



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2012

Comparative modular analysis of gene expression in vertebrate development

Barbara PIASECKA

Piasecka, 2012, Comparative modular analysis of gene expression in vertebrate development

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive.
<http://serval.unil.ch>

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'écologie et évolution

COMPARATIVE MODULAR ANALYSIS OF GENE EXPRESSION
IN VERTEBRATE DEVELOPMENT

Thèse de doctorat ès sciences de la vie (PhD)

présentée a la

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

par

Barbara PIASECKA

Master en Mathématiques de l'Université Adam Mickiewicz de Poznań, Pologne
Master en Biotechnologie de l'Université Adam Mickiewicz de Poznań, Pologne

Jury

Prof. Jan Roelof van der Meer, Président
Prof. Marc Robinson-Rechavi, Directeur de thèse
Prof. Sven Bergmann, co-Directeur de thèse
Prof. Alexandre Reymond, expert
Prof. Günter Wagner, expert

Lausanne, 2012

Imprimatur

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

Président	Monsieur Prof. Jan Roelof Van der Meer
Directeur de thèse	Monsieur Prof. Marc Robinson-Rechavi
Co-directeur de thèse	Monsieur Prof. Sven Bergmann
Experts	Monsieur Prof. Günter Wagner
	Monsieur Prof. Alexandre Reymond

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

Madame Barbara Piasecka

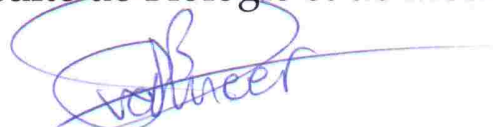
master of Science de "Adam Mickewicz University in Poznan", Pologne

intitulée

**COMPARATIVE MODULAR ANALYSIS OF GENE EXPRESSION
IN VERTEBRATE DEVELOPMENT**

Lausanne, le 9 novembre 2012

pour Le Doyen
de la Faculté de Biologie et de Médecine



Prof. Jan Roelof Van der Meer

Contents

Acknowledgements	viii
Abstract	x
Résumé de la thèse	xi
Introduction	1
1 Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human	14
1.1 Introduction	16
1.2 Methods	18
1.2.1 Gene expression data	18
1.2.2 Human–mouse orthologous genes	19
1.2.3 Normalization procedures	19
1.2.4 Pearson’s and Euclidean distances	20
1.2.5 Organ specificity of gene expression	20
1.2.6 τ -group composition	21
1.2.7 Randomization procedures	21
1.3 Results and Discussion	22
1.3.1 Correlation between Pearson’s and Euclidean distances depends on data normalization	22
1.3.2 Commonly used measures of gene expression similarity depend on the organ specificity of the genes	23
1.3.3 The rate of neutral expression evolution estimated with randomly permuted gene pairs depends on the organ specificity of the genes	24
1.3.4 A large fraction of broadly expressed genes leads to an underestimation of expression conservation	26
1.3.5 An alternative construction of random gene pairs improves the estimation of expression conservation	29
1.3.6 Results of the comparative study of human and mouse gene expression differ strongly according to the choice of randomization method	30
1.4 Conclusions	32

2 Comparative modular analysis of gene expression in vertebrate organs	38
2.1 Background	40
2.2 Results	41
2.2.1 The Ping-Pong Algorithm	41
2.2.2 Co-modules based on orthologous genes contain homologous organs	42
2.2.3 Co-modules based on homologous organs are organ- or system-specific	44
2.2.4 Genes belonging to co-modules are enriched in functions relevant to the corresponding organs	46
2.2.5 Organ-specific gene expression is often related to organ-specific hypomethylation of regulatory elements	46
2.2.6 Constraint on gene sequence is organ-specific	48
2.2.7 Genes' essentiality, duplicability, and age are weakly related to organ-specificity	49
2.2.8 Gene expression is conserved between mouse and human organs	50
2.2.9 Gene expression is conserved between mammalian and fish organs	51
2.3 Discussion	53
2.4 Conclusions	58
2.5 Methods	58
2.5.1 Gene expression data	58
2.5.2 Mapping probe sets to Ensembl genes	60
2.5.3 Mouse–human orthologous genes	60
2.5.4 Ping-Pong Algorithm	60
2.5.5 Post-processing of the PPA results	62
2.5.6 Enrichment analysis of hypomethylated regulatory regions	63
2.5.7 Gene sequence analysis	64
2.5.8 GO enrichment analysis	64
2.5.9 Zebrafish–mouse orthologous genes	65
2.5.10 Organ enrichment analysis	65
3 The hourglass and the early conservation models — co-existing evolutionary patterns in vertebrate development	68
3.1 Introduction	70
3.2 Results	73
3.2.1 Modules	73
3.2.2 Functional annotation	74
3.2.3 Sequence conservation	75
3.2.4 Gene age	76
3.2.5 Gene family size	77
3.2.6 Expression conservation	78
3.2.7 Regulatory regions	79

CONTENTS	VII
3.3 Discussion	83
3.4 Methods	86
3.4.1 Gene expression data	86
3.4.2 Mapping probe sets to Ensembl genes	87
3.4.3 Iterative Signature Algorithm (ISA)	87
3.4.4 GO enrichment analysis	88
3.4.5 Gene sequence analysis	89
3.4.6 Gene age analysis	90
3.4.7 Zebrafish–mouse orthologous genes	90
3.4.8 Gene expression conservation	91
3.4.9 Highly conserved non-coding elements	91
3.4.10 Transposon-free regions	92
3.4.11 Transcription factors	92
Outlook	96
Appendix A Supporting Information	101
Appendix B Poster	119
Appendix C Curriculum Vitae	121
Bibliography	124

Acknowledgements

It would not have been possible to complete this PhD without the support and guidance that I received from many people, to only some of whom it is possible to give particular mention here.

Above all, I would like to thank my principal supervisor, Prof. Marc Robinson-Rechavi, for our perfect collaboration during last four years. When needed he was always there - to discuss the science, to proofread the manuscripts (in amazingly short time), or to fight with editors. In the same time, he gave me a lot of independence, which made the work with him a very positive and fruitful experience. It is not easy when two stubborn people (like two of us) collaborate, luckily we have always engaged in discussions, and never in quarrels.

I would also like to thank my second supervisor, Prof. Sven Bergmann, for being a perfect complement of my first supervisor. While it was sometimes challenging to arrange a meeting with him, once we met it was always a great experience that involved digging into mathematical details of my work. More than once, it led me to a much deeper understanding of the general context of my research. Also, I would like to thank him for all encouraging words and support.

I'm very thankful to my thesis committee, Prof. Jan Roelof van der Meer, Prof. Alexandre Reymond, and Prof. Günter Wagner, for accepting to review this PhD dissertation, and for their interesting comments and advices during my private defense. Particularly, I would like to thank Prof. Günter Wagner, for accepting our invitation to Lausanne. His thoughtful remarks during our discussions allowed me to look at my work from a broader perspective.

I would like to thank all my colleagues, from both groups, who made these four years a great time: Julien, Fernando, Vidhya, Frédéric, Sacha, Marta, Nadja, Patricia, Hannes, Walid, Aurélie, Balazs, Patrick, Namrata, Joséphine, Sébastien, Anne, Romain, Sascha, Tim, Andrea, Aurélien, Armand, Nadya, Aitana, Diana, Karen, Zoltán, Tanguy, Rico, Micha, David, and Gábor. Thank you for sharing with me the experience of being a PhD student, and for the wise advices of my postdoc colleagues. Thank you for your help and encouragement. And, thank you for all beer outings, vodka parties, ski days, tobogganing days, and all other activities I

shared with you.

I would also like to thank all members of the Department of Ecology and Evolution, and especially Oksana, Eyal, Yannick, Anna, Charlotte, Maryam, Nicolas, Paweł, Tomasz, and Sylwester, as well as the members of SIB PhD training network. Thank you all for the joyful moments we shared together on very different occasions.

I'm very grateful to my parents, Krystyna and Franciszek, who have always given me an enormous support - both in private and scientific life. They have not only encouraged me to broaden my knowledge in any direction I wanted, but also supported me financially, as long as I needed it. Without their help I would never go to Poznań, Montpellier, Lyon, Lausanne, and I would never have the opportunity to write these words. My dear parents, I thank you for everything I got from you, which is impossible to list here, but the most I thank you for always letting me make my own decisions, and for respecting them.

I'm also grateful to my two brothers, Dominik and Marek, without whom I would not be the same person. Thank you for being so supportive in really important moments of my life, and so light-hearted in all the others. Special thank to Dominik for all our hiking and hitch-hiking trips. That was unforgettable.

The most grateful I'm to my husband, Paweł. Since I know him, he has never ceased to amaze me. I truly can't imagine the last four years without him by my side. He has been a great supporter and has unconditionally loved me during my good and bad times. He has never allowed the problems to overwhelm me, and he has always given me a helping hand. He was the first and the most critical reviewer of my work, which helped a lot to shape my thesis. Every day I have spent with him has been a great intellectual challenge, and I have been constantly learning new things. There are simply not enough words to express how much I owe him.

Abstract

The focus of my PhD research was the concept of modularity. In the last 15 years, modularity has become a classic term in different fields of biology. On the conceptual level, a module is a set of interacting elements that remain mostly independent from the elements outside of the module.

I used modular analysis techniques to study gene expression evolution in vertebrates. In particular, I identified “natural” modules of gene expression in mouse and human, and I showed that expression of organ-specific and system-specific genes tends to be conserved between such distance vertebrates as mammals and fishes.

Also with a modular approach, I studied patterns of developmental constraints on transcriptome evolution. I showed that none of the two commonly accepted models of the evolution of embryonic development (“evo-devo”) are exclusively valid. In particular, I found that the conservation of the sequences of regulatory regions is highest during mid-development of zebrafish, and thus it supports the “hourglass model”. In contrast, events of gene duplication and new gene introduction are most rare for genes expressed during early development, which supports the “early conservation model”.

In addition to the biological insights on transcriptome evolution, I have also discussed in detail the advantages of modular approaches in large-scale data analysis. Moreover, I re-analyzed several studies (published in high-ranking journals), and showed that their conclusions do not hold out under a detailed analysis. This demonstrates that complex analysis of high-throughput data requires a cooperation between biologists, bioinformaticians, and statisticians.

Résumé de la thèse

Le concept de modularité était au centre d'intérêt de ma thèse de doctorat. Au cours des 15 dernières années, la modularité est devenue un terme classique dans les différents domaines de la biologie. Sur le plan conceptuel, un module est un ensemble d'éléments qui interagissent entre eux, et en même temps restent indépendants des éléments extérieurs au module.

J'ai utilisé des techniques d'analyse modulaire pour étudier l'évolution de l'expression des gènes chez les vertébrés. En particulier, j'ai identifié les modules "naturelles" de l'expression des gènes chez la souris et l'homme, et j'ai montré que l'expression de gènes organe-spécifiques et système-spécifiques est conservée entre les vertébrés aussi distants que les mammifères et les poissons.

Egalement avec une approche modulaire, j'ai étudié les modèles de contraintes de développement sur l'évolution du transcriptome. J'ai montré qu'aucun des deux modèles de l'évolution du développement embryonnaire ("évo-dévo") qui sont communément admis n'est exclusivement valable. En particulier, j'ai trouvé que la conservation des séquences des régions de régulation est la plus élevée au cours du milieu du développement du poisson zèbre, et donc il suit le "hourglass model". Au contraire, les événements de duplication des gènes et de l'introduction de gènes nouveaux sont les plus rares pour les gènes exprimés au début du développement, qui soutient le "early conservation model".

En plus des aperçus biologiques sur l'évolution du transcriptome, j'ai également examiné en détail les avantages des approches modulaires de l'analyse des données à grande échelle. De plus, j'ai ré-analysé plusieurs études (publiées dans des

journaux de haut rang), et j'ai montré que leurs conclusions ne résistent pas à une analyse détaillée. Cela démontre que l'analyse complexe de données à grande échelle nécessite une coopération entre biologistes, bio-informaticiens et statisticiens.

Introduction

It is unusual for a single idea to underpin such distant fields as biological sciences, enterprise management, or the toy industry. Modularity is exactly this kind of concept. Its meaning varies slightly depending on the context, but in general a module is a part of the system that can be easily recombined with another module to create a novel structure. One of the most illustrative example of modules are the popular LEGO bricks (<http://www.lego.com>). The number of forms which we can create with them is limited only by our imagination (and by the number of bricks parents are willing to buy). The modular structure is also becoming increasingly popular in the organization of enterprises. Especially in the computer and apparel industries where integrated hierarchical organizations are replaced by loosely coupled organizational forms. This allows for the flexible recombination of organizational components into different configurations (Schilling and Steensma, 2001). While these and several other examples of modularity in product design (Gershenson and Prasad, 1997), software design (Gamma, 1995), or even art (Jablan, 2002), are human-derived, the most sophisticated modules were created during the evolution of living organisms. Nevertheless, it is only in the last 15 years that the concept of modularity has attracted the attention of biologists from different fields, including molecular biology, systems biology, evolutionary developmental biology, or even cognitive psychology (Wagner *et al.*, 2007).

From a biological point of view, a module is a set of elements that interact preferentially between each other, and remain largely independent from the elements outside of the module. It can be viewed as a semi-autonomous entity that evolves,

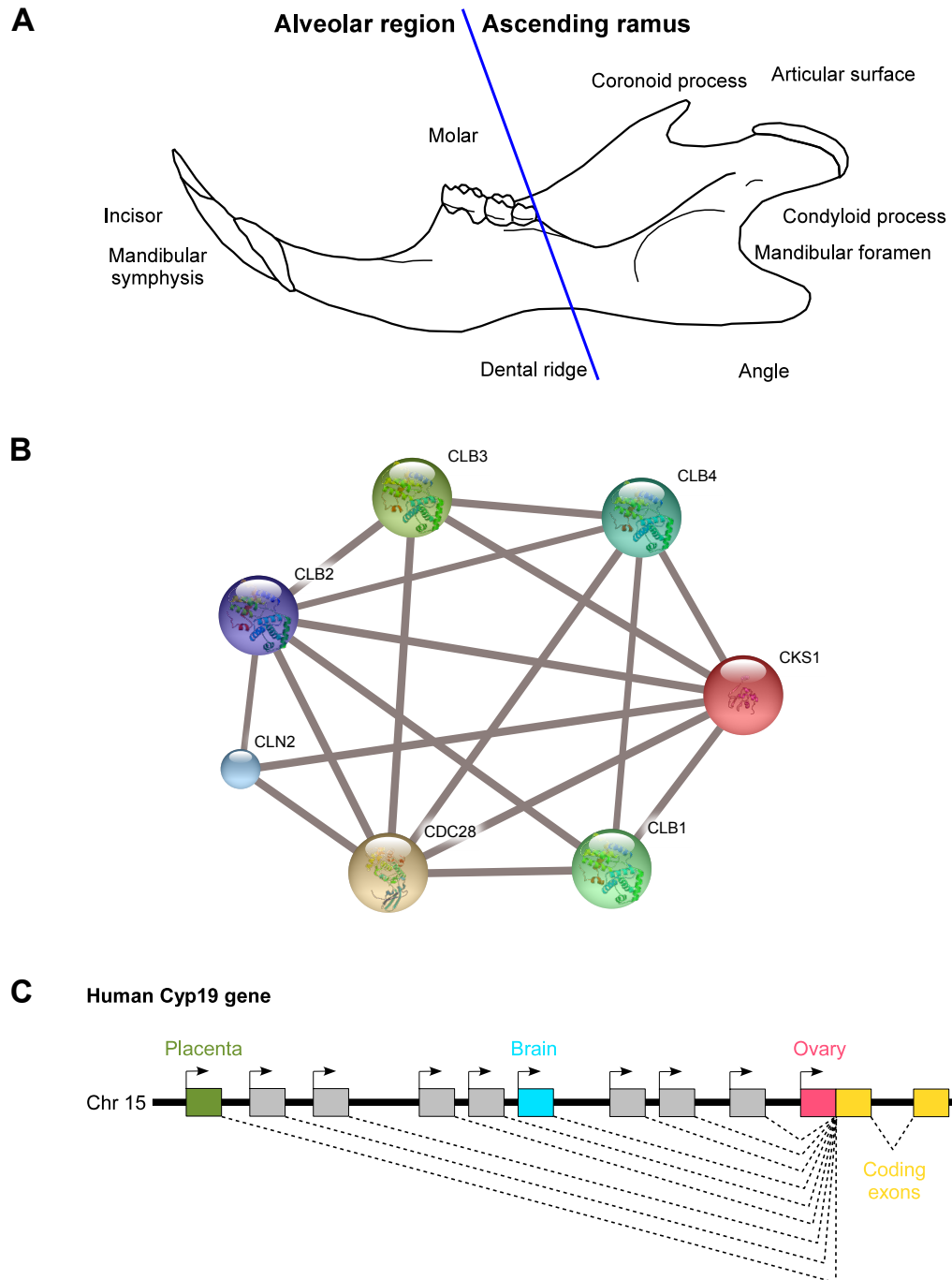


Figure 1: Examples of modules. (A) Mouse mandible consists of two variational modules — the alveolar part and the ascending ramus (separated by the blue line). Figure adapted from [Cook *et al.* \(1965\)](#). (B) The interaction network of cyclins and cyclin-dependent kinases is a functional module responsible for yeast cell cycle progression. Figure generated with STRING version 9 ([Jensen *et al.*, 2009](#)). (C) *cis*-regulatory region of *Cyp19* gene in which each promoter is a module driving tissue-specific expression. Figure adapted from [Rawn and Cross \(2008\)](#).

functions or participates in given processes relatively independently from other modules (Espinosa-Soto and Wagner, 2010). A more precise definition of modularity depends on the specific organismal level to which it applies. Several kinds of modules are commonly recognized in biology, e.g. variational, functional, and developmental (Wagner *et al.*, 2007). A variational module is a set of phenotypic features that vary together due to the pleiotropic effects of the genes involved in their regulation, but remain independent of other such features due to the lack of pleiotropic effects between them (Schlosser and Wagner, 2004). The two main parts of the mouse mandible — the ascending ramus and the alveolar region — are good examples of variational modules (figure 1A). Cheverud *et al.* (1997) has shown that most of the QTLs influencing mandibular morphology affect either the components of the ascending ramus or the components of the alveolus. A functional module is a discrete unit that performs a biological function which is relatively independent from the function of other modules (Hartwell *et al.*, 1999). For example, a ribosome is a module responsible for the synthesis of proteins, and the interaction network of cyclins and cyclin-dependent kinases (figure 1B) is a module responsible for yeast cell-cycle progression (Spirin and Mirny, 2003). A developmental module is a part of an embryo that is autonomous in its differentiation process (Schlosser and Wagner, 2004). It means that it can develop also outside its normal context, e.g., in a different body location or even outside the body in a tissue culture. The insect compound eye is a well known example of developmental module. Its development has been induced on such different structures as wings, legs and antennae of *Drosophila* (Halder *et al.*, 1995). Importantly, the different kinds of modules described above are not necessarily exclusive. For example, the two parts of the mouse mandible are both variational and functional modules; the alveolus is a teeth-bearing part and the ascending ramus is a muscle attachment site, which

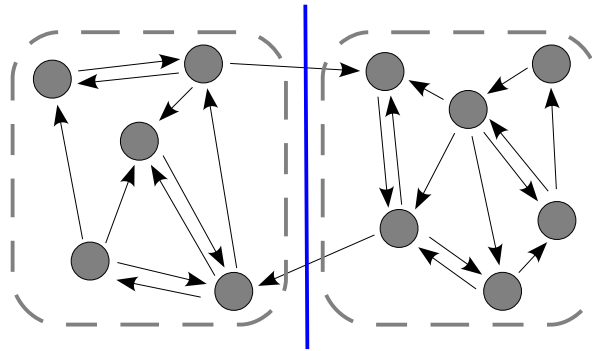


Figure 2: Integration of modules. Number of interactions between elements within the modules is much higher, than the number of interactions between elements from different modules. Dashed rectangles denote two modules. Figure adapted from [Klingenberg \(2008\)](#).

also articulates with the skull.

Modularity has been long ago recognized as an important feature for organismal evolvability ([Needham, 1933](#); [Gould, 1977](#); [Raff *et al.*, 1991](#); [Bonner, 1988](#); [Wagner and Altenberg, 1996](#)). Due to the semi-independence of the modules, the changes inside one module do not perturb the function of the other modules, and thus modularity facilitates evolutionary processes. For example, *cis*-regulatory regions show a highly modular structure, thanks to which adding a new promoter does not disrupt gene expression but adds new tissue-specific function (e.g., in the primate *Cyp19* gene, which encodes for a key enzyme of estrogen biosynthesis, has evolved a placenta-specific promoter; [Bulun *et al.*, 2004](#) [figure 1C]). Also, existing modules can be re-deployed to perform new functions. Ancestral regulatory networks are often re-used in a new context. For example, the gene network in which the hedgehog protein induces the engrailed protein to pattern the insect wing is also used to determine the localization of eyespots in butterflies wings ([Keys *et al.*, 1999](#))

Depending on the level at which one studies modularity, and on the accessible data, different methods are applied to detect the modules. Nevertheless, the final result can always be pictured as shown on figure 2. Whatever the elements and

the interactions are, the integration is always much stronger within than between the modules. Here, only three examples of biological data and the methodologies used for studying modularity will be discussed. First, one can analyze molecular networks, such as protein–protein interaction networks or metabolic networks. One of the common approaches to search for modular structures in networks is based on the [Girvan and Newman \(2002\)](#) concept of edges betweenness, that allows to distinguish inter- and intra-module edges, and thus divide the network into modules. Briefly, one needs to compute all-against-all shortest paths of a network and calculate the number of times each edge is traveled. The assumption is that inter-modules edges are more often on some shortest path than intra-modules edges. Second, one can analyze morphological data in order to detect variational modules. High covariation among morphological traits allows to infer pleiotropic effects of the genes and to delimit the modules. To this end all-against-all correlation coefficients of the traits are calculated, and the sets of highly correlated traits define the modules. Third, one can analyze gene expression data to identify regulatory modules, i.e., sets of co-regulated genes that share a common function. These modules are expected to consist of genes with coherent patterns of expression. The most common approach to detect regulatory modules is based on clustering methods that group together the genes with similar expression patterns. This type of modules, their detection methods, and their study in an evolutionary context are the focus of my work.

When I started my PhD, in 2008, microarray technology was in its heyday. It allowed, for the first time ever, to study the function of thousands of genes simultaneously. But also for the first time ever it faced biologists with the challenges of analyzing such complex and nonindependent data sets. Consequently the help of statisticians and bioinformaticians became inevitable in molecular biological

sciences. Having my degree both in biotechnology and mathematics I found it exciting to contribute to the field.

The size of microarray data itself suggested the use of modular analysis. Dividing the large-scale data into modules consisting of similarly expressed genes had two important advantages. First, studying a limited number of modules, instead of thousands of genes separately, simply made the analysis more feasible. Second, the measure of expression level is more robust when averaged within the module, than when considered separately for each gene, because fluctuations tend to cancel each other out.

Initially, the most popular algorithms to partition gene expression data into modules were hierarchical clustering and k -means clustering. In hierarchical clustering, every gene starts in its own cluster, and the two most similar clusters (according to the selected distance metric) are merged. The process is repeated until a single cluster remains. As a result, the data are arranged in a tree structure that can be divided into the desired number of clusters by cutting along the tree at a given height. In k -means clustering, one needs to first specify the desired number of clusters, k , and start with k data points (centroids of an initial set of clusters) chosen either randomly or arbitrary. Then, all samples are partitioned into the k clusters based on the selected distance metric. Next, the centroids are adjusted to represent the new clusters' center points and the partition of genes is repeated. The procedure stops when the assignment of the genes to the clusters no longer changes. Both hierarchical and k -means clustering methods were successfully used to detect expression modules in very different contexts, e.g., functional clusters of genes in the time course data of different processes in yeast (Eisen *et al.*, 1998); groups of genes differentially regulated in human tumor tissues (Alon *et al.*, 1999; Perou *et al.*, 2000; Bittner *et al.*, 2000); or transcriptional regulatory sub-networks

in yeast (Tavazoie *et al.*, 1999).

When the scale of the conditions under which gene expression was analyzed changed from tens to hundreds the two clustering methods became of limited use. The reason was twofold. First, standard clustering methods assign each gene to a single cluster, while it is well known that thanks to the modular structure of *cis*-regulatory elements a gene can perform more than one function in the organism (e.g., Yuh *et al.*, 1998; Pilpel *et al.*, 2001). Thus, it would be desirable to allow for partial overlap between the identified modules. Second, in standard clustering methods the distance between genes (typically, Pearson's correlation coefficient or Euclidean distance) is calculated based on their expression across all experimental conditions. This is problematic, because many genes are expressed only in a limited number of conditions, and thus taking into account also the irrelevant conditions introduces unwanted noise. This can hamper the identification of genes co-regulated over small subsets of conditions.

The first method that took into consideration the two limitations of standard clustering methods was the biclustering algorithm of Cheng and Church (2000). The concept of bicluster introduced by the authors corresponded to a subset of genes and a subset of conditions with a high similarity score (e.g., low mean squared residue of bicluster elements). The algorithm was based on deletion and addition of genes and conditions in order to iteratively improve the score of biclusters. Biclusters which were discovered were masked to allow the detection of other clusters in the next runs. Other algorithms that clustered simultaneously the genes and their conditions of expression are briefly described in Ihmels and Bergmann (2004), e.g., Coupled Two-Way Clustering (Getz *et al.*, 2000), SAMBA (Tanay *et al.*, 2002), Fuzzy *k*-means (Gasch and Eisen, 2002), etc.

In 2003, Bergmann *et al.* proposed a new algorithm for large-scale data analysis,

the Iterative Signature Algorithm (ISA). It aimed at discovering “transcription modules” in gene expression data. A transcription module consisted of a set of co-regulated genes, and the set of their regulating conditions. Thus, the module genes show the most coherent expression patterns under the module conditions. And vice versa, the module conditions are those that induce the most similar expression of the genes in the module. The criterion for a gene (resp. condition) to be assigned to a module is to have a gene (condition) score beyond the gene (condition) threshold (figure 3A). A range of thresholds can be applied in a single ISA run, which decomposes the data into the modules at different resolutions. The higher the thresholds used, the smaller the modules obtained. Typically, the algorithm starts from a set of randomly selected genes and through the iterations it refines the genes and conditions until they match the definition of transcription module. If the number of initial sets is large enough, all modules corresponding to the given pair of thresholds can be recovered. An important advantage of the ISA is its computation time that scales only linearly with the number of genes multiplied by the number of conditions.

While the ISA was suitable to study the structure of a single data set at a time, it soon became desirable to deal with more than one large-scale data set in the studies of cellular phenotypes. For example, expression profiles of genes (Staunton *et al.*, 2001), proteins (Shankavaram *et al.*, 2007), and microRNAs (Gaur *et al.*, 2007) were measured for 60 human cancer cell lines (NCI-60), along with drug response profiles (Scherf *et al.*, 2000). Thus, it was very tempting to integrate these different kinds of high-throughput data, and thus shed light on their interconnections on the molecular level. Initially, algorithms that aimed at integrating different phenotypic data did it in a sequential manner. For instance, groups of genes assigned to a cluster were tested for enrichment in genes from other

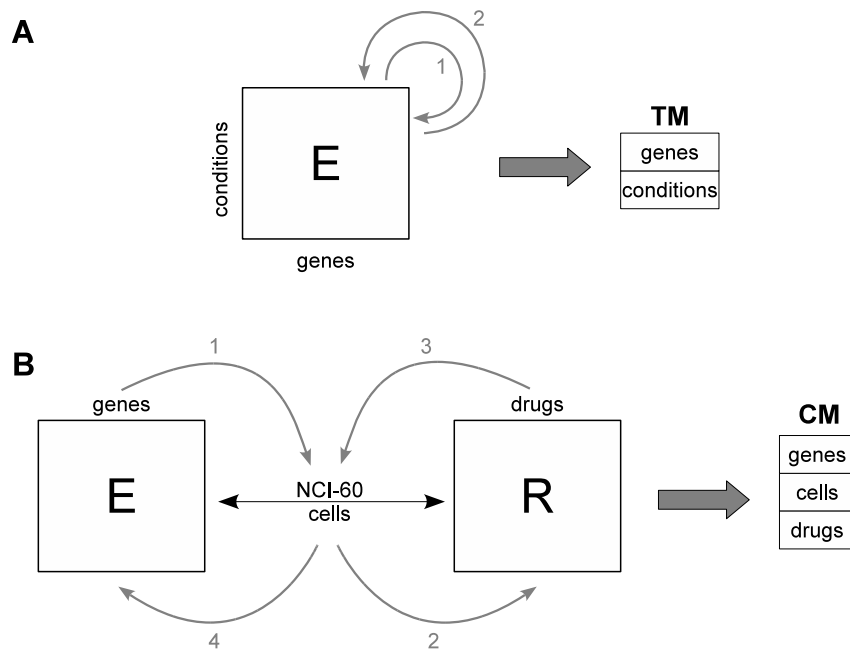


Figure 3: Schematic representation of two modular algorithms. (A) The Iterative Signature Algorithms discovers transcription modules (TM) by iteratively refining: (1) conditions in which genes are expressed, and (2) genes which are expressed in conditions. (B) The Ping-Pong Algorithm discovers co-modules (CM) by iteratively refining: (1) cells in which genes are expressed, (2) drugs which affect cells, (3) cells responsive to drugs, and (4) genes that are expressed in cells. E — expression data, R — drug response data.

predefined groups (such as those belonging to a cluster of a different data set). In 2008, [Kutalik *et al.*](#) proposed an extension of the ISA, a Ping-Pong Algorithm (PPA), which allowed the simultaneous analysis of two large-scale data sets that share one common dimension. Originally, it was used to combine the gene expression data and drug-response data for NCI-60 cell lines. The PPA partitioned the data into “co-modules” which consisted of genes, drugs and cell lines, such that the genes were expressed in a similar way and the drugs induced similar response these cell lines. Analogously to the ISA implementation, the PPA iteration also starts from a random set of genes and uses thresholds when assigning genes, drugs and cells to a co-module (figure 3B). Importantly (for my work), the PPA can be applied to any large-scale data sets sharing one common dimension.

An important question in modern biology is whether gene expression is con-

served through evolution between species (e.g., [Yanai *et al.*, 2004](#); [Khaltovich *et al.*, 2005](#); [Liao and Zhang, 2006a](#); [Zheng-Bradley *et al.*, 2010](#)). There are two essential methodological issues with the majority of studies aiming to answer this question. First, they use the Pearson's correlation coefficient or the Euclidean distance to measure the similarity between gene expression patterns of two species. As I discuss in [chapter 1](#), both metrics depend on the gene expression specificity, i.e., if one compares a pair of broadly expressed genes vs. a pair of specifically expressed genes, both pairs highly conserved, one pair will seem to be more conserved than the other. Moreover, the results will be opposite for the Pearson's correlation coefficient and for the Euclidean distance. According to the Pearson's correlation coefficient, specifically expressed gene pair will seem more conserved, and the contrary will be true for the Euclidean distance. Second, in order to assess whether the expression has been conserved by selection one needs to refer to an expectation for expression similarity under neutral evolution. For species that have diverged for a long time, and evolved under no selective pressure, no similarity in expression is expected to remain. [Jordan *et al.* \(2005\)](#) suggested that such large neutral divergence could be well approximated by calculating the distance between expression profiles of genes with permuted orthology relations between them. In [chapter 1](#) I show that this approach is of very limited use when broadly expressed genes are abundant in the data, which is a common case. Thus, I propose a novel randomization procedure that is not biased by overrepresentation of any expression profiles in the data set. Still, as discussed in [section 1.4](#), the method is not free from some drawbacks.

In a further stage of my PhD research I abandoned the standard, but problematic approach, and used only the modular approach to study transcriptome evolution. At that time, a modular approach to study gene expression evolution was rather innovative. Only few studies in the field used modular approach,

e.g., [Oldham *et al.* \(2006\)](#) analyzed modules of co-expressed genes in discrete brain regions of human and chimpanzee, [Yang and Su \(2010\)](#) analyzed tissue-related co-expression modules in mouse and human, and [Cai *et al.* \(2010\)](#) analyzed co-expression modules of mouse and human stem cells. Although the modular approach was not yet well established in the studies of gene expression evolution, it led my research to very interesting findings.

In chapter 2, I present the results of my study on organ-specific and system-specific genes and their expression conservation between vertebrates. Several studies already reported some evidence for conservation of gene expression between homologous organs of vertebrates ([Liao and Zhang, 2006a](#); [Zheng-Bradley *et al.*, 2010](#); [McCall *et al.*, 2011](#)). However, they computed organs' similarity using Pearson's or Euclidean distances that capture only global similarity across samples. Specifically, these measures do not allow for detection of between-species units of conservation, i.e., modules of organs and their specific genes that have remained largely unchanged since the speciation event. To overcome this limitation, I used the PPA to analyze mouse and human gene expression data. In this particular case, the PPA requirement for data having a common dimension was fulfilled twice, i.e., through one-to-one orthologous genes, and through homologous organs. Thus, the resulting co-modules consisted of orthologous genes and the mouse and human organs in which these genes were overexpressed; or they consisted of sets of homologous organs and sets of mouse and human genes with coherent overexpression in these organs. In the PPA run with genes on the common dimension, I recovered the information of organ homology based only on orthologous genes expression patterns. In the PPA run with organs on the common dimension, I found organs grouped into homologous systems (between mouse and human), and their functional genes in both species. These genes were often orthologous between

species, i.e., with expression conserved through evolution. I also found that genes with expression conserved between mammals have their orthologs expressed in the corresponding homologous organs in zebrafish, and thus are conserved within vertebrates. In conclusion, I found conserved modularity of gene expression in vertebrates that is clearly related to anatomical modularity.

In chapter 3, I present the results of my study on patterns of developmental constraints acting on vertebrate evolution. Two main hypotheses of the evolution of embryonic development have been put forward so far. First, the early conservation model predicts that the highest conservation occurs at the beginning of embryogenesis (von Baer, 1828). Second, the hourglass model predicts that the highest conservation can be found during mid-embryogenesis (Duboule, 1994; Raff, 1996). In recently published studies the hourglass model has been favored (Domazet-Lošo and Tautz, 2010; Irie and Kuratani, 2011). Usually, authors have compared descriptive statistics of all genes across all developmental time points. Such an approach introduces dependencies between the sets of compared genes, and may lead to results biased by constantly expressed genes. To overcome this limitation, I used the ISA to study the evolution of zebrafish development. I identified modules of genes co-expressed specifically in consecutive stages of zebrafish development. Next, I performed a detailed comparison of several gene properties between modules. I detected the hourglass pattern only at the regulatory level, where sequences of regulatory regions were most conserved for genes expressed in mid-development. In contrast to some previous studies, I did not detect the hourglass model at the level of gene sequence, gene age or gene expression. Gene duplication and birth were most rare in early development, supporting the early conservation model. Finally, all gene properties displayed the least conservation in late development and adult, consistent with both models of developmental constraints. Overall,

different levels of molecular evolution follow different patterns of developmental constraints, and thus neither the early conservation nor the hourglass model seems exclusively valid.

An important part of chapter 3 (in my humble opinion) is a detailed discussion of work of Domazet-Lošo and Tautz (2010) who showed that the age of the transcriptome expressed over zebrafish development reflects the hourglass pattern. The existence of the hourglass model has been debated in the evo-devo community over last 25 years. And any published evidence supporting this model was always welcome with enthusiasm by the community. In 2010, the hourglass problem has attracted considerable attention even outside the evo-devo field thanks to two landmark papers in Nature (Domazet-Lošo and Tautz, 2010; Kalinka *et al.*, 2010) that were granted a cover page. The paper of Domazet-Lošo and Tautz was widely commented also on general public venues, such as the Panda's Thumb blog (<http://pandasthumb.org/archives/2010/12/its-just-a-stag-2.html>). The comments mostly referred to the finally discovered proof of the hourglass model. It remained unnoticed that the methodology applied in this work was far from the standard and widely accepted statistical methods for microarray data analysis. In chapter 3, I show that after a detailed analysis of the data from Domazet-Lošo and Tautz (2010), the authors' conclusion does not hold out.

If I had to summarize my four years experience of being a graduate student, I would say it reminded me of looking for a needle in a haystack. Moreover, I could never be sure if I faced the right haystack. Nevertheless, I hope that the reading of this dissertation will convince the reader that my work was not that hopeless as it seems from purely probabilistic point of view.

1

Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human

Barbara Piasecka, Marc Robinson-Rechavi, Sven Bergmann

Abstract

Comparative analyses of gene expression data from different species have become an important component of the study of molecular evolution. Thus methods are needed to estimate evolutionary distances between expression profiles, as well as a neutral reference to estimate selective pressure. Divergence between expression profiles of homologous genes is often calculated with Pearson's or Euclidean distance. Neutral divergence is usually inferred from randomized data. Despite being widely used, neither of these two steps has been well studied. Here, we analyze these methods formally and on real data, highlight their limitations, and propose improvements.

It has been demonstrated that Pearson's distance, in contrast to Euclidean distance, leads to underestimation of the expression similarity between homologous genes with a conserved uniform pattern of expression. Here, we first extend this study to genes with conserved, but specific pattern of expression. Surprisingly, we

find that both Pearson's and Euclidean distances used as a measure of expression similarity between genes depend on the expression specificity of those genes. We also show that the Euclidean distance depends strongly on data normalization. Next, we show that the randomization procedure that is widely used to estimate the rate of neutral evolution is biased when broadly expressed genes are abundant in the data. To overcome this problem, we propose a novel randomization procedure that is unbiased with respect to expression profiles present in the data sets. Applying our method to the mouse and human gene expression data suggests significant gene expression conservation between these species.

This article was published in *Bioinformatics* (2012) **28** (14): 1865–1872.

doi: [10.1093/bioinformatics/bts266](https://doi.org/10.1093/bioinformatics/bts266)

1.1 Introduction

Changes in gene expression have been suggested to underlie many differences in gene function or in phenotype. More generally, expression is an important component of gene function, and studying the evolution of gene expression is a key step in evolutionary genomics. While there has been a great deal of research concerning the primary treatment of expression data in general (see [Quackenbush, 2002](#), and [Garber *et al.*, 2011](#) for reviews), there has been little investigation into the methods used more specifically to quantify expression evolution ([Pereira *et al.*, 2009](#)). This can make it difficult to critically assess contradictory results, such as the reports that broadly expressed genes are more conserved ([Khaitovich *et al.*, 2005](#)) or less conserved ([Liao and Zhang, 2006b](#); [Liao *et al.*, 2010](#)) than specifically expressed genes.

In order to assess whether and how much expression has been conserved between two orthologous genes by selection, we need an expectation for expression similarity under neutral evolution. Thus, the estimation of gene expression conservation requires two components: i) a measure of gene expression similarity, and ii) the expected value of the divergence level under neutrality.

The two most common measures of similarity between expression profiles of orthologous genes are Pearson's correlation coefficient ([Yanai *et al.*, 2004](#); [Yang *et al.*, 2005](#); [Liao and Zhang, 2006a,b](#); [Xing *et al.*, 2007](#); [Chan *et al.*, 2009](#); [Zheng-Bradley *et al.*, 2010](#)) and Euclidean distance ([Yanai *et al.*, 2004](#); [Jordan *et al.*, 2005](#); [Liao and Zhang, 2006a](#)). The results obtained with Pearson's and Euclidean distances have been reported to be poorly correlated ([Liao and Zhang, 2006a](#); [Pereira *et al.*, 2009](#)). This poses the question which of these measures provides a better description of expression similarity. It has been demonstrated that Pearson's correlation coefficient, in contrast to Euclidean distance, underestimates the expression similarity

between orthologous genes with a conserved uniform pattern of expression. In consequence, use of the Euclidean distance has been encouraged ([Pereira *et al.*, 2009](#)).

For neutral evolution, one expects that similarity between expression profiles of orthologous genes gradually decreases with time. For species that have diverged for sufficiently long time no detectable similarity in expression is expected to remain; this has been postulated to be the case between mouse and human (~ 100 million years; [Jordan *et al.*, 2005](#)). It has been suggested that such large neutral divergence could be approximated by calculating the distance between expression profiles of randomly chosen pairs of genes from the species compared. The standard approach used to generate random pairs of genes is to permute the orthology relationship between them ([Liao and Zhang, 2006a,b](#); [Xing *et al.*, 2007](#); [Chan *et al.*, 2009](#); [Zheng-Bradley *et al.*, 2010](#)).

Here, we show formally and empirically that, in contrast to previous reports ([Liao and Zhang, 2006a](#); [Pereira *et al.*, 2009](#)), there exists a relationship between the Pearson's correlation coefficient and the Euclidean distance, which depends on the data normalization. We also extend the previous study of [Pereira *et al.* \(2009\)](#) by considering more than just the uniform pattern of expression. We demonstrate that in fact both distance measures depend on the expression specificity of analyzed genes. Next, we discuss these observations in the context of the assessment of gene expression conservation. We show that the comparison of expression profiles for randomly permuted gene pairs is biased when broadly expressed genes are abundant in the data, a distribution characteristic of many datasets. To overcome this problem, we propose a novel procedure to generate random gene pairs. This procedure is not biased by the over- or underrepresentation of any expression profile in the data sets. Finally, we use our approach to provide clear evidence for

constrained evolution of gene expression between mouse and human.

1.2 Methods

1.2.1 Gene expression data

We used the human and mouse gene expression data from the GNF Gene Expression Atlas of [Su *et al.* \(2004\)](#) as a case study. This study was performed on the Affymetrix HG-U133A array as well as on the custom array GNF1H for human, and on the custom array GNF1M for mouse. In total, expression profiles for 79 human and 61 mouse organs were measured, with 44,928 probe sets for human and 36,182 probe sets for mouse. We only took into account organs belonging to the homologous organ groups (HOGs) defined in the Bgee database ([Bastian *et al.*, 2008](#)). Using the mapping available in the Bgee database we could connect 36 human organs and 30 mouse organs to 27 HOGs. See Supplementary table S1 for the list of HOGs and their corresponding organs. Microarray data were normalized with the *gcrma* R package ([Wu *et al.*, 2004](#)).

To assign the probe sets to their corresponding human or mouse genes we used the mapping available in Bgee. We kept only probe sets which matched to a unique Ensembl gene. A total of 15,121 probe sets corresponding to 13,853 mouse genes, and 23,920 probe sets corresponding to 15,338 human genes were found.

To estimate the expected values of distances for gene pairs with conserved expression patterns, we used data from replicated experiments, performed in each species. Thus, for each probe set we had two vectors of values representing its expression over the organs. The data sets contained 36 organs and 23,920 probe set pairs for human, and 30 organs and 15,121 probe set pairs for mouse. The results of the study on mouse gene expression data are presented in the Supplementary

Materials.

To study gene expression evolution between mouse and human we merged human and mouse organs into 27 HOGs. For every probe set in each HOG the arithmetic mean of the gcRMA normalized expression values was calculated (each HOG was represented by at least two microarrays). We used a subset of 8,942 one-to-one orthologous gene pairs (see [Human–mouse orthologous genes](#)). If the gene was matched by more than one probe set on the microarray, we randomly picked one probe set to represent that gene.

1.2.2 Human–mouse orthologous genes

Homology information of human and mouse genes was retrieved from Ensembl release 55 ([Hubbard *et al.*, 2009](#)), using BioMart ([Smedley *et al.*, 2009](#)). A total of 8,942 pairs of human–mouse one-to-one orthologous genes had expression information in the data sets we used.

1.2.3 Normalization procedures

For a given gene we consider a vector \mathbf{x} of expression intensities x_i across n different organs indexed by $i = 1, \dots, n$. The Manhattan normalization of \mathbf{x} is calculated by dividing it by its L^1 norm:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|.$$

In some studies ([Liao and Zhang, 2006a](#); [Pereira *et al.*, 2009](#)) this normalization is called relative abundance. The Euclidean normalization of vector \mathbf{x} is calculated by dividing the vector by its L^2 norm:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}.$$

Finally, we introduce a so-called z -like normalization of \mathbf{x} which corresponds to the Euclidean normalization of \mathbf{x} minus its mean value:

$$\tilde{\mathbf{z}}_{\mathbf{x}} = \frac{\mathbf{x} - \bar{x}}{\|\mathbf{x} - \bar{x}\|_2}.$$

1.2.4 Pearson's and Euclidean distances

The Pearson's distance (d_P) between two expression profiles is defined as $1 - r$, where

$$r = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_{\mathbf{x}}s_{\mathbf{y}}} = \tilde{\mathbf{z}}_{\mathbf{x}}^T \tilde{\mathbf{z}}_{\mathbf{y}} = \frac{1}{n} \mathbf{z}_{\mathbf{x}}^T \mathbf{z}_{\mathbf{y}} \quad (1.1)$$

is the Pearson's correlation coefficient between vectors \mathbf{x} and \mathbf{y} . Here the vector elements x_i and y_i are the expression signal intensities of two genes in the condition i , \bar{x} and \bar{y} are the sample means, $s_{\mathbf{x}}$ and $s_{\mathbf{y}}$ are the sample standard deviations. $\mathbf{z}_{\mathbf{x}}$ and $\mathbf{z}_{\mathbf{y}}$ are the z -scores of vectors \mathbf{x} and \mathbf{y} .

The Euclidean distance (d_E) between two expression profiles is defined as

$$d_E = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1.2)$$

with notations as for equation 1.1.

1.2.5 Organ specificity of gene expression

In order to measure the expression specificity of human genes we used the organ specificity index τ (Yanai *et al.*, 2005). The τ of a given gene with an expression vector \mathbf{x} is defined as follows:

$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n - 1}, \quad (1.3)$$

where

$$\hat{x}_i = \frac{x_i}{\|\mathbf{x}\|_\infty} = \frac{x_i}{\max_{1 \leq i \leq n}(x_i)}.$$

The value of τ varies between 0 and 1, with higher values indicating higher organ specificity.

1.2.6 τ -group composition

In order to study the relation between d_E and τ we used replicated expression data for human genes (36 organs, 23,920 probe sets). We sorted the probe set pairs according to the organ specificity index τ (equation 1.3) of the first replicate, and we divided the probe set pairs into three τ -groups of equal size (e.g., the first group contained 1/3 of the probe set pairs with the first replicate having lowest τ). For each group we recorded the minimum and maximum τ value of the first replicate, and used these values to filter out probe sets with the two replicates having τ values from different groups. The resulting τ -groups were of similar, but not equal, size (table 1.1). An alternative τ -group composition, with a more balanced distributions of τ values (first group containing genes with $\tau \in [0, 0.2)$; second group with $\tau \in [0.2, 0.6)$; and third group with $\tau \in [0.6, 1)$) leads to unbalanced sizes of three groups. Nevertheless, for both approaches the results are qualitatively the same (Supplementary figures S6 and S7).

1.2.7 Randomization procedures

Changes in gene expression patterns between randomly chosen genes from two species have been suggested as an approximation for the result of neutral expression evolution (Jordan *et al.*, 2005). We used two different randomization procedures to create such sets of random gene pairs. First, we permuted the

gene order within replicates (or within species). We refer to these as *randomly permuted pairs*. Second, we performed what we refer to as “ τ -uniform sampling”. We first randomly chose an organ specificity index (τ), uniformly from the interval of (τ_{min}, τ_{max}) , where τ_{min} and τ_{max} are the lowest and the highest values of the observed τ , respectively. Next, we picked the gene with the value of τ closest to the randomly chosen τ within one data set (i.e., within one replicate, or one species). Then, independently, we repeated the procedure for the second data set. Thus, we obtained two randomly chosen genes which form a new random pair. Repeating the procedure provides the “ τ -uniform” random gene pairs.

1.3 Results and Discussion

1.3.1 Correlation between Pearson’s and Euclidean distances depends on data normalization

To compare gene expression between species, over many different conditions, it is important to normalize the expression levels between the conditions in order to obtain a common scale between species. This is distinct from the preprocessing normalization (within condition), which is typically done using methods such as LOESS (Yang *et al.*, 2002b) or gcRMA (Wu *et al.*, 2004), and is not specific to inter-species evolutionary studies. In the following, we only consider the impact of the between conditions normalization on the evolutionary comparisons. We discuss three normalization procedures commonly used for evolutionary studies: Manhattan normalization (also referred to as “relative abundance”; Liao and Zhang, 2006a), Euclidean normalization and z -like normalization (see [Normalization procedures](#) for mathematical definition of all three normalizations).

One can use any of these normalizations before calculating the Pearson’s or

Euclidean distance between two gene expression profiles. However, the choice of normalization can affect the results. Pearson's distance (d_P) between two expression profiles remains the same, regardless of whether and how the data are normalized, and it ranges between 0 and 2. The reason is that r is defined on the z -scores (see equation 1.1 in Methods), which are invariant with respect to linear transformation. In contrast, the Euclidean distance between two expression profiles (d_E) changes its value depending on the normalization used, even though the interval of possible d_E values is always between 0 and 2.

The correlation between d_P and d_E is poor for Manhattan (Supplementary figure S1A; see also [Liao and Zhang, 2006a](#); [Pereira et al., 2009](#)) and Euclidean normalizations (Supplementary figure S1B). In contrast, z -like normalization leads to an interdependent relationship between d_P and d_E , defined by

$$d_E^2 = 2d_P \quad (1.4)$$

(see Theoretical Analysis in Supplementary Material, and Supplementary figure S1C). As d_P gives the same results for all three normalizations, and for z -like normalization it is equal to $\frac{1}{2}d_E^2$, we focused on the Euclidean distance. If not stated otherwise, the Euclidean distance was calculated for all three normalizations: Manhattan, Euclidean and z -like, referred to as d_E^M , d_E^E and d_E^Z , respectively.

1.3.2 Commonly used measures of gene expression similarity depend on the organ specificity of the genes

Intuitively, one might assume that the distance between two orthologous genes which have conserved the expression profile of their last common ancestor should be close to zero, and that this should hold regardless of the gene expression pattern. In

order to assess if this is indeed the case, we performed an empirical study. We used human microarray data with the expression information from 36 different organs in two replicates (Su *et al.*, 2004). The replicates were used to “simulate” pairs of genes with conserved expression profiles. We calculated the organ specificity index τ (equation 1.3) for each pair of replicates, and then divided them into three τ -groups of similar size (see τ -group composition for details). The first two groups contained broadly expressed genes ($\tau \leq 0.295$), and only the third group consisted of genes with more specific expression patterns ($\tau > 0.295$) (table 1.1).

Table 1.1: Composition of three τ -groups of human probe set (ps) pairs

	Organ specificity (τ)	Number of ps pairs
τ -group 1	$0.003 \leq \tau \leq 0.117$	6348
τ -group 2	$0.117 < \tau \leq 0.295$	5280
τ -group 3	$0.295 < \tau \leq 0.879$	6692

We measured the Euclidean distances (d_E^M , d_E^E and d_E^Z) for probe set pairs within each τ -group. The resulting levels of expression similarity between replicates strongly depended on the organ specificity level. Values of d_E^M and d_E^E were significantly lower for broadly expressed genes than for organ-specific genes ($p < 10^{-16}$, Mann–Whitney U test, figure 1.1A,B; Supplementary figure S5A,B). In contrast, values of d_E^Z were significantly higher for broadly expressed genes than for organ-specific genes ($p < 10^{-16}$, Mann–Whitney U test; figure 1.1C; Supplementary figure S5C). See Supplementary figure S3 for the correlation analysis between the Euclidean distances and organ specificity index.

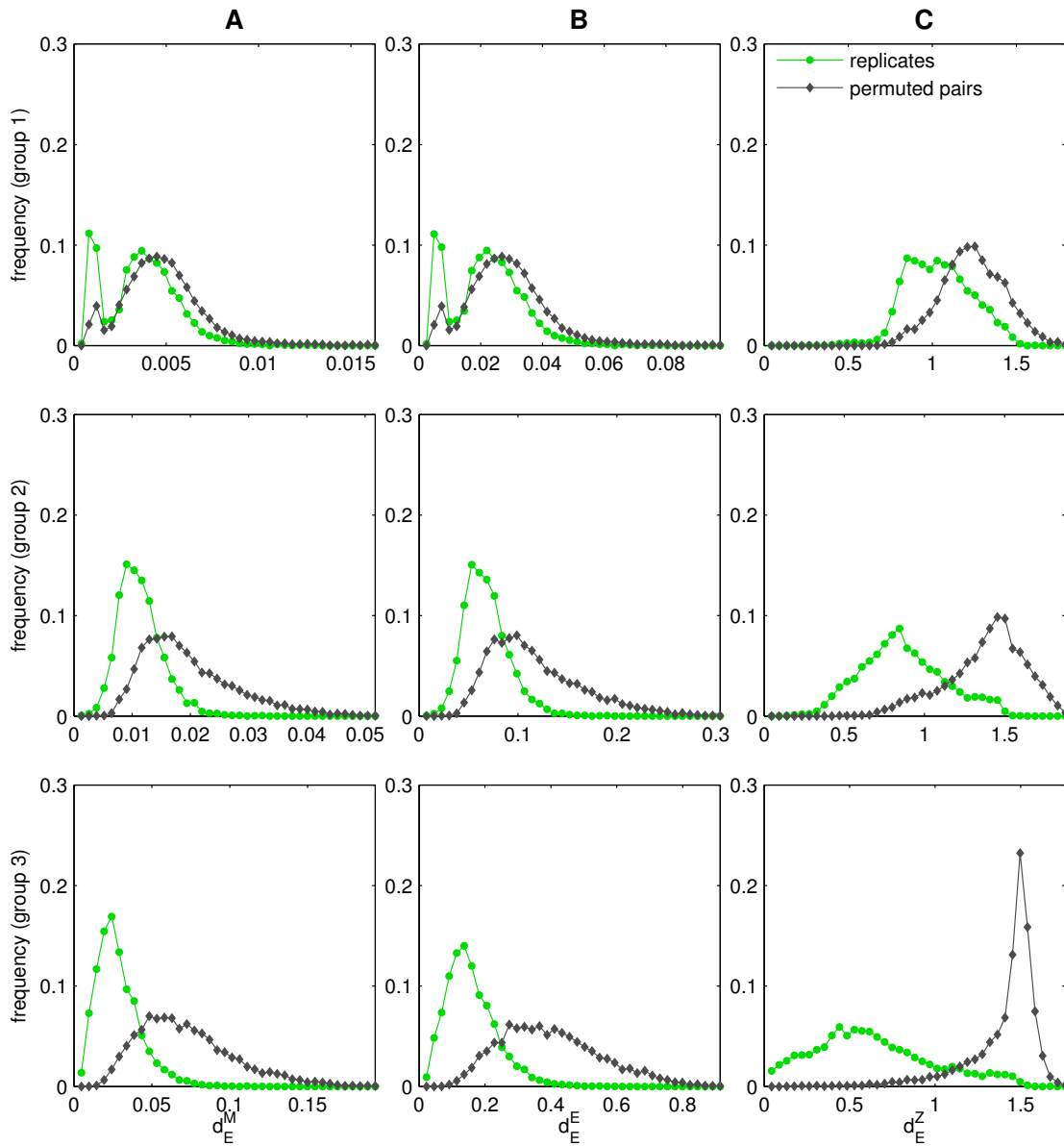


Figure 1.1: The distribution of expression similarity between human replicates depends on their organ specificity. (A) d_E^M and (B) d_E^E are significantly lower for broadly expressed genes (group 1) than for organ-specific genes (group 3). For randomly permuted gene pairs d_E^M and d_E^E also differ between the three τ -groups. They are significantly lower for random pairs in group 1 than in group 3. (C) d_E^Z is significantly higher for broadly expressed genes (group 1) than for organ-specific genes (group 3). d_E^Z for randomly permuted pairs is high in all three groups, even in the first τ -group, where random pairs consist of two broadly expressed genes (this is a consequence of low r for uniformly expressed genes). Note that the scale of the x-axis differs strongly between graphs.

1.3.3 The rate of neutral expression evolution estimated with randomly permuted gene pairs depends on the organ specificity of the genes

The rate of neutral expression evolution is typically approximated by calculating the distance between expression profiles of randomly paired genes. The random choice of the genes is assumed to remove any similarity between them (Jordan *et al.*, 2005). The standard approach to generate random gene pairs is to permute the ortholog relationship between the genes in the data sets. We created random probe set pairs by permuting the probe set order within each of the three τ -groups separately, and we then calculated the Euclidean distances (d_E^M , d_E^E and d_E^Z) between their expression profiles. We found that d_E^M and d_E^E were significantly lower for random pairs from the first τ -group, than for random pairs from the third τ -group ($p < 10^{-16}$, Mann–Whitney U test; figure 1.1A,B; Supplementary figure S5A,B). This is because the first τ -group consisted of broadly expressed genes. Consequently, even the randomly matched probe set pairs tended to have similar expression patterns and thus low distances. In contrast, the third τ -group consisted of genes with more specific expression patterns, and so the random pairs were truly different.

d_E^Z between random pairs was not affected by organ specificity, in the sense that in all three τ -groups the median d_E^Z was around 1.4 (figure 1.1C; Supplementary figure S5C). Values of d_E^Z were high even in the first τ -group, although it consisted of random pairs with similar, broad patterns of expression. The reason is that $d_E^Z = \sqrt{2(1-r)}$ is a decreasing function of r , which for broadly expressed gene pairs reflects mainly the noise of the measurement and is close to 0 (see Pereira *et al.*, 2009 for details, and Supplementary figure S2). Thus, random gene pairs from the first τ -group tend to have high d_E^Z values (around $\sqrt{2}$).

1.3.4 A large fraction of broadly expressed genes leads to an underestimation of expression conservation

Our analysis shows that if the fraction of broadly expressed genes is large, the level of gene expression conservation is likely to be underestimated. This is especially important if we consider the fact that housekeeping genes (broadly expressed) are more frequent than organ-specific genes (Ramsköld *et al.*, 2009). We found such skewed distributions not only in the human data considered here (figure 1.3A), but also in several other data sets, e.g., most mouse genes are broadly expressed over different organs, most Arabidopsis genes are broadly expressed over different light conditions, and most zebrafish genes are broadly expressed over different developmental stages (Supplementary figure S4).

To illustrate the extent to which the abundance of broadly expressed genes affects measures of gene expression conservation, we re-analyzed all the human probe set pairs, without dividing them into τ -groups. We created random probe set pairs by permuting the probe set order within both replicates, and we calculated the Euclidean distances (d_E^M , d_E^E and d_E^Z) both for the pairs of replicates and for the random pairs. Ideally, one would expect to detect very high similarity between replicates, and very low similarity between random pairs.

For Manhattan and Euclidean normalizations, distances for most human random pairs were very small, indistinguishable from the distances between replicates (figure 1.2A,B; Supplementary figure S8A,B). This contradicts the assumption that differences between randomly paired genes are to approximate well the rate of neutral divergence, with very low similarity (i.e., high distance) expected (Jordan *et al.*, 2005). For the z -like normalization, distances between random pairs were high, which is consistent with the assumption of pseudo-neutrality (Jordan *et al.*, 2005). However the d_E^Z values for the replicates were similarly high (figure 1.2C;

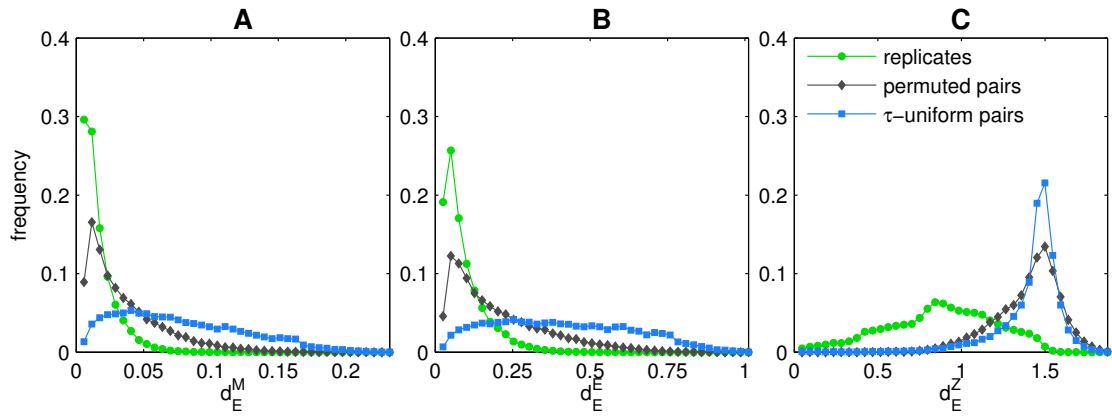


Figure 1.2: Overrepresentation of broadly expressed human genes causes underestimation of the conservation of expression when randomly permuted pairs are used to approximate the neutral evolution rate. (A, B) For most randomly permuted pairs (grey) the distance (d_E^M and d_E^E) is small, indistinguishable from the distances between replicates (green). For τ -uniform random pairs (blue) d_E^E and d_E^M are higher, which is more consistent with the assumption about neutral evolution (Jordan *et al.*, 2005). (C) d_E^Z is high both for randomly permuted gene pairs and for the group of replicates. The distribution of d_E^Z does not change with the new random pairs set.

Supplementary figure S8C), whereas they are expected to be low. Thus, the presence of numerous broadly expressed genes causes systematically low values of d_E^M and d_E^E between randomly paired genes, and systematically high values of d_E^Z between conserved gene pairs. The first is a consequence of the fact that it is easier to randomly choose two broadly expressed genes, and thus to get a low value of d_E^M or d_E^E . The second is a consequence of low values of r for uniformly expressed genes, leading to the high values of d_E^Z (as discussed in subsection 1.3.3). In all cases, the level of gene expression conservation is underestimated.

Although we show this effect using a specific set of human microarray data, our conclusions are very general and hold for any study in which a significant fraction of the genes is uniformly expressed over conditions (see figure S2 and its caption for a mathematical explanation).

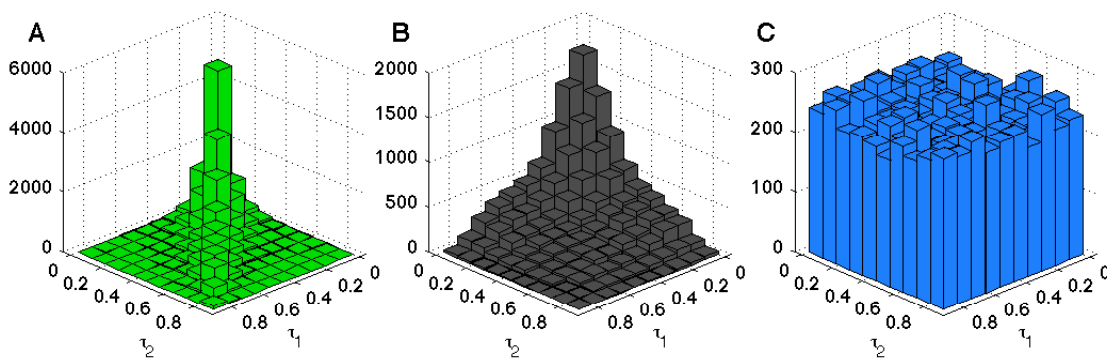


Figure 1.3: Random gene pairs have their τ values differently distributed depending on the randomization procedure used. (A) τ distribution for human replicates. The τ pairs are distributed along the diagonal, which is expected for replicates. (B) τ distribution for randomly permuted gene pairs. The τ pairs are biased towards low values, which are the most frequent values in human data sets. (C) τ distribution for τ -uniform random pairs. The τ pairs are uniformly distributed, and not biased towards the low values.

1.3.5 An alternative construction of random gene pairs improves the estimation of expression conservation

To overcome the limitation of using randomly permuted gene pairs to estimate the expression divergence under neutrality, we propose a new procedure to create random gene pairs. This procedure is unbiased regardless of over- or underrepresentation of any expression profiles in the data sets. Consequently, it provides a better approximation of the expression divergence under neutral evolution between distant species. In order to generate a single random pair of genes, one randomly chooses two expression specificity values, τ_1 and τ_2 , uniformly from the interval of (τ_{min}, τ_{max}) , where τ_{min} and τ_{max} are the lowest and the highest values of the observed τ , respectively. Next, one picks the two genes from the two data sets that have the closest τ values to τ_1 and τ_2 , respectively. The resulting pairs of genes have the two τ values uniformly distributed, and not biased as for randomly permuted gene pairs (figure 1.3B,C).

We applied our new procedure 23,920 times to create as many random probe set pairs for human data sets. Then, we calculated the Euclidean distances (d_E^M ,

d_E^E , and d_E^Z) both for replicates and random probe set pairs. We found that, relative to classical randomly permuted pairs, the distribution of d_E^E and d_E^M for τ -uniform random pairs differs strongly from that for replicates (figure 1.2A,B), with a high frequency of large distance values, as expected for very divergent pairs. Of note, d_E^M and d_E^E give the same shape of distribution (figure 1.1A,B and figure 1.2A,B). While both of these measures could be combined with τ -uniform sampling to estimate gene expression conservation, for mathematical consistency we prefer the use of d_E^E .

The estimation of gene expression conservation with d_E^Z cannot be corrected by creating the set of random gene pairs differently, because d_E^Z varies significantly with organ specificity for replicates, i.e., for conserved genes, and not for random gene pairs. Thus, we do not recommend using d_E^Z , and consequently the Pearson's correlation coefficient, in any study which aims to detect similarity between genes expressed uniformly over all conditions.

Of note, neither the standard procedure used to generate random pairs, nor our new proposed approach takes into consideration the time passed since the divergence of two organisms. Therefore, the estimated “neutral” divergence will be the same for closely related species (e.g., human and chimp) and more distant species (e.g., human and mouse).

1.3.6 Results of the comparative study of human and mouse gene expression differ strongly according to the choice of randomization method

To demonstrate the importance of our novel approach, we investigated how much evidence of selectively constrained gene expression evolution we can detect between human and mouse. We selected 8,942 one-to-one orthologous gene pairs from the

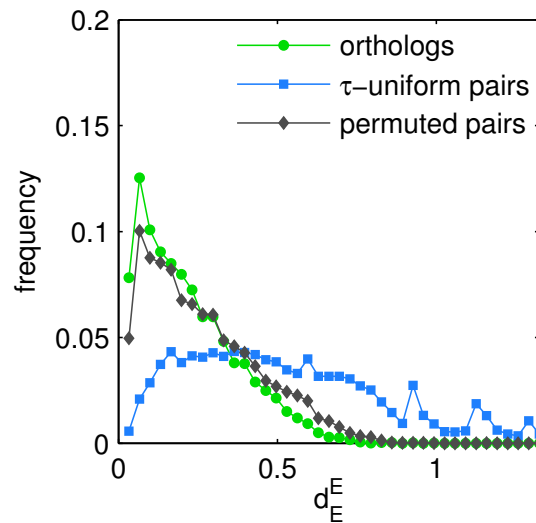


Figure 1.4: The choice of the randomization method changes the conclusions about gene expression evolution between mouse and human. There is no clear evidence for constrained evolution if we compare the distribution of d_E^E for orthologous (green) and randomly permuted gene pairs (grey). Whereas, comparison of d_E^E distribution for orthologous (green) and τ -uniform random pairs (blue) suggest that expression evolution is far from neutral.

human and mouse data sets (Su *et al.*, 2004). We created two sets of random gene pairs, using both random permutation and the procedure of τ -uniform sampling, and we calculated the Euclidean distance (d_E^E) for orthologous gene pairs and for both sets of random pairs (see Figure S9 for analogous analysis with d_E^M). If the d_E^E value for a human–mouse orthologous gene pair is smaller than the 5th percentile of d_E^E for randomly paired genes, there is some evidence that the expression evolution of this pair has been constrained (Liao and Zhang, 2006a). Using randomly permuted gene pairs did not provide clear evidence for constrained evolution (figure 1.4). Only 8% of orthologous pairs were identified to have a conserved expression pattern, which was close to the random expectation of 5%. In contrast, with τ -uniform random pairs, 29% of orthologous genes were identified to have conserved expression (figure 1.4).

The number of detected genes with conserved expression pattern may seem surprisingly low in comparison to Liao and Zhang (2006a), who reported that as

much as 84% of genes showed conserved expression between human and mouse. However, we note that [Liao and Zhang \(2006a\)](#) used two different metrics to calculate the distance between orthologous genes and between randomly paired genes — the so called net distance and the Euclidean distance, respectively. We show that this inconsistency caused an overestimation of the expression conservation between human and mouse (see Supplementary Materials and Supplementary figure S10). Consequently, we believe that correcting for the randomization process yields more accurate results than a one-sided correction of the distance.

We are aware that the alternative way of creating random gene pairs proposed in this paper has some weaknesses, such as visible artificial peaks in the d_E^E distribution (figure 1.4), which are the consequence of the non uniform distribution of τ between 0 and 1. This is because with the τ -uniform sampling one chooses the genes with less frequent τ values more often than genes with more frequent τ values. For example here, the number of narrowly expressed genes was increased at the expense of decreasing the number of broadly expressed genes. Consequently, when only a few genes have a τ value in some non-negligible range, these few genes might repeat many times in the randomized set, and discrete effects may manifest themselves causing artificial peaks. Note that the peaks would disappear if τ values were uniformly distributed between 0 and 1, but then there would be no need for τ -uniform sampling of gene pairs at all. Note also that the peaks do not affect the analysis, as they do not change the overall shape of the distribution of distance values between the randomized gene pairs (figure 1.4).

Finally, one may argue that the τ -uniform sampling contradicts the very purpose of randomization because it makes a probability of choosing a gene higher, if its τ -value is underrepresented in the data set. But the aim of the set of randomized gene pairs is not to be “just random”, but to display maximal divergence between

gene pairs, i.e., to simulate the neutral evolution defined in [Jordan *et al.* \(2005\)](#). In contrast to the standard approach, the τ -uniform sampling makes the distribution of distance values between gene pairs actually independent of the τ distribution observed in the analyzed data set. Thus, we believe that the distance between τ -uniform random gene pairs approximates better a large neutral divergence.

1.4 Conclusions

The Euclidean distance should be used with caution as an estimator of gene expression conservation because it varies as a function of expression specificity. Our results strongly suggest that to assess whether gene expression evolves neutrally, one should use d_E^E (Euclidean distance preceded by Euclidean normalization) and compare its distribution for orthologous and τ -uniform random pairs. Importantly, we validated this approach on real data, and recovered clear evidence for gene expression conservation between mouse and human. Previous small differences reported between real and random gene pairs were likely caused by the way the random pairs were constructed ([Liao and Zhang, 2006a,b](#)). Although in this study we applied our approach to microarray data analysis, the issues highlighted here are also relevant to data acquired with RNA-seq technology ([Mortazavi *et al.*, 2008](#)).

We would like to emphasize that while it is possible to verify whether the expression of a given set of genes was under selective pressure, there is no straightforward way to compare the strength of selection acting on two groups of genes with different expression patterns. Indeed, if we compare a group of broadly expressed genes with a group of narrowly expressed genes, with similar high conservation of expression, the latter will always have higher d_E^E values (and lower d_E^Z values). This methodological problem suggests a need to re-interpret results from previous

evolutionary studies comparing the evolution of broadly and narrowly expressed genes. In particular, studies which have reported higher conservation of organ-specific genes (Liao and Zhang, 2006b; Liao *et al.*, 2010; Movahedi *et al.*, 2011) could have been biased by the fact of using the Pearson's correlation coefficient (equivalent to d_E^Z) as a measure of conservation.

In this paper, we thoroughly analyzed, formally and experimentally, the common measures of expression conservation, and we showed the superiority of the Euclidean distance paired with the Euclidean normalization. We also highlighted the limitation of using randomly permuted pairs to approximate neutrally evolving genes, and proposed a new methodology to better estimate the rate of neutral evolution. With the increase of expression data for many species, our work is likely to become very useful for evolutionary studies of gene expression.

Acknowledgements

We thank P. Lichocki for fruitful discussion. We thank F. Bastian, N. Galtier, and O. Riba-Grognuz for critical comments on the manuscript. We acknowledge funding from Etat de Vaud; Swiss National Science Foundation (ProDoc grant 1206624/1); Swiss Institute of Bioinformatics (to S.B.)

Supporting Information

Supporting materials can be downloaded from:

<http://bioinformatics.oxfordjournals.org/content/28/14/1865/suppl/DC1>

Figure S1: Interdependence between r and d depends on data normalization mode. Both for Manhattan (A) and Euclidean (B) normalizations the

correlation between r and d^2 is low (0.12 and 0.09, respectively). (C) For z -like normalization there is linear dependence between r and d^2 .

Figure S2: Euclidean distance between two genes with conserved expression patterns (A,B) depends on data normalization mode and specificity of genes expression. For Euclidean normalization the distance is lower for genes expressed over all conditions (C) than for specifically expressed genes (D). For z -like normalization the distance is higher for genes expressed over all conditions (E) than for specifically expressed genes (F). Regression line is plotted in red. Identity line ($y = x$) is plotted in blue. Note that d_E^2 can be estimated by summing squared distances (in the y -direction) from the points to the blue line.

Figure S3: Euclidean distance between replicates (simulating genes with conserved expression patterns) depends on data normalization and specificity of genes expression. For Manhattan and Euclidean normalization (A,B) the distance is positively correlated with the expression specificity, whereas for z -like normalization (C) this correlation is negative. Top: human replicates. Spearman correlation coefficients: for d_E^M : 0.89, for d_E^E : 0.88, for d_E^Z : -0.56. Bottom: mouse replicates. Spearman correlation coefficients: for d_E^M : 0.68, for d_E^E : 0.64, for d_E^Z : -0.81.

Figure S4: τ distribution is not uniform in the real data. (A,B) τ distribution for human replicates. (C,D) τ distribution for mouse replicates. (E) τ distribution for zebrafish genes expressed during the ontogeny (Domazet-Lošo and Tautz, 2010). (F) τ distribution for Arabidopsis genes expressed in different light conditions (NASC 2007, GEO accession number GSE5617).

Figure S5: The distribution of expression similarity between mouse replicates depends on their organ specificity. (A) d_E^M and (B) d_E^E are significantly lower for broadly expressed genes (group 1) than for organ specific genes

(group 3). For randomly permuted pairs of genes d_E^M and d_E^E also differ between the three τ -groups. They are significantly lower for random pairs in group 1 than in group 3. (C) d_E^Z is significantly higher for broadly expressed genes (group 1) than for organ specific genes (group 3). d_E^Z for randomly permuted pairs is high in all three groups even in the first τ -group, where random pairs consist of two broadly expressed genes (this is a consequence of low r for uniformly expressed genes) Note that scale of x -axis differs strongly between graphs.

Figure S6: The distribution of expression similarity between human replicates depends on their organ specificity. Presented 3 groups of gene pairs have balanced τ distribution. Group 1: τ : 0–0.2; 10,723 gene pairs; Group 2: τ : 0.2–0.6; 7,551 gene pairs; Group 3: τ : 0.6–1; 1,514 gene pairs. For the explanation of the figure please refer to figure S5.

Figure S7: The distribution of expression similarity between mouse replicates depends on their organ specificity. Presented 3 groups of gene pairs have balanced τ distribution. Group 1: τ : 0–0.2; 5,424 gene pairs; Group 2: τ : 0.2–0.6; 5,688 gene pairs; Group 3: τ : 0.6–1; 2,303 gene pairs. For the explanation of the figure please refer to figure S5.

Figure S8: Overrepresentation of broadly expressed mouse genes causes underestimation of the conservation of expression when randomly permuted pairs are used to approximate the neutral evolution rate. (A,B) For noticeable number of randomly permuted pairs the distances (d_E^M and d_E^E) are small, indistinguishable from the distances for replicates. (C) d_E^Z is high both for permuted gene pairs and for the group of replicates. (A,B) For τ -uniform random pairs d_E^E and d_E^M are higher, which is more consistent with the assumption about neutral evolution from Jordan et al. (2005). (C) distribution of d_E^Z does not change with the new random pairs set.

Figure S9: The choice of the randomization method changes the conclusions about gene expression evolution between mouse and human.

There is no clear evidence for constrained evolution if we compare the distribution of d_E^M for orthologous (green) and randomly permuted gene pairs (grey). Whereas, comparison of d_E^M distribution for orthologous (green) and τ -uniform random pairs (blue) suggest that expression evolution is far from neutral.

Figure S10: One-sided correction of the Euclidean distance lead to different distributions of distance values for two sets of randomly paired genes. Using 3,193 human–mouse orthologous gene pairs (all human genes covered by multiple probe sets), we generated two sets of randomly permuted gene pairs. For the first set (simulating the set of genes with non-conserved expression profiles) we calculated the net distance, for the second set (used to estimate neutral evolution) we calculated the Euclidean distance. Because both sets were “equally random”, one should not expect any differences between them. However, as much as 20% of gene pairs from the first random set (green) was detected to be more conserved than gene pairs from the second random set (grey).

Table S1: List of homologous organ groups (HOGs) and their corresponding organs (sample names) in mouse and human.

Table S2: Composition of three τ -groups of mouse probe set (ps) pairs.

2

Comparative modular analysis of gene expression in vertebrate organs

Barbara Piasecka, Zoltán Kutalik, Julien Roux, Sven Bergmann,
Marc Robinson-Rechavi

Abstract

The degree of conservation of gene expression between homologous organs largely remains an open question. Several recent studies reported some evidence in favor of such conservation. Most studies compute organs' similarity across all orthologous genes, whereas the expression level of many genes are not informative about organ specificity.

Here, we use a modularization algorithm to overcome this limitation through the identification of inter-species co-modules of organs and genes. We identify such co-modules using mouse and human microarray expression data. They are functionally coherent both in terms of genes and of organs from both organisms. We show that a large proportion of genes belonging to the same co-module are orthologous between mouse and human. Moreover, their zebrafish orthologs also tend to be expressed in the corresponding homologous organs. Notable exceptions to the general pattern of conservation are the testis and the olfactory bulb. Inter-

estingly, some co-modules consist of single organs, while others combine several functionally related organs. For instance, amygdala, cerebral cortex, hypothalamus and spinal cord form a clearly discernible unit of expression, both in mouse and human.

Our study provides a new framework for comparative analysis which will be applicable also to other sets of large-scale phenotypic data collected across different species.

This article was published in *BMC Genomics* (2012) **13**: 124.

doi: 10.1186/1471-2164-13-124

2.1 Background

Specific over-expression of a gene in an organ is often taken to imply a function of the gene in that organ. If so, and if orthologous genes have conserved function, we would expect a slow rate of organ-specific expression evolution. Some early comparisons of microarray data between species suggested the opposite. The most studied data set in this regard is the GNF gene atlas of human and mouse organs (Su *et al.*, 2002, 2004). Yanai *et al.* (2004) used an early version of these data (Su *et al.*, 2002), and reported that the expression profiles of orthologous genes differed remarkably between two mammalian species. Moreover, comparing the expression profiles of 16 tissues (for both species), they found that human tissues were more similar to each other than to their corresponding mouse tissues. In contrast, Liao and Zhang (2006a), based on a more recent version of the data (Su *et al.*, 2004), and correcting for systematic error, found that human–mouse orthologous gene pairs had significantly lower expression divergence than random gene pairs. Additionally, they found that gene expression profiles of homologous tissues between species are more similar to each other than expression profiles of non-homologous tissues. Two recent studies (Zheng-Bradley *et al.*, 2010; McCall *et al.*, 2011) have confirmed that gene expression profiles of mouse and human homologous organs are indeed more similar than expression profiles between two different organs within a species, at least for the limited number of samples studied (immune system, heart and muscle, skin and gastrointestinal organs, liver and brain in Zheng-Bradley *et al.*, 2010; kidney, liver, brain, spleen, skeletal muscle and lung in McCall *et al.*, 2011).

In many of these studies the Pearson’s correlation coefficient or Euclidean distance were used as estimators of gene expression conservation, either when calculating the distance between expression profiles of orthologous genes, or when

clustering homologous organs from two species. These measures depend strongly on data normalization (Piasecka *et al.*, 2012), and only capture global similarity across all samples. Specifically, none of these measures allows discovering between-species units of conservation, i.e., modules of organs and their specific genes that have remained largely unchanged since the speciation event. To facilitate gene expression studies, McCall *et al.* (2011) have created a database of gene expression states in different conditions. It allows finding groups of co-expressed genes, but only for manually chosen conditions. Consequently, discovering modules of organs and their specific genes, requires an a priori guess about the potential groups of organs that express the same set of genes.

In this work, we take an alternative approach that automatically discovers such modules. We use the Ping-Pong Algorithm (that was originally developed for the unsupervised simultaneous modularization of gene expression and drug response data (Kutalik *et al.*, 2008)) to co-analyze microarray gene expression data from mouse and human. Using the resulting co-modules, that contain genes and organs in which these genes are coherently expressed, we address several questions: 1) Are there any “natural” modules of mammalian organs, meaning groups of organs with very similar sets of co-expressed genes? 2) Which genes are module-specific? 3) Are these genes conserved between species?

2.2 Results

2.2.1 The Ping-Pong Algorithm

The Ping-Pong Algorithm (PPA; Kutalik *et al.*, 2008) is an algorithm for the integrative analysis of two large-scale data sets sharing one dimension. When applied to gene expression data from two species, it identifies, simultaneously in both

data sets, subsets of samples for which certain sets of genes are coherently overexpressed. We refer to the combined subsets of samples and genes as co-modules. The dimensions shared by our data sets are twofold: orthology relation between genes (figure 2.1A) and organ homology (figure 2.1B). First, we ran the PPA on the data sets matched through one-to-one orthologous gene pairs. Thus, the co-modules consisted of orthologous genes and the mouse and human organs in which these genes were overexpressed. Second, we ran the PPA on the data sets matched through homologous organ groups (HOGs; [Parmentier *et al.*, 2010](#); [Niknejad *et al.*, 2012](#)). The resulting co-modules consisted of sets of homologous organs and (potentially different) sets of mouse and human genes with coherent overexpression in these organs. Each organ and gene received a score indicating their membership (if non-zero) and contribution to a given co-module. The further the score for a gene or organ is from zero, the stronger the association to the co-module.

Representing coherent features across both data sets in terms of co-modules reduces the complexity of the data and facilitates the study of its biological properties. There are only a few dozen co-modules to study, instead of thousands of genes. Moreover, the mean expression level of genes in a co-module is more robust than the expression measure for a single gene, as measurement noise tends to cancel out.

2.2.2 Co-modules based on orthologous genes contain homologous organs

We applied the PPA to the mouse–human data sets matched through 8,942 one-to-one orthologous genes, containing the expression signal from 27 organs of both species. We ran the PPA starting from 10,000 different seeds consisting of random homologous organ groups. We obtained 25 distinct co-modules consisting

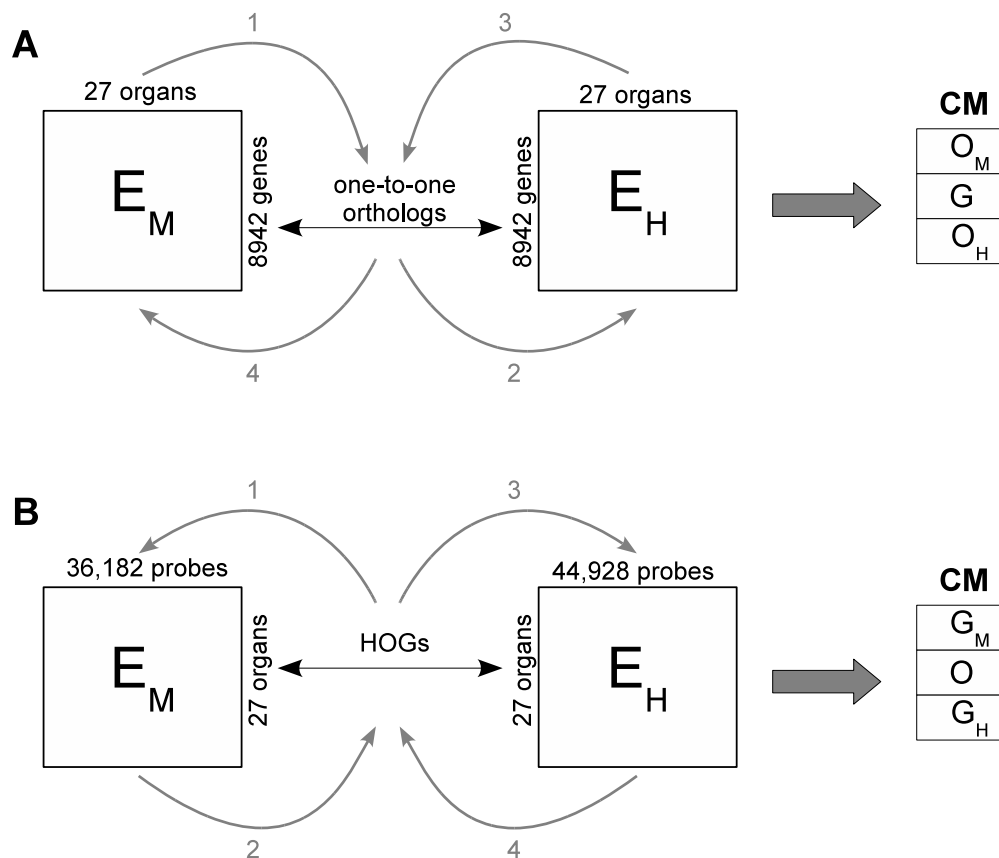


Figure 2.1: Schematic representation of the Ping-Pong Algorithm. (A) The PPA run for two data sets with orthologous genes on the common dimension. (B) The PPA run for two data sets with homologous organs on the common dimension. E_H — human expression data, E_M — mouse expression data, O_M — mouse organs, O_H — human organs, G — human and mouse one-to-one orthologs, G_H — human genes, G_M — mouse genes, O — homologous organs, CM — co-module.

of orthologous genes and the mouse and human organs where these genes were expressed.

Importantly, this analysis allowed us to recover the information about organ homology: co-modules contained mouse and human organs that are known to be homologous. The mouse organs which were grouped together with their human homolog were the following: lymph node, cerebellum, hypothalamus, tongue, testis, pancreas, liver and kidney. Moreover, we recovered information about functional groups of organs, which are conserved between mouse and human. In particular, we found a muscle co-module containing heart, skeletal muscle and tongue, a

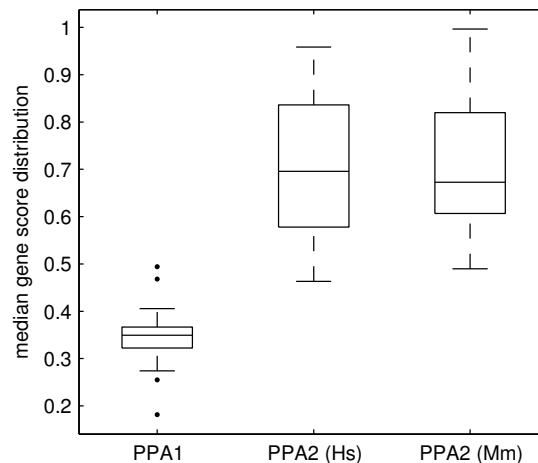


Figure 2.2: Median gene score of co-modules from both Ping-Pong Algorithm runs. (i) median gene scores for 25 co-modules detected in the PPA run on data matched through orthologous genes. (ii) median human gene scores for 98 co-modules detected in the PPA run on data matched through homologous organs. (iii) median mouse gene scores for 98 co-modules detected in the PPA run on data matched through homologous organs.

central nervous system (CNS) co-module with amygdala, and cerebral cortex, and an immune system co-module containing both lymph node and thymus. Genes and organs belonging to the same co-module were coherent in terms of functional annotation. For example, the muscle co-module was enriched in genes involved in glycolysis, the immune co-module in immune response, the testis co-module in sperm motility, and the liver co-module in catabolic processes (see Additional file 1).

The median gene score for each co-module varied between 0.18 and 0.49 (figure 2.2), which suggests that the contribution of individual genes to co-modules was rather weak.

2.2.3 Co-modules based on homologous organs are organ- or system-specific

Above, we applied the PPA to the data sets matched through one-to-one orthologous genes. This recovered the information about organ homology and thus validated

our approach, but limited it to one-to-one orthologous genes only. In a second step, in order to broaden the analysis, we applied the PPA to the data sets matched through 27 homologous organ groups. In contrast to the first run, here we used the expression signal coming from all 36,182 mouse probe sets and 44,928 human probe sets. We ran the PPA starting from 10,000 seeds consisting of random homologous organ groups. We obtained 98 distinct co-modules consisting of homologous organ groups and mouse and human probe sets carrying the signal specific for these HOGs. Next, the probe sets were mapped to their corresponding genes, and those which did not map unambiguously to a gene were excluded from further analysis.

First, for every single organ we detected a co-module containing this organ and its specific genes from mouse and human (e.g., figure 2.3A), which confirms that organs are “natural” modules of gene expression in mammals. We refer to these co-modules as *organ-specific* co-modules. The median numbers of mouse and human genes assigned to these co-modules were 117 and 264.5, respectively.

Second, we confirmed and extended the discovery of co-modules containing several functionally related organs. We refer to them as *system-specific* co-modules. These notably include ovary and uterus; lung and trachea; lymph node and thymus; and liver and kidney (figure 2.3B). The median numbers of mouse and human genes assigned to these co-modules were 257 and 281, respectively.

Third, the central nervous system (CNS) emerged as a particular case of a system-specific co-module. For instance, we found co-modules consisting of: amygdala, cerebellum and cerebral cortex; amygdala, hypothalamus and spinal cord; or cerebellum, hypothalamus, and spinal cord. After closer analysis of these co-modules we found that four central nervous system organs were connected more tightly than the others. These organs were: amygdala, cerebral cortex, hypothalamus and spinal cord. Whenever a co-module detected by the PPA contained one

of these four CNS organs (e.g., cerebral cortex and olfactory bulb, figure 2.3C), the genes from that co-module were also expressed in the three other CNS organs, although sometimes just below the threshold level that PPA used to add the organ into co-modules (see Methods). The median number of mouse and human genes assigned to these co-modules were 336 and 149, respectively.

The median gene score for each co-module varied from 0.46 to 0.96 for human, and 0.49 to 0.99 for mouse (figure 2.2). The genes' contribution to co-modules was stronger than in the analysis with genes on the common dimension, which indicates that these co-modules are more reliable. This is probably due to the larger data sets used.

2.2.4 Genes belonging to co-modules are enriched in functions relevant to the corresponding organs

Functional annotation analysis confirmed that genes belonging to each co-module were enriched in functions relevant to the respective organs, for both mouse and human. For example, the testis co-module was enriched in genes involved in spermatogenesis and sperm motility, the heart co-module in those involved in regulation of heart contraction, the lymph node co-module in those involved in immune response, and the nervous system co-modules were enriched in genes important during nervous system development (see Additional file 2). This confirms the functional coherence of the organ- or system-specific co-modules detected.

2.2.5 Organ-specific gene expression is often related to organ-specific hypomethylation of regulatory elements

Recently, [Nagae *et al.* \(2011\)](#) reported a strong association between hypomethylated CpG-poor promoters and tissue-specific patterns of gene expression. We found

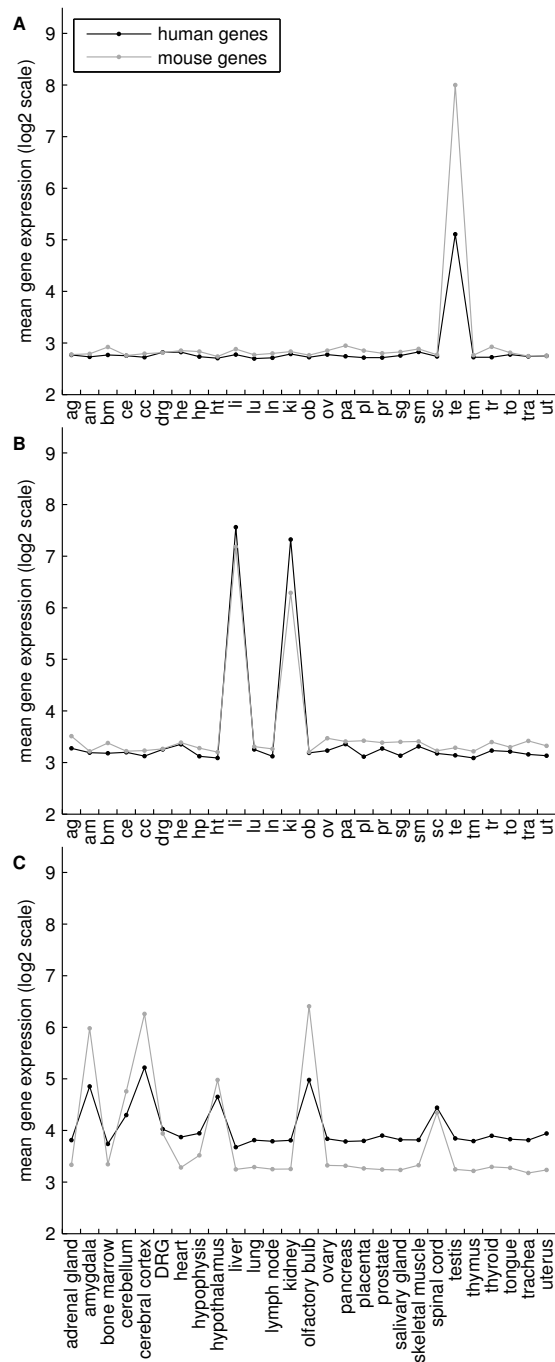


Figure 2.3: Mean expression of genes belonging to three exemplary co-modules. (A) testis-specific co-module; (B) liver and kidney co-module; (C) co-module with two CNS organs assigned: cerebral cortex and olfactory bulb, but with an evidence for the gene expression also in amygdala, hypothalamus, and spinal cord.

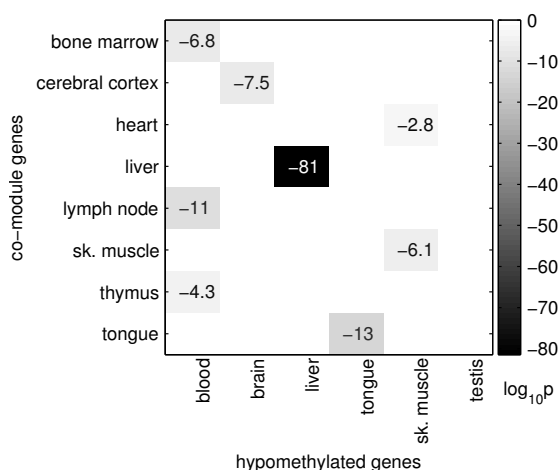


Figure 2.4: Relation between organ-specific expression of genes and organ-specific hypomethylation of their regulatory regions. For every organ-specific co-module we calculated the overlap between human genes belonging to the co-module and genes reported to be hypomethylated specifically in blood, brain, liver, tongue, skeletal muscle, and testis (Nagae *et al.*, 2011). Only co-modules with significant overlap are presented on the heat map. The shade of grey corresponds to the corrected P-values of hypergeometric test in log₁₀ scale.

a very significant overlap between our results and these of Nagae *et al.*, for five out of six common tissues between both studies. For instance, genes that were hypomethylated in a brain-specific manner were over-represented in our cerebral cortex-specific co-module ($p = 3.1 \times 10^{-8}$), and genes hypomethylated specifically in the liver, were overrepresented in liver-specific co-module ($p = 3.6 \times 10^{-82}$). See figure 2.4 for a summary of the results.

2.2.6 Constraint on gene sequence is organ-specific

To check whether sequences of genes assigned to different co-modules evolve under different selective pressure, we computed their nonsynonymous to synonymous substitution ratios (d_N/d_S). For most co-modules the selective pressure did not differ from a random expectation (see Methods for test details). However, genes belonging to CNS-specific co-modules had significantly lower d_N/d_S , and genes from co-modules related to lymph node, liver, and testis had significantly higher

d_N/d_S , than expected by chance (Additional file 3).

2.2.7 Genes' essentiality, duplicability, and age are weakly related to organ-specificity

Looking for other gene characteristics that may be related to different co-modules, we also studied: 1) gene essentiality, 2) gene duplicability, and 3) gene age (for details see Additional file 4). First, we did not detect any significant relation between the co-modules and essentiality of the genes. Second, we found that CNS-related co-modules are significantly enriched in duplicated genes. Further studies are needed to investigate the causality of this relation. Third, we found that human genes from four co-modules and mouse genes from fourteen co-modules had an age distribution significantly different than expected. Importantly, only two co-modules were consistent in the age distribution for mouse and human genes, i.e., the tongue–trachea co-module showed an overrepresentation of young genes (Euteleostomi and later taxonomic levels), and the cerebellum–olfactory bulb co-module showed an overrepresentation of old genes (Bilateria). A few other CNS-related co-modules showed a similar age distribution, but only for mouse genes (figure S1 in Additional file 4). In addition, we found that testis-related genes in mouse were enriched in genes from the Chordate level, and tongue-related genes were particularly young (Euteleostomi and later taxonomic levels). For human only we found that thymus-related genes were enriched in very old genes (Fungi/Metazoa). While these observations were significant in terms of statistics, they were not supported by consistent evidence from both mouse and human. This makes the interpretation of any relationship between gene age and co-modules difficult. Like for duplicability, we believe that further studies with more data will be necessary.

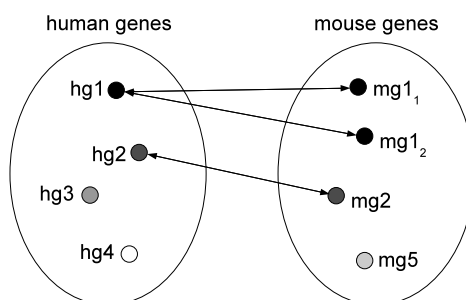


Figure 2.5: Estimating the expression conservation rate (γ) of co-modules. For every co-module we calculated the following numbers: n_{og} — number of orthologous groups, n_{fam_h} — number of all human gene families, n_{fam_m} — number of all mouse gene families, and γ — the expression conservation rate. Here, $n_{og} = 2$, $n_{fam_h} = 4$, $n_{fam_m} = 3$, $\gamma = n_{og}/\min(n_{fam_h}, n_{fam_m}) = 2/3$.

2.2.8 Gene expression is conserved between mouse and human organs

In order to study gene expression evolution between mouse and human, we calculated the rate of expression conservation (γ) for all co-modules resulting from the PPA run on data sets matched through homologous organs. We defined γ (equation 2.1, Methods) as the ratio between the actual number of orthologous groups in a given co-module, and the maximal possible number of orthologous groups, i.e., the minimum of the number of human gene families and the number of mouse gene families present in this co-module (figure 2.5). Thus the values of γ ranged from 0 to 1, with higher values indicating higher gene expression conservation in a given co-module. To assess if γ was significantly higher than expected by chance, we calculated it also for randomly paired mouse and human genes. The median γ for mouse–human orthologous genes was equal to 0.20, while for randomly paired genes the median γ was equal to 0.03 (figure 2.6). Thus, the conservation of gene expression in mammals was significantly higher than expected by chance ($p = 6.5 \times 10^{-6}$, Mann–Whitney U test). γ values for all co-modules are shown in figure 2.7 and in Additional file 5. To determine the upper bound of the expression

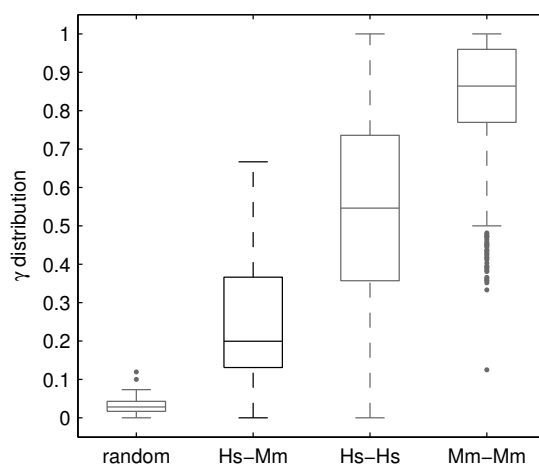


Figure 2.6: Distribution of expression conservation rate (γ). Value of γ was estimated in four different cases: (i) for co-modules containing randomly paired human–mouse genes; (ii) for co-modules containing human–mouse orthologous genes; and for co-modules containing replicated human probe sets (iii) and replicated mouse probe sets (iv).

conservation rate that can be detected by our method with these data, we applied the PPA also to mouse–mouse and human–human data sets constructed by distributing the technical replicates from [Su et al. \(2004\)](#) into two disjoint sets. If these replicates had given identical expression profiles, we would observe $\gamma = 1$. However, due to experimental noise even using the replicate data one expects smaller values for γ . Indeed, this was the case for both comparisons, with a median γ of 0.86 for mouse replicates and a median γ of 0.55 for human replicates. Such low values of γ for data sets with identical underlying biological gene expression suggests that the values of γ which we obtained for human–mouse comparison probably underestimate the actual expression conservation.

2.2.9 Gene expression is conserved between mammalian and fish organs

Given the conservation of expression between mouse and human organs, we asked if this is also true for more distant vertebrates. Using a modified version of the topGO R package ([Alexa et al., 2006](#); Roux and Robinson-Rechavi, unpublished), we

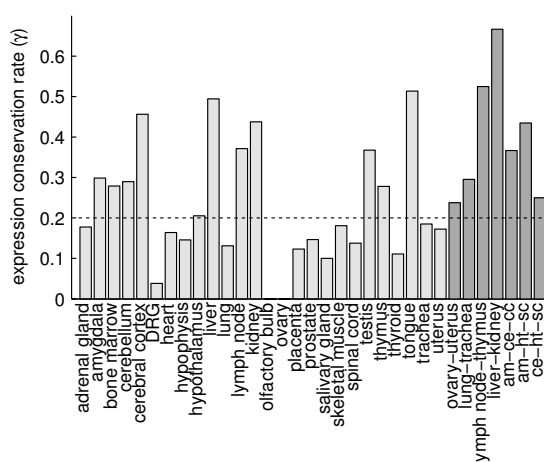


Figure 2.7: Expression conservation rate (γ) for organ-specific and selected system-specific co-modules. The median γ for all co-modules is marked with dotted line. Abbreviation for CNS-specific co-modules: am — amygdala, ce — cerebellum, cc — cerebral cortex, ht — hypothalamus, sc — spinal cord.

assessed organ expression enrichment for the zebrafish orthologs of genes, which belonged to the co-modules detected within the PPA run on data sets matched through organs, and were conserved between mouse and human. In other words, we measured in which zebrafish organs these orthologs were expressed more often than expected by chance. We found conservation of gene expression both for organ-specific co-modules, such as heart or liver, and for nervous system co-modules. For example, genes conserved in the co-module consisting of amygdala, cerebral cortex, hypothalamus, olfactory bulb and spinal cord in mammals, were found to be expressed in the following nervous system organs in zebrafish: retinal ganglion cell, trigeminal placode, cranial ganglion, and spinal cord in fishes. An exception to the general pattern of conservation was that the zebrafish orthologs of mammalian testis-specific genes seemed to be expressed in a wider variety of organs, including Kupffer’s vesicle, the peripheral olfactory organ or the pronephric duct, but not including the zebrafish testis (see Additional file 6).

2.3 Discussion

Our methodology has allowed us to find “natural” modules of mammalian gene expression. In the first PPA run, with genes on the common dimension, we were able to recover the information of organ homology based only on orthologous genes expression patterns. In the second PPA run, with organs on the common dimension, we found organs grouped into homologous systems (between mouse and human), and their functional genes in both species; the latter were enriched in, but not limited to, orthologous genes.

According to our results the whole nervous system, and amygdala, cerebral cortex, hypothalamus and spinal cord in particular, forms a clearly discernible module both in mouse and human. Co-clustering of amygdala, hypothalamus and spinal cord was also reported in [Liao and Zhang \(2006a\)](#). We found several other functionally related co-modules, for instance a co-module containing kidney and liver, a co-module related to the immune system (including lymph node and thymus), a female reproductive system co-module (ovary and uterus), or a respiratory system co-module (lung and trachea). A recent study of [Brawand *et al.* \(2011\)](#) also showed that neural tissues (brain and cerebellum), and kidney and liver, form expression modules in amniotes. Grouping of some of the nervous system organs, was also reported in [Zheng-Bradley *et al.* \(2010\)](#), but it was not possible to know exactly which CNS organs group together, as their annotation was simplified to “brain + nerve”. The only other system reported in [Zheng-Bradley *et al.* \(2010\)](#) combines heart and muscle. In their PCA results heart and muscle formed two distinguishable units, which were then grouped by the authors. Here, we found heart and muscle in a single co-module with the PPA run on data matched by orthologous genes, and in two separate co-modules with the PPA run on data matched by homologous organs. In the latter case the gene scores were higher (figure 2.2),

which suggests that heart and skeletal muscle, although similar, compose two distinct units of expression.

In addition to system-specific co-modules, we also found organ-specific co-modules. Thus with the PPA it is possible to simultaneously detect genes specific for a certain organ and genes shared between organs which form a system. On average, in mouse, there were less organ-specific genes than system-specific genes. No significant difference was found for human genes. All co-modules contained genes whose function was clearly related to the respective organs, justifying our notion of organ/system-specificity for the co-modules. This also confirms that an overexpressed gene has an important role in a given organ or organ system, in agreement with common expectations.

We explored the cause of organ-specific patterns of expression. One possible explanation was proposed by [Nagae *et al.* \(2011\)](#). They discovered that genes with CpG-poor regulatory regions hypomethylated in an organ-specific manner tend to be expressed in an organ-specific manner. Indeed, for all but one of the organs that were included in their study we found that a significant fraction of the genes from our corresponding organ-specific co-module was hypomethylated. The only exceptions were testis-specific genes, for which we did not find evidence of hypomethylation in their promoter regions. However, genes that are specifically hypomethylated in testis tend to have CpG-rich promoters ([Nagae *et al.*, 2011](#)). Thus, further work is needed to understand the regulation of testis-specific expression.

Our analysis of protein-coding gene sequences shows that the selection pressure on gene sequence is organ-specific. In particular, genes of CNS-related co-modules evolve slower in sequence, and genes from the co-modules related to lymph node, liver, and testis evolve faster, than expected by chance. These results are consistent with other reports that compared sequence evolutionary rate between human and

chimpanzee (Khaitovich *et al.*, 2005), and between human and mouse (Gu and Su, 2007). A possible explanation for slower evolution of neural related genes was given by Drummond and Wilke (2008). These authors suggested that the structure and lifetime of tissues composed of neurons make them extremely sensitive to protein misfolding, and thus selection against protein sequence mutations is higher in these tissues. Possibly, the conserved protein sequence might be also related to the higher duplication rate of the genes expressed in CNS. However, this hypothesis needs to be addressed by a more specifically tailored study.

We found that co-module-specific genes are often orthologous between mammals. On average about 20% of the genes present in a given co-module had their orthologs in the same co-module. Note that the co-module-specific gene expression conservation rate (γ) from our analysis is rather underestimated, because of the noise present in the data. Even for human and mouse replicates only 55% and 86% of the genes present in a co-module had their replicate in the same co-module. The latter figures indicate that higher quality data (e.g., RNA-seq) are needed to improve our knowledge of gene expression evolution in mammals (e.g., Brawand *et al.*, 2011; preferably with more organs).

Interestingly, we discovered two organ-specific co-modules with no detectable signs of expression conservation ($\gamma = 0$), i.e., the ovary-specific and the olfactory-specific co-module. The observed lack of expression conservation between mouse and human ovaries might simply be the effect of differences in sampling from two species: the mouse samples came from young, sexually mature individuals, whereas human samples were mostly taken from elderly people (Su *et al.*, 2004). Ovary function varies strongly with age, independently of evolutionary conservation. For the olfactory bulb co-module such an explanation is less likely (even though olfactory sensitivity decreases with age). Rather, the absence of any detectable

sign of expression conservation in this co-module suggests that different genes are involved in olfactory function in mouse and human. Indeed, it has been reported that the olfactory sense genes were shaped by different evolutionary processes in rodents and primates (Young *et al.*, 2002; Zhang and Firestein, 2002; Niimura and Nei, 2007). This shows that with the modular approach it is not only possible to discover “natural” modules of expression, but also to address questions about their evolutionary history since the divergence of two species.

To further study the extent of gene expression conservation, we contrasted the mammalian conserved genes with the expression data from zebrafish. We found that genes expressed in the brains of both mammals were also expressed in the brain of zebrafish. Similarly, genes expressed in mammalian heart or liver were found to be expressed also in their zebrafish homologs. This is a remarkable result indicating that indeed organ/system-specific gene expression evolution is rather slow. The exception was that zebrafish genes orthologous to mammalian testis-specific genes appear to be expressed in a wider variety of organs. This is consistent with previous reports of fast evolution of genes expressed in testis (Xu *et al.*, 2002; Gu and Su, 2007; Voolstra *et al.*, 2007; Brawand *et al.*, 2011).

Our comparative study of homologous organs between mouse and human has several advantages, relative to previous approaches (Yanai *et al.*, 2004; Jordan *et al.*, 2005; Yang *et al.*, 2005; Liao and Zhang, 2006a; Xing *et al.*, 2007; Zheng-Bradley *et al.*, 2010; McCall *et al.*, 2011). First, we analyzed a larger data set than most previous studies, with 27 homologous organs of mouse and human. Second, using the PPA instead of hierarchical clustering of organs, we were able to distinguish homologous modules at different levels of resolution — single organ or organ systems. Third, it is straightforward from our analysis to identify organ-specific or system-specific genes and to further analyze their features, while in most studies

only the Pearson's correlation coefficient between organs is reported. Fourth, in all studies concerning the comparison of orthologous genes or homologous organs expression profiles, one had to decide how to represent gene expression values if a gene is targeted by more than one probe set. Because it is not possible to say which probe set most accurately measures the real expression level of a given gene, some arbitrary choice must be made [e.g., calculating the mean over all probe sets (Yang *et al.*, 2005; Pereira *et al.*, 2009), picking a random probe set (Liao and Zhang, 2006a; Xing *et al.*, 2007), taking the probe set with the highest expression level (Jordan *et al.*, 2005; Gu and Su, 2007), or removing genes covered by multiple probe sets (Yanai *et al.*, 2004; Wang and Rekaya, 2009)]. In the case of the PPA on data sets with organs on a common dimension all probe sets are used. Thus, if at least one of the multiple probe sets mapped to a gene carries an informative signal, the PPA can detect it and automatically find the group of similar probes representing other genes. This is impossible with any of the methods of probe sets pre-processing mentioned before. Notably, as many as 34.9% of human genes and 8.4% of mouse genes were mapped to multiple probe sets. And around half of the multiple probe sets mapped to a given gene were not together in the same co-module (48.6% of human genes and 52% of mouse genes had half or less probe sets together in the same co-module), which is a strong indication that these probe sets do not all correctly represent a gene, or possibly that they represent alternatively spliced forms, which code for different protein isoforms in different organs.

Two recent studies also applied modularization as a mean for cross-species comparative analysis of gene expression data. Yang and Su (2010) used our Iterative Signature Algorithm (ISA; Bergmann *et al.*, 2003; a precursor of the PPA) to identify and compare organ-related modules in human and mouse. Contrary to the PPA, the ISA discovers modules for a single species only. To conduct an

inter-species study, Yang and Su compared modules from two independent ISA runs. They found fewer and smaller modules than we did with the PPA. This may have been a consequence of using only a single threshold for genes and organs. Importantly, they observed little cross-species overlap between the modules both in the organ and gene dimension. Consequently, they concluded that the content of modules in mouse and human diverged extensively. However, they found that modules with corresponding organs in mouse and human usually were enriched for genes of the same biological function. [Brawand *et al.* \(2011\)](#) used the ISA to analyze RNA-seq data from six tissues and ten species, limited to one-to-one orthologous genes. This allowed the identification of several modules, which confirm the correspondence between organ-specific expression and functional annotation of genes. This study did not investigate the evolutionary conservation of organ-specific gene expression, and the detection of functional systems was limited by the few organs studied (i.e., brain and cerebellum, kidney and liver). On the other hand, using ten species allowed the detection of changes of expression in amniote evolution. These examples, and our analysis, illustrate the power of the modular approach to answer diverse questions in evolutionary biology.

2.4 Conclusions

In conclusion, gene expression defines organ-specific or system-specific co-modules. These co-modules contain functionally related genes that are conserved between species. Thus there does exist a conserved modularity of gene expression in vertebrates, and it is related to anatomical modularity (i.e., organs).

2.5 Methods

2.5.1 Gene expression data

We used human and mouse gene expression data of [Su *et al.* \(2004\)](#). This study was performed on the Affymetrix HG-U133A array as well as the custom array GNF1H for human, and on the custom array GNF1M for mouse. In total, expression profiles from 79 human and 61 mouse organs were measured, with 44,928 probe sets for human and 36,182 probe sets for mouse. We only took into account organs belonging to the homologous organ groups (HOGs) defined in the Bgee database ([Bastian *et al.*, 2008](#)) (see <http://bgee.unil.ch/bgee/bgee?page=documentation#sectionHomologyRelationships>). Using the mapping available in the Bgee database we could map 36 human organs and 30 mouse organs to 27 HOGs. See Additional file 7 for the list of HOGs and their corresponding organs. Microarray data were normalized with the `gcrma` package ([Wu *et al.*, 2004](#)) of Bioconductor ([Gentleman *et al.*, 2004](#)).

Before we applied the PPA to the human–mouse data we merged human and mouse organs into 27 HOGs. For every probe set in each HOG the arithmetic mean of the `gcrma` normalized expression values was calculated (each HOGs was represented by at least two microarrays).

To study if it is possible to recover the information about organ homology based on the expression patterns of orthologous genes, we applied the PPA to the data sets consisting of a subset of 8,942 one-to-one orthologous gene pairs (see Mapping Probe sets to Ensembl genes in Methods) and their expression patterns in 27 homologous organ groups in mouse and human. If a gene was matched by more than one probe set on the microarray, we randomly picked one probe set to represent that gene.

To study organ expression conservation between human and mouse we applied

the PPA to the data consisting of expression values for 27 homologous organ groups, 44,928 probe sets for human and 36,182 probe sets for mouse. This time, the probe sets were mapped to their corresponding Ensembl genes after the PPA run.

To estimate the expected values of co-module expression conservation (γ , equation 2.1) when the gene pairs show conserved expression patterns we used replicated experiments as two different data sets, both for mouse and human. Therefore, for each probe set in mouse data and for each probe set in human data we had two vectors of values representing its expression over the organs. We applied the PPA to the data sets that contained 36 organs and 44,928 replicated probe sets for human and 30 organs and 36,182 replicated probe sets for mouse. We did not merge the organs into HOGs, because it was straightforward to pair the organs between replicated experiments.

2.5.2 Mapping probe sets to Ensembl genes

To assign the probe sets to their corresponding mouse or human genes we used the mapping available in Bgee release 6, based on Ensembl release 55. We kept only probe sets which matched to a unique Ensembl gene. A total of 15,123 probe sets corresponding to 13,855 mouse genes, and 23,921 probe sets corresponding to 15,338 human genes, were taken into account in our analysis.

2.5.3 Mouse–human orthologous genes

Homology information of mouse and human genes was retrieved from Ensembl release 55 (Hubbard *et al.*, 2009), using BioMart (Smedley *et al.*, 2009). A total of 10,321 pairs of mouse–human orthologous genes had expression information in the data sets we used (9,982 mouse genes and 9,883 human genes). One-to-one orthologous pairs account for 86.6% (8,942/10,321) of all pairs.

2.5.4 Ping-Pong Algorithm

A detailed description of the algorithm in the general case is given in [Kutalik *et al.* \(2008\)](#). In this specific study, the algorithm starts with ten thousand candidate seeds consisting of randomly chosen homologous organ groups (HOGs), for both runs. Further steps are presented on figure [2.1](#). Here, we only detail the PPA applied to the mouse–human data matched through HOGs: (step 1) the mouse expression data are used to identify the genes that exhibit similar expression in a given set of HOGs. (step 2): this set of genes is then used to refine the set of HOGs by excluding those which have an incoherent expression profile and adding others that behave similarly relative to genes. (step 3): in the next step the human expression data are used to find human genes that exhibit similar expression in a given set of organs. (step 4): similarly to step 2 the set of human genes is used to further refine the set of HOGs. Finally, this refined set of HOGs is used to look for mouse genes that are co-expressed in these HOGs (step 1). This procedure is reiterated until it converges to stable sets of HOGs and mouse and human genes (so-called co-modules). Every HOG and every mouse and human gene in a given co-module have a score assigned (between 0 and 1). The closest the HOG/gene score is to 1, the stronger the association between the HOG/gene and the rest of the co-module.

The PPA was applied to the mouse–human data sets twice. First, the two data sets shared the gene dimension. Second, the two data sets shared the organ dimension. The second experiment was coupled with the control experiment, which aimed to compare two matrices of replicated data within a species. The control experiment was done both on mouse–mouse and human–human data, with organs on the common dimension. We repeated this experiment ten times for each species. In every run the two replicates for each organ were randomly distributed between

the two matrices and a thousand seeds consisting of random HOGs were created.

In every run of the PPA (both types) we used various thresholds for genes and organs, ranging from 2.5 to 6, and from 1 to 4.5, respectively. The thresholding is done by calculating the mean and standard deviation of the gene/organ scores vector and keeping only the elements that are t standard deviation above the mean, where t correspond to the value of the threshold. If the gene threshold is high, then the co-modules will have very similar genes. If it is low, then co-modules will be bigger, with less similar genes. The same applies to the organ threshold and the organs belonging to the co-modules (see http://www2.unil.ch/cbg/homepage/downloads/ISA_tutorial.pdf for detailed explanation).

2.5.5 Post-processing of the PPA results

The procedure described below was applied to the co-modules resulting from the PPA run on data matched through homologous organ groups. As we ran the PPA with different sets of thresholds, redundant modules were obtained. Before further analysis we eliminated this redundancy. For each pair of co-modules we calculated the correlation c_h between human gene scores in the first and in the second co-module, and the correlation c_m between mouse gene scores in the first and in the second co-module. If $c_h \cdot c_m > 0.8$, which implies that the pair of co-modules had a very similar content for both species, the co-module with a higher sum of the two thresholds for human and mouse genes was kept, and the other co-module was disregarded. This procedure reduced the number of co-modules from 556 to 414. Next, we eliminated co-modules that had less than 10 probe sets assigned for at least one species. This procedure reduced the number of co-modules further, to 231. Still, many sets of organs were represented by several overlapping co-modules. Consider two co-modules containing H_1 and H_2 sets of human genes, respectively.

We say that two modules have fully overlapping sets of human genes H_1 and H_2 , if either $H_1 \subseteq H_2$ or $H_2 \subseteq H_1$. For each set of co-modules with fully overlapping sets of human genes the biggest co-module was chosen for the further analysis, and the rest were disregarded. The size of a co-module was defined as the minimum of the two values: 1) the number of human genes in a co-module and 2) the number of mouse genes in a co-module. After this final step, there were 98 co-modules used in further analysis.

In order to assess the rate of gene expression conservation we used only orthologous gene pairs with corresponding probe sets present on both the human and mouse microarrays. The rate of the expression conservation in a co-module was calculated as

$$\gamma = \frac{n_{og}}{\min(n_{fam_h}, n_{fam_m})}, \quad (2.1)$$

where n_{og} is the number of orthologous groups in a given co-module, n_{fam_h} is the number of human gene families in a given co-module for which ortholog(s) are present on the mouse microarray (but not necessarily in the same co-module) and n_{fam_m} is the number of mouse gene families in a given co-module for which ortholog(s) are present on the human microarray (but not necessarily in the same co-module) (figure 2.5). The same procedure was applied to calculate the γ for co-modules from mouse–mouse and human–human comparison, with n_{og} being the number of probe sets present in replicates in a given co-module, and n_{fam_h} , and n_{fam_m} being the total number of probe sets from the first and second experiment present in a given co-module.

To verify if the results of our analysis were different than expected by chance we created lists of random pairs of mouse–human genes. This was done ten times by reshuffling the list of 10,321 mouse–human orthologous pairs, in a way that

kept the same number of one-to-one and many-to-many gene pairs. For every co-module and every list of random gene pairs we recalculated the γ . Finally, for every co-module the mean γ was calculated.

2.5.6 Enrichment analysis of hypomethylated regulatory regions

To determine if hypomethylated regions are over-represented in genes belonging to organ-specific co-modules we used data from the work of [Nagae *et al.* \(2011\)](#). They provided the lists of genes specifically hypomethylated in: brain, tongue, liver, blood, skeletal muscle, and testis. We used these sets of tissue-specific hypomethylated genes, and intersected them with the genes from our organ-specific co-modules. We performed the hypergeometric test to verify if the genes reported in [Nagae *et al.* \(2011\)](#) were overrepresented in any of our co-modules. To correct for multiple testing we applied the Bonferroni correction.

2.5.7 Gene sequence analysis

The one-to-one orthology relationship between mouse and human genes, and the values of d_N (rate of nonsynonymous substitution per codon) and d_S (rate of synonymous substitution per codon) were retrieved from Ensembl version 55 ([Hubbard *et al.*, 2009](#)), using BioMart ([Smedley *et al.*, 2009](#)). We used the set of 12,248 human genes with d_N , d_S , and microarray expression data. To assess whether the genes belonging to a given co-module have d_N/d_S ratios significantly different than expected by chance, we performed a Wilcoxon rank sum test comparing the median d_N/d_S from a co-module to the median d_N/d_S for all human genes. After the Bonferroni correction the significance level was set at $p = 0.0005$. We repeated the same procedure for 10,540 mouse genes.

2.5.8 GO enrichment analysis

Gene ontology (GO) association for all genes mapped to mouse and human probe sets were downloaded from Ensembl release 55, using BioMart. GO enrichment was tested by Fisher's exact test, using the Bioconductor package topGO (Alexa *et al.*, 2006) version 1.12.0. The reference set consisted of all Ensembl genes mapped to probe sets of the microarray used. The "elim" algorithm of topGO was used to eliminate the (tree-like) hierarchical dependency of the GO terms. To correct for multiple testing (98 co-modules tested) the Bonferroni correction was applied. For every co-module only GO categories with corrected P-value lower than 0.05 were reported.

2.5.9 Zebrafish–mouse orthologous genes

Homology information of zebrafish and mouse genes was retrieved from Ensembl release 55 (Hubbard *et al.*, 2009), using BioMart (Smedley *et al.*, 2009). Only mouse genes with expression conserved in mouse–human co-modules were used to find their zebrafish orthologs. A total of 1,892 pairs of zebrafish–mouse orthologous genes was found (1,560 zebrafish genes and 1,026 mouse genes).

2.5.10 Organ enrichment analysis

Associations of zebrafish genes to anatomical ontologies were downloaded from the Bgee database, release 6. Association between genes and organs was based on expression patterns detected in *in situ* hybridization experiments (see Bgee documentation at <http://bgee.unil.ch/bgee/bgee?page=documentation> for more information). Enrichment of expression in organs was tested using a modified version of the topGO package (Alexa *et al.*, 2006; Roux and Robinson-Rechavi,

unpublished). To correct for multiple testing (82 co-modules tested) the Bonferroni correction was applied. For every co-module only zebrafish organs with corrected P-value lower than 0.05 were reported.

Authors' contributions

BP, SB and MRR contributed to the research design. BP and JR gathered the data. ZK and JR contributed analysis tools. BP performed the analysis and wrote the original manuscript. ZK, JR, SB and MRR provided critical comments about the statistical analyses and revised thoroughly the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank P. Lichocki and all members of MRR and SB labs for helpful discussion. We thank J. Daub and R. Studer for comments on the manuscript. We acknowledge funding from Etat de Vaud, and Swiss National Science Foundation ProDoc grant 1206624/1. SB was supported by the Swiss National Science Foundation (grant 31003A_130691/1) and the Swiss Institute of Bioinformatics. MRR was supported by the Swiss National Science Foundation (grant 31003A_133011/1).

Supporting Information

Supporting materials can be downloaded from:

<http://www.biomedcentral.com/1471-2164/13/124/additional>

Additional file 1: The list of co-modules obtained with the PPA run on data matched by orthologous genes. For every co-module we show number of

genes, list of human and mouse organs assigned to the co-module, and the list of enriched GO categories for genes belonging to these co-modules.

Additional file 2: The list of co-modules obtained with the PPA run on data matched by homologous organs. For every co-module we show list of homologous organs, number of human and mouse genes assigned to the co-module, and the lists of enriched GO categories for human and mouse genes belonging to these co-modules.

Additional file 3: The list of co-modules obtained with the PPA run on data matched by homologous organs. For every co-module we show the median d_N/d_S value for mouse and human genes separately. We also show a P-value of Wilcoxon signed rank test. In bold black we marked modules with median d_N/d_S significantly lower than global median, and in bold red we marked co-modules with median d_N/d_S significantly higher than global median.

Additional file 4: Analysis of the relationship between the co-modules and genes' age, duplicability, or essentiality.

Additional file 5: Expression conservation rate (γ) for other system-specific co-modules. The median γ for all co-modules is marked with dotted line. Abbreviations for organ names: ag — adrenal gland; am — amygdala; bm — bone marrow; ce — cerebellum; cc — cerebral cortex; drg — dorsal root ganglion; he — heart; hp — hypophysis; ht — hypothalamus; li — liver; lu — lung; ln — lymph node; ki — kidney; ob — olfactory bulb; ov — ovary; pl — placenta; pr — prostate; sc — spinal cord; te — testis; tm — thymus; to — tongue; tra — trachea; ut — uterus.

Additional file 6: The list of co-modules obtained with the PPA run on data matched by homologous organs. For every co-module we show list of homologous organs assigned to the co-module and the list of the zebrafish organs

enriched in expression of genes orthologous to mouse genes assigned to the co-module.

Additional file 7: The list of homologous organ groups and their corresponding sample names in human and mouse expression data sets.

3

The hourglass and the early conservation models — co-existing evolutionary patterns in vertebrate develop- ment

Barbara Piasecka, Paweł Lichocki, Sven Bergmann,
Marc Robinson-Rechavi

Abstract

Developmental constraints have been postulated to limit the space of feasible phenotypes and thus shape animal evolution. These constraints have been suggested to be the strongest during either early or mid-embryogenesis, which corresponds to the early conservation model or to the hourglass model, respectively. Apparently conflicting results have been reported, but in recent studies of vertebrate transcriptomes the hourglass model has been favored. Studies usually report descriptive statistics calculated for all genes over all developmental time points. This introduces dependencies between the sets of compared genes, and may lead to biased results. Here we overcome this problem using an alternative approach based on a modular analysis. We used the Iterative Signature Algorithm to identify distinct sets of genes (modules) co-expressed specifically in consecutive stages of zebrafish development. We then performed a detailed comparison of several gene properties

between modules, allowing for a less biased and more powerful analysis. Notably, our analysis corroborated the hourglass pattern only at the regulatory level, with sequences of regulatory regions being most conserved for genes expressed in mid-development, but not at the level of gene sequence, gene age or gene expression, in contrast to some previous studies. The early conservation model was supported at the level of gene family size evolution, with gene duplication and introduction being most rare for genes expressed in early development. Finally, for all studied gene properties we observed the least conservation for genes expressed in late development or adult, consistent with both models. Overall, with the modular approach, we showed that different levels of molecular evolution follow different patterns of developmental constraints, and thus that neither the early conservation nor the hourglass model is exclusively valid.

This article was submitted to *PLOS Genetics*.

3.1 Introduction

Developmental constraints have been suggested to play an important role in shaping the evolution of embryonic development in animals. Briefly, the concept of developmental constraints assumes that the scope of developmental mechanisms limits the set of phenotypes that may evolve. Thus, morphological similarities between embryos of different species could reflect these underlying constraints (Poe and Wake, 2004). Two main models of embryonic developmental constraints have been put forward. The *early conservation* model predicts that the highest developmental constraints occur at the beginning of embryogenesis. This corresponds to von Baer's third law (von Baer, 1828), postulating that embryos of different species progressively diverge from one another during ontogeny. However, in modern times, the highest morphological similarity between embryos of different species was observed in the *phylotypic stage* (i.e., mid-embryogenesis) (Seidel, 1960; Sander, 1983; Elinson, 1987). Consequently, Duboule (1994) and Raff (1996) proposed the so-called *hourglass* model, which has since become widely accepted (see, e.g., Prud'homme and Gompel, 2010; Kalinka and Tomancak, 2012). It predicts the highest developmental constraints during mid-embryogenesis.

At the genomic level, the hourglass model was originally linked to the expression of HOX genes in vertebrates (Duboule, 1994). More recently, the emphasis has shifted to the relation, if any, between developmental constraints and the evolution and function of the genome (reviewed in Kalinka and Tomancak, 2012). Different studies have reported several characteristics supporting the hourglass model in vertebrates on the genomic level, e.g.: higher protein sequence similarity (Hazkani-Covo *et al.*, 2005), higher expression conservation (Irie and Kuratani, 2011), and older age (Domazet-Lošo and Tautz, 2010) of genes expressed in the mid-development when compared to the genes expressed early or late in the devel-

opment. However, some of these results do not hold out under a detailed analyses (see Box 3.1 and Supplementary Materials). For example, applying a standard log-transformation (McDonald, 2009; Speed, 2000) to microarray signal intensities used in Domazet-Lošo and Tautz (2010) changes the reported pattern such that it no longer supports the hourglass model (figure 3.1). Moreover, other studies have also found genetic patterns supporting an early conservation model (Roux and Robinson-Rechavi, 2008; Comte *et al.*, 2010).

In most of the studies of developmental constraints the authors compared descriptive statistics of all genes across all developmental time points [e.g., median expression (Roux and Robinson-Rechavi, 2008), weighted mean age (Domazet-Lošo and Tautz, 2010), mean expression correlation (Irie and Kuratani, 2011)]. Such an approach introduces dependencies between the sets of genes which are compared, and consequently can produce results biased by genes expressed at many time points. For example, housekeeping genes contribute to the average gene expression at all time points, and hence dilute trends. To overcome this essential problem, we have used a *modularization* approach, which we applied to the recently published transcriptome data of zebrafish development (Domazet-Lošo and Tautz, 2010). We decomposed the genes into independent sets, i.e., *modules*, that contained genes overexpressed solely in one of seven developmental stages: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile and adult. This decomposition allowed us to compare only sets of genes that have specific functions during embryonic development. For each of the seven modules, we studied five properties of its genes: 1) gene sequence conservation, 2) gene age, 3) gene expression conservation, 4) gene orthology relationships, and 5) regulatory elements conservation.

Here, we show that different levels of molecular evolution follow different

patterns of developmental constraints. First, the regulatory elements are most conserved for transcription factors expressed at mid-development, consistent with the hourglass model. Contrary to what has been reported previously ([Hazkani-Covo *et al.*, 2005](#); [Domazet-Lošo and Tautz, 2010](#); [Irie and Kuratani, 2011](#)), we did not detect the hourglass pattern for gene sequence, age and expression. Second, constraints on gene duplication and on new gene introduction are the strongest in early development, supporting the early conservation model (consistent with [Roux and Robinson-Rechavi, 2008](#)). Finally, all gene properties displayed the least conservation in late development and adult, which is in agreement with both models of developmental constraints.

Box 3.1: Transcriptome Age Index

Recent results of [Domazet-Lošo and Tautz \(2010\)](#) suggest that the oldest transcriptome set is expressed at the phylotypic stage, and that younger sets are expressed during early and late development, which supports the hourglass model. To study the relationship between gene expression, ontogeny and phylogeny, the authors proposed a measure called the “transcriptome age index”, or TAI. The TAI was defined as the mean of the phylogenetic ranks (“phylostrata”) across genes, weighted by their microarray signal intensity values at each developmental stage. Note that the microarray signal intensity values used in [Domazet-Lošo and Tautz \(2010\)](#) displayed a log-normal distribution and spanned from 1 to 10^5 (Supplementary figure A.1). Using these values to calculate TAI made the weights of phylogenetic ranks differ by five orders of magnitude between lowly and highly expressed genes. Consequently, only the most expressed genes (Supplementary figure A.2), and potentially outliers (Supplementary figure A.3), contributed to the hourglass pattern discovered with TAI. We found that applying a standard log-transformation to the intensity values changes the pattern, which then indicates older genes being expressed preferentially in early development (figure 3.1). The use of log-transformed data for microarray intensities is generally encouraged ([McDonald, 2009](#); [Speed, 2000](#)) because it keeps the biological signal, while removing dependency between variance and intensity of the analyzed signals. We present a more detailed re-analysis of the study of [Domazet-Lošo and Tautz \(2010\)](#) in Supplementary Materials.

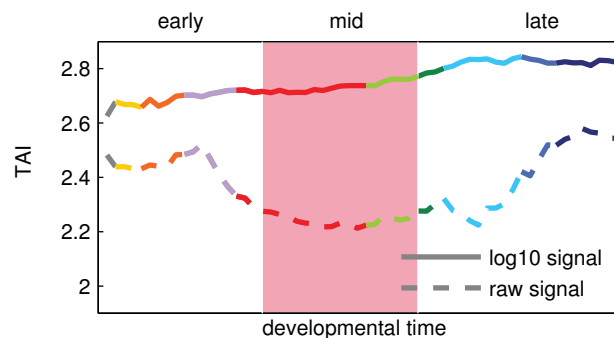


Figure 3.1: Transcriptome age index (TAI) using raw and log-transformed expression signal intensities. A higher TAI value implies that evolutionary younger genes are preferentially expressed at the corresponding time point. The pink shaded area indicates the phylotypic stage. Colors of the curves reflect the main developmental periods and correspond to the colors used in [Domazet-Lošo and Tautz \(2010\)](#).

3.2 Results

3.2.1 Modules

Our goal was to analyze the developmental constraints acting on different gene properties. To this end we identified and analyzed groups of genes co-expressed during distinct developmental stages. We applied the Iterative Signature Algorithm (ISA; Bergmann *et al.*, 2003; Ihmels and Bergmann, 2004) to the zebrafish expression data published by Domazet-Lošo and Tautz (2010), which measured the dynamics of the transcriptome during development with a resolution of 60 time points. The ISA is a modularization algorithm that finds genes with similar expression profiles and groups them into so-called transcription modules. In order to detect modules of genes with specific expression during the zebrafish development, we initialized the ISA with seven idealized expression profiles that corresponded to successive developmental stages (see Supplementary Materials and Supplementary figure A.8).

We obtained seven modules, each containing genes overexpressed during one of the following developmental stages: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile, and adult.

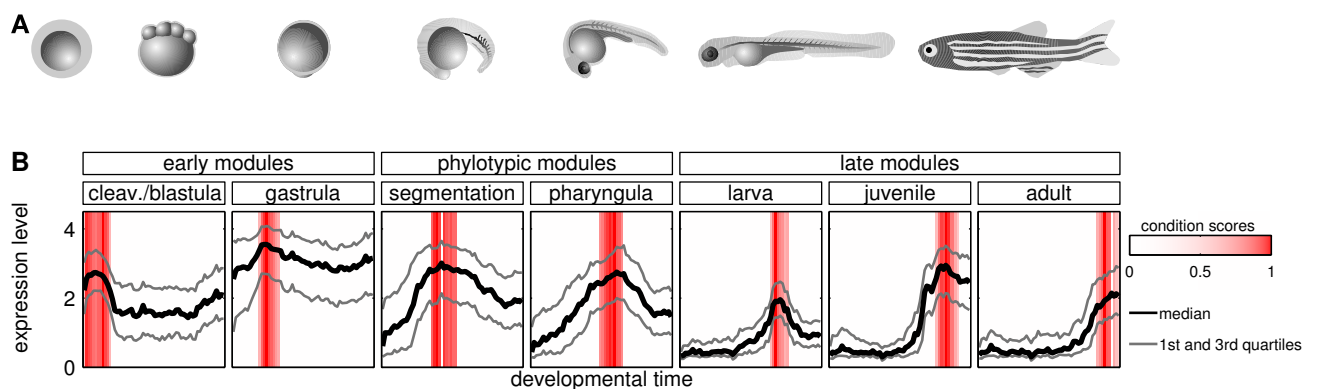


Figure 3.2: Modules of genes with time-specific expression during zebrafish development. A) Zebrafish ontogeny (drawings of the embryos are based upon sketches and photographs from Kimmel *et al.*, 1995). B) Median, 25th and 75th percentiles of expression value of genes in modules. Red bars denote the condition scores assigned to developmental points by the ISA.

pharyngula, larva, juvenile and adult (figure 3.2). Overall, the modules covered the entire development. The phylotypic stage in which the hourglass model predicts the highest evolutionary constraints corresponds to the segmentation and pharyngula modules. We will refer to these two modules as phylotypic modules. The cleavage/blastula and gastrula modules will be referred to as early modules, and larva, juvenile and adult modules as late modules.

The adjacent modules partially overlapped in their gene content. In order to allow for unbiased cross-module comparisons, genes belonging to two modules were kept in the one with the highest ISA gene score (see Methods); this concerned 534 genes in total. The seven modules, i.e., cleavage/blastula, gastrula, pharyngula, segmentation, larva, juvenile and adult, contained 444, 820, 487, 414, 415, 290 and 207 genes, respectively. Overall, 3,077 different genes were present in these modules, which implies a significant reduction of the number of genes being analyzed in comparison to the original data (14,293 genes on the microarray). In particular, the ISA removed the bias related to the genes expressed uniformly across development (i.e., housekeeping genes).

3.2.2 Functional annotation

We verified the function of genes in modules detected by the ISA by comparing them to relevant known lists of genes. We found that the cleavage/blastula module was significantly enriched in maternal genes identified in [Aanes *et al.* \(2011\)](#) (hypergeometric test, $p = 0.01$, see Methods for details of this, and all other statistical tests), and the gastrula module was highly significantly enriched in post-midblastula transition (post-MBT) genes identified in [Aanes *et al.* \(2011\)](#) (hypergeometric test, $p = 2.8 \times 10^{-18}$). We confirmed the relevance of the pharyngula and segmentation modules by verifying that they were enriched in HOX genes, which is consist-

ent with their role in mid-development (Krumlauf, 1994) (hypergeometric test, $p = 5.6 \times 10^{-16}$ and 2.9×10^{-4} , respectively). We did not have any gold standard for genes expressed at the late stages of development. However, since the early and phylotypic modules were enriched in genes with relevant functions, we are confident that the same is true for the late modules.

Moreover, GO enrichment analysis confirmed that genes from the modules were enriched in functions relevant to the respective developmental stages. For example, the cleavage/blastula module was enriched in genes involved in protein phosphorylation and dephosphorylation processes, which is consistent with kinase-dependent control of cell cycle and regulation of mid-blastula transition (MBT) in vertebrates (Hartley *et al.*, 1996; Yarden and Geiger, 1996). The pharyngula module was enriched in genes associated with cell differentiation, and anatomical structure development. Finally, the adult module was enriched in genes involved in responses to environment, although not significantly (Supplementary table A.2).

3.2.3 Sequence conservation

We checked whether the sequences of genes from different modules evolved under different selective pressure. To this end, we calculated the non-synonymous to synonymous substitution ratios (d_N/d_S) for genes in the modules and asked if the ratio was significantly lower for any of them. With the early conservation model, we would expect the lowest d_N/d_S values for genes from early modules. Whereas with the hourglass model, we would expect the lowest d_N/d_S values for genes from the phylotypic modules. In the first five modules, covering whole embryonic development from zygote to larva, the median d_N/d_S was lower than the median d_N/d_S for all genes, but the difference was significant only for the larva module (figure 3.3A, randomization test, $p < 7 \times 10^{-4}$). In the juvenile module, the median

d_N/d_S was higher than the median d_N/d_S for all genes, but the difference was not significant. In the adult module, the median d_N/d_S was significantly higher than the median d_N/d_S for all genes (randomization test, $p = 4.2 \times 10^{-3}$).

These results were consistent with the study by [Roux and Robinson-Rechavi \(2008\)](#), who also reported equally low d_N/d_S values during the entire zebrafish embryogenesis, and a small increase in mid-larva, juvenile and adult. In contrast, [Hazkani-Covo et al. \(2005\)](#) reported an hourglass pattern for protein distance between mouse and human genes expressed during development. However, the trend was not significant. In [Roux and Robinson-Rechavi \(2008\)](#) some evidence for early conservation was reported in mouse. Projecting the genes from zebrafish modules to mouse–human orthologs, we found equal conservation across development (Supplementary figure [A.9](#)). Overall, data analyses support similar evolutionary constraints on sequences of genes expressed during whole embryogenesis of zebrafish, while for mouse more developmental data is needed to be conclusive.

3.2.4 Gene age

The differences in age of genes expressed during different stages of the development have been suggested to be a good indicator of evolutionary constraints ([Irie and Sehara-Fujisawa, 2007](#); [Domazet-Lošo and Tautz, 2010](#)). Thus, we investigated the age of genes belonging to different modules. We dated each gene by its first appearance in the phylogeny and assigned it to one of the five age groups: 1) Fungi/Metazoa, 2) Bilateria, 3) Coelomata+Chordata, 4) Euteleostomi and 5) Clupeocephala+*Danio rerio*. Next, for each module we calculated the age distribution of its genes, i.e., the number of genes belonging to each age group, and compared it with the age distribution of all genes.

For all but the cleavage/blastula module we detected significant age variations

(chi-square goodness of fit test, all $p < 1.3 \times 10^{-5}$), which differed across modules. The oldest genes were overrepresented in the gastrula module, the Bilateria genes were overrepresented in the phylotypic modules, and the youngest genes were overrepresented in the late modules (figure 3.3B). In contrast, Domazet-Lošo and Tautz (2010) reported that genes expressed in early and late development tend to be younger than genes expressed in mid-development, supporting the hourglass model. Yet, that result does not hold for log-transformed gene expression levels (Box 3.1), and is not recovered with measures of gene age other than the transcriptome age index (see Supplementary Materials and Supplementary figure A.6). With the modular approach we observed that the age of expressed genes decreased throughout ontogeny. This pattern suggests that the oldest evolutionary stages tend to express the oldest genes.

3.2.5 Gene family size

Both gene duplication and gene loss can impact phenotypic evolution (Ohno *et al.*, 1970; Zhang, 2003; Nei, 2007; Wang *et al.*, 2006; Demuth and Hahn, 2009). The outcome of these events can be summarized by the resulting gene family size. Consequently, constrained developmental stages should display less changes in gene family size than other stages. To test this hypothesis, for each zebrafish module we calculated the number of its genes that were in 1) one-to-one, 2) one-to-many, 3) many-to-many, and 4) no orthology relation to mouse genes (i.e., no ortholog detectable by the criteria used in Ensembl Compara; Vilella *et al.*, 2009).

We compared the observed distributions with the distribution of the ortholog relationships for all genes. We detected significant variations of the ortholog relationship for the cleavage/blastula module and for all three late modules (chi-square goodness of fit test, all $p < 9 \times 10^{-5}$). Moreover, the pattern of variation itself

differed across different modules. The number of one-to-one orthologs decreased throughout development, and was significantly higher than expected only in the cleavage/blastula module (figure 3.3C). In contrast, the number of genes with no orthologous relationship increased throughout development. It was significantly higher than expected only in the juvenile and adult modules (figure 3.3C), consistent with the excess of “young” genes. A similar pattern was observed for many-to-many orthologs. Finally, the number of one-to-many orthologs was higher than expected only in the larva module, and did not differ from expectation in all other modules.

These results were consistent with Roux and Robinson-Rechavi (2008) in which the genes retained in duplicates after the fish-specific whole genome duplication were reported to have low expression early in the development. Here, we recovered an analogous pattern with the modular approach, showing that the genes expressed early in the development are retained in duplicates less often than genes expressed later. Note that our observation is not limited to whole genome duplication. In addition, we detected the highest number of novel genes amongst genes expressed late in the development.

3.2.6 Expression conservation

Changes in gene expression are one of the main sources of morphological variation (King and Wilson, 1975; Preuss *et al.*, 2004; Carroll, 2005). The developmental constraints on gene expression might differ from those on the gene sequence (Jordan *et al.*, 2004; Yanai *et al.*, 2004; Jordan *et al.*, 2005). Thus, for each module, we compared the mean expression profile of its genes with the mean expression profile of their one-to-one orthologs in mouse. We used two different data sets (Wang *et al.*, 2004; Irie and Kuratani, 2011) with expression values of mouse genes

during the development. The use of two data sets was necessary, because there does not exist a single experiment covering the entire mouse development. The incompatibility of the two microarrays impaired the statistical strength of the analysis. For this reasons the results reported here should be regarded rather as qualitative than quantitative.

Since homology cannot be defined for individual developmental stages between zebrafish and mouse, we first mapped every time point to its broad metastage defined in Bgee ([Bastian *et al.*, 2008](#)) (figure 3.4). Next, we calculated the mean expression level in every metastage. This resulted in six expression values for each gene during the development of mouse and zebrafish: zygote, cleavage, blastula, neurula, organogenesis, and post-embryonic stage. Note that the mouse microarrays did not cover the gastrula stage at all. For each module we calculated the Pearson's correlation between the mean expression of its genes and their mouse orthologs across the six metastages. For the cleavage/blastula module no correlation was detected, probably due to the incompatibility of the two mouse microarrays. For other modules the correlation was positive (figure 3.3D), however due to the low number of data points in the analysis, no correlation values were significant (all $p > 0.01$).

These results stood in contrast with the report by [Irie and Kuratani \(2011\)](#) who showed the highest conservation of gene expression in mid-development. However, a re-analysis of their data suggested that this observation was not significant (see Supplementary Materials and Supplementary figure A.7). Also, both their and our studies shared problems related to the use of two data sets from different sources to cover mouse development. This and the lack of a straightforward homology between ontogenies of different species make it difficult to conclude on the conservation of gene expression during vertebrate development.

3.2.7 Regulatory regions

The *cis*-regulatory hypothesis asserts that most morphological evolution is due to changes in *cis*-regulatory sequences (Stern, 2000; Wray, 2007; Carroll, 2008). A reasonable prediction of this hypothesis is slower *cis*-element turnover in morphologically conserved developmental periods. We examined the presence of highly conserved non-coding elements (HCNEs; Engström *et al.*, 2008) and of transposon-free regions (TFRs; Simons *et al.*, 2007) in the proximity of genes from each module. In the analysis of HCNEs, we counted their number between zebrafish and mouse (detected with 70% identity) in regions of 500 base pairs upstream from the transcription start site. We found that only genes from the phylotypic modules were significantly enriched in HCNEs (hypergeometric test, $p = 8 \times 10^{-6}$, and $p = 1.1 \times 10^{-4}$ for segmentation and pharyngula modules, respectively). We tested the sensitivity of the results by changing the analyzed regions' length to 200 and 1,000 base pairs upstream from the transcription start site, by looking for HCNEs in introns, and using HCNEs detected with identity of 90%. In all cases, we obtained similar results (see Supplementary table A.1). In the analysis of TFRs, we counted the number of genes from each module that have been associated with TFRs in zebrafish. Importantly, these TFRs were reported to be conserved between vertebrates as distant as zebrafish and human. We found that only genes from the pharyngula module were significantly enriched in TFRs (hypergeometric test, $p = 5.7 \times 10^{-7}$).

The highly conserved non-coding elements and transposon-free regions are often associated with developmental regulatory genes, and with transcription factors (TFs) in particular (Sandelin *et al.*, 2004; Woolfe *et al.*, 2005; Vavouri *et al.*, 2007; Engström *et al.*, 2008; Simons *et al.*, 2007). In order to confirm this association, we calculated the fractions of genes with HCNEs or with TFRs in their proximity.

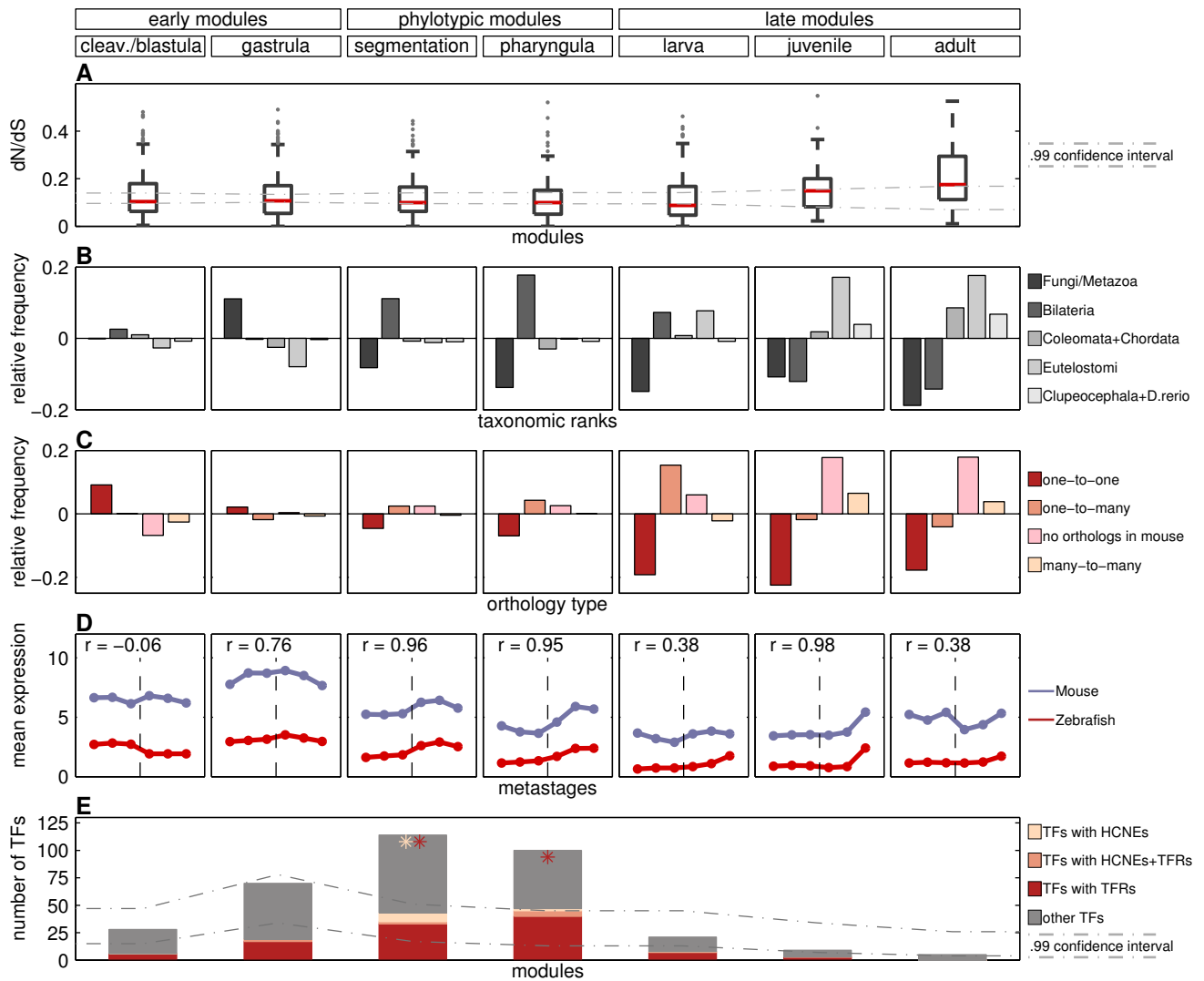


Figure 3.3: Measures of developmental constraints for various gene properties. A) Box and Whisker plot showing non-synonymous to synonymous substitution ratios (d_N/d_S) for genes in the modules. The dash-dotted lines denote confidence interval for the median. B) Observed minus expected age distribution of genes in modules. C) Observed minus expected distribution of orthology type (between zebrafish and mouse) for genes in modules. D) Mean expression level of zebrafish genes in modules, and their one-to-one orthologs in mouse in six developmental metastages. The transition between the two mouse data sets is denoted with the vertical dashed line. The Pearson's correlation coefficients for zebrafish and mouse expression profiles are reported for every module. E) The number of transcription factors (TFs) in modules (whole bar) and their enrichment in highly conserved non-coding elements (HCNEs) and transposon-free regions (TFRs). The stars denote significant enrichment ($p < 0.01$) of TFs in HCNEs (yellow) and in TFRs (red). The dash-dotted lines denote confidence interval for the expected number of TFs in modules.

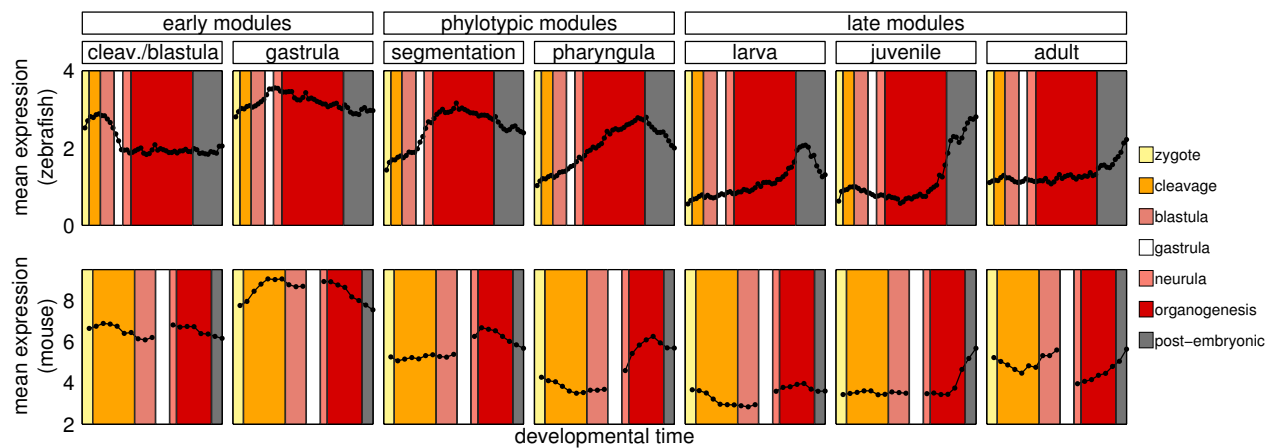


Figure 3.4: Developmental metastages. Mean expression level of zebrafish genes in modules, and of their one-to-one orthologs in mouse. The same colors denote corresponding developmental metastages in zebrafish and mouse.

We observed that for both features this fraction was higher for TFs than for all genes. Importantly, we observed that only the phylotypic modules were enriched in TFs (figure 3.3E). This partially explained the enrichment in HCNEs and TFRs for genes expressed in mid-development. In addition, HCNEs were more often present in the proximity of TFs from the pharyngula module than in the proximity of TFs in general (figure 3.3E; 8.8% of TFs from the pharyngula module had at least one HCNE in their proximity, and only 3.7% of all TFs had at least one HCNEs in their proximity). Also TFRs were more often present in the proximity of TFs from the phylotypic modules than in the proximity of TFs in general (figure 3.3E; 31% and 45% of TFs from the segmentation and pharyngula modules, respectively, had TFRs in their proximity, and only 26% of all TFs had TFRs in their proximity). Consequently, the enrichment in HCNEs and TFRs for genes expressed in the phylotypic stage seems to be related to the regulation of developmental processes.

In addition, we checked for genes that preserved their specific ancestral order in the genome across metazoans (so called conserved ancestral microsyntenic pairs; Irimia *et al.*, 2012) and are known to be involved in the regulation of development. We found that they were slightly overrepresented in the segmentation module, but

only at the limit of statistical significance (see Supplementary Materials).

Finally, we checked for core developmental genes in each module (see [Vavouri et al., 2007](#) for the list of genes). These genes are known to be involved in the regulation of development, and to have highly conserved regulatory regions within different taxa, including, nematodes, insects and vertebrates ([Vavouri et al., 2007](#)). We detected a significant enrichment in these genes only in the pharyngula module (20 core genes; hypergeometric test, $p = 6.9 \times 10^{-19}$), supporting the hourglass model.

3.3 Discussion

Our goal was to study developmental constraints acting on various gene properties. To this end we identified distinct sets of genes with time-specific expression in zebrafish development, i.e., genes expressed in one of the seven consecutive stages: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile and adult. Overall, we analyzed and compared five gene characteristics, namely the conservation of gene sequence, gene expression, and regulatory elements, as well as age and orthology relationships.

Several features do not show any significant pattern over embryonic development, often in contradiction to previous reports. There is notably no evidence for change in selective pressure acting on sequences of protein-coding genes (i.e., d_N/d_S) over development (in contrast to [Hazkani-Covo et al., 2005](#)). Unfortunately, the available data does not allow a strong conclusion concerning the conservation of expression (in contrast to [Irie and Kuratani, 2011](#)), despite the probable importance of this feature in the evolution of development. In this respect, the situation in vertebrates stands in contrast to the relatively clear results in flies ([Kalinka et al., 2010](#)), where the evolution of expression has been shown to be most

constrained in mid-development.

Gene orthology relations support the early conservation model. We show that early stages are less prone to tolerate both gene duplication (consistent with [Roux and Robinson-Rechavi, 2008](#)) and gene introduction. The interpretation of transcriptome age is less straightforward. Our observations suggest that the oldest evolutionary stages tend to express of the oldest genes. It is possible that early stages are evolutionarily oldest, and that this is why they are enriched in oldest genes. Consequently, it is the presence of young genes in a module that would mark relaxed developmental constraints during the corresponding stage. However, neither early nor phylotypic modules are enriched in young genes (Euteleostomi and Clupeocephala+*Danio rerio*), which suggests similar developmental constraints in early and mid-ontogeny. In any case, we do not find any support for the hypothesis that the phylotypic stage would be characterized by the oldest transcriptome (in contrast to [Domazet-Lošo and Tautz, 2010](#)).

While the modularization approach does not support several previous hypotheses of genomic traces of the phylotypic stage, it allows us to distinguish a strong signal of conservation of gene regulation in mid-development. While this had not yet been reported in genomic studies, it is consistent with early descriptions of the phylotypic stage as characterized by HOX genes body patterning activity ([Duboule, 1994](#)). We observed an excess of HCNEs only for genes expressed in the pharyngula module, and an excess of TFRs only for genes expressed in the phylotypic modules. The enrichment in HCNEs and TFRs has been related to developmental regulatory genes, and to transcription factors (TFs) in particular ([Sandelin et al., 2004](#); [Woolfe et al., 2005](#); [Vavouri et al., 2007](#); [Engström et al., 2008](#)). Indeed, we observed that more TFs were expressed in mid-development than in other stages. Also, we showed that a significant proportion of TFs expressed in mid-development

had conserved regulatory regions (i.e., HCNEs and TFRs), in contrast to TFs expressed early or late. Consequently, the enrichment in HCNEs and TFRs for genes expressed in mid-development can be explained by both a higher number of TFs and a higher number of HCNEs and TFRs for these TFs, than for genes expressed earlier or later. Moreover, the pharyngula module was associated with core developmental genes. Overall, these results suggest that mid-developmental processes have extremely high conservation of regulation. This conservation could translate into observed common traits of the phylum expressed at the phenotypic level during mid-development. In addition, core developmental genes are known to be present in different taxa (e.g., nematodes, insects and vertebrates), in each of which they have a conserved regulation that evolved in parallel (Vavouri *et al.*, 2007). This could explain why the phylotypic stage is observed not only in vertebrates (Kimmel *et al.*, 1995), but also in other phyla, e.g., in arthropods (Sander, 1983; Kalinka *et al.*, 2010).

Finally, for all of the features which we have considered there is at least some trend towards weaker evolutionary constraints in the latest stages: d_N/d_S is significantly higher in adults; correlation of expression is lowest for maternal, larval and adult genes; young genes and genes with duplications in fishes or other vertebrates are overrepresented in late modules; and genes expressed in juveniles and adults have the less HCNEs and TFRs. Although not all of these trends are significant, no feature shows stronger conservation in late development or adult. Thus, while different aspects of gene evolution show constraints at different times of development, there appears to be a generally faster evolution of all aspects of larval, juvenile and adult genes. Whether this is due to lower constraints (i.e., less purifying selection) or to stronger involvement in adaptation (i.e., more diversifying selection), remains an open question.

In summary, we studied evidence for, or against, any particular pattern of developmental constraints by considering sets of genes with time-specific expression patterns. Comparing such independent sets of genes with a clear function during embryogenesis resulted in cleaner and more fine-grained characterization of evolutionary patterns than previously reported. Notably, we showed that different levels of molecular evolution follow different patterns of developmental constraints. The sequence of regulatory regions is most conserved for genes expressed in mid-development, consistent with the hourglass model. Gene duplication and new gene introduction is most constrained during early development, supporting the early conservation model. Whereas, all gene properties coherently show the least conservation for the latest stages, consistent with both the early conservation and the hourglass models.

3.4 Methods

3.4.1 Gene expression data

Microarray data of zebrafish development (GSE24616) were downloaded from NCBI's Gene Expression Omnibus (Edgar *et al.*, 2002). This study was performed on the Agilent Zebrafish (V2) Gene Expression Microarray. In total, expression profiles for 60 developmental stages (from unfertilized egg to adults stages) were measured. The last ten stages (55 days–1 year 6 months) were measured separately for male and female. Two replicates were made per time point, resulting in $(50 + 2 \times 10) \times 2 = 140$ microarrays in total. For each microarray, values of `gProcessedSignal` were \log_{10} transformed and normalized as follows. Separately for each replicate, we equalized the expression signals between microarrays using the spike-ins reference, to account for different amounts of RNA present throughout

development. To this aim, we first quantile normalized the expression signal of all spike-ins from all microarrays. Then, for each spike-in level we took the median value of expression signal before and after quantile normalization. This resulted in 10 pairs of expression signals (original signal vs. normalized signal). With linear interpolation between these points, we obtained a piecewise linear curve that defined a mapping from original to normalized expression signals, which we used to equalize the expression signals from all microarrays. This was done by projecting each expression signal onto the piecewise linear curve and calculating the corresponding normalized value. Finally, we quantile normalized the data within replicates and computed the mean value for each gene within replicates. Expression values measured separately for males and females were averaged for each time point.

Microarray data of mouse development were downloaded from Array Express (E-MEXP-51 and E-MTAB-368). The E-MEXP-51 study was performed on (C57BL/6×CBA)F1 mice using Affymetrix GeneChip Murine Genome U74Av2. In total, expression profiles for 10 early developmental stages (zygote, early 2-cell, mid 2-cell, late 2-cell, 4 cell, 8 cell, 16 cell, early blastocyst, mid-blastocyst, late blastocyst) were measured. 2–4 replicates were made per time point. The data were normalized using gcRMA package.

The E-MTAB-368 study was performed on C57BL/6 mice using Affymetrix GeneChip Mouse Genome 430 2.0. In total, expression profiles for 8 mid and late developmental stages (E7.5, E8.5, E9.5, E10.5, E12.5, E14.5, E16.5, E18.5) were measured. 2–3 replicates were made per time point. The data were normalized using gcRMA package.

3.4.2 Mapping probe sets to Ensembl genes

Agilent probe sets were mapped to their corresponding zebrafish genes (Ensembl release 63; [Hubbard *et al.*, 2009](#)) using BioMart ([Smedley *et al.*, 2009](#)). Probe sets which did not map unambiguously to an Ensembl gene were excluded from the analysis. A total of 19,049 probe sets corresponding to 14,293 zebrafish genes were taken into account in our analysis.

Affymetrix probe sets were mapped to their corresponding mouse genes (Ensembl release 63; [Hubbard *et al.*, 2009](#)) using BioMart ([Smedley *et al.*, 2009](#)). Probe sets which did not map unambiguously to an Ensembl gene were excluded from the analysis. For genes that were mapped by several probe sets we used the signal averaged across the probe sets. A total of 2,883 mouse genes mapped by probe sets present on both mouse microarrays were taken into account in the gene expression analysis.

3.4.3 Iterative Signature Algorithm (ISA)

The ISA identifies modules by an iterative procedure. A detailed description of the algorithm in the general case is given in [Bergmann *et al.* \(2003\)](#) (see also http://www2.unil.ch/cbg/homepage/downloads/ISA_tutorial.pdf). In this specific study, the algorithm was initialized with seven candidate seeds, each consisting of one artificial expression profile corresponding to one of the zebrafish developmental stages (see Supplementary Materials for details). Next, these seeds were refined through iterations by adding or removing genes and developmental time points until the processes converge to stable sets, which are referred to as (transcription) modules. Each developmental time point and gene received a score indicating their membership (if non-zero) and contribution to a given module. The closer the score for a gene or developmental time point was to one, the stronger the association

between the gene/developmental time point and the rest of the module.

The ISA was run twice with the following sets of thresholds: 1) $t_g = 1.8$ and $t_c = 1.2$, and 2) $t_g = 1.8$ and $t_c = 1.4$, for genes and developmental time points, respectively. We obtained the pharyngula module only in the case of $t_c = 1.2$, and all other modules with both $t_c = 1.2$ and $t_c = 1.4$. All the modules contained their corresponding idealized profile. For further analysis, we kept a single module per developmental stage. From the pair of modules, we chose the one in which the idealized profile had a higher gene score. Overall, segmentation, pharyngula and juvenile modules were obtained with $t_c = 1.2$, and cleavage/blastula, gastrula, larva, and adult modules were obtained with $t_c = 1.4$.

3.4.4 GO enrichment analysis

Gene ontology (GO) association for all genes mapped by zebrafish probe sets were downloaded from Ensembl release 63 (Hubbard *et al.*, 2009), using BioMart (Smedley *et al.*, 2009). GO enrichment was tested by Fisher's exact test, using the Bioconductor package topGO (Alexa *et al.*, 2006) version 2.2.0. The reference set consisted of all Ensembl genes mapped by probe sets of the microarray used. The "elim" algorithm of topGO was used to eliminate the (tree-like) hierarchical dependency of the GO terms. To correct for multiple testing the Bonferroni correction was applied. For every module GO categories with corrected P-value lower than 0.01 were reported, if less than ten GO categories were significant we reported the top ten (see Supplementary table A.2).

3.4.5 Gene sequence analysis

The orthology relationships, and the values of d_N (number of non-synonymous substitutions per non-synonymous site) and d_S (number of synonymous substi-

tutions per synonymous site) were obtained from Ensembl version 63 (Hubbard *et al.*, 2009). We retrieved zebrafish genes with one-to-one orthologs in *Tetraodon nigroviridis* and *Takifugu rubripes* (the estimated divergence time is 32 million years ago (MYA) between the two pufferfish species and 150 MYA with *Danio rerio*; Benton and Donoghue, 2007) and the pairwise d_N and d_S between *Tetraodon* and *Takifugu* using Biomart (Smedley *et al.*, 2009). We used the set of 7,854 genes having d_N and d_S for Tetraodon–Fugu, and having the expression measured on the zebrafish microarray. For every module we calculated the median d_N/d_S ratio of its k genes, where k was the number of genes having one-to-one relationship with *Tetraodon* and *Fugu* genes. Next, we generated 10,000 sets of k randomly chosen genes. For each set we calculated the median d_N/d_S ratio. Thus, we constructed a sampling distribution of the median d_N/d_S values for a set of k genes. Then we calculated the probability that the median d_N/d_S of the original module was sampled from the constructed distribution. It allowed us to assess if the observed median d_N/d_S ratio was significantly different from the expected median value. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level. We repeated the same procedure for mouse–human genes (see Supplementary Materials).

3.4.6 Gene age analysis

To study the age of genes belonging to different modules we dated the genes by their first appearance in the phylogeny. This consisted of retrieving the age of the oldest node of their Gene tree in Ensembl release 63 (Hubbard *et al.*, 2009). Genes' age was described with one of the following categories: Fungi/Metazoa, Bilateria, Coelomata, Chordata, Eutelostomi, Clupeocephala, and *Danio rerio*. To fit the chi-square test requirements (more than 5 elements in a group) we merged the

genes into five age categories: Fungi/Metazoa, Bilateria, Coelomata + Chordata, Eutelostomi, Clupeocephala + *Danio rerio*. Next, for every module we calculated the age distribution of its genes. We performed chi-square goodness of fit test to compare the observed and expected distributions of age classes in the modules. The expected distribution was estimated by classifying all zebrafish genes into one of the five age categories. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

3.4.7 Zebrafish–mouse orthologous genes

Homology information of zebrafish and mouse genes was retrieved from Ensembl release 63 (Hubbard *et al.*, 2009), using BioMart (Smedley *et al.*, 2009). A total of 17,482 pairs of zebrafish–mouse orthologous genes had expression information in the zebrafish microarray data (14,293 zebrafish genes and 11,322 mouse genes). Among them there were 6,441 one-to-one orthologous pairs, 5,048 one-to-many orthologous pairs, and 2,993 many-to-many orthologous pairs. 2,901 zebrafish genes showed no orthology relationship with mouse genome. From further analysis we excluded 99 “apparent-one-to-one” gene pairs. For every module we calculated the number of genes that were in one-to-one, one-to-many, many-to-many and no orthology relation to mouse genes. Next, we performed chi-square goodness of fit test to compare the observed and expected distributions of orthology classes in the modules. The expected distribution was estimated by classifying all zebrafish genes into one of the four orthology categories. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

3.4.8 Gene expression conservation

To study expression conservation between zebrafish genes assigned to the modules and their mouse one-to-one orthologs, we used gene expression data for 2,883 orthologous gene pairs (the limiting factor being the mapping to both mouse microarrays). For genes that were mapped by several probe sets we averaged their signal across the probe sets for both species. In order to compare gene expression between two species, we first calculated the mean expression for zebrafish genes present in the modules and their one-to-one mouse orthologs. Due to the incompatibility of two mouse microarray data used it was difficult to provide a meaningful comparison of expression for the two species. To calculate the correlation between expression profiles between zebrafish and mouse we reduced their expression profiles to six metastages: zygote, cleavage, blastula, neurula, organogenesis, and post-embryonic stage (see [Bastian et al., 2008](#) for detailed definition of metastage). For every module and every metastage we calculated the mean expression level for zebrafish genes and their mouse one-to-one orthologs, and next we calculated the Pearson correlation coefficient between them.

3.4.9 Highly conserved non-coding elements

Location data for highly conserved non-coding elements (HCNE) between zebrafish and mouse (70% of identity) was retrieved from Ancora ([Engström et al., 2008](#); http://ancora.genereg.net/downloads/danRer7/vs_mouse).

The file *HCNE_danRer7_mm9_70pc_50col.bed.gz* was downloaded and used in the analysis. For each of the 14,293 Ensembl genes considered in our analysis, we calculated the number of HCNE in regions of 500 base pairs upstream from the transcription start site. Next, for every module we performed a hypergeometric test to assess if they were significantly enriched in genes with HCNE. To correct

for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level. In additional analyses, we calculated the number of HCNE in regions of 200 and 1000 base pairs upstream from the transcription start site, as well as in introns. Also, we repeated the analysis with HCNEs of 90% identity (see Supplementary Materials).

3.4.10 Transposon-free regions

Location data for transposon-free regions (TFRs) in zebrafish was retrieved from (Simons *et al.*, 2007; <http://www.biomedcentral.com/content/supplementary/1471-2164-8-470-S1.txt>). First, each TFR was associated with Ensembl ID of its closest transcript from genome assembly Zv6. Then for each Ensembl transcript ID we retrieved an Ensembl gene ID from genome assembly Zv9 (Ensembl release 63; Hubbard *et al.*, 2009). For every module we performed a hypergeometric test to assess if they were significantly enriched in genes with TFRs in their proximity. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

3.4.11 Transcription factors

The set of transcription factors was defined based on GO category annotation: GO:0006355, regulation of transcription, DNA-dependent. Among 14,293 Ensembl genes, 957 were annotated as transcription factors. For every module we performed a hypergeometric test to assess if they were significantly enriched in TFs. Next, we performed a hypergeometric test to assess if the TFs present in the modules were enriched in HCNEs and TFRs. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

Authors' contributions

Conceived and designed the experiments: BP PL MRR. Performed the experiments: BP PL. Analyzed the data: BP PL. Contributed reagents/materials/analysis tools: BP PL SB. Wrote the paper: BP PL SB MRR.

Acknowledgments

We thank Tim Hohm, Anna Kostikova, Zoltan Kutalik, Eyal Privman, and Pavan Ramdya for useful comments on the manuscript. We thank Julien Roux and all members of MRR and SB labs for helpful discussion. We acknowledge the funding from Etat de Vaud, and Swiss National Science Foundation (ProDoc grant 1206624/1). SB was supported by the Swiss National Science Foundation (grant 31003A 130691/1) and the Swiss Institute of Bioinformatics. MRR was supported by the Swiss National Science Foundation (grant 31003A 133011/1) and the Swiss Institute of Bioinformatics.

Supporting Information

Supporting materials are presented in [Appendix A](#).

Figure A1. Total distribution of signal intensity from all 140 microarrays.

Figure A2. TAI hourglass pattern is driven by the subset of most expressed genes. TAI calculated using untransformed (black) and log₁₀-transformed (red) gene expression intensities across zebrafish development. In both cases, TAI is calculated using the entire data sets (dotted line), or using the 25% highest partial concentrations¹ chosen separately for each stage (solid line).

Figure A3. Sensitivity to outliers. (A) Raw expression signal of probe A_15_P161596 across zebrafish development. (B) TAI calculated on non-transformed data across zebrafish development without this probe (red) and the effect of this probe on TAI pattern (grey). (C) TAI calculated on log10-transformed data across zebrafish development without this probe (red) and the effect of this probe on TAI pattern (grey).

Figure A4. TAI calculated using expression intensities of genes, instead of probes, across zebrafish development. For each gene we averaged the signal intensity from all corresponding probes. After this process 16,188 probes' intensities values were reduced to 12,892 genes' intensities values, which were used to weight the phylogenetic ranks of genes (if two different phylostrata were assigned to the same gene, the older one was chosen). (A) non-transformed data was used. (B) log10-transformed data was used.

Figure A5. TAI calculated using genes recoded as present-absent across zebrafish development. At a given stage of development, if the log10-intensity value of a gene is above one (LeProust, 2008), its expression is set to 1, otherwise it is set to 0. Other notations as in figure 3.1.

Figure A6. Alternative measures of transcriptome age. (A) Mean age of genes expressed across zebrafish development; age estimated with the TimeTree database (www.timetree.org). A gene is considered expressed at a given stage of development if its log10-intensity is above one (LeProust, 2008). (B) Difference between median expression profiles of old genes and young genes across zebrafish development. Here, the genes that have emerged before the evolution of Metazoa are considered old and the genes that have emerged since the ancestor of Euteleostomi are considered young. The difference between the two groups is always positive, reflecting that old genes tend to be more expressed than young genes

(Wolf *et al.*, 2009). The results are robust to the choice of cutoffs used to define old and young genes (data not shown). Red dashed line — female data, blue dashed line — male data. Other notations as in figure 3.1.

Figure A7. Correlation between expression levels of genes across developmental time points of mouse, chicken and zebrafish. Field A denotes the early stages, field B denotes the phylotypic stages, and field C denotes the late stages of development.

Figure A8. Artificial expression profiles used to initialize the ISA: pre-MBT, post-MBT, “middle”, pharyngula, larva, “late”, adult. These profiles resulted in modules containing genes expressed specifically in: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile, and adult, respectively.

Figure A9. d_N/d_S ratio for human–mouse one-to-one orthologs. The orthologs were obtained by projecting the genes expressed in the zebrafish modules to their one-to-one orthologs in mouse and human.

Table A1. P-values from HCNE enrichment analyses.

Table A2. The list of modules and their enriched GO categories (biological process).

Outlook

I trust that I have convinced the reader that the modular approach is a very fruitful method of data analysis when applied to evolutionary studies. Here, I would like to briefly summarize my results and discuss possible extension of my work in the field of evo-devo.

In chapter 1, I showed how to improve standard, non-modular approach to study gene expression evolution. While I overcame some of the previous limitations, there are still some to be resolved. First, the proposed new method to estimate the rate of neutral divergence is not time-dependent. One can safely assume that it does not pose a problem for the analysis of distant species (e.g., mouse and human, as suggested by [Jordan *et al.*, 2005](#)), but it certainly biases the calculations when two closely related species are being compared (e.g., human and chimpanzee). Second, the use of the Euclidean distance as a measure of expression similarity, although encouraged, does not totally eliminate the dependence of the metric on the expression specificity. Thus, the question raised in several studies ([Liao and Zhang, 2006b](#); [Liao *et al.*, 2010](#); [Movahedi *et al.*, 2011](#)), whether the broadly or narrowly expressed genes tend to be more conserved in evolution, remains unanswered with this methodology. Both of these limitations should be kept in mind, when gene expression conservation is evaluated with the standard metrics.

In chapter 2, I used a comparative modular approach to study gene expression in vertebrate organs. I found that organs form “natural” modules of expression in mouse and human. Thanks to the PPA I could directly identify organ-specific or system-specific genes and further analyze their features. In contrast, in typical

studies only the Pearson's correlation coefficient between organs is reported. I found that organ-specific and system-specific genes are often orthologous between mammals. Having identified the mammalian modules of expression, I then confirmed their evolutionary conservation also in zebrafish. As discussed in chapter 2, the detected level of gene expression conservation between mammals was rather underestimated due to the high level of noise present in the data. It will be worthwhile in the future to re-evaluate the expression conservation in vertebrates using more and better quality data. Fast development of RNA-Seq technology should allow this in a near future. My results are already partly confirmed by the study of [Brawand *et al.* \(2011\)](#), where the authors used such data and detected some modules of organ-specific genes conserved across ten species (9 mammals and 1 bird). It will be also of interest to extend the study to vertebrate outgroup species, such as *Drosophila*, to look for evolutionary ancient modules of expression.

Another interesting perspective to look at the evolution of gene expression is to study it in relation to the development of an organism. The hourglass and the early conservation models have been proposed to describe patterns of developmental constraints acting on evolution. Recently, the hourglass model has been clearly favored ([Domazet-Lošo and Tautz, 2010](#); [Irie and Kuratani, 2011](#); [Kalinka and Tomancak, 2012](#)). To verify this hypothesis I studied gene expression data from zebrafish development. I identified distinct modules of genes expressed in consecutive stages of zebrafish development and compared their properties. I found that each of the two models explains in part developmental constraints, thus none of them is exclusively valid. In particular, I detected the hourglass pattern only in the conserved regulatory regions, and the early conservation pattern in events of gene duplication and birth. These patterns would have been difficult, if not impossible, to observe without the modular analysis. Other studies that aimed at

discovering patterns of developmental constraints compared descriptive statistics of all genes across all developmental time points. This introduced dependencies between the sets of compared genes, and might resulted in biases due to genes expressed at many time points. While the modular approach gave an interesting and fresh insight into the important evo-devo question, I could only apply it to one vertebrate species, zebrafish. It would be of great interest to compare these results with analogous studies of developmental data of other vertebrates. Only then one could confidently say that my findings refer to vertebrates in general. Both, microarray and RNA-Seq data would be of great use to deepen this study. I hope that such data will be available soon.

I would like to conclude with a comment on the theme that runs through the entire thesis. In all three chapters, I faced methodological issues related to large scale data analysis. In chapter 1, I showed that in the analysis of large scale data many nuances of the common measures of similarity escaped the attention of the researchers. For example, Pearson's correlation coefficient yields zero for two broadly expressed genes (thus similar), contrary to naive expectation. With small scale data, this discrepancy would have been easily noticed with a glance over the results, which is virtually impossible in the case of the analysis of thousands of genes. Consequently, large scale data analysis requires a good understanding of the methodology which is being applied. In chapter 2, I was forced to rely on the GNF human and mouse expression data [Su et al. \(2004\)](#), which consisted of only two replicates (many of them being technical replicates of pooled RNA samples). The low number of replicates potentially limited the number of significant patterns which I could observe with the modular analysis. Of note, the GNF gene atlas is the only publicly available data set of human and mouse expression measured in many organs. Thus, despite the low number of replicates, it has been widely used in

many different studies (out of 1,803 citations of the [Su *et al.*, 2004](#) study, 538 refer directly to the GNF expression data; retrieved from Google Scholar on 27th August 2012). In chapter 3 and appendix A, I discussed the work that was published in Nature ([Domazet-Lošo and Tautz, 2010](#)) even though it dismissed the most standard procedure in microarray data analysis, which is the log-transformation of the signal. I demonstrated the large impact of this omission on the final conclusions — the title hourglass pattern disappeared when log-transformed data were being analyzed. I also showed that the presented pattern was highly influenced by an outlier probe, whose effect was enormous with non-transformed data. The lack of appreciation for rigorous statistical analysis displayed by a journal such as Nature is a bit surprising, since the importance of good experimental design and applying adequate statistical tools have been discussed many times ([Fisher, 1971](#); [Nadon and Shoemaker, 2002](#); [Yang *et al.*, 2002a](#); [Kathleen Kerr, 2003](#); [Allison *et al.*, 2006](#)).

There is no doubt that high-throughput transcriptomic studies have a great potential to deepen our understanding of evolutionary processes. Nevertheless, it will not be fully exploited as long as the co-operation between biologists, bioinformaticians, and statisticians is not well established.

APPENDICES

A

Supporting Materials: The hourglass and the early conservation models — co-existing evolutionary patterns in vertebrate development

Barbara Piasecka, Paweł Lichocki, Sven Bergmann,
Marc Robinson-Rechavi

Re-analysis of previous studies

Domazet-Lošo and Tautz (2010)

In a recent paper, [Domazet-Lošo and Tautz \(2010\)](#) suggested that “*the phylotypic stage does express the oldest transcriptome set and that younger sets are expressed during early and late development*”. To study the relationship between gene expression, ontogeny and phylogeny, the authors proposed a measure called the “transcriptome age index” (TAI). In [Box 3.1](#) we show that the transcriptome age measured with TAI ([Domazet-Lošo and Tautz, 2010](#)) differs strongly if the log10-transformation of the data is applied. Here, we first discuss the advantages of log-transformation, and next we show that also applying several other measures and transformations of the data never reproduces the results reported by [Domazet-Lošo and Tautz \(2010\)](#). On the contrary, we find always that the age of the

transcriptome decreases during development.

The microarray signal intensity values that were used in [Domazet-Lošo and Tautz \(2010\)](#) display a log-normal distribution and span from 1 to 10^5 (figure [A.1](#)). If one uses non-transformed data to calculate TAI, then the five orders of magnitude of difference between expressions of highly and lowly expressed genes translates into five orders of magnitude of difference of the weights of the phylogenetic ranks. In practice, this means that highly expressed genes are given a very high importance, whereas lowly expressed genes are given almost none (figure [A.2](#)). It is disputable whether this is a correct interpretation of the biological reality, because even lowly expressed genes (which are a large majority) do play a role in the development and are shaped by evolutionary forces. Thus, one should not neglect them, if one wishes to interpret the TAI profile in the context of evolutionary constraints or evolutionary adaptation on the whole transcriptome, as in [Domazet-Lošo and Tautz \(2010\)](#). It can also be legitimate to study only a subset of genes, but then this should be done explicitly, and the properties of this subset should be well defined. In order to take into account all genes having a function during development, the data must be transformed, so that the weights of the phylogenetic ranks span a more comparable range. Of note, X-fold difference in signal intensity does not necessarily imply X-fold difference in RNA concentration ([Kahn, 2008](#)).

Moreover, non-log-transformed data are very sensitive to outliers. We identified the probe A_15_P161596 as an outlier (figure [A.3A](#)) which strongly distorts the TAI profile reported in [Domazet-Lošo and Tautz \(2010\)](#). If this single outlier is removed, a TAI peak during gastrulation — which in [Domazet-Lošo and Tautz \(2010\)](#) was given an evolutionary interpretation and linked to the action of the group of genes that emerged in Metazoa — disappears and leaves the gastrulation trend less marked (figure [A.3B](#)). In contrast, the presence of the outlier has little, if

any, influence on the TAI profile calculated on log-transformed data (figure A.3C), showing how the log-transformation leads to a more robust analysis.

Also, in [Domazet-Lošo and Tautz \(2010\)](#), the authors used all 16 188 probes to calculate TAI. Since some of them map to the same gene, this results in signal multiplication for some phylostrata. To overcome this problem, we calculated TAI on data with averaged signal from probes mapped to the same gene. This changes the TAI pattern observed by the authors ([Domazet-Lošo and Tautz](#)): the oldest transcriptome now seem to be expressed in mid-larval stage, instead of the phylotypic stage (figure A.4A). In contrast, the TAI profile calculated on log-transformed data is more robust, as the pattern remains unchanged and does not depend on mapping to probes or genes (figure A.4B).

Another approach to reduce the effect of highly expressed genes is to treat all expressed genes as equally important, i.e., recode as present–absent. This recovers the same pattern as log-transformation (figure A.5). Of note, this approach was suggested in [Domazet-Lošo and Tautz \(2010\)](#), without discussion of the results.

Finally, we searched for alternative measures of the evolutionary age of the transcriptome over ontogeny. We computed: (i) the difference in median expression profile of old genes vs. young genes (figure A.6A; similar to [Roux and Robinson-Rechavi, 2008](#)); and (ii) the mean age of expressed genes (figure A.6B). Both measures recover the decreasing trend over ontogeny. Moreover, measure (i) confirms that the male transcriptome is younger than the female one, consistent with the known fast evolution of male-specific genes ([Ellegren and Parsch, 2007](#)), whereas the original analysis ([Domazet-Lošo and Tautz, 2010](#)) indicated the opposite — younger female transcriptome.

Overall, it seems that the transcriptomic hourglass pattern reported previously ([Domazet-Lošo and Tautz, 2010](#)) is not robust to different methods of analysis.

Irie and Kuratani (2011)

Another analysis suggested that expression diverges less between vertebrate species in the phylotypic stage (Irie and Kuratani, 2011). The authors calculated Spearman correlations between expression profiles of genes of four species: mouse, zebrafish, chicken and frog. They calculated these correlations for all possible pairs of stages, because it was not obvious how to map developmental stages between species. The correlations between expression profiles of genes were reported to be strongest on average at mid-development, supporting the hourglass model.

Here, we reproduced these results for three species: mouse, zebrafish and chicken. We did not re-analyze the frog data, because the expression was measured for tetraploid *Xenopus laevis*, whereas genome annotation available in Ensembl comes from diploid *Xenopus tropicalis*.

We first divided the development of three species into three general stages: early, middle and late (figure A.7). The middle stage contained the time points from the phylotypic stage. The early stage contained the time points preceding the phylotypic stage. And, the late stage contained the time points following the phylotypic stage. We excluded from the analysis the first time point of mouse and zebrafish development, as they had no corresponding time point in the chicken development.

We verified if the middle stage displayed a higher expression similarity than the early stage. To this aim, for each pair of species, we compared the Spearman correlation values between all time points from the early stages of the two species with the Spearman correlation values between all time points from the middle stages of the two species (field A vs. field B on figure A.7). We detected a statistically significant difference only for mouse and chicken (Mann–Whitney U test, $p = 0.018$). However, because the time points from mid and late mouse stages displayed high

correlation with almost any chicken time point, we performed a randomization test to confirm the significance of our observation. We permuted the order of chicken time points and compared again the correlation values between early and middle stages. Notably, among the 100 randomizations as many as 43 comparisons had P-value lower than the previously observed $p = 0.018$. Overall, the pattern of presumably conserved gene expression in middle development, reported in [Irie and Kuratani \(2011\)](#), was not significant for any pair of species.

Supplementary materials and methods

Artificial expression profiles for the ISA

We initialized the ISA with seven artificial expression profiles corresponding to consecutive developmental stages. Our main goal was to compare genes expressed in early, mid and late development. The early genes are known to divide into maternal genes (pre-MBT) and zygotic genes (post-MBT) ([Aanes *et al.*, 2011](#)). Consequently, we originally envisioned four artificial expression profiles: pre-MBT, post-MBT, middle and late. During the ISA run, these profiles resulted in four modules containing genes with expression limited to cleavage/blastula, gastrula, segmentation and juvenile stages, respectively. To cover the entire development we added three other artificial profiles corresponding to the missing stages (pharyngula, larva and adult) and we run ISA again. The seven profiles used to run the ISA are shown on the figure [A.8](#).

Sequence conservation between mouse and human

In the main text, we investigated the conservation level of sequences of protein-coding genes in fishes. Here, we repeated this analysis by projecting the genes

expressed in the seven modules to mouse–human orthologs. The orthology relationships, and the d_N and d_S values were obtained from Ensembl version 63 (Hubbard *et al.*, 2009). We retrieved 6,039 zebrafish genes with one-to-one orthologs in mouse and human (the estimated divergence time is 61.5 MYA between the two mammalian species and 416 MYA with *Danio rerio*; Benton and Donoghue, 2007), and the pairwise d_N/d_S between mouse and human genes using Biomart (Smedley *et al.*, 2009). Other settings and the statistical analysis were the same as in the main text (see [Methods](#)). We found a good agreement between results reported in the main text and for mouse–human orthologs (compare figure 3.3A with figure A.9).

Highly conserved non-coding elements

We tested the sensitivity of the observed enrichment of HCNEs for genes expressed in mid-development, reported in the main text. To this aim, for each of the 14 293 Ensembl genes considered in our analysis, we calculated the number of HCNEs (70% identity) in regions of 200, and 1000 base pairs upstream from the transcription start site (TSS), as well as in the intronic regions. Also, we repeated the analysis looking for HCNEs in regions of 500 bp upstream from the transcription start site (as in the main text), but for HCNEs of 90% identity. To this aim we downloaded and used the file *HCNE_danRer7_mm9_90pc_50col.bed.gz*. Other settings and the statistical analysis were the same as in the main text (see [Methods](#)). The results of all four additional analyses are in a good agreement with the results reported in the main text (table A.1).

Table A.1: P-values from HCNE enrichment analyses.

	200bp	500bp	1000bp	intron	500bp(90%)
segmentation module	4.0e-3	8.0e-6	2.2e-7	2.5e-5	7.9e-1
pharyngula module	1.4e-3	1.1e-4	2.3e-7	2.0e-4	6.5e-4

The column in bold corresponds to the case reported in the main text.

Microsynteny conservation

We checked for modules' enrichment in genes belonging to conserved ancestral microsyntenic pairs (CAMPs; [Irimia *et al.*, 2012](#). From the list of 260 zebrafish CAMPs (Irimia, private communication) we selected 75 gene pairs involved in developmental regulation, i.e., “bystander gene + trans-dev gene”. Both, bystander and trans-dev genes were reported to have conserved introns sequences. Thus, the trans-dev genes could potentially overlap with genes for which we detected enrichment in HCNEs in introns, as well as in the regions 1000 bp upstream from the TSS (CAMPs were shown to have very short intergenic regions, in some cases < 1kb). We crossed the list of trans-dev genes with the list of genes from each module. We performed hypergeometric test to assess if the overlap between genes was significant. To correct for multiple testing we applied the Bonferroni correction. The number of CAMP-trans-dev genes in the seven modules were the following: 1, *n.s.*; 6, *n.s.*; 7, $p = 0.018$; 4, *n.s.*; 2, *n.s.*; 1, *n.s.*; 0, *n.s.* The overrepresentation of trans-dev genes in the segmentation module stays in agreement with enrichment in HCNE detected in introns and in regions 1000 bp upstream from the TSS for genes belonging to this module. We also checked for enrichment in the remaining 185 CAMPs. Although they were reported to often be co-expressed, we did not find any such pair in our modules.

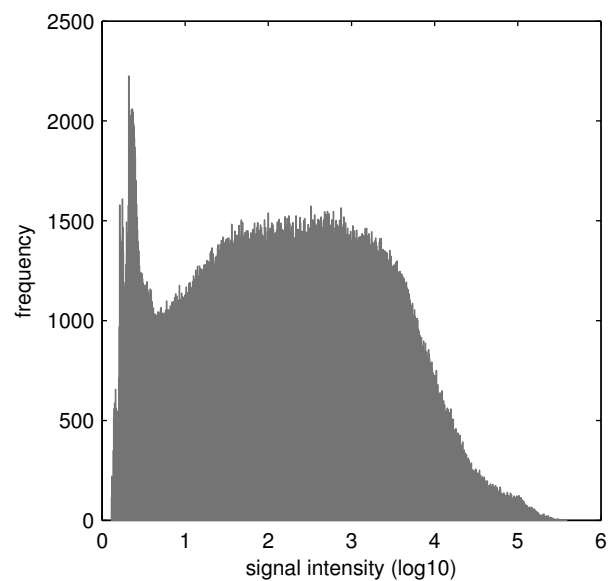


Figure A.1: Total distribution of signal intensity from all 140 microarrays.

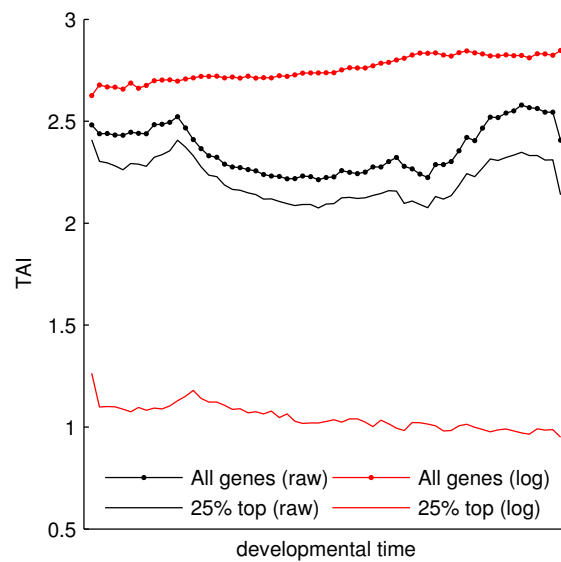


Figure A.2: TAI hourglass pattern is driven by the subset of most expressed genes. TAI calculated using untransformed (black) and log₁₀-transformed (red) gene expression intensities across zebrafish development. In both cases, TAI is calculated using the entire data sets (dotted line), or using the 25% highest partial concentrations¹ chosen separately for each stage (solid line).

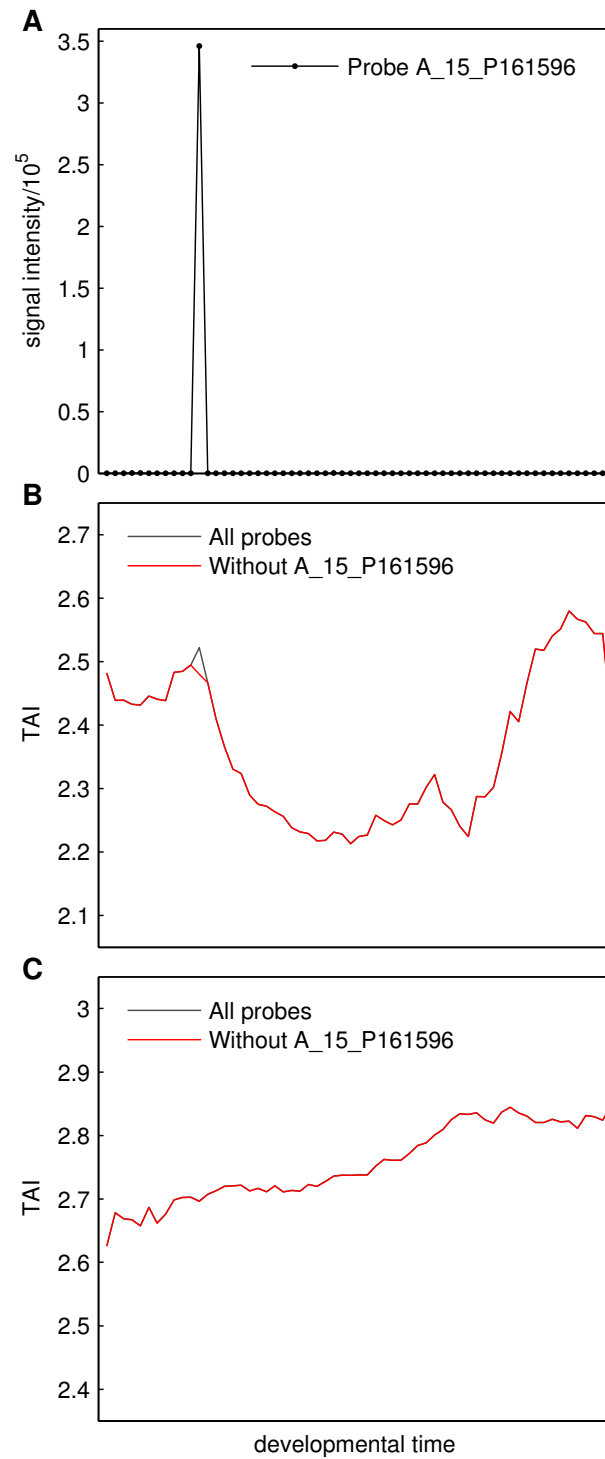


Figure A.3: Sensitivity to outliers. (A) Raw expression signal of probe A_15_P161596 across zebrafish development. (B) TAI calculated on non-transformed data across zebrafish development without this probe (red) and the effect of this probe on TAI pattern (grey). (C) TAI calculated on log₁₀-transformed data across zebrafish development without this probe (red) and the effect of this probe on TAI pattern (grey).

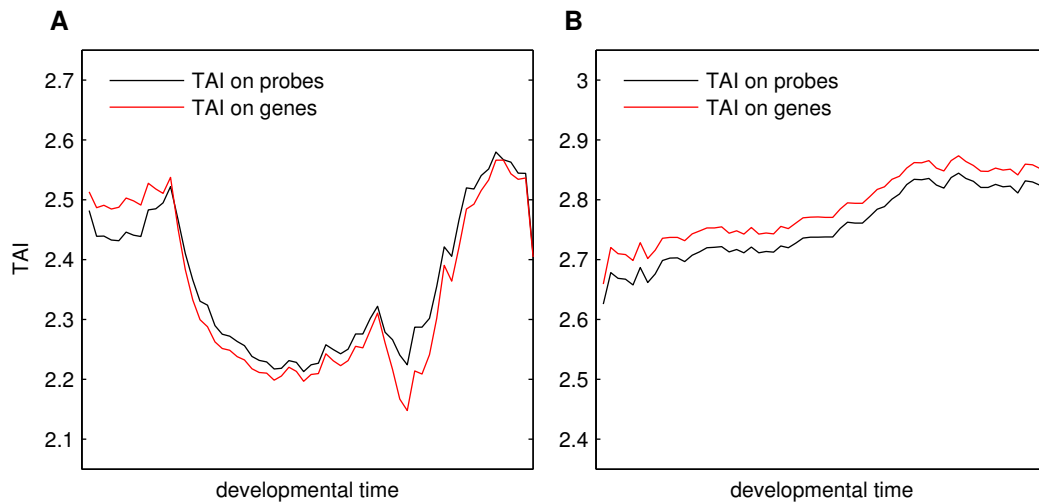


Figure A.4: TAI calculated using expression intensities of genes, instead of probes, across zebrafish development. For each gene we averaged the signal intensity from all corresponding probes. After this process 16 188 probes' intensities values were reduced to 12 892 genes' intensities values, which were used to weight the phylogenetic ranks of genes (if two different phylostrata were assigned to the same gene, the older one was chosen). (A) non-transformed data was used. (B) log10-transformed data was used.

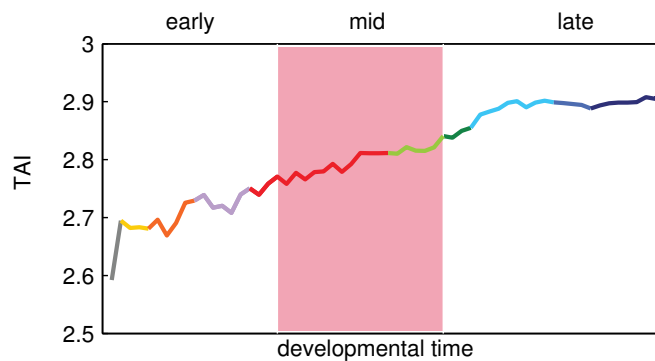


Figure A.5: TAI calculated using genes recoded as present-absent across zebrafish development. At a given stage of development, if the log10-intensity value of a gene is above one (LeProust, 2008), its expression is set to 1, otherwise it is set to 0. Other notations as in figure 3.1.

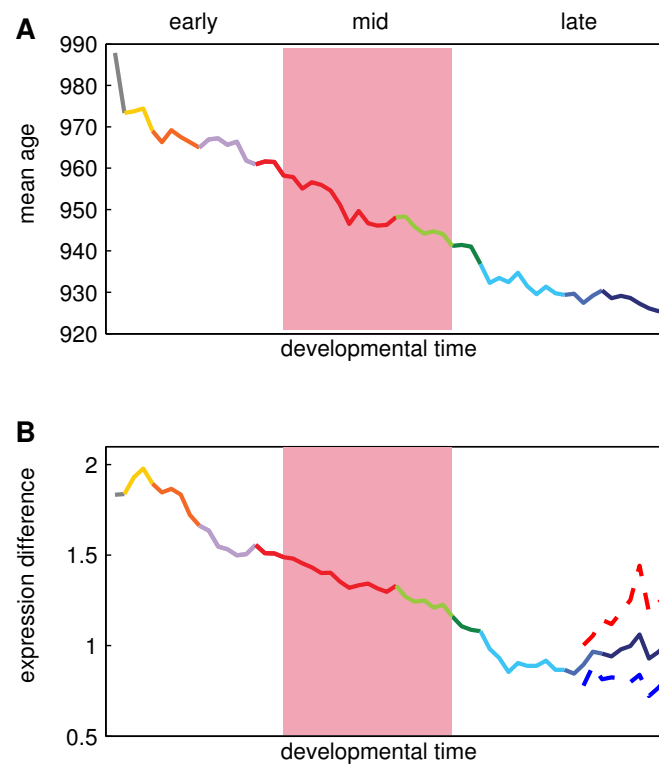


Figure A.6: Alternative measures of transcriptome age. (A) Mean age of genes expressed across zebrafish development; age estimated with the TimeTree database (www.timetree.org). A gene is considered expressed at a given stage of development if its log₁₀-intensity is above one (LeProust, 2008). (B) Difference between median expression profiles of old genes and young genes across zebrafish development. Here, the genes that have emerged before the evolution of Metazoa are considered old and the genes that have emerged since the ancestor of Euteleostomi are considered young. The difference between the two groups is always positive, reflecting that old genes tend to be more expressed than young genes (Wolf *et al.*, 2009). The results are robust to the choice of cutoffs used to define old and young genes (data not shown). Red dashed line — female data, blue dashed line — male data. Other notations as in figure 3.1.

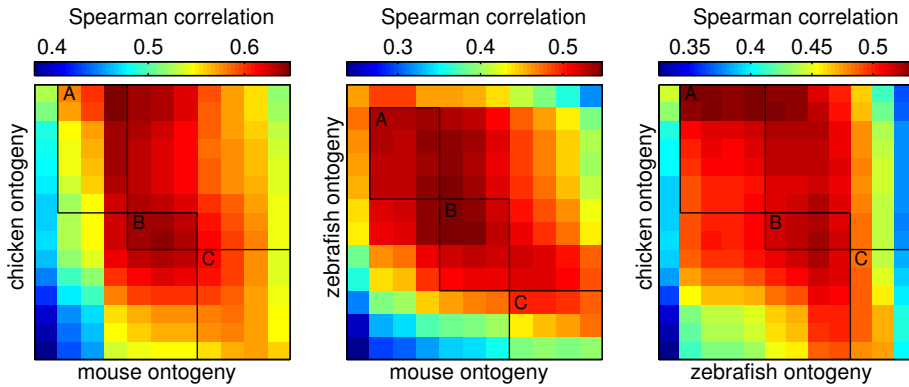


Figure A.7: Correlation between expression levels of genes across developmental time points of mouse, chicken and zebrafish. Field A denotes the early stages, field B denotes the phylotypic stages, and field C denotes the late stages of development.

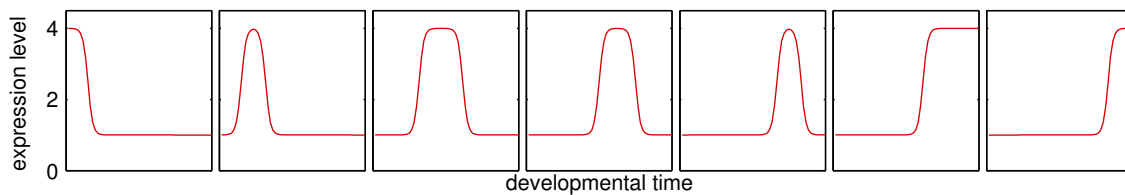


Figure A.8: Artificial expression profiles used to initialize the ISA: pre-MBT, post-MBT, “middle”, pharyngula, larva, “late”, adult. These profiles resulted in modules containing genes expressed specifically in: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile, and adult, respectively.

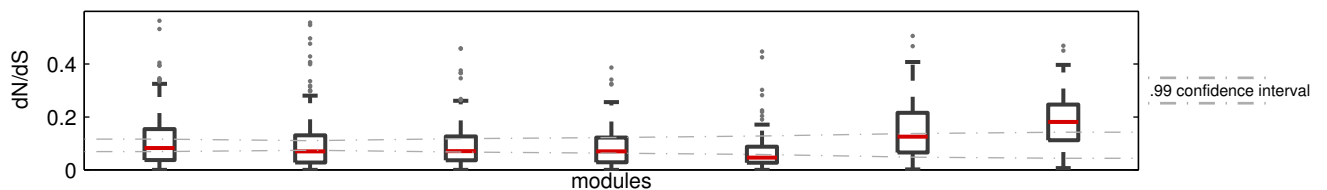


Figure A.9: d_N/d_S ratio for human-mouse one-to-one orthologs. The orthologs were obtained by projecting the genes expressed in the zebrafish modules to their one-to-one orthologs in mouse and human.

Table A.2: The list of modules and their enriched GO categories (biological process).

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
Module 1						
GO:0006468	protein amino acid phosphorylation	446	29	13.7	1.00E-04	7.00E-04
GO:0090244	Wnt receptor signaling pathway involved in somitogenesis	2	2	0.06	9.40E-04	6.58E-03
GO:0006470	protein amino acid dephosphorylation	98	9	3.01	3.09E-03	2.16E-02
GO:0031290	retinal ganglion cell axon guidance	24	4	0.74	5.70E-03	3.99E-02
GO:0043149	stress fiber assembly	5	2	0.15	8.84E-03	6.19E-02
GO:0090090	negative regulation of canonical Wnt receptor signaling pathway	5	2	0.15	8.84E-03	6.19E-02
GO:0021915	neural tube development	31	4	0.95	1.43E-02	9.98E-02
GO:0042451	purine nucleoside biosynthetic process	8	2	0.25	2.33E-02	1.63E-01
GO:0042455	ribonucleoside biosynthetic process	8	2	0.25	2.33E-02	1.63E-01
GO:0046129	purine ribonucleoside biosynthetic process	8	2	0.25	2.33E-02	1.63E-01
Module 2						
GO:0016055	Wnt receptor signaling pathway	80	14	4.43	1.10E-04	7.70E-04
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	6	4	0.33	1.30E-04	9.10E-04
GO:0042664	negative regulation of endodermal cell fate specification	6	4	0.33	1.30E-04	9.10E-04
GO:0035468	positive regulation of signaling pathway	30	8	1.66	1.70E-04	1.19E-03
GO:0010159	specification of organ position	3	3	0.17	1.70E-04	1.19E-03
Continued on next page						

Table A.2 – continued from previous page

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
GO:0035050	embryonic heart tube development	70	21	3.88	2.00E-04	1.40E-03
GO:0001706	endoderm formation	18	9	1	2.80E-04	1.96E-03
GO:0060218	hemopoietic stem cell differentiation	7	4	0.39	2.90E-04	2.03E-03
GO:0007420	brain development	149	21	8.26	4.10E-04	2.87E-03
GO:0030903	notochord development	31	10	1.72	5.00E-04	3.50E-03
GO:0014028	notochord formation	4	3	0.22	6.50E-04	4.55E-03
GO:0001522	pseudouridine synthesis	9	4	0.5	9.40E-04	6.58E-03
GO:0045893	positive regulation of transcription, DNA-dependent	47	9	2.6	9.50E-04	6.65E-03
Module 3						
GO:0009952	anterior/posterior pattern formation	91	19	3.45	1.10E-04	7.70E-04
GO:0048741	skeletal muscle fiber development	14	5	0.53	1.10E-04	7.70E-04
GO:0030510	regulation of BMP signaling pathway	15	5	0.57	1.70E-04	1.19E-03
GO:0043049	otic placode formation	15	5	0.57	1.70E-04	1.19E-03
GO:0030901	midbrain development	15	5	0.57	1.70E-04	1.19E-03
GO:0021523	somatic motor neuron differentiation	4	3	0.15	2.10E-04	1.47E-03
GO:0042694	muscle cell fate specification	4	3	0.15	2.10E-04	1.47E-03
GO:0021508	floor plate formation	9	4	0.34	2.20E-04	1.54E-03
GO:0033334	fin morphogenesis	57	9	2.16	2.60E-04	1.82E-03
Continued on next page						

Table A.2 – continued from previous page

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
GO:0007156	homophilic cell adhesion	58	9	2.2	3.00E-04	2.10E-03
GO:0007517	muscle organ development	59	16	2.23	3.00E-04	2.10E-03
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	71	10	2.69	3.10E-04	2.17E-03
GO:0009888	tissue development	329	41	12.46	3.40E-04	2.38E-03
GO:0031016	pancreas development	39	7	1.48	5.70E-04	3.99E-03
GO:0030182	neuron differentiation	156	27	5.91	5.90E-04	4.13E-03
GO:0009953	dorsal/ventral pattern formation	65	9	2.46	7.10E-04	4.97E-03
GO:0007399	nervous system development	326	60	12.34	8.40E-04	5.88E-03
GO:0007223	Wnt receptor signaling pathway, calcium modulating pathway	21	5	0.8	9.30E-04	6.51E-03
GO:0021984	adenohypophysis development	9	5	0.34	9.70E-04	6.79E-03
GO:0001708	cell fate specification	36	9	1.36	1.06E-03	7.42E-03
Module 4						
GO:0030154	cell differentiation	422	39	13.18	1.20E-04	8.40E-04
GO:0030902	hindbrain development	57	11	1.78	2.30E-04	1.61E-03
GO:0050769	positive regulation of neurogenesis	12	4	0.37	3.80E-04	2.66E-03
GO:0048663	neuron fate commitment	13	4	0.41	5.30E-04	3.71E-03
GO:0048593	camera-type eye morphogenesis	47	9	1.47	8.60E-04	6.02E-03
GO:0030900	forebrain development	51	7	1.59	9.60E-04	6.72E-03
Continued on next page						

Table A.2 – continued from previous page

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
GO:0051091	positive regulation of transcription factor activity	7	3	0.22	9.60E-04	6.72E-03
GO:0030901	midbrain development	15	4	0.47	9.70E-04	6.79E-03
GO:0021915	neural tube development	31	5	0.97	2.50E-03	1.75E-02
GO:0002043	blood vessel endothelial cell proliferation involved in sprouting angiogenesis	3	2	0.09	2.86E-03	2.00E-02
Module 5						
GO:0007602	phototransduction	10	4	0.28	1.10E-04	7.70E-04
GO:0006813	potassium ion transport	79	9	2.18	3.10E-04	2.17E-03
GO:0018298	protein-chromophore linkage	13	4	0.36	3.40E-04	2.38E-03
GO:0007156	homophilic cell adhesion	58	7	1.6	1.03E-03	7.21E-03
GO:0006836	neurotransmitter transport	36	8	1	1.62E-03	1.13E-02
GO:0006814	sodium ion transport	52	6	1.44	2.96E-03	2.07E-02
GO:0007267	cell-cell signaling	41	8	1.13	3.14E-03	2.20E-02
GO:0007194	negative regulation of adenylate cyclase activity	5	2	0.14	7.21E-03	5.05E-02
GO:0007268	synaptic transmission	21	6	0.58	1.04E-02	7.25E-02
GO:0006208	pyrimidine base catabolic process	6	2	0.17	1.06E-02	7.43E-02
Module 6						
GO:0006805	xenobiotic metabolic process	3	2	0.05	9.20E-04	6.44E-03
GO:0006584	catecholamine metabolic process	6	2	0.11	4.44E-03	3.11E-02
Continued on next page						

Table A.2 – continued from previous page

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
GO:0019882	antigen processing and presentation	20	3	0.35	4.95E-03	3.47E-02
GO:0006022	aminoglycan metabolic process	22	3	0.39	6.52E-03	4.56E-02
GO:0046686	response to cadmium ion	8	2	0.14	8.10E-03	5.67E-02
GO:0009607	response to biotic stimulus	47	4	0.83	9.29E-03	6.50E-02
GO:0000272	polysaccharide catabolic process	9	2	0.16	1.03E-02	7.21E-02
GO:0006026	aminoglycan catabolic process	9	2	0.16	1.03E-02	7.21E-02
GO:0055114	oxidation reduction	409	14	7.23	1.35E-02	9.42E-02
GO:0006144	purine base metabolic process	11	2	0.19	1.54E-02	1.08E-01
Module 7						
GO:0043687	post-translational protein modification	748	16	8.26	7.70E-03	5.39E-02
GO:0050896	response to stimulus	622	16	6.87	9.40E-03	6.58E-02
GO:0051707	response to other organism	40	3	0.44	9.60E-03	6.72E-02
GO:0006950	response to stress	329	9	3.63	1.04E-02	7.28E-02
GO:0006508	proteolysis	391	10	4.32	1.10E-02	7.70E-02
GO:0051715	cytolysis of cells of another organism	1	1	0.01	1.10E-02	7.70E-02
GO:0044403	symbiosis, encompassing mutualism through parasitism	1	1	0.01	1.10E-02	7.70E-02
GO:0051801	cytolysis of cells in other organism involved in symbiotic interaction	1	1	0.01	1.10E-02	7.70E-02
GO:0031640	killing of cells of another organism	1	1	0.01	1.10E-02	7.70E-02
Continued on next page						

Table A.2 – continued from previous page

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
GO:0070193	synaptonemal complex organization	1	1	0.01	1.10E-02	7.70E-02

Annot. — total number of genes annotated with a given GO category; Sign. — number of (significant) genes in the module annotated with a given GO category; Expect. — expected number of genes in the module annotated with a given GO category; elim p — P-value from “elim” algorithm of topGO, bonf p — P-value after Bonferroni correction.

B

Poster awarded the Best Poster Prize of SIB days 2012
and the 2nd Best Poster Award at ECCB 2012

In search of lost hourglass

Developmental constraints on transcriptome evolution vary for different molecular features

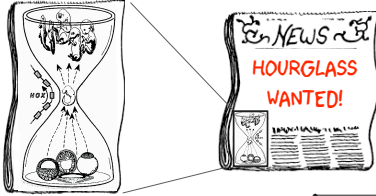
Barbara Piasecka^{1,2,4}, Paweł Lichocki³, Marc Robinson-Rechavi^{1,4}

¹Department of Ecology and Evolution, University of Lausanne, Switzerland

²Department of Medical Genetics, University of Lausanne, Switzerland

³Laboratory of Intelligent Systems, École Polytechnique Fédérale de Lausanne, Switzerland

⁴Swiss Institute of Bioinformatics, Lausanne, Switzerland



"TWO MAIN HYPOTHESES OF THE EVOLUTION OF EMBRYONIC DEVELOPMENT HAVE BEEN PUT FORWARD SO FAR. FIRST, AN EARLY CONSERVATION MODEL PREDICTS THAT THE HIGHEST CONSERVATION OCCURS AT THE BEGINNING OF EMBRYOGENESIS. IT DATES BACK TO KARL VON BAER WHO POSTULATED THAT EMBRYOS OF DIFFERENT SPECIES PROGRESSIVELY DIVERGE FROM ONE ANOTHER DURING ONTOGENY. SECOND, AN HOURGLASS MODEL PREDICTS THAT THE HIGHEST CONSERVATION CAN BE FOUND DURING MID-EMBRYOGENESIS. IT HAS BEEN PROPOSED WHEN THE MORPHOLOGICAL VARIATION IN THE EARLY STAGES OF DEVELOPMENT WAS OBSERVED. NOWADAYS, THE HOURGLASS MODEL IS COMMONLY ACCEPTED, ALTHOUGH ITS MOLECULAR SIGNATURE HAS BEEN ELUSIVE."

LET'S FIND IT!



SO, LET'S COMPARE STATISTICS OF ALL EXPRESSED GENES OVER ALL DEVELOPMENTAL TIME-POINTS?

NO, IT'S A COMMON APPROACH, BUT IT'S BIASED BY HOUSE-KEEPING GENES. WE SHOULD DECOMPOSE THE GENES INTO MODULES ACCORDING TO WHEN THE GENES ARE EXPRESSED, AND THEN COMPARE GENES FROM THESE MODULES.

OK, I RUN ISA ALGORITHM AND I HAVE THE MODULES. FIRST, I'LL CHECK DN/DS ...

IT SEEMS THE GENE SEQUENCE IS LESS CONSERVED IN ADULTS, BUT EQUALLY CONSERVED DURING ENTIRE DEVELOPMENT.

OK, I CHECKED THE GENES' AGE. OLDER GENES ARE EXPRESSED EARLIER IN THE DEVELOPMENT, EXCEPT FOR CLEAVAGE WHEN THE MATERNAL GENES ARE STILL ACTIVE.

NOT REALLY, YOU EXPECT THIS KIND OF PATTERN, WHEN ALL STAGES ARE EQUALLY CONSERVED, BECAUSE THE STAGES THEMSELVES DIFFER IN AGE

HUH, ANY MORE IDEAS?

YUP, I'VE JUST ANALYZED FOR EACH MODULE THE GENES THAT HAVE ORTHOLOGS IN MOUSE

AND? EARLIER IN THE DEVELOPMENT THE MAJORITY OF ORTHOLOGS REMAIN IN ONE-TO-ONE RELATIONSHIP. WHICH MEANS THE GENOME EVOLUTION IS MORE CONSTRAINED THEN. SO, STILL NO HOURGLASS BUT WE FOUND SOME EVIDENCE FOR EARLY CONSERVATION.

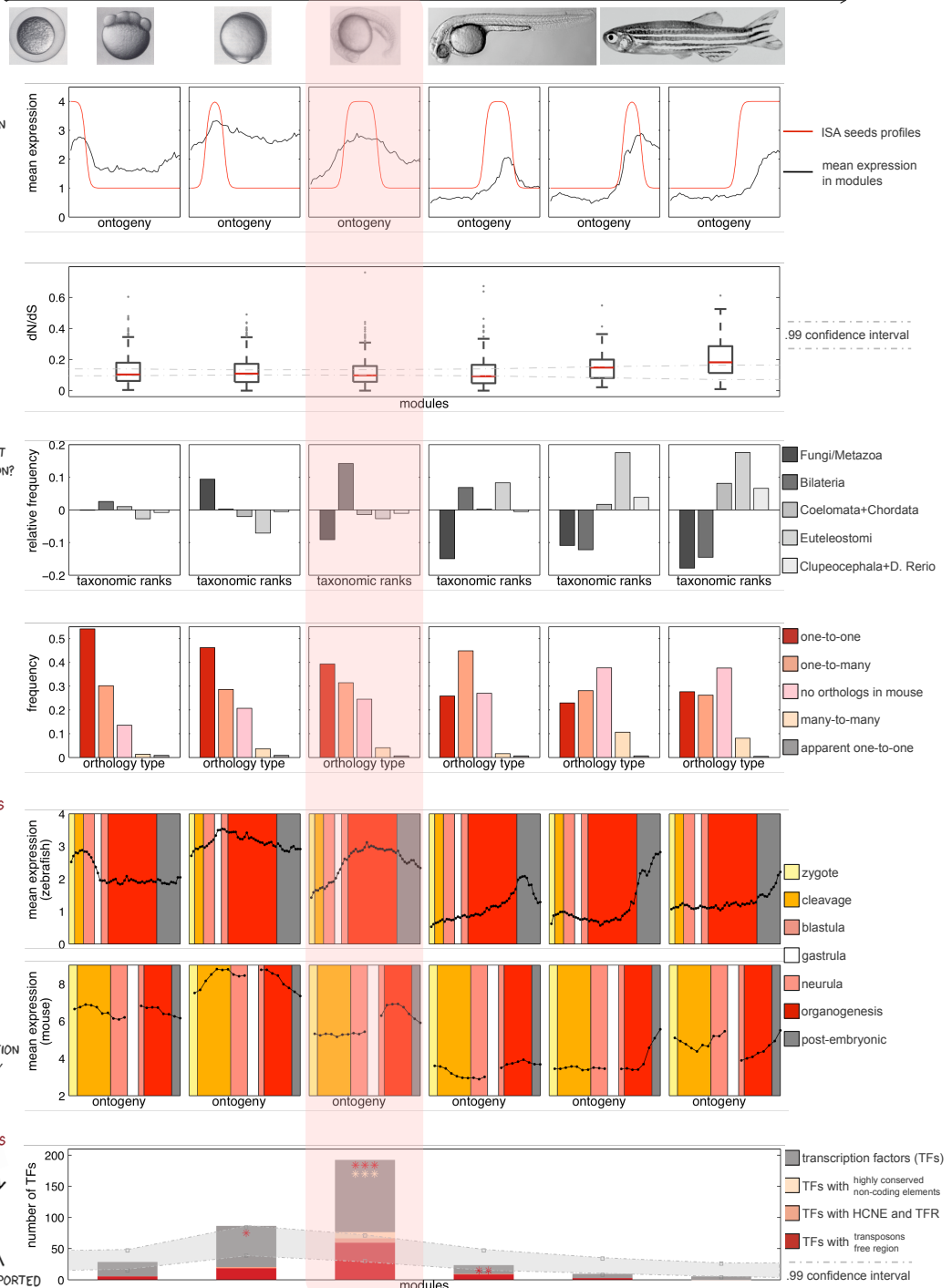
USING THE ONE-TO-ONE ORTHOLOGS BETWEEN ZEBRAFISH AND MOUSE I ALSO COMPARED THEIR EXPRESSION PROFILES

YES, BUT LOOK, IT'S HARD TO DO A PROPER QUANTITATIVE COMPARISON, DUE TO DATA MISMATCHES

WE CAN'T FIND ANY EVIDENCE FOR HOURGLASS MODEL!

WHO! REGULATORY ELEMENTS OF TRANSCRIPTION FACTORS ARE HIGHLY CONSERVED EXACTLY IN MID-DEVELOPMENT!

IN CONCLUSION, THE EARLY CONSERVATION MODEL IS SUPPORTED BY CONSTRAINT ON GENE DUPLICATION WHICH DECREASES DURING ONTOGENY. WHEREAS THE HOURGLASS MODEL IS SUPPORTED BY THE CONSERVATION OF REGULATORY ELEMENTS, WHICH IS THE HIGHEST DURING MID-DEVELOPMENT.



C

Curriculum Vitae

Barbara Piasecka

CONTACT INFORMATION	Biophore Department of Ecology and Evolution University of Lausanne CH-1015 Lausanne, Switzerland	<i>Phone:</i> +4121 692 4221 <i>Fax:</i> +4121 692 4165 <i>E-mail:</i> barbara.piasecka@unil.ch
RESEARCH INTERESTS	gene expression evolution, modular approach in transcriptomics, statistical methods for large-scale data	
EDUCATION	University of Lausanne , Lausanne, Switzerland <i>Faculty of Biology and Medicine</i> Ph.D. Candidate, Bioinformatics, October 2008 <ul style="list-style-type: none">• Dissertation Topic: “Comparative Modular Analysis of Gene Expression in Vertebrate Development”• Advisors: Marc Robinson-Rechavi and Sven Bergmann Adam Mickiewicz University , Poznań, Poland <i>Faculty of Mathematics and Computer Science</i> M.S., Applied Mathematics, October 2003 - September 2008 <i>Faculty of Biology</i> M.S., Biotechnology, October 2004 - October 2006 B.A., Biotechnology, October 2001 - June 2004	
HONORS AND AWARDS	SIB Competitive Fund for Conference Attendance, 2012 SIB Days Best Poster Award, 2012 Scholarship of the Adam Mickiewicz University for outstanding academic accomplishments, 2007/2008	
ACADEMIC EXPERIENCE	University of Lausanne , Lausanne, Switzerland <i>Graduate Student</i> October 2008 - present Includes current Ph.D. research, Ph.D. coursework and research/consulting projects. <i>Teaching Assistant</i> October 2008 - present Assisting in the courses of: Introduction to Bioinformatics, Bioinformatics for Genomics, Statistics for Biologists, and Experimental Design.	
INTERNSHIPS AND EXCHANGES	ProfileXpert Platform , Bron, France <i>Lifelong Learning Programme/Erasmus exchange student</i> February 2008 - July 2008 Biostatistics internship. My duties included statistical data analysis coming from microarray technology with the aim of identifying differentially expressed genes or detecting polymorphism sites. Ecole Normale Supérieure , Lyon, France <i>Socrates-Erasmus exchange student</i> February 2006 - July 2006 Work in the INSERM U404 Laboratory of Immunology of Viral Infection. I was involved in the project of development of the transgenic murine model of HHV-6 infection.	

- PUBLICATIONS Piasecka B., Lichocki P., Bergmann S., Robinson-Rechavi M., Submitted. The hourglass and the early conservation models — co-existing patterns of developmental constraints in vertebrates. *PLOS Genetics*.
- Piasecka B., Robinson-Rechavi M., Bergmann S., 2012. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics* **28**(14): 1865-1872.
- Piasecka B., Kutalik Z., Roux J., Bergmann S., Robinson-Rechavi M., 2012. Comparative modular analysis of gene expression in vertebrate organs. *BMC Genomics* **13**:124.
- CONFERENCE PRESENTATIONS Piasecka B., Robinson-Rechavi M., Bergmann S. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. European Student Council Symposium, Basel, Switzerland, September 2012.
- Piasecka B., Bergmann S., Robinson-Rechavi M.. Conservation of tissue-specific gene expression in vertebrates. Society for Molecular Biology and Evolution (SMBE) Conference, Lyon, France, July 2010.
- LANGUAGES
- Polish: Native
 - English: Very good
 - French: Very good
- COMPUTER SKILLS
- Statistical Packages: R, Bioconductor.
 - Languages: Matlab, BASH scripts.
 - Applications: \LaTeX , Microsoft Office, iWorks, Inkscape.
 - Operating Systems: Mac OS, Windows.

Bibliography

- Aanes, H., Winata, C. L., Lin, C. H., Chen, J. P., Srinivasan, K. G., Lee, S. G. P., Lim, A. Y. M., Hajan, H. S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Aleström, P., and Mathavan, S. (2011). Zebrafish mrna sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res*, **21**(8), 1328–38.
- Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**(13), 1600–1607.
- Allison, D., Cui, X., Page, G., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**(1), 55–65.
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, **96**(12), 6745–50.
- Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: Integrating and comparing heterogeneous transcriptome data among species. In A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, editors, *Data Integration in the Life Sciences*, volume 5109 of *Lecture Notes in Computer Science*, pages 124–131. Springer Berlin / Heidelberg.
- Benton, M. J. and Donoghue, P. C. J. (2007). Paleontological evidence to date the tree of life. *Mol Biol Evol*, **24**(1), 26–53.
- Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys*, **67**(3 Pt 1), 031902.
- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**(6795), 536–40.
- Bonner, J. (1988). *The evolution of complexity by means of natural selection*. Princeton Univ Pr.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, **478**(7369), 343–348.
- Bulun, S. E., Takayama, K., Suzuki, T., Sasano, H., Yilmaz, B., and Sebastian, S. (2004). Organization of the human aromatase p450 (cyp19) gene. *Semin Reprod Med*, **22**(1), 5–9.
- Cai, J., Xie, D., Fan, Z., Chipperfield, H., Marden, J., Wong, W. H., and Zhong, S. (2010). Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells. *PLoS Comput Biol*, **6**(3), e1000707.

- Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS Biol*, **3**(7), e245.
- Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, **134**(1), 25–36.
- Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J. H., Wilde, A., Brudno, M., Morris, Q. D., and Hughes, T. R. (2009). Conservation of core gene expression in vertebrate tissues. *J Biol*, **8**(3), 33.
- Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, **8**, 93–103.
- Cheverud, J., Routman, E., and Irschick, D. (1997). Pleiotropic effects of individual gene loci on mandibular morphology. *Evolution*, **51**(6), 2006–2016.
- Comte, A., Roux, J., and Robinson-Rechavi, M. (2010). Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evolution & development*, **12**(2), 144–156.
- Cook, M. *et al.* (1965). The anatomy of the laboratory mouse. *The anatomy of the laboratory mouse*.
- Demuth, J. P. and Hahn, M. W. (2009). The life and death of gene families. *Bioessays*, **31**(1), 29–39.
- Domazet-Lošo, T. and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, **468**(7325), 815–8.
- Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**(2), 341 – 352.
- Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*, pages 135–42.
- Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1), 207–10.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**(25), 14863–8.
- Elinson, R. (1987). Change in developmental patterns: Embryos of amphibians with large eggs. In R. E. Raff RA, editor, *Development as an Evolutionary Process*, pages 1–21. New York: Alan R. Liss.
- Ellegren, H. and Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics*, **8**(9), 689–698.
- Engström, P. G., Fredman, D., and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol*, **9**(2), R34.
- Espinosa-Soto, C. and Wagner, A. (2010). Specialization can drive the evolution of modularity. *PLoS Comput Biol*, **6**(3), e1000719.
- Fisher, R. (1971). *The design of experiments*. New York: Hafner, 8th edition.
- Gamma, E. (1995). *Design patterns: elements of reusable object-oriented software*. Addison-Wesley Professional.

- Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using rna-seq. *Nat Methods*, **8**(6), 469–77.
- Gasch, A. P. and Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*, **3**(11), RESEARCH0059.
- Gaur, A., Jewell, D. A., Liang, Y., Ridzon, D., Moore, J. H., Chen, C., Ambros, V. R., and Israel, M. A. (2007). Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. *Cancer Res*, **67**(6), 2456–68.
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10), R80.
- Gershenson, J. and Prasad, G. (1997). Modularity in product design for manufacturability. *International Journal of Agile Manufacturing*, **1**(1), 99–110.
- Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A*, **97**(22), 12079–84.
- Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, **99**(12), 7821–6.
- Gould, S. (1977). *Ontogeny and phylogeny*. Belknap press.
- Gu, X. and Su, Z. (2007). Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc Natl Acad Sci U S A*, **104**(8), 2779–2784.
- Halder, G., Callaerts, P., and Gehring, W. (1995). Induction of ectopic eyes by targeted expression of the eyeless gene in drosophila. *Science*, **267**(5205), 1788–1792.
- Hartley, R. S., Rempel, R. E., and Maller, J. L. (1996). In vivo regulation of the early embryonic cell cycle in *Xenopus*. *Dev Biol*, **173**(2), 408–19.
- Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, **402**(6761 Suppl), C47–52.
- Hazkani-Covo, E., Wool, D., and Graur, D. (2005). In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von baer’s third law. *J Exp Zool B Mol Dev Evol*, **304**(2), 150–8.
- Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Res*, **37**(Database issue), D690–7.
- Ihmels, J. H. and Bergmann, S. (2004). Challenges and prospects in the analysis of large-scale gene expression data. *Brief Bioinform*, **5**(4), 313–27.

- Irie, N. and Kuratani, S. (2011). Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun*, **2**, 248.
- Irie, N. and Sehara-Fujisawa, A. (2007). The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol*, **5**, 1.
- Irimia, M., Tena, J. J., Alexis, M., Fernandez-Miñan, A., Maeso, I., Bogdanovic, O., de la Calle-Mustienes, E., Roy, S. W., Gomez-Skarmeta, J. L., and Fraser, H. B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res*.
- Jablan, S. (2002). *Symmetry, ornament and modularity*, volume 30. Imperial College Pr.
- Jensen, L., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., *et al.* (2009). String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, **37**(suppl 1), D412–D416.
- Jordan, I. K., Mariño-Ramírez, L., Wolf, Y. I., and Koonin, E. V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol*, **21**(11), 2058–70.
- Jordan, I. K., Marino-Ramirez, L., and Koonin, E. V. (2005). Evolutionary significance of gene expression divergence. *Gene*, **345**(1), 119–126.
- Kahn, K. (2008). Tutorial: Introduction to DNA microarrays. http://www.chem.ucsb.edu/~kalju/chem162/public/genechip_intro.html.
- Kalinka, A. and Tomancak, P. (2012). The evolution of early animal embryos: conservation or divergence? *Trends in Ecology & Evolution*, **27**(7), 385–393.
- Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., Ohler, U., Bergman, C. M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature*, **468**(7325), 811–4.
- Kathleen Kerr, M. (2003). Design considerations for efficient and effective microarray studies. *Biometrics*, **59**(4), 822–828.
- Keys, D. N., Lewis, D. L., Selegue, J. E., Pearson, B. J., Goodrich, L. V., Johnson, R. L., Gates, J., Scott, M. P., and Carroll, S. B. (1999). Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science*, **283**(5401), 532–4.
- Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., and Pääbo, S. (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**(5742), 1850–4.
- Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., and Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Dev Dyn*, **203**(3), 253–310.
- King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, **188**(4184), 107–16.
- Klingenberg, C. (2008). Morphological integration and developmental modularity. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 115–132.
- Krumlauf, R. (1994). Hox genes in vertebrate development. *Cell*, **78**(2), 191–201.

- Kutalik, Z., Beckmann, J. S., and Bergmann, S. (2008). A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat Biotechnol*, **26**(5), 531–539.
- LeProust, E. (2008). Agilent's microarray platform: How high-fidelity dna synthesis maximizes the dynamic range of gene expression measurements. Application note - Agilent technologies 5989-9159EN. http://www.chem.agilent.com/en-US/Search/Library/_layouts/Agilent/PublicationSummary.aspx?whid=56080&liid=2024.
- Liao, B.-Y. and Zhang, J. (2006a). Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol*, **23**(3), 530–540.
- Liao, B.-Y. and Zhang, J. (2006b). Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol*, **23**(6), 1119–1128.
- Liao, B.-Y., Weng, M.-P., and Zhang, J. (2010). Contrasting genetic paths to morphological and physiological evolution. *Proc Natl Acad Sci U S A*, **107**(16), 7353–7358.
- McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., and Irizarry, R. A. (2011). The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res*, **39**(Database issue), D1011–5.
- McDonald, J. H. (2009). *Handbook of Biological Statistics (2nd ed.)*. Sparky House Publishing, Baltimore, Maryland.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, **5**(7), 621–628.
- Movahedi, S., Van de Peer, Y., and Vandepoele, K. (2011). Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant Physiol*, **156**(3), 1316–30.
- Nadon, R. and Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics*, **18**(5), 265–271.
- Nagae, G., Isagawa, T., Shiraki, N., Fujita, T., Yamamoto, S., Tsutsumi, S., Nonaka, A., Yoshida, S., Matsusaka, K., Midorikawa, Y., Ishikawa, S., Soejima, H., Fukayama, M., Suemori, H., Nakatsuji, N., Kume, S., and Aburatani, H. (2011). Tissue-specific demethylation in cpg-poor promoters during cellular differentiation. *Hum Mol Genet*, **20**(14), 2710–2721.
- Needham, J. (1933). On the dissociability of the fundamental processes in ontogenesis. *Biological Reviews*, **8**(2), 180–223.
- Nei, M. (2007). The new mutation theory of phenotypic evolution. *Proc Natl Acad Sci U S A*, **104**(30), 12235–42.
- Niimura, Y. and Nei, M. (2007). Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE*, **2**(8), e708.
- Niknejad, A., Comte, A., Parmentier, G., Roux, J., Bastian, F. B., and Robinson-Rechavi, M. (2012). vhog, a multispecies vertebrate ontology of homologous organs groups. *Bioinformatics*, **28**(7), 1017–20.
- Ohno, S. *et al.* (1970). *Evolution by gene duplication*. Berlin, Heidelberg and New York: Springer-Verlag.

- Oldham, M. C., Horvath, S., and Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A*, **103**(47), 17973–8.
- Parmentier, G., Bastian, F. B., and Robinson-Rechavi, M. (2010). Homolonto: generating homology relationships by pairwise alignment of ontologies and application to vertebrate anatomy. *Bioinformatics*, **26**(14), 1766–71.
- Pereira, V., Waxman, D., and Eyre-Walker, A. (2009). A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*, **183**(4), 1597–1600.
- Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–52.
- Piasecka, B., Robinson-Rechavi, M., and Bergmann, S. (2012). Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics*, **28**(14), 1865–72.
- Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, **29**(2), 153–9.
- Poe, S. and Wake, M. H. (2004). Quantitative tests of general models for the evolution of development. *Am Nat*, **164**(3), 415–22.
- Preuss, T. M., Cáceres, M., Oldham, M. C., and Geschwind, D. H. (2004). Human brain evolution: insights from microarrays. *Nat Rev Genet*, **5**(11), 850–60.
- Prud'homme, B. and Gompel, N. (2010). Evolutionary biology: Genomic hourglass. *Nature*, **468**(7325), 768–9.
- Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet*, **32 Suppl**, 496–501.
- Raff, R., Kaufman, T., *et al.* (1991). *Embryos, genes, and evolution: the developmental-genetic basis of evolutionary change*. Indiana University Press.
- Raff, R. A. (1996). *The shape of life: genes, development, and the evolution of animal form*. Chicago; London: University of Chicago Press.
- Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, **5**(12), e1000598.
- Rawn, S. and Cross, J. (2008). The evolution, regulation, and function of placenta-specific genes. *Annual review of cell and developmental biology*, **24**, 159–181.
- Roux, J. and Robinson-Rechavi, M. (2008). Developmental constraints on vertebrate genome evolution. *PLoS Genet*, **4**(12), e1000311.
- Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**(1), 99.

- Sander, K. (1983). The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In W. C. Goodwin BC, Holder N, editor, *Development and evolution*, pages 137–159. Cambridge University Press.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. (2000). A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, **24**(3), 236–44.
- Schilling, M. A. and Steensma, H. K. (2001). The use of modular organizational forms: An industry-level analysis. *The Academy of Management Journal*, **44**(6), pp. 1149–1168.
- Schlosser, G. and Wagner, G. (2004). *Modularity in development and evolution*. University of Chicago Press.
- Seidel, F. (1960). Körpergrundgestalt und keimstruktur. eine erörterung über die grundlagen der vergleichenden und experimentellen embryologie und deren gültigkeit bei phylogenetischen überlegungen. *Zool. Anz.*, **164**, 245–305.
- Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U., Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., and Weinstein, J. N. (2007). Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. *Mol Cancer Ther*, **6**(3), 820–32.
- Simons, C., Makunin, I. V., Pheasant, M., and Mattick, J. S. (2007). Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics*, **8**, 470.
- Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). Biomart–biological queries made easy. *BMC Genomics*, **10**, 22.
- Speed, T. (2000). Always log spot intensities and ratios. <http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html>.
- Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, **100**(21), 12123–8.
- Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. (2001). Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, **98**(19), 10787–92.
- Stern, D. L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution*, **54**(4), 1079–91.
- Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., and Hogenesch, J. B. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A*, **99**(7), 4465–4470.
- Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**(16), 6062–6067.
- Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18 Suppl 1**, S136–44.

- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat Genet*, **22**(3), 281–5.
- Vavouri, T., Walter, K., Gilks, W. R., Lehner, B., and Elgar, G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol*, **8**(2), R15.
- Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, **19**(2), 327–35.
- von Baer, K. E. (1828). *Ueber Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion*. Königsberg: Bornträger.
- Voolstra, C., Tautz, D., Farbrotter, P., Eichinger, L., and Harr, B. (2007). Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Res*, **17**(1), 42–49.
- Wagner, G. and Altenberg, L. (1996). Perspective: Complex adaptations and the evolution of evolvability. *Evolution*, pages 967–976.
- Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat Rev Genet*, **8**(12), 921–31.
- Wang, Q. T., Piotrowska, K., Ciemerych, M. A., Milenkovic, L., Scott, M. P., Davis, R. W., and Zernicka-Goetz, M. (2004). A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev Cell*, **6**(1), 133–44.
- Wang, X., Grus, W. E., and Zhang, J. (2006). Gene losses during human origins. *PLoS Biol*, **4**(3), e52.
- Wang, Y. and Rekaya, R. (2009). A comprehensive analysis of gene expression evolution between humans and mice. *Evol Bioinform Online*, **5**, 81–90.
- Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A*, **106**(18), 7273–80.
- Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., and Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, **3**(1), e7.
- Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, **8**(3), 206–16.
- Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, **99**(468), 909–917.
- Xing, Y., Ouyang, Z., Kapur, K., Scott, M. P., and Wong, W. H. (2007). Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol Biol Evol*, **24**(6), 1283–1285.
- Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*, **30**(17), 3754–3766.

- Yanai, I., Graur, D., and Ophir, R. (2004). Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS*, **8**(1), 15–24.
- Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**(5), 650–659.
- Yang, J., Su, A. I., and Li, W.-H. (2005). Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Mol Biol Evol*, **22**(10), 2113–2118.
- Yang, R. and Su, B. (2010). Characterization and comparison of the tissue-related modules in human and mouse. *PLoS One*, **5**(7), e11730.
- Yang, Y., Speed, T., *et al.* (2002a). Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, **3**(8), 579–588.
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30**(4), e15.
- Yarden, A. and Geiger, B. (1996). Zebrafish cyclin E regulation during early embryogenesis. *Dev Dyn*, **206**(1), 1–11.
- Young, J. M., Friedman, C., Williams, E. M., Ross, J. A., Tonnes-Priddy, L., and Trask, B. J. (2002). Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Human Molecular Genetics*, **11**(5), 535–546.
- Yuh, C. H., Bolouri, H., and Davidson, E. H. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**(5358), 1896–902.
- Zhang, J. (2003). Evolution by gene duplication - an update. *Trends Ecol Evol*, **18**, 292–298.
- Zhang, X. and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. *Nat Neurosci*, **5**(2), 124–133.
- Zheng-Bradley, X., Rung, J., Parkinson, H., and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*, **11**(12), R124.