford, R., Henderson, R., Malaker, R., Shakoor, S., Grassi Almeida, S. C., Saha, S. K., Doiphode, S., Madhi, S. A., Devi Sekaran, S., Srifuengfung, S., Obaro, S., Clarke, S. C., Nzenze, S. A., Kastrin, T., Ochoa, T. J., Balaji, V., Hryniewicz, W., & Urban, Y. (2019). Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *The Lancet Infectious Diseases, 19*(7), 759-769. doi:10.1016/s1473-3099(19)30297-x

Loughran, A. J., Orihuela, C. J., & Tuomanen, E. I. (2018). Streptococcus pneumoniae: Invasion and Inflammation. *Microbiology Spectrum, 7*(2). doi:10.1128/microbiolspec

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Mahdi, L. K., Van der Hoek, M. B., Ebrahimie, E., Paton, J. C., & Ogunniyi, A. D. (2015). Characterization of Pneumococcal Genes Involved in Bloodstream Invasion in a Mouse Model. *PLoS One, 10*(11), e0141816. doi:10.1371/journal.pone.0141816

Majchrzykiewicz, J. A., Kuipers, O. P., & Bijlsma, J. J. (2010). Generic and specific adaptive responses of Streptococcus pneumoniae to challenge with three distinct antimicrobial peptides, bacitracin, LL-37, and nisin. *Antimicrob Agents Chemother, 54*(1), 440-451. doi:10.1128/AAC.00769-09

Man, W. H., Clerc, M., de Steenhuijsen Piters, W. A. A., van Houten, M. A., Chu, M., Kool, J., Keijser, B. J. F., Sanders, E. A. M., & Bogaert, D. (2019). Loss of Microbial Topography between Oral and Nasopharyngeal Microbiota and Development of Respiratory Infections Early in Life. *Am J Respir Crit Care Med, 200*(6), 760-770. doi:10.1164/rccm.201810-1993OC

Margolis, E. (2009). Hydrogen peroxide-mediated interference competition

by Streptococcus pneumoniae has no significant effect on Staphylococcus aureus nasal colonization of neonatal rats. *J Bacteriol, 191*(2), 571-575. doi:10.1128/JB.00950-08

Margolis, E., Yates, A., & Levin, B. R. (2010). The ecology of nasal colonization of Streptococcus pneumoniae, Haemophilus influenzae and Staphylococcus aureus: the role of competition and interactions with host's immune response. *BMC Microbiology, 10*(59).

Marquart, M. E. (2021). Pathogenicity and virulence of Streptococcus pneumoniae: Cutting to the chase on proteases. *Virulence, 12*(1), 766-787. doi:10.1080/21505594.2021.1889812

Marrer, E., Schad, K., Satoh, A. T., Page, M. G., Johnson, M. M., & Piddock, L. J. (2006). Involvement of the putative ATP-dependent efflux proteins PatA and PatB in fluoroquinolone resistance of a multidrug-resistant mutant of Streptococcus pneumoniae. *Antimicrob Agents Chemother, 50*(2), 685-693. doi:10.1128/AAC.50.2.685-693.2006

Martin, B., García, P., Castanié, M.-P., & Claverys, J.-P. (1995). The recA gene of Streptococcus pneumoniae is part of a competence-induced operon and controls lysogenic induction. *Mol Micro, 15*(2), 367-379.

Martins, A., Spengler, G., Rodrigues, L., Viveiros, M., Ramos, J., Martins, M., Couto, I., Fanning, S., Pages, J. M., Bolla, J. M., Molnar, J., & Amaral, L. (2009). pH Modulation of efflux pump activity of multi-drug resistant Escherichia coli: protection during its passage and eventual colonization of the colon. *PLoS One, 4*(8), e6656. doi:10.1371/journal.pone.0006656

Matthew S. Kelly, Michael G. Surette, Marek Smieja, Jeffrey M. Pernica, Laura Rossi, Kathy Luinstra, Andrew P. Steenhoff, Kristen A. Feemster, David M. Goldfarb, Tonya Arscott-Mills, Sefelani Boiditswe, Ikanyeng Rulaganyang, Charles Muthoga, Letang Gaofiwe, Tiny Mazhani, John F. Rawls, Coleen K. Cunningham, Samir S. Shah, & Seed, P. C. (2017). The Nasopharyngeal Microbiota of Children With Respiratory Infections in Botswana. *The Pediatric Infectious Disease Journal, 36*(9).

Mlacha, S. Z. K., Romero-Steiner, S., Hotopp, J. C. D., Kumar, N., Ishmael, N., Riley, D. R., Farooq, U., Creasy, T. H., Tallon, L. J., Liu, X., Goldsmith, C. S., Sampson, J., Carlone, G. M., Hollingshead, S. K., Scott, J. A. G., & Tettelin, H. (2013). Phenotypic, genomic, and transcriptional characterization of Streptococcus pneumoniae interacting with human pharyngeal cells. *BMC Genomics, 14*(383).

Neto, A. S., Lavado, P., Flores, P., Dias, R., Pessanha, M. A., Sousa, E., Palminha, J. M., Caniça, M., & Esperança-Pina, J. (2003). Risk Factors for the Nasopharyngeal Carriage of Respiratory Pathogens by Portuguese Children: Phenotype and Antimicrobial Susceptibility of Haemophilus influenzae and Streptococcus pneumoniae. *Microbial Drug Resistance, 9*(1), 99 - 108.

Novick, R. (1967). Properties of a cryptic high-frequency transducing phage in Staphylococcus aureus. *Virology, 33*, 155-166.

Novick, S., Shagan, M., Blau, K., Lifshitz, S., Givon-Lavi, N., Grossman, N., Bodner, L., Dagan, R., & Mizrachi Nebenzahl, Y. (2017). Adhesion and invasion of Streptococcus pneumoniae to primary and secondary respiratory epithelial cells. *Mol Med Rep, 15*(1), 65-74. doi:10.3892/mmr.2016.5996

Olaniyi, R., Pozzi, C., Grimaldi, L., & Bagnoli, F. (2017). Staphylococcus aureus-Associated Skin and Soft Tissue Infections: Anatomical

Localization, Epidemiology, Therapy and Potential Prophylaxis. *Curr Top Microbiol Immunol, 409*, 199-227. doi:10.1007/82_2016_32

Orihuela, C. J., Radin, J. N., Sublett, J. E., Gao, G., Kaushal, D., & Tuomanen, E. I. (2004). Microarray analysis of pneumococcal gene expression during invasive disease. *Infect Immun, 72*(10), 5582-5596. doi:10.1128/IAI.72.10.5582-5596.2004

Park, B., Nizet, V., & Liu, G. Y. (2008). Role of Staphylococcus aureus catalase in niche competition against Streptococcus pneumoniae. *J Bacteriol, 190*(7), 2275-2278. doi:10.1128/JB.00006-08

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pericone, C. D., Overweg, K., Hermans, P. W. M., & Weiser, J. N. (2000). Inhibitory and bactericidal effects of hydrogen peroxide production by Streptococcus pneumoniae on other inhabitants of the upper respiratory tract. *Infection and Immunity, 68*(7), 3990-3997.

Pettigrew, M. M., Gent, J. F., Revai, K., Patel, J. A., & Chonmaitree, T. (2008). Microbial interactions during upper respiratory tract infections. *Emerg Infect Dis, 14*(10), 1584-1591. doi:10.3201/eid1410.080119

Ratner, A. J., Lysenko, E. S., Paul, M. N., & Weiser, J. N. (2005). Synergistic proinflammatory responses induced by polymicrobial colonization of epithelial surfaces. *Proceedings of the National Academy of Sciences, 102*(9), 3429-3434.

Reddinger, R. M., Luke-Marshall, N. R., Sauberan, S. L., Hakansson, A. P., & Campagnaria, A. A. (2018). Streptococcus pneumoniae Modulates Staphylococcus aureus Biofilm Dispersion and the Transition from Colonization to Invasive Disease. *MBio, 9*(1), e02089-02017.

Regev-Yochay, G., Dagan, R., Raz, M., Carmeli, Y., Shainberg, B., Derazne, E., Rahav, G., & Rubinstein, E. (2004). Association Between Carriage of Streptococcus pneumoniae and Staphylococcus aureus in Children. *JAMA, 292*(6).

Regev-Yochay, G., Malley, R., Rubinstein, E., Raz, M., Dagan, R., & Lipsitch, M. (2008). In vitro bactericidal activity of Streptococcus pneumoniae and bactericidal susceptibility of Staphylococcus aureus strains isolated from cocolonized versus noncocolonized children. *J Clin Microbiol, 46*(2), 747-749. doi:10.1128/JCM.01781-07

Regev-Yochay, G., Trzcinski, K., Thompson, C. M., Malley, R., & Lipsitch, M. (2006). Interference between Streptococcus pneumoniae and Staphylococcus aureus: In vitro hydrogen peroxide-mediated killing by Streptococcus pneumoniae. *J Bacteriol, 188*(13), 4996-5001. doi:10.1128/JB.00317-06

Reiss-Mandel, A., & Regev-Yochay, G. (2016). Staphylococcus aureus and Streptococcus pneumoniae interaction and response to pneumococcal vaccination: Myth or reality? *Hum Vaccin Immunother, 12*(2), 351-357. doi:10.1080/21645515.2015.1081321

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nat Methods, 9*(7), 676-682. doi:10.1038/nmeth.2019

Shak, J. R., Vidal, J. E., & Klugman, K. P. (2013). Influence of bacterial interactions on pneumococcal colonization of the nasopharynx. *Trends Microbiol, 21*(3), 129-135. doi:10.1016/j.tim.2012.11.005

Siegel, S. J., & Weiser, J. N. (2015). Mechanisms of Bacterial Colonization of the Respiratory Tract. *Annu Rev Microbiol, 69*, 425-444. doi:10.1146/annurev-micro-091014-104209

Slager, J., Aprianto, R., & Veening, J. W. (2018). Deep genome annotation of the opportunistic human pathogen Streptococcus pneumoniae D39. *Nucleic Acids Res, 46*(19), 9971-9989. doi:10.1093/nar/gky725

Sorg, R. A., Kuipers, O. P., & Veening, J. W. (2015). Gene expression platform for synthetic biology in the human pathogen Streptococcus pneumoniae. *ACS Synth Biol, 4*(3), 228-239. doi:10.1021/sb500229s

Subramanian, K., Henriques-Normark, B., & Normark, S. (2019). Emerging concepts in the pathogenesis of the Streptococcus pneumoniae: From nasopharyngeal colonizer to intracellular pathogen. *Cell Microbiol, 21*(11), e13077. doi:10.1111/cmi.13077

Tonkin-Hill, G., Ling, C., Chaguza, C., Salter, S. J., Hinfonthong, P., Nikolaou, E., Tate, N., Pastusiak, A., Turner, C., Chewapreecha, C., Frost, S. D. W., Corander, J., Croucher, N. J., Turner, P., & Bentley, S. D. (2022). Pneumococcal within-host diversity during colonization, transmission and treatment. *Nature Microbiology*. doi:10.1038/s41564-022-01238-1

Turner, P., Hinds, J., Turner, C., Jankhot, A., Gould, K., Bentley, S. D., Nosten, F., & Goldblatt, D. (2011). Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray. *J Clin Microbiol, 49*(5), 1784-1789. doi:10.1128/JCM.00157-11

Valente, C., Dawid, S., Pinto, F. R., Hinds, J., Simoes, A. S., Gould, K. A., Mendes, L. A., de Lencastre, H., & Sa-Leao, R. (2016). The blp Locus of Streptococcus pneumoniae Plays a Limited Role in the Selection of Strains That Can Cocolonize the Human Nasopharynx. *Appl Environ Microbiol, 82*(17), 5206-5215. doi:10.1128/AEM.01048-16

van Belkum, A., Verkaik, N. J., de Vogel, C. P., Boelens, H. A., Verveer, J., Nouwen, J. L., Verbrugh, H. A., & Wertheim, H. F. (2009). Reclassification of Staphylococcus aureus nasal carriage types. *J Infect Dis, 199*(12), 1820-1826. doi:10.1086/599119

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods, 6*(10), 767-772. doi:10.1038/nmeth.1377

van Opijnen, T., & Camilli, A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res, 22*(12), 2541-2551. doi:10.1101/gr.137430.112

van Opijnen, T., & Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol, 11*(7), 435-442. doi:10.1038/nrmicro3033

van Opijnen, T., Lazinski, D. W., & Camilli, A. (2014). Genome-Wide Fitness and Genetic Interactions Determined by Tn-seq, a High-Throughput Massively Parallel Sequencing Method for Microorganisms. *Curr Protoc Mol Biol, 106*, 7 16 11-24. doi:10.1002/0471142727.mb0716s106

Wang, C. Y., Patel, N., Wholey, W. Y., & Dawid, S. (2018). ABC transporter content diversity in Streptococcus pneumoniae impacts competence

regulation and bacteriocin production. *Proc Natl Acad Sci U S A, 115*(25), E5776-E5785. doi:10.1073/pnas.1804668115

Weight, C. M., Venturini, C., Pojar, S., Jochems, S. P., Reine, J., Nikolaou, E., Solorzano, C., Noursadeghi, M., Brown, J. S., Ferreira, D. M., & Heyderman, R. S. (2019). Microinvasion by Streptococcus pneumoniae induces epithelial innate immunity during colonisation at the human mucosal surface. *Nat Commun, 10*(1), 3060. doi:10.1038/s41467-019-11005-2

Weiser, J. N. (2010). The pneumococcus: why a commensal misbehaves. *J Mol Med (Berl), 88*(2), 97-102. doi:10.1007/s00109-009-0557-x

Wholey, W. Y., Abu-Khdeir, M., Yu, E. A., Siddiqui, S., Esimai, O., & Dawid, S. (2019). Characterization of the Competitive Pneumocin Peptides of Streptococcus pneumoniae. *Front Cell Infect Microbiol, 9*, 55. doi:10.3389/fcimb.2019.00055

Yang, H. B., Hou, W. T., Cheng, M. T., Jiang, Y. L., Chen, Y., & Zhou, C. Z. (2018). Structure of a MacAB-like efflux pump from Streptococcus pneumoniae. *Nat Commun, 9*(1), 196. doi:10.1038/s41467-017-02741-4

Zhang, J. R., Idanpaan-Heikkila, I., Fischer, W., & Tuomanen, E. I. (1999). Pneumococcal licD2 gene is involved in phosphorylcholine metabolism. *Mol Microbiol, 31*(5), 1477-1488. doi:10.1046/j.1365-2958.1999.01291.x

Zhang, Y., Hong, Z., Zhou, L., Zhang, Z., Tang, T., Guo, E., Zheng, J., Wang, C., Dai, L., Si, T., & Wang, H. (2022). Biosynthesis of Gut-Microbiota-Derived Lantibiotics Reveals a Subgroup of S8 Family Proteases for Class III Leader Removal. *Angew Chem Int Ed Engl, 61*(6), e202114414. doi:10.1002/anie.202114414

This page is intentionally left blank

Chapter 6

General Discussion

The oldest question in genetics

Throughout this thesis we have approached one of the oldest standing topics in genetics: what are the minimal requirements for life. In the general introduction we described how the history of genetics is intertwined with mutagenesis analysis, and how mutants are used to not only map genetic interactions, but also to characterize phenotypic adaptations. 95 years after the official discovery of the first mutagen (Muller, 1927), such methodologies remain paramount. With the advent of NGS, mutagenesis-based techniques have exploded in use, having since then ushered in an era of high-throughput genetics. Such large scaling testing, however, generates what is commonly known as 'big data' (Dolinski & Troyanskaya, 2015). In chapters 2, 3 and 4, we described and developed new computational pipelines capable of efficiently mining such data, and of returning all the essential genes of any bacterium, under any condition, when Tn-Seq or CRISPRi-Seq methods are used.

In chapter 3, using Tn-seq, we determined the strain-level influence on both the core and pan essentialome of *Escherichia coli*, and how both essentialomes are defined by the conservation of essential functions, not by the conservation of gene homology *per se*. In chapter 5, we focused on a single strain, *Streptococcus pneumoniae* D39V, and leveraged CRISPRi-seq to demonstrate the impact that changing environments have on gene essentiality, and how essentiality arises and disappears in regards to specific transient stimuli, thus creating conditional essentiality.

A semantics issue: Are not all genes either essential, or conditionally essential?

Genes persist in a population if they display a net neutral or positive gain, that is, if they cause a cell lineage to survive equally well, or better, than its pears. Such relates with the current accepted definition of fitness, ultimately derived from Darwin: the ability to survive and reproduce in a given environment (Barker, 2009; Darwin, 1869).

Considering that an actively expressed, non-mobile, gene requires a certain amount of resources for processing (i.e. for transcription/translation), any originating gene activity must then be balanced out by the cost of acquiring said resources by the cell. In a situation where the used resources are higher than the acquired resources, a decreased cell fitness will ensue, and eventually force the cell/gene out of the

population if other cells are not incurring in similar, or worse, losses. In this case, the responsible gene would reduce the cell's relative growth potential (reflected as a slower growth in the prevalent environment), and ultimately cause the cell to be outcompeted.

If genes disappear from a population when their respective overall resource net gain is negative, such implies that actively expressed genes do not decrease overall bacterial resources in regards to a competitor. This non net negative concerted genetic action further implies that actively expressed, non-mobile genes are, under any one of the multitude of relevant natural encountered environments, essential in the way that they allow a cell to not be outcompeted (Fang *et al.*, 2005; Rancati *et al.*, 2018). All genes under these circumstances are thus biologically required (needed for the proliferation of individual cells), as a bacterium's overall competitive fitness is increased by their presence, allowing it to outcompete its pears and assure its own survival. With perhaps the exception of mobile or non-active genes, bacterial genes are thus either essential, if they assure bacterial reproduction in all environments, or conditionally essential, as they have been selected to ensure its host outcompetes others under either general or very specific conditions. Genes are thus either advantageous or deleterious under different environments. With this simplistic model, an always non-essential gene would thus either be existing in a transitory fading-out-of-existence state (frequency decreases within a population), or not existing at all.

Such genes, however, often persist within a population for longer than expected by mathematical modeling, suggesting that other forces are also at play. Negative frequency-dependent selection has been called the major driver behind this phenomenon. In this case, the relative fitness of a gene decreases as its frequency increases within a population, due to exactly its rareness. In bacteria, such can apply to genes related with resource competition/production, and/or resistance to invading elements (bacteriophages/plasmids). These circumstances would favor scenarios where genes that have become deleterious in a certain condition, would become advantageous again over time (Brisson, 2018; Kazancıoğlu *et al.*, 2014; Mitchell-Olds *et al.*, 2007).

Such observations highlight the dangers of extrapolating data from the artificial biological setups routinely used by biology laboratories, including the ones in this thesis, to a bacterium's natural environment. Indeed, in chapter 5 we observed how slightly different environments can dramatically change gene essentiality, with a gene

not essential in most tested conditions, *SPV 686/7/8* (renamed *arpABC*), becoming conditionally essential at pH 6 in the presence of *Staphylococcus aureus*.

A pooling issue: When the method is a condition by itself

Both Tn-seq and CRISPRi-seq methods used in this thesis rely on the existence of a pool of distinct mutants, whose relative frequency will change when submitted to any particular environment over the course of generations, according to the individual fitness of each strain. The biggest advantage of these methods is therefore the easy simultaneous evaluation of the relative fitness of each strain in a pooled population, within a single tube. However, when these pooled libraries of mutants are used, individual strains will not only display a fitness in regards to the test condition, but also to the presence of their pool competitors. For example, community factor producing strains can artificially enhance the fitness of non-producing strains. Concomitantly, phenomena such as frequency-dependent selection, bet-hedging, and labor division can also enable the persistence of sub fit strains within a test pool (Thibault *et al.*, 2019; Veening *et al.*, 2008). Despite such mechanisms being at play in both Tn-seq and CRISPRi-seq, due to the typically larger randomly obtained mutant pool obtained by Tn-seq, it is possible that more mutants would display this kind of behavior. It is feasible that this would result in increased noise upon essentiality determination, as deleterious transposon insertions would remain in the genome and be propagated over time within the population, skewering any insertion frequency dependent essentiality determination, especially when evaluating non-essential sub gene domains. The ramifications of such processes were explored in chapter 3, where a general Tn-seq analysis pipeline was developed.

In chapter 4, we adapted a pooled library arraying method known as SUDOKU (Anzai *et al.*, 2017; Baym *et al.*, 2016; Erlich *et al.*, 2009), and applied it to the UTI89 *Escherichia coli* Tn-seq library (chapter 3 and 4). We further developed this method to be able to operate with CRISPRi-seq libraries. Effectively, SUDOKU is then now able to array, from any pooled library size, either Tn-seq or CRISPRi-seq libraries. SUDOKU relies on the random picking of mutant strains from the pool (following pool plating), followed by their individual arraying. As both laboratorial complexity and cost rises with the increase in required arrayed mutants, it is advantageous to array libraries whose relative mutant frequency is as homogenous as possible as a skewed mutant

distribution will result in more repeated (unnecessary) mutants being picked. Smaller libraries are thus at an advantage in this regard, with our described D39V CRISPRi library with a single guide RNA per genetic feature (chapter 5) requiring comparatively less resources and time for arraying into characterized single occupancy wells than either transposon or multi guide CRISPRi libraries. The large-scale implementation of such high-throughput arraying methods are still in their infancy, but its application will yield invaluable insight into single gene mutant phenotypes in the absence of any compensatory community effects. Such methods can also further enhance the field of multi omics. Indeed, several –omics approaches can then be readily performed, without the need for library purification or temporary re-arraying every time a new experiment is attempted.

Pooled community effects have been recently clearly observed by Thibault *et al.*. By encapsulating individual Tn-seq *S. pneumoniae* generated mutants within microscopic agar droplets, the authors demonstrated a difference in the fitness of 1-3% of all mutants (Thibault *et al.*, 2019). An example effect was shown for the *nagA* and *nagB* genes, which display severe growth defect in droplet culture, but not in pooled culture, when AGP, a sugar, is added. Co-culture with wild type abolished all growth defects, as AGP could be metabolized into its sugar subunits, and then be released as a community metabolite.

In chapter 5, we approached *S. pneumoniae* interaction with *S. aureus*. Community effects might be of particular importance in these settings, as both species, especially the first, have been associated with bacteriocin and quorum-sensing molecules production (Dawid *et al.*, 2007; Kjos *et al.*, 2016; Kvich *et al.*, 2022; Potter *et al.*, 2014). Our top hit, ArpABC (SPV 686/7/8), is a general-purpose efflux pump capable of exporting several antibiotics and antimicrobial peptides (Yang *et al.*, 2018). Therefore, we hypothesized about the involvement of molecules active in the extracellular milieu, and how the true negative fitness of some strains might be masked by ‘community-friendly’ strains. We are thus currently developing droplet CRISPRi-seq, where each single droplet will carry a single CRISPRi mutant strain. The same competition assays between *S. pneumoniae* and *S. aureus* described in chapter 5 will then be, once again, performed, with *S. pneumoniae* being encapsulated with different CFU ratios of *S. aureus*. Any ‘cheater’ strains surviving in the pool collection due to a protective effect from other strains will then become apparent, with their phenotype further elucidating the nature of both species’ interactions. Moreover, our recent

discovery of pH influence on these strains (indicated by SPV 686/7/8 conditional essentiality) will also prompt us to attempt another CRISPRi-seq screen at pH 6, which we have so far not performed.

A systematic approach to a system's problem

Despite the advances described in the thesis, the nature of the complex interaction between *S. pneumoniae* and *S. aureus* is, to this day, still shrouded in uncertainty. Interestingly, a negative carriage correlation is observed for both species. Some reports have attributed such to H_2O_2 production by Sp, however, the physiological impact of H_2O_2 on this interaction is still unclear, with several works having either demonstrated or refuted this mechanism (Lijek & Weiser, 2012; Margolis, 2009; Pericone *et al.*, 2000; Gili Regev-Yochay *et al.*, 2004; G. Regev-Yochay *et al.*, 2008; G. Regev-Yochay *et al.*, 2006). Currently, the most prevailing explanation for this seems to be dependent on the host. Indeed, *S. pneumoniae* antibody has been described to cross-react with *S. aureus*, and proven sufficient to inhibit *S. aureus* nasal colonization (Lijek *et al.*, 2012; McNally *et al.*, 2006; Melles *et al.*, 2007). Despite these factors, co-colonization has been shown to still persist both *in vivo* and *in vitro*, with both species forming stable dual-species biofilms (Reddingera *et al.*, 2018). Similar conflicting scenarios have also been seen for *S. pneumoniae* interacting with other nasopharynx commensals, such as *Haemophilus influenzae* or *Moraxella catarrhalis* (Pericone *et al.*, 2000), or even between *S. aureus* and *Pseudomonas aeruginosa* (Kvich *et al.*, 2022). Such are only a few examples of a larger remark: bacteria are optimally adapted to handle an almost infinite combination of parameters, and small environmental variations result in distinct degrees of adaptations. Therefore, the only way to obtain a clear overall picture of how, or even why, bacteria (will) behave, is to obtain as much as possible different data, using as distinct as possible conditions.

Biologists have, since the beginning of genetics, mostly described how organisms adapt to different conditions, and what causes such adaptations. This rationale has changed little in 100 years, mostly due to technological and data constrictions. Such logic dramatically differs from that of classical physics, where the standard relies mostly on predicting and simulating physical systems, not on their description (Freddolino & Tavazoie, 2012). There is, however, a justification for such a difference: lack of systematic data. Unlike Physics' case, our understanding of life

has always been too limited and fragmented to attempt any accurate inference on how life adapts and interacts with both itself and the environment. In the past decade, however, biologists have not only started cataloging more types of cellular processes to an unprecedented detail, but are also doing so in thousands of distinct conditions (Dolinski & Troyanskaya, 2015; Krassowski *et al.*, 2020; Leshchiner *et al.*, 2022). In this thesis we mostly focused on the unidimensional characterization of genome-wide gene essentiality at the strain and environmental level (chapter 4 and 5). It would be interesting to further this description and collect data, under the same parameters, for the proteome, transcriptome, and metabolome. This assortment of data could then be used for a multi-omics approach, similar to ones already attempted for other organisms. For example, such multi-dimensional and abundant biological data are already available for some bacteria, and are currently creating one of the most demanding and exciting challenges the field has ever faced: how to integrate such distinct data types, and simulate bacteria. In effect, a few attempts at full cell simulations have so far been made using the simple organism *Mycoplasma genitalium*, with another model having been recently published for *Escherichia coli* (Freddolino & Tavazoie, 2012; Karr *et al.*, 2012; Macklin *et al.*, 2020). These simulations were performed by manually integrating and tweaking vast amounts of different biological data into classic mathematical models. Recent advances with Artificial Intelligence (AI), however, might now finally enable the escalation of these approaches into fully automated systems. AlphaFold2 is one of such examples, where new protein structures can be inferred from described structures (Jumper *et al.*, 2021). These examples demonstrate how a systematic collection of coherent complete datasets is the way to eventually instate system biology as the biological paradigm. It would then not be surprising for systems biology to finally catch up to the Physics rational of prediction, not just observation.

A question of function, not sequence

Even for the well characterized *E. coli* model organism, around 30% of all genes lack experimental functional evidence (Ghatak *et al.*, 2019). This figure gets worse considering that less than 1% of all currently predicted proteins across all domains of life have been experimentally validated, and 24% have a completely unknown gene function (Chang *et al.*, 2016). Such genes could harbor important metabolic functions,

possibly capable of functional replacement when known essential genes are absent, or by providing unknown support to known pathways (Shields & Jensen, 2019). Indeed, an attempt at synthetizing a minimal bacterial genome, JCVI-syn3.0, revealed that 149 out of the 473 genes had an unknown function, with the cell not being viable in their absence (Hutchison *et al.*, 2016).

In chapter 3 we reported a pan- and core-essentialome concomitant increase and decrease in size with the number of considered strain essentialomes. Gene functional replacement seems to, at least partly, explain this phenomenon (Coe *et al.*, 2019; Fang *et al.*, 2005; Narayanan *et al.*, 2017; Rosconi *et al.*, 2022; Rousset *et al.*, 2021; Shields & Jensen, 2019). Functional replacement can be the result of non-orthologous gene displacement, that is, how genes that are unrelated or paralogous can perform the same function in different organisms (Forterre, 1999; Koonin *et al.*, 1996). For example, *E. coli* and *Bacillus subtilis* have different non-homologous systems for DNA recombination, the RecBCD and AddAB/RexAB systems, respectively (Forterre, 1999; Karoui *et al.*, 1998). Similar situations can thus occur not only at the species level, but also in individual strains. Indeed, in their natural environment, cells are exposed to several mobile elements such as plasmids and/or viruses. The integration/excision of these elements, often able of infecting different lineages of organisms, can result in the carriage of extra non-self-DNA, and thus create plentiful chances for non-orthologous gene displacement.

We reported how genes related to translation and ribosome biogenesis/structure were the only ones significantly enriched in the 8-bacteria-wide core essentialome. Such could be due to several factors, albeit a mix of several is probably likely. Most non-orthologues genes are involved in metabolic functions (Dessimoz *et al.*, 2006), and thus, most of the non-homologs essential genes will not be in the core essentialome due to lack of homology, resulting in these categories being absent. Such a scenario gains strength when considering the pan-essentialome, where these same metabolic categories are enriched, due to, in this case, non-orthologs being considered. Experimentally, another factor would be library saturation, where some genes could be categorized as non-essential or, most likely, be too small to be significantly assessed, and thus removing them from the comparison pool. Considering the range of gene sizes, it is still unlikely this scenario would result in only one COG category being enriched. Another possibility is errors arising from the gene annotation prediction, and pan-genome standardization across all used strains.

Homologous, or partially homologous genes could be incorrected labelled as distinct. As we used standard benchmarked methods, such a scenario is also unlikely (Galardini *et al.*, 2017). The most likely possibility is related to these results originating from a true biological effect. Indeed, several cases of essential genes involved in DNA information processing (such as replication and transcription) have been attributed to also occur in viral/plasmid sequences. These elements, on occasion, can encode and/or require their own DNA processing proteins, thus submitting any genes in these functional categories to high mobility and evolutionary pressure (Forterre, 1999). Genes not carried or directly used by mobile elements, such as those involved in synthesizing proteins, like ribosomes, would thus be the functional category least submitted to non-orthologous gene displacement. This seems to be the observed case, possibly exacerbated by the mentioned experimental and data processing procedures. Expectedly, careful in-depth analysis of the tolerance needed for considering any gene the same as another gene based on homology would thus be required to validate any findings. Strain specific essentiality could also be validated using the CRISPRi system. Ultimately, applications of such strain level discrepancies can potentially be applied for species, or even strain, specific therapy.

Ecce, fortis novum mundum

Currently, biologists typically generate more data than can humanly be interpreted in regards to both time availability and complexity. In an effort to increase the throughput of such data analysis, computer science has been brought to the forefront of biology, prompting the appearance of the interdisciplinary field known as computational biology: the application of analytical processes in modeling biological systems. Bioinformatics is thus a subset of this latter, being mostly described as the application of informatics to the understanding of biological data (Gibas & Jambeck, 2001). In this thesis we dwelled into both fields, with 2FAST2Q being a bioinformatics tool, and TnSeeker closer to a computational biology pipeline.

In chapter 2 we described 2FAST2Q, a program capable of translating NGS raw data into an organized human readable format of feature counts. 2FAST2Q solved a recent inconvenience in bioinformatics, exacerbated by the rise in CRISPRi-seq usage, the non-existence of a single easy to use program capable of easily extracting, filtering, and counting features from .fastq files. In chapter 3, we developed TnSeeker,

a Tn-seq analysis pipeline. TnSeeker could be considered a computational biology method more than ‘just bioinformatics’ due to its modeling and predictive nature. Indeed, the inference of gene essentiality by use of a self-optimizing statistical thresholding based on recalling a gene gold set is an example of statistical bootstrapping and cross validation. Similar to AI based methods, such reliance on previous data, possibly significantly different from the one being analyzed, could also be argued to be one of the program’s biggest weaknesses. TnSeeker depends on the existence of not only homologous essential genes between the gold set and a new sample, which we have described to be limited, but also those genes must have similar names to the ones in the database. These issues, however, can be mitigated by using standardized annotation and pan-genome assembly methods (Galardini *et al.*, 2017). Such requires the careful annotation of any *de novo* sequenced organism, making sure all annotation is up to date. Considering the fast-paced environment of current biology, and the lack of standardizations regarding data analysis, especially concerning NGS and genbank annotations, it is likely that in the future these comparisons will be rendered obsolete. Future iterations of TnSeeker should therefore include homology-based comparisons for gene essentiality, where gold set essential and non-essential genes are compared directly by sequence, and not by database name. However, as described before, sequence comparison might also be a sub optimal mechanism, as, due to non-orthologous gene displacement, essentiality is primarily linked to function, not sequence. Gold set comparison could thus be performed based on functional category annotations. Such implementation would render TnSeeker a more powerful essentiality predictive tool, especially in regards to non-bacterial genomes, or organisms distant from the ones used for implementing the gold set.

In the –omics era, where several experimental hypotheses are already derived from ‘big data’, computational biology is starting to step in as the next big paradigm shift. As fields become more and more interdisciplinary, and as AI becomes more advanced and sips further into society, it is only a matter of time until AI driven data hypothesis derived from direct integration of all known biological databases becomes a lab standard. Such large-scale integration of data, beyond the comprehension of any human mind, would result in the pursuit of now unforeseeable biological questions. Indeed, AI interdisciplinary versatility has been recently demonstrated, with models developed for social platforms having been used for protein structure prediction (Lin

et al., 2022). Moreover, advances by OpenAI have shown AI can interact with humans, and even derive images, scientific texts, and hypothesis on demand. AI has become exceedingly good at explaining the bigger picture, however, the responsibility of what is worth pursuing or developing is still very human. For now.

In 100 years, we have progressed from describing natural mutations, to inducing precise alterations where needed. From observing phenotypes, to predicting them. From assuming proteins were the base of heredity, to storing digital information in DNA (Ceze *et al.*, 2019). It is not possible to know what the future will bring, however it is expected to be bewildering and unimaginable if the same knowledge growth rate continues.

[5]

References

Anzai, I. A., Shaket, L., Adesina, O., Baym, M., & Barstow, B. (2017). Rapid curation of gene disruption collections using Knockout Sudoku. *Nat Protoc, 12*(10), 2110-2137. doi:10.1038/nprot.2017.073

Barker, J. S. F. S. (2009). *Defining Fitness in Natural and Domesticated Population*. Dordrecht: Springer.

Baym, M., Shaket, L., Anzai, I. A., Adesina, O., & Barstow, B. (2016). Rapid construction of a whole-genome transposon insertion collection for Shewanella oneidensis by Knockout Sudoku. *Nat Commun, 7*, 13270. doi:10.1038/ncomms13270

Brisson, D. (2018). Negative Frequency-Dependent Selection Is Frequently Confounding. *Front Ecol Evol, 6*. doi:10.3389/fevo.2018.00010

Ceze, L., Nivala, J., & Strauss, K. (2019). Molecular digital data storage using DNA. *Nat Rev Genet, 20*(8), 456-466. doi:10.1038/s41576-019-0125-3

Chang, Y. C., Hu, Z., Rachlin, J., Anton, B. P., Kasif, S., Roberts, R. J., & Steffen, M. (2016). COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res, 44*(D1), D330-335. doi:10.1093/nar/gkv1324

Coe, K. A., Lee, W., Stone, M. C., Komazin-Meredith, G., Meredith, T. C., Grad, Y. H., & Walker, S. (2019). Multi-strain Tn-Seq reveals common daptomycin resistance determinants in Staphylococcus aureus. *PLoS Pathog, 15*(11), e1007862. doi:10.1371/journal.ppat.1007862

Darwin, C. R. (1869). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (5th ed.). London.

Dawid, S., Roche, A. M., & Weiser, J. N. (2007). The blp bacteriocins of Streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo. *Infect Immun, 75*(1), 443-451. doi:10.1128/IAI.01775-05

Dessimoz, C., Boeckmann, B., Roth, A. C., & Gonnet, G. H. (2006). Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res, 34*(11), 3309-3316. doi:10.1093/nar/gkl433

Dolinski, K., & Troyanskaya, O. G. (2015). Implications of Big Data for cell biology. *Mol Biol Cell, 26*(14), 2575-2578. doi:10.1091/mbc.E13-12-0756

Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M., & Hannon, G. J. (2009). DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res, 19*(7), 1243-1253. doi:10.1101/gr.092957.109

Fang, G., Rocha, E., & Danchin, A. (2005). How essential are nonessential genes? *Mol Biol Evol, 22*(11), 2147-2156. doi:10.1093/molbev/msi211

Forterre, P. (1999). Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Micro, 33*(3), 457-465.

Freddolino, P. L., & Tavazoie, S. (2012). The dawn of virtual cell biology. *Cell, 150*(2), 248-250. doi:10.1016/j.cell.2012.07.001

Galardini, M., Koumoutsi, A., Herrera-Dominguez, L., Cordero Varela, J. A., Telzerow, A., Wagih, O., Wartel, M., Clermont, O., Denamur, E., Typas, A., & Beltrao, P. (2017). Phenotype inference in an Escherichia coli strain panel. *Elife, 6*. doi:10.7554/eLife.31035

Ghatak, S., King, Z. A., Sastry, A., & Palsson, B. O. (2019). The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids*

Res, 47(5), 2446-2454. doi:10.1093/nar/gkz030

Gibas, C., & Jambeck, P. (2001). *Developing Bioinformatics Computer Skills: An Introduction to Software Tools for Biological Applications*: O'Reilly Media, Inc.

Hutchison, C. A., 3rd, Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z. Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., Glass, J. I., Merryman, C., Gibson, D. G., & Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science, 351*(6280), aad6253. doi:10.1126/science.aad6253

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583-589. doi:10.1038/s41586-021-03819-2

Karoui, M. E., Ehrlich, D., & Gruss, A. (1998). Identification of the lactococcal exonuclease/recombinase and its modulation by the putative Chi sequence. *Proc Natl Acad Sci, 95*(2), 626–631.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr., Assad-Garcia, N., Glass, J. I., & Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell, 150*(2), 389-401. doi:10.1016/j.cell.2012.05.044

Kazancıoğlu, E., Arnqvist, G., & Ebert, D. (2014). The maintenance of mitochondrial genetic variation by negative frequency-dependent selection. *Ecology Letters, 17*(1), 22-27. doi:10.1111/ele.12195

Kjos, M., Miller, E., Slager, J., Lake, F. B., Gericke, O., Roberts, I. S., Rozen, D. E., & Veening, J. W. (2016). Expression of Streptococcus pneumoniae Bacteriocins Is Induced by Antibiotics via Regulatory Interplay with the Competence System. *PLoS Pathog, 12*(2), e1005422. doi:10.1371/journal.ppat.1005422

Koonin, E. V., Mushegian, A. R., & Bork, P. (1996). Non-orthologous gene displacement. *Trends Genet, 12*(9), 334-336.

Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front Genet, 11*, 610798. doi:10.3389/fgene.2020.610798

Kvich, L., Crone, S., Christensen, M. H., Lima, R., Alhede, M., Alhede, M., Staerk, D., & Bjarnsholta, T. (2022). Investigation of the Mechanism and Chemistry Underlying Staphylococcus aureus' Ability to Inhibit Pseudomonas aeruginosa Growth In Vitro. *Journal of Bacteriology, 10.1128/jb.00174-22*.

Leshchiner, D., Rosconi, F., Sundaresh, B., Rudmann, E., Ramirez, L. M. N., Nishimoto, A. T., Wood, S. J., Jana, B., Bujan, N., Li, K., Gao, J., Frank, M., Reeve, S. M., Lee, R. E., Rock, C. O., Rosch, J. W., & van Opijnen, T. (2022). A genome-wide atlas of antibiotic susceptibility targets and pathways to tolerance. *Nat Commun, 13*(1), 3165. doi:10.1038/s41467-022-30967-4

Lijek, R. S., Luque, S. L., Liu, Q., Parker, D., Bae, T., & Weiser, J. N. (2012).

Protection from the acquisition of Staphylococcus aureus nasal carriage by cross-reactive antibody to a pneumococcal dehydrogenase. *Proc Natl Acad Sci U S A, 109*(34), 13823-13828. doi:10.1073/pnas.1208075109

Lijek, R. S., & Weiser, J. N. (2012). Co-infection subverts mucosal immunity in the upper respiratory tract. *Curr Opin Immunol, 24*(4), 417-423. doi:10.1016/j.coi.2012.05.005

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. doi:10.1101/2022.07.20.500902

Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., Sun, G., Agmon, E., DeFelice, M. M., Maayan, I., Lane, K., Spangler, R. K., Gillies, T. E., Paull, M. L., Akhter, S., Bray, S. R., Weaver, D. S., Keseler, I. M., Karp, P. D., Morrison, J. H., & Covert, M. W. (2020). Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science, 369*(6502). doi:10.1126/science.aav3751

Margolis, E. (2009). Hydrogen peroxide-mediated interference competition by Streptococcus pneumoniae has no significant effect on Staphylococcus aureus nasal colonization of neonatal rats. *J Bacteriol, 191*(2), 571-575. doi:10.1128/JB.00950-08

McNally, L. M., Jeena, P. M., Gajee, K., Sturm, A. W., Tomkins, A. M., Coovadia, H. M., & Goldblatt, D. (2006). Lack of association between the nasopharyngeal carriage of Streptococcus pneumoniae and Staphylococcus aureus in HIV-1-infected South African children. *J Infect. Dis., 194*(3), 385-390.

Melles, D. C., Bogaert, D., Gorkink, R. F. J., Peeters, J. K., Moorhouse, M. J., Ott, A., van Leeuwen, W. B., Simons, G., Verbrugh, H. A., Hermans, P. W. M., & van Belkum, A. (2007). Nasopharyngeal co-colonization with Staphylococcus aureus and Streptococcus pneumoniae in children is bacterial genotype independent. *Microbiology (Reading), 153*(Pt 3), 686-692. doi:10.1099/mic.0.2006/002279-0

Mitchell-Olds, T., Willis, J. H., & Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet, 8*(11), 845-856. doi:10.1038/nrg2207

Muller, H. J. (1927). Artificial transmutation of the gene. *Science, 66*(1699), 84-87.

Narayanan, A. M., Ramsey, M. M., Stacy, A., Whiteley, M., & Drake, H. L. (2017). Defining Genetic Fitness Determinants and Creating Genomic Resources for an Oral Pathogen. *American Society for Microbiology, 83*(14), 1098-5336. doi:10.1128/AEM

Pericone, C. D., Overweg, K., Hermans, P. W. M., & Weiser, J. N. (2000). Inhibitory and bactericidal effects of hydrogen peroxide production by Streptococcus pneumoniae on other inhabitants of the upper respiratory tract. *Infection and Immunity, 68*(7), 3990-3997.

Potter, A., Ceotto, H., Coelho, M. L. V., Guimaraes, A. J., & Bastos, M. (2014). The gene cluster of aureocyclicin 4185: the first cyclic bacteriocin of Staphylococcus aureus. *Microbiology (Reading), 160*(Pt 5), 917-928. doi:10.1099/mic.0.075689-0

Rancati, G., Moffat, J., Typas, A., & Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. *Nat Rev Genet, 19*(1), 34-49. doi:10.1038/nrg.2017.74

Reddingera, R. M., Luke-Marshalla, N. R., Sauberana, S. L., Hakanssom, A. P., & Campagnari, A. A. (2018). Streptococcus pneumoniae Modulates Staphylococcus aureus Biofilm Dispersion and the Transition from Colonization to Invasive Disease. *MBio, 9*(1).

Regev-Yochay, G., Dagan, R., Raz, M., Carmeli, Y., Shainberg, B., Derazne, E., Rahav, G., & Rubinstein, E. (2004). Association Between Carriage of Streptococcus pneumoniae and Staphylococcus aureus in Children. *JAMA, 292*(6).

Regev-Yochay, G., Malley, R., Rubinstein, E., Raz, M., Dagan, R., & Lipsitch, M. (2008). In vitro bactericidal activity of Streptococcus pneumoniae and bactericidal susceptibility of Staphylococcus aureus strains isolated from cocolonized versus noncocolonized children. *J Clin Microbiol, 46*(2), 747-749. doi:10.1128/JCM.01781-07

Regev-Yochay, G., Trzcinski, K., Thompson, C. M., Malley, R., & Lipsitch, M. (2006). Interference between Streptococcus pneumoniae and Staphylococcus aureus: In vitro hydrogen peroxide-mediated killing by Streptococcus pneumoniae. *J Bacteriol, 188*(13), 4996-5001. doi:10.1128/JB.00317-06

Rosconi, F., Rudmann, E., Li, J., Surujon, D., Anthony, J., Frank, M., Jones, D. S., Rock, C., Rosch, J. W., Johnston, C. D., & van Opijnen, T. (2022). A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat Microbiol*. doi:10.1038/s41564-022-01208-7

Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernandez-Rodriguez, J., Clermont, O., Denamur, E., Rocha, E. P. C., & Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the Escherichia coli species. *Nat Microbiol, 6*(3), 301-312. doi:10.1038/s41564-020-00839-y

Shields, R. C., & Jensen, P. A. (2019). The bare necessities: Uncovering essential and condition-critical genes with transposon sequencing. *Mol Oral Microbiol, 34*(2), 39-50. doi:10.1111/omi.12256

Thibault, D., Jensen, P. A., Wood, S., Qabar, C., Clark, S., Shainheit, M. G., Isberg, R. R., & van Opijnen, T. (2019). Droplet Tn-Seq combines microfluidics with Tn-Seq for identifying complex single-cell phenotypes. *Nat Commun, 10*(1), 5729. doi:10.1038/s41467-019-13719-9

Veening, J. W., Smits, W. K., & Kuipers, O. P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. *Annu Rev Microbiol, 62*, 193-210. doi:10.1146/annurev.micro.62.081307.163002

Yang, H. B., Hou, W. T., Cheng, M. T., Jiang, Y. L., Chen, Y., & Zhou, C. Z. (2018). Structure of a MacAB-like efflux pump from Streptococcus pneumoniae. *Nat Commun, 9*(1), 196. doi:10.1038/s41467-017-02741-4