Evolution of spatiotemporal organization of biological systems: Origins and phenotypic impact of duplicated genes

Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
Kalliopi Trachana
born in
Heraklion, Crete

Referees: Prof. Dr. Detlev Arendt

Prof. Dr. Joachim Wittbrodt

Ithaka

As you set out for Ithaka
hope the voyage is a long one,
full of adventure, full of discovery.
Laistrygonians and Cyclops,
angry Poseidon—don't be afraid of them:
you'll never find things like that on your way
as long as you keep your thoughts raised high,
as long as a rare excitement
stirs your spirit and your body.
Laistrygonians and Cyclops,
wild Poseidon—you won't encounter them
unless you bring them along inside your soul,
unless your soul sets them up in front of you.

Hope the voyage is a long one.
May there be many a summer morning when,
with what pleasure, what joy,
you come into harbors seen for the first time;
may you stop at Phoenician trading stations
to buy fine things,
mother of pearl and coral, amber and ebony,
sensual perfume of every kind—
as many sensual perfumes as you can;
and may you visit many Egyptian cities
to gather stores of knowledge from their scholars.

Keep Ithaka always in your mind.
Arriving there is what you are destined for.
But do not hurry the journey at all.
Better if it lasts for years,
so you are old by the time you reach the island,
wealthy with all you have gained on the way,
not expecting Ithaka to make you rich.

Ithaka gave you the marvelous journey.
Without her you would not have set out.
She has nothing left to give you now.

And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you will have understood by then what these Ithakas mean.

(C.P. Cavafy, *Collected Poems.*
Translated by Edmund Keeley and Philip Sherrard.
Princeton University Press, 1992)

Acknowledgments

Foremost, I would like to thank my supervisor Peer Bork, for giving me the opportunity to do research in his lab and for his support and ideas throughout the PhD. It has been a great privilege to work under his supervision and gain much experience, –not only scientifically, but also in many other aspects of academic life.

All members of the Bork lab are especially thanked for an amazing atmosphere and for being so collaborative and open to discussing ideas and helping me whenever needed. Especially, I would like to thank the orthology team (Tobias Doerks, Tomas Larsson, Sean Powell and Wei-Hua Chen) for the many brainstorming meetings, ending with funny drawings and ideas. Moreover, Lars Jensen and Chris Creevey, previous members of the Bork group, introduced me to the orthology/paralogy and phylogeny world, with the first of these topics being the primary interest of my thesis. Mani Arumugam, Julien Tap, Vera van Noort and Georg Zeller have always given important statistical tips, while Takuji Yamada has taught me how to deal with the KEGG database and design figures with Japanese accuracy. I would like also to specially thank Pablo Minguez Paniagua, who has always been helpful when my computer skills dragged me down and Tomas Larsson, who has supported me so many times – and now with the thesis writing - and has had to deal with my expansive/invasive tendencies towards his desk every day. Tomas and other members of Arendt lab - in particular Maria-Antonietta Tosches and Oleg Simikov - who with their zoological and evolutionary background helped nurse many exciting conclusions and creative moments. Of course, my dearest affiliation with the Arendt lab has been Foteini Christodoulou, who gave me the opportunity to work with her, Oleg and Detlev Arendt on an amazing and really innovative project. Moreover, I would like to thank Alison Waller and Daniel Mende for the English and German touch of this thesis.

This brings me to the acknowledgment of my thesis advisory committee: Detlev Arendt, Joachim Wittbrodt and Jan Korbel. They have been always insightful with their comments and have encouraged me to study what I was most passionate about. I would like to thank, in particular, Detlev, for giving me the opportunity to work with his team. Moreover, I would like to thank Rob Russel, who accepted to be part of my defense committee.

I wish to thank my friends at the EMBL, without whom my PhD would have been less fun. First of all, my personal psychoanalysts in the Bork group Mani, Pablo, Ivica and Tobias, who have been patient enough to hear many greek dramatic monologues. The Bork coffee team has been always composed of very nice people, leaving me with many nice memories. I would like also to thank Nelly van der Jagt González, our super secretary, who I am allowed to talk to about strictly non-scientific subjects and who takes care all of us. Generally, I have to thank many people at the Computational Biology Unit and members of the EMBL “family”, who are always open-minded and ready to seize

the day. However, I would like to acknowledge a few of them specially, like the so-called greek mafia (Evangelia Petsalaki, Yiorgos Pavlopoulos, Yannis Legouras, Konstantina Diamantara, Foteini Christodoulou, Maria Papatriantafyllou and Lefteris Touros), who really adopted me immediately and took care not to let me forget my greek habits. Among them, Tina and my other "internationally-selected" friend Jacopo Lucci have been my fellow travelers in this journey, sharing many vivid discussions, future dreams and dinners (with a lot of wine).

The person, who probably influenced me (personality-wise) more during my time here at EMBL, is Foteini Christodoulou. Fay with her gifted personality and tireless energy has proven a friend for life. I am really looking forward to living new exciting moments and sharing new ideas with her.

Last but not least, I wish to thank all my friends and my family from Greece who have been so patient with me in the last 4 years. Especially, I would like to thank Pantelis Hatzis for his constant understanding and his many many advices. As a proud Cretan, I cannot find other words than those below to thank them for their love.

"Παίρνω τση μάνας μου ευχή, κι απ' την ακλή μου δάκρυ
και φεύγω να 'βρω το θεριό στων σκοταδιών την άκρη."
(Το δρακοδοντι, Χαϊνηδες)

Summary

A deluge of genomic sequences and other functional large-scale datasets has allowed the description and comparison of the functional components and their interactions for a large number of species. This information can be further evaluated and transferred across newly sequenced species through orthology (homology derived via speciation) to provide insights into the evolutionary aspect of function and organization. Robust orthology prediction is a prerequisite for accurate phylogenomic and comparative analyses. Despite the advances in the field, orthology prediction is still conflicting and uncertain. Therefore, quality control tests should be established to deal with this issue. In the course of this thesis, a phylogeny-based benchmark dataset for orthology prediction for the animal clade was established. This dataset has been used to evaluate the orthology predictions for five publicly available repositories and estimate the impact of several technical and biological factors.

At the same time, the large number of fully sequenced genomes has enabled the formulation of interesting hypothesis on the mechanisms of functional evolution. For instance, paralogs, defined as homologs derived via gene/genome duplication, has been associated with expansion or division of functionality. A plethora of studies have investigated how duplicated genes that related to morphological innovations have diverged their expression in different tissues. Thus far, it has not been examined in a large scale, if the regulatory divergence of the paralogs is favored in certain patterns and how these patterns have emerged. To study this, the expression data of 31 human tissues were used to identify the preferable tissue combinations of sub(neo)functionalized paralogs. Interestingly, it has been revealed that paralogs related to chordate-vertebrate transition and belong to protein families that predate the vertebrate origin, are often diverged between brain and non-brain tissue. This suggests that the elaborated brain of vertebrates might have been developed by adapting one paralog in a brain-specific manner. In contrast to the rich literature on tissue evolution and paralogy, the role of duplication in the temporal regulation of biological systems has been understudied. Again, by combining orthology and transcriptomic data, we identified that cell cycle and other cellular periodic processes (namely circadian and ultradian regulation) tend to be orchestrated through paralogs. The functional repertoires of periodically diverged paralogs are different for three eukaryotic species (Arabidopsis, human and budding yeasts), implying that the temporal organization of cells through paralogs has evolved independently in the three lineages.

To conclude, the greatest challenge in the postgenomic era is to effectively integrate functionally relevant genomic data in order to determine how complex traits have emerged. To accomplish this we have to study the dynamic changes of gene inventories with respect to orthologous (common origin) and paralogous (potential of divergence) relationships.

Zusammenfassung

Eine Schwemme von Genomsequenzen sowie weitere groß angelegte Studien zur Charakterisierung von molekularen Funktionen hat Forschern erlaubt, komparative Studien der funktionellen Komponenten und ihrer Interaktionen für eine große Anzahl von Spezies durchegeführt werden. Die so gewonnen Erkenntnisse können weiter untersucht werden und mithilfe von Orthologie (Homologie abgeleitet durch Artenbildung) auf neu-sequenzierte Spezies übertragen werden um Erkenntnisse über die Evolution von molekularen Funktionen und ihrer Organisation zu gewinnen.

Eine robuste Orthologie ist Voraussetzung für akkurate phylogenomische und komparative Analysen. Obwohl sich das Forschungsfeld der Orthologie Fortschritte gemacht hat, ist die Orthologie-Voraussage noch immer von widersprüchlich und unsicher. Aus diesem Grund sollten Tests zur Qualitätskontrolle eingeführt werden. Im Rahmen dieser Arbeit wurde ein Phylogenie-basierter Datensatz entwickelt, mit dem die Orthologie Voraussage in den Animalia überprüft werden kann. Dieser Datensatz wurde benutzt um die Orthologie-Voraussagen von fünf öffentlich zugänglichen Repositorien zu evaluieren und die Auswirkungen von einer Anzahl von technischen und biologischen Faktoren zu untersuchen.

Gleichzeitig hat die große Anzahl von komplett sequenzierten Genomen zur Formulierung von interessanten Hypothesen über die Mechanismen der Evolution von molekularen Funktionen geführt. Zum Beispiel wurden Paraloge, Homologe die durch Gen- oder Genomduplikation entstanden sind, mit der Erweiterung und Teilung von molekularen Funktionen assoziiert. Eine Vielzahl von Studien wurden durchgeführt um herauszufinden, wie siche duplizierte Gene, die mit morphologischen Veränderungen assoziert werden, ihre Genexpressionsraten in unterschiedlichen Geweben ändern. Es wurde jedoch noch nicht großflächig untersucht, ob die regulatorische Divergenz von Paralogen bestimmte Muster bevorzugt und wie diese Muster entstanden sind. Um dies zu untersuchen wurden die Expressionsdaten von 31 menschlichen Geweben benutzt und bevorzugte Gewebekombinationen von sub(neo)funktionalisierten Paralogen identifiziert. Interessanterweise stellte sich heraus, dass Paraloge die mit dem Choradata- Wirbeltiere Übergang im Zusammenhang stehen und bereits vor dem Ur-Wibeltier vorhanden waren, häufig zwischen Gehirn und nicht-Gehirngeweben divergieren. Im Kontrast zur weitreichenden Literatur über die Evolution von Geweben und Paralogie, ist die Rolle von Genduplikation in der temporalen Regulation von biologischen Systemen schlechter untersucht. Um dies zu untersuchen wurden Orthologie und Genexpressionsdaten kombiniert. Wir konnten herausfinden, dass der Zell-Zyklus und andere periodische Prozesse (wie der Circadianen und Ultradianen Rhythmik) von Paralogen reguliert werden. Das funktionelle Repertoire dieser Paraloge unterscheidet sich in 3 eukaryotischen Spezies (Arabidopsis, Mensch und Hefe), was impliziert, dass sich die temporale Regulation der Zellen durch Paraloge sich in den drei Organismen unabhängig

voneinander entwickelt hat. Zusammenfassend ist die größte Herausvorderung der postgenomischen Ära eine effektive Integration von funktionell relevanten genomischen Daten um herauszufinden, wie komplexe Eigenschaften sich entwickelt haben. Um dieses Ziel zu erreichen sollten die dynamischen Veränderungen der Gen-Inventare unter Beachtung von der Beziehung von Orthologen (gleicher Ursprung) und Paralogen (Potenzial für Divergenz) untersucht werden.

"What characterizes the living world is both its diversity and its underlying unity."

(Jacob, 1977)

Introduction

The theory of evolution, as published in "*On the Origin of Species"* by Charles Darwin, profoundly revolutionized our understanding of biodiversity. The integration of Mendelian genetics, and later on of molecular biology and genomics, into evolutionary biology unified several previously isolated fields (*Pigliucci M, 2009*) and shaped one of the most debatable questions: "how do genomes evolve to generate biological diversity". The exponentially increasing number of full genome sequences and the emerging tools for their analysis have allowed the systematic analysis of the inter- (*Koonin, Aravind & Kondrashov, 2000; Koonin EV, 2005; Koonin EV, 2010; Srivastava et al, 2010; Prochnik et al, 2010; Colbourne et al, 2011*) and intra-species evolution (*Sudmant et al, 2010; Gravel et al, 2011*). Despite the multiple pieces of empirical evidence demonstrating that phenotypic diversity is not the pure outcome of vertical transmission (inheritance through the genome), but is influenced by many environmental (vertically inherited or not) factors (*Pigliucci M, 2007; Danchin et al, 2011*), comparative genomics is still the most comprehensive field to understand different evolutionary mechanisms.

Establishing the homologous – shared due to common ancestry – parts between the compared species has been the ground step of all comparative studies. In the modern biology era, genomics have revealed more fine-grained evolutionary relationships between genes. Orthology, for instance, which is homology derived via speciation (*Fitch WM, 1970*), has been the most appropriate way to compare the genomic content of different species (*Koonin EV, 2000; Koonin EV, 2005*). On the other hand, paralogy, which is homology derived via duplication (*Fitch WM, 1970*), is considered to be the major source of functional novelty (*Ohno S, 1970; Scanell et al, 2006; Hittinger & Caroll, 2007; Ames et al, 2010*). Thus, to understand the ancestry of modern organisms and their functional divergence through evolutionary time, we have to study their gene inventories with respect to their orthologous (common origin) and paralogous (potential of divergence) relationships. To decipher phenotypes, besides the evolution of gene families, it is essential to discover "when, where and how" genes are expressed and translate their regulatory evolution into evolution of function. In fact, current studies endeavor to integrate data for multiple cellular components and their interactions (*Kuhner et al, 2009; Yus et al, 2009; Guell et al, 2009; Costanzo et al, 2010; Schwanhausser et al, 2011; Maier et al, 2011*). These studies have elucidated a complex interplay of DNA, RNA, proteins and metabolites inside cells, emphasizing the importance of post-transcriptional and post-translational regulation, protein turnover and epistasis –among other mechanisms- in phenotypic readout. However, these mechanisms are beyond the scope of this thesis.

Herein, I will briefly introduce the field of comparative genomics and how its integration with other high-throughput methods has revealed important functional insights of cellular organization and species biology. As the primary interest of this thesis has been to understand the

functional repertoire of eukaryotes, and more specifically of multi-cellular animals, the examples that are outlined herein are related to the evolution of this clade. Moreover, I will give an overview of the orthology field: the caveats of orthology prediction and a short classification of the methodologies developed to detect orthology. Finally, I will focus on the functional insights of gene/genome duplication and its impact on animal evolution.

A. "Omics": powerful fields to understand the functional diversity

An ever-quickening pace of scientific and technological developments has transformed the modern biology era. Rapid sequencing technologies have helped to accumulate a huge amount of genomic sequences (*Pareek, Smoczynski & Tretyn, 2011*) that provide the raw material for understanding phenotypic diversity. At the same time, the development of various genome-scale experimental techniques – referred as "omics" approaches– has linked the genomic information to the levels of RNA (e.g. microarrays, SAGE, RNA-Seq), proteins (e.g. ChIP-chip, ChIP-Seq, yeast 2H, mass spectrometry) and metabolites (e.g. isotopic tracing) providing additional functional understanding of the studied system (*Joyce and Palsson, 2006* and all references within).

1. Insights from comparative genomics for animal evolution

Perhaps the most fascinating aspect of the genomic era is the radical change it has brought to evolutionary biology. Although comparisons of partial genomes have been carried out for decades (*Bork P, 1989; Bork P, 1991; Adams et al, 1992*), the completion of the first whole bacterial genomes, followed by archaeal and eukaryotic genomes, revealed the genomic architecture of all living organisms on an unbiased approach (*Koonin EV, 2000*). By identifying what has been conserved, and vertically transmitted, throughout the evolutionary history from the last universal common ancestor (LUCA), scientists were able to root the "Tree of Life" (*Ciccarreli et al, 2006*) - the representation of evolutionary relationships since Darwin and Haeckel.

Apart from the huge progress in the field of phylogeny (*Delsuc, Brinkmann & Philippe, 2005; Telford & Copley, 2011* and all references within), comparative analyses and integration of "omics" data have also provided many functional insights in the organization of biological systems (*Nurse & Hayles, 2011*). However, for large systems with multiple levels of complexity, such as animals and plants, it is hard to directly associate the functional repertoire of the studied system with their phenotype. For instance, the presence of many shared gene families among long-diverged animals, such as sponges or cnidaria and human (*Putnam et al, 2007; Chapman et al, 2010; Srivastava et al, 2010*), is controversial to the huge phenotypic diversity observed between them (Figure 1). This raises the question: If animals share a large set of genes, how have biodiversity

evolved? Two major hypotheses have been put forward: emergence and expansion of functionality through i) gene/genome duplication or ii) differential gene expression. Of course, both mechanisms can contribute to functional novelty at the same time; for instance, it has been reported that duplicated genes usually expand their functionality through differential expression (*Khaitovich et al, 2005; Wapinski et al, 2007a; Conant & Wolfe, 2008*). And genomics has provided evidence above mechanisms. For instance, several studies have supported that modifications in cis-regulatory sequences lead to morphological and behavioral adaptation (*Prud'homme, Gompel & Carroll, 2007; Davidson & Erwin, 2006; Wray et al, 2007* and the references within it), while *S.B. Carroll et al* in their book "From DNA to diversity" present a series of examples that expansion of the developmental toolkit (a core set of developmental genes that are involved in body plan formation across all bilaterian species, e.g. Hox genes) through gene and genome duplication correlates with increased animal complexity. Within the next lines, I will outline the importance of gene expression in the animal evolution, while the impact of duplication will be reviewed thoroughly towards the end of the Introduction.

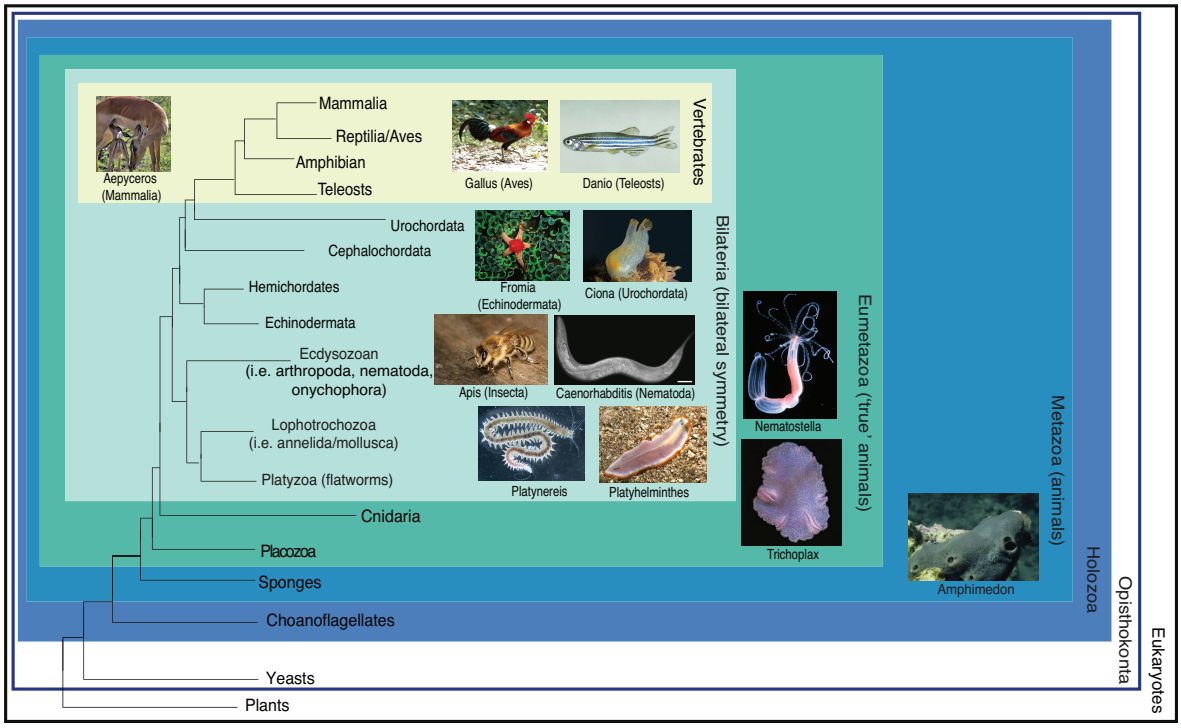


Figure 1: Schematic representation of the phylogeny of Eukaryotes. The topology of the species represents the view of the recent literature (reviewed by *Telford & Copley, 2011; the topology of the basal animals is based on Srivastava et al, 2010)*. Branch lengths are relative to rates of evolution; for instance, ecdysozoa and urochordata include fast-evolving, derived animals. As is obvious different animal clades accommodate many diverse body plans, while the complexity of the latter increases as we move from sponges to mammalia. Sponges are an ancient group of animals (blue square) that diverged from other metazoans (green squares) over 600 million years ago. Bilaterian animals can be further subdivided into Proteostomia and Deuterostomia; the first group includes species such as annelids, nematodes and insects, while vertebrates and other chordate and non-chordates species belong to the second one. Chordata is composed of three subphyla, Vertebrata, Cephalochordata, and Urochordata. Even within the Chordata

clade, which spans 520 millions years of evolution (*Shu, Morris & Zhang, 1996*), there is a large diversity of animal body plans.

2. Insights from comparative transcriptomics for animal evolution

The large effect of gene expression differences on phenotype is evident from the range of cell types seen in a single organism, all of which share the same genome. Generally, cell type number has been used as an index of complexity. Indeed, two periods of major morphological novelty: i) transition to bilaterian animals and ii) transition to vertebrates have been associated with the emergence of novel cell types (*Arendt D, 2008*). Comparative studies of cell type inventories can elucidate the evolutionary diversification of cell types (cell typogenesis) and pinpoint similarities and differences between their gene expression patterns (*Arendt D, 2008; Denes et al, 2007; Tessmar-Raible et al, 2007; Tomer et al, 2010*). Similarly, evolution of tissues and patterns of animal body plans have been better understood by exploring the timing and location of the expression of species-specific transcript repertoires (*Averof & Patel, 1997; Caroll SB, 2000; Prud'homme et al, 2003; Jeong et al, 2008*). In the majority of the reported cases, the regulatory changes occur on duplicated genes, suggesting that gene duplication and gene expression should be studied carefully together. Even though rich information has been collected by small-scale studies, high-throughput experiments allow the study of development and its evolution in a systematic manner.

Microarray experiments have been conducted to study tissue-specificity across species (*Su et al, 2002; Khaitovich et al, 2005; Vaquerizas et al, 2009; Lukk et al, 2010; Chan et al, 2009; Zheng-Bradley et al, 2010*), as well as, developmental timing (*Wardle et al, 2006; Roux & Robinson-Rechavi, 2008; Domazet-Lošo & Tautz, 2010; Paxton et al, 2010; Fang et al, 2010*). Below, I outline few important functional insights of these studies. *Zheng-Bradley* et al. (2010) and *Chan* et al. (2009) show that similar tissues share significant expression patterns across mammalian and vertebrate evolutionary history, respectively. In addition, both studies have reported that genes expressed on a restricted patter show a greater similarity of expression patterns between species. This is quite surprising, even controversial with other studies. *Huminiecki & Wolfe* (2004) have reported that orthologs that have undergone recent duplication are less likely to have strongly correlated expression profiles than those that remain in a one-to-one relationship between human and mouse. *Zheng-Bradley* et al. (2010) and *Chan* et al. (2009) have drawn their conclusions based on single copy orthologs, which are usually conserved over large evolutionary distances (*Ciccarelli et al, 2006*). Given that the rate of evolution is associated with the breadth and intensity of gene expression; for instance, in eukaryotes, the breadth of expression is known to constrain the rate of protein evolution (*Duret & Mouchiroud 2000; Pal et al. 2001; Krylov et al. 2003; Khaitovich et al, 2006*), while *Subramanian and Kumar* (2004) have demonstrated that slow-evolving genes tend to be highly

expressed; it would be expected that they are biased towards broadly, highly-expressed genes. On the other hand, it is clear that fast-evolving genes, which are usually species- or taxon- specific, are expressed in tissues that have adapted to species biology, like reproductive or digestive tissues (*Khaitovich et al, 2006; Axelsson et al, 2008; Khalturin et al, 2009; Sunagawa et al, 2009*) or under species-specific eco-responses (*Colbourne et al, 2011*).

Expression is a dynamic and continuous variable changes with developmental and physiological states. Several studies have used microarray experiments to understand the temporal organization of transcriptome during development. There are several studies that try to identify the developmental timing of expressed genes with their phylogenetic origin (when was arisen in evolution) (*Domazet-Lošo & Tautz, 2010*), their duplication status (*Roux & Robinson-Rechavi, 2008*) and their rate of evolution. Recently, *Fang et al* (2010) have investigated the transcriptome of human embryos from Carnegie stages 9 to 14, covering an important part of human organogenesis, using microarrays. They could identify different clusters of co-expressed genes that progressively regulate the transformation of embryonic stem cells to differentiated organs. At the same time using protein-protein interaction data, they were able to define functional modules of stemness and organ differentiation. Further studies have combined evidence from the field of proteomics and network interactions with transcriptomic data to gain insights on cellular and tissue organization. For instance, it has also been reported that highly interacting transcription factors are broadly expressed across tissues and that roughly half of the measured interactions are conserved among human and mouse (*Ravasi et al, 2010*). As detectable expression differences between species or individuals are not always related to observable phenotypic differences, it might be able to predict more accurately the functional importance of expression variation on phenotype by combining multiple experimenrtal evidence on different evolutionary distances; there are, however, still major challenges to be addressed regarding the data integration and the quality of the datasets (*Nurse & Hayles, 2011*).

B. The quest for orthologs

1. The definition of orthology and its functional implications

To study the evolution of molecular components, one first must establish the correspondence between them across different species and set the framework on which similarity and diversity can be estimated. All entities encoded in genomes (genes, miRNAs, repeated elements etc) can be described using key concepts of evolutionary biology, primarily, the definitions of homologs, orthologs and paralogs. Homology, the most general definition, designates a common origin for the compared entities without specifying an evolutionary scenario. The further – and very essential- classification of

homolgs in orthologs and paralogs was initially presented in the classic work of *Walter Fitch* (1970) (Figure 2). Multi-species comparative studies have revealed the complex evolutionary relationships between the studied genes and the importance of introducing new terms to designate the complete set of phylogenetic relationships. During the first large-scale orthology assignment project of multiple species (*Tatusov et al, 1997*), the concept of clusters of orthologous groups (COGs) was established. A COG consists of proteins that have evolved from a single ancestral sequence existing in the last common ancestor (LCA) of the species that are being compared, through a series of speciation and duplication events. Other newly introduced terms are: i) co-orthologs - lineage-specific expansions produced by duplications of orthologs -, ii) outparalogs - paralogs resulting from a duplication event preceding a given speciation event- and iii) inparalogs – paralogs resulting after a given speciation event (*Sonnhammer & Koonin, 2002*). *E.V. Koonin (2001)* responding on an editorial commentary which questioned the importance of above classification, wrote "These are not just words, after all: they are new memes for the science of a new age". He wanted to point out the necessity of understanding the evolutionary impact of the aforementioned concepts and translating them into functional terms.

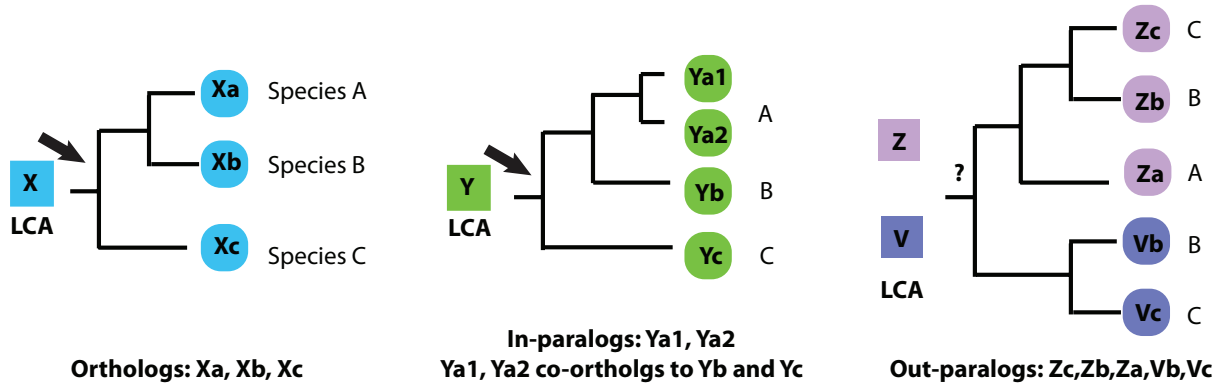


Figure 2: Schematic representation of important evolutionary terms: i) orthologs; genes originating from a single ancestral gene in the LCA of the compared species, ii) Ya1 and Ya2 are in-paralogs; paralogs, meaning genes that arose via duplication, resulting from a lineage-specific duplication subsequent a speciation event (speciation of A, B and C). These in-paralogs are co-orthologs to Yb and Yc; two or more genes in one lineage (A) that are collectively, orthologs to one or more genes in another lineage. Iii) Z and V are paralogs on the ancestral genome and Zc and Vc are out-paralogs in species C; paralogs resulting from duplication(s) preceding a speciation event. All definitions have been adapted by *Koonin EV, 2005*.

Orthology and paralogy, despite being evolutionary designations, have been essential tools of the field of functional genomics. The annotation of newly sequenced genomes and their function prediction depend on robust orthology assignment (*Eisen JA, 1998; Huynen MA et al, 2003; von Mering et al, 2005*). The essential role of orthology lies in the crucial property of orthologs that they often perform equivalent or similar functions in the respective organisms. Of course, there are a number of cases where orthologs carry out different functions (i.e. gene sharing) (*Piatigorsky et al, 1988; Kuhner et al, 2009*). On the other hand, paralogs are often functionally diverged (reviewed

extensively by *Koonin EV, 2005*). The functional differentiation of paralogs is a complex subject that is properly introduced in a subsequent section, as is the major aim of this thesis. At this point, I would like only to note why it is important to distinguishing between orthologs and paralogs by using a frequently occuring evolution scenario, as lineage-specific gene loss. Pseudoorthologs are actually paralogs that appear to be orthologs due to differential, lineage-specific gene loss (*Koonin EV, 2005*); thus, any functional inference between the pseudoorthologs would be inappropriate, as they might have diverged in function already in the ancestral genome. Similar scenarios can be caused by further factors that are summarized in the following section.

2. Caveats of orthology prediction

Accurate orthology prediction is challenged by a number of biological and technical factors. For instance, I have already outlined above that duplications, especially, if they have taken place on internal branches of a phylogenetic tree and/or followed by lineage-specific losses, create complex evolutionary scenarios. Mucins, an animal-specific family exemplifies nicely a few of the caveats of orthology prediction (Figure 3), and of phylogenetic analyses in general (*Lang, Hansson & Samuelsson, 2007*). The phylogenetic tree of mucins resolves the orthlogous relationships among the members of the family in every pair of species (*Hydra*, fruitfly, *Ciona*, zebrafish, chicken, mouse and human). The decision as to how mucins are grouped into orthologous groups (OGs) depends on the phylogenetic range of the species compared. In general, to define the state of LCA (single vs duplicated sequence) and set the boundaries of an OG, there should be taken into consideration an outgroup species. Herein, *Hydra* sequences reveal the existence of two paralogous sequences in the LCA of bilaterians; thus, the descendants of each sequence should be clustered in two different OGs. In a similar manner, each bilaterian OG should be further separated in more fine-grained vertebrate-specific OGs (Figure 3). Thus, analyzing the OGs at different taxonomic levels (e.g. vertebrates vs. bilaterians) sheds light on the evolutionary history of the family.

Additionally, mucins have a very complicated protein structure, introducing further obstacles to the quest of orthology annotation (Figure 3), as the multiple domain architectures exist across the different members of the family, which are not always conserved across species. The latter is one of the most important problems in orthology prediction, especially for the eukaryotic genomes. Generally, the vast majority of proteins contain only one domain and the most common multi-domain proteins tend to have few (2 or 3) domains. In metazoans, however, there is a larger fraction of larger multi-domain proteins, which has been associated with animal complexity (*Lehner & Fraser, 2004; Tordai et al, 2005*). Due to a variety of genetic processes (duplication, inversion, recombination, retrotransposition etc.) (*Copley, Letunic & Bork, 2002; Ciccarelli et al, 2005; Koonin EV, 2005; Campillos et al, 2006(b)*), these proteins, usually, consist of domains with independent evolutionary

origins. The latter leads to conceptual but also practical challenges (e.g. alignment) in orthology prediction, as the domains have followed distinct evolutionary trajectories.

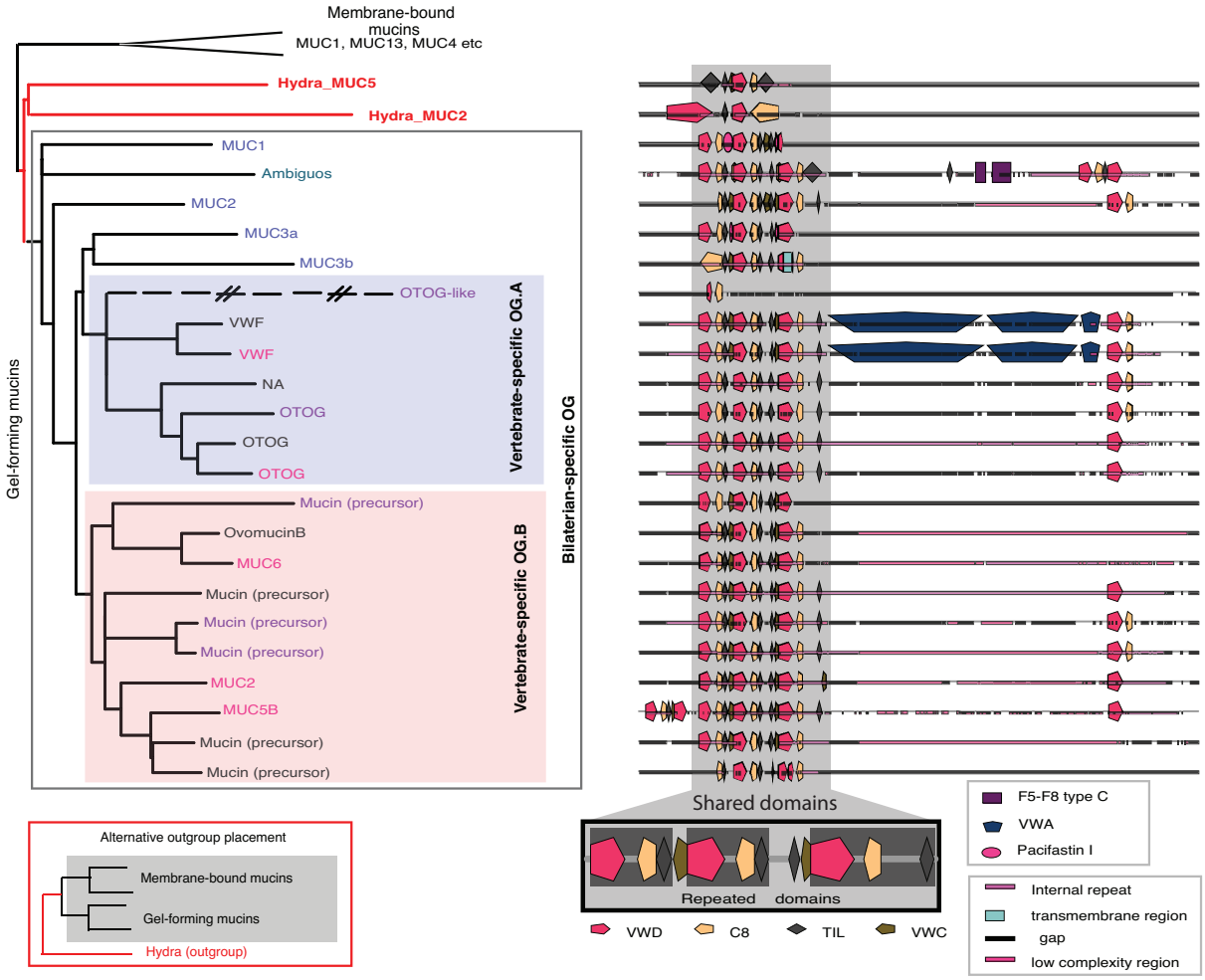


Figure 3: Mucins: a challenging family for orthology prediction. This figure shows the phylogenetic tree and domain architecture of aligned mucins. The identification of cnidarian (an outgroup for bilaterians) mucin2/5 orthologs separates the gel-forming mucis from other mucins, defining a bilaterian-specific OG (grey box). An alternative topology of Hydra in respect to the LCA of bilaterian species (shown schematically in the red box) would propose that those two different classes of mucins should be clustered together at the bilaterian level. The bilaterian OG can be further resolved at the vertebrate-level into OG.A (blue) and OG.B (red), illustrating the hierarchical nature of OGs. This family, besides its large size due to vertebrate-specific duplications, exemplify 5 additional problems that often lead to orthology miss-assignment: 1) Uneven evolutionary rate illustrated as branch lengths, lowering the sequence similarity among members of the family, 2) Quality of genome annotation: the particular zebrafish protein can be either a derived member of the mucin family or a erroneous gene prediction, 3) Repeated domains: the domain combination VWD-C8-VWC (Von Willebrand factor C), which is the core of the family, is repeated multiple times within the protein, 4) Complexity of domain architectures: there are multiple unique domain combinations (e.g. the VWD (Von Willebrand factor D) domain is combined with the F5-F8 type C only in the Drosophila ortholog) and 5) Low complexity regions: internal repeats within the amino acid sequences and other low complexity features impede the correct sequence alignment of the mucins. *Possible orthologous sequence at the LCA of cnidarians-bilaterians.

To conclude, a proper phylogenetic analysis including outgroup species and the subsequent phylogenetic tree of a gene family are the most appropriate method for disentangling orthologs and paralogs, as all pairwise relationships are evident and complicated scenarios that have arisen via duplications followed by species- or lineage-specific losses can be better resolved. Although the operational definitions to describe these relationships exist, there are additional factors that affect the OG accuracy starting from the phylogenetic range of species to protein domain architecture.

3. Orthology detection methodologies: advantages and disadvantages

A plethora of methods that automatically predict orthologs among organisms has been developed (*Muller et al, 2010; Waterhouse et al, 2011; Altenhoff et al, 2011; Huerta-Cepas et al, 2011; Vilella et al, 2009; Ruan et al, 2008*). Despite the fact that tree representation is the most appropriate approach to orthology prediction, not all prediction methods decipher orthology via tree topology. As a result, they can be classified into a) tree-based and b) graph-based methods (reviewed extensively in *Kuzniar et al, 2008; Gabalodon T, 2008; Krinstensen et al, 2011*). In all methods, homology detection is the first step. Tree-based methods collect homologs, which use to construct an MSA and phylogenetic tree and their further discrimination to orthologs and paralogs is based either on reconciliation with a species-tree (*Huerta-Cepas et al, 2011; Viella et al, 2009; Ruan et al, 2008*) or on presence/absence of species (*van der Heijden et al, 2007; Datta et al, 2009*). On the other hand, in graph-based methods, the homology detection step is followed by a clustering step. The clustering algorithm varies among the different methodologies. For instance, COG/KOG (the first multiple species orthology pipeline) infers orthology using congruent 'triangles' of best reciprocal hits BRHs (*Tatusov et al, 2003*), while OrthoMCL applies a Markov Cluster algorithm approach (*Chen et al, 2006*) (Figure 4). A few of the methods to refine the orthology assignment use synteny. In fact, the inclusion of synteny information limits the errors due to low sequence similarity and increases orthology accuracy (*Goodstadt & Ponting, 2006; Byrne & Wolfe, 2005*). However, this requires a certain level of synteny conservation among the compared species.

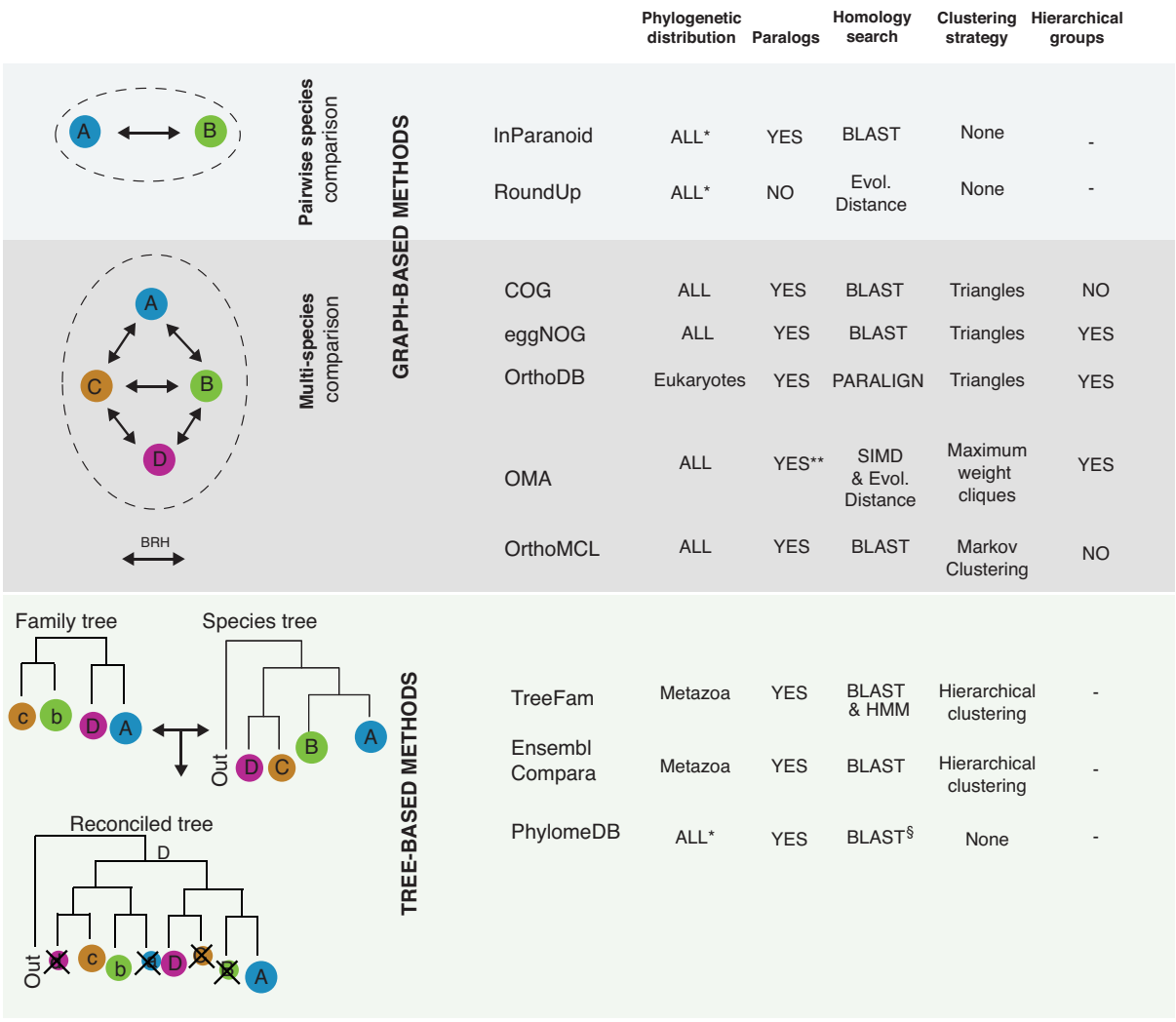


Figure 4: Comparison of a few orthology prediction methods. In general, the methods of orthology prediction can be classified into (i) graph-based (infer orthology using clustering algorithms) and (ii) tree-based methods (infer orthology through tree topology). Different graph-based methods are designed to assign orthologous relationships for two (pairwise) or more (multiple) species. Usually, they cluster proteins into Orthologous Groups (OGs) based on their similarity scores. Phylogenetic Distribution describes the species range of each database. Homology Search shows the heuristics or other approaches that each resource applies to recruit orthologs. **: Supplies OGs that their members share only orthologous relationships. *: The user can compare any two genomes spanning a phylogenetic distance from bacteria to animals.

In the modern biology era, graph-base methods have taken the lead; eggNOG (*Muller et al, 2010*) and OMA (*Altenhoff et al, 2011*), the biggest multi-species repositories, accumulate orthology information for more than 1000 species. In comparisons including large numbers of species, graph-based methods achieve a better trade-off between speed and accuracy than tree-based methods (*Kristensen et al, 2011*). Additionally, tree-based methods are computationally expensive and at times fail due to the complexity of the family or the substantial number of species in the comparison

(*Kristensen et al, 2011; Pryszcz, Huerta-Cepas & Gabaldón, 2010)*. Finally, orthology via tree-reconciliation requires a species tree that might not be consensually accepted for certain phylogenetic lineages (*Telford & Copley, 2011),* such as metazoans, or might even be highly questionable, as in bacteria that might have evolved through a phylogenetic net (*Doolittle WF, 1999*).

Taking into consideration all the above differences, it would be expected that the resulting orthology predictions vary considerably among the different repositories, as is indeed the case (*Pryszcz, Huerta-Cepas & Gabaldón, 2010*). Several studies conducted in recent years have dealt with the comparison and quality assessment of orthology predictions (*Hulsen et al, 2006; Chen et al, 2007; Altenhoff & Dessimoz, 2009*). Functional consistency of the predicted orthologs is the most common evaluation method. However, orthology is an evolutionary term and functional equivalences are not always inferable, as gene sharing that was mentioned above (for more examples please see *Koonin EV, 2005*). Moreover, the functional divergence between orthologs and paralogs (sub-/neo-functionalization of paralogs) or alteration of function during long evolutionary distances suggests that those tests are biased towards single copy genes or conserved families and less suited for large diversified families. Thus, a new evaluation approach is required, which is one of the aims of this thesis.

C. Functional novelties through functional divergence: the role of duplication

The concept of evolutionary innovation via gene duplication was coherently developed in Ohno's famous book "Evolution by gene Duplication" (1970). He proposed that gene duplication leads to functional novelty during evolution as one of the newborn paralogs escapes the selective constraints and become free to evolve a new function. Ohno also suggested that big leaps in evolution should happen through duplication of whole genomes (polyploidization); in particular, he hypothesized that at least two whole genome duplications had occurred in the ancestral vertebrates. The availability of completely sequenced genomes has sparked renewed attention on this subject. Nowadays, there are several hundreds of finished or ongoing metazoan genome projects in public repositories (Genomes OnLine Database v3, July 2011) (*Liolios et al, 2009*). Comparative analyses of the vertebrate genomes, including species that are basal (most deeply branching) members of the clade or outgroups (e.g. Amphioxus or ascidians), have inferred the state of the ancestral genome before the evolution and radiation of more recent groups (*Dehal et al, 2002; Sodergren et al, 2006; Putnam et al, 2008*). The presence of multiple copies of many transcription factors and other developmental genes (*Garcia-Fernandez & Holland, 1994; Holland et al., 1994; Meyer & Schartl,*

1999; Larroux et al, 2008; *Putnam et al. 2008*) compared to the single copy orthologs in the outgroup species, as well as, large syntenic regions in vertebrate genomes (*Postlethwait et al, 2000; Jaillon et al, 2004; Nakatani et al, 2007; Catchen, Conery & Postlethwait, 2009; Denoeud et al, 2010*) support Ohno's hypothesis of whole genome duplication events during the evolution of vertebrates. Additionally, an increasing amount of evidence has supported the mechanism of duplication as a major source of adaptation to new environments and speciation (*Scanell et al, 2006; Sémon & Wolfe, 2007; Conant & Wolfe, 2007 (a), Hittinger & Caroll, 2007; Ames et al, 2010; Colbourne et al, 2011; Gonzales-Vigil et al, 2011; Jiao et al, 20110*). Below I will outline recent insights in the mechanism of duplication, the fates of the duplicated genes and the consequences of the functional divergence.

1. Mechanisms of duplication: from single genes to whole genomes

Many genomic studies of gene duplication have focused on the mechanisms responsible for generating duplicate genes. Genetic mechanisms such as unequal crossing over or retrotransposition are the most common source of duplicated genes (*Edlund & Normark, 1981; Plaitakis et al, 2003; Cusack & Wolfe, 2007; Jun et al, 2009*). The former usually results in paralogs with similar intron-exon structures, while the latter produces intronless genes that make the detection of their common evolutionary history difficult (*Zhang J, 2003*). The aforementioned mechanisms commonly involve a single or a few genes, and hence are called small-scale duplications (SSDs); in some cases, however, duplications have been detected spanning several genes at the same time or even a whole chromosomal segment (segmental duplications) (*Gaudieri S et al, 1997; Venter et al, 2001; Wong et al, 2004; She et al, 2008*). Finally, the most radical mechanism of duplication is the whole genome duplication (WGD), or otherwise known as polyploidy, which has been detected in all eukaryotic lineages (*Wolfe and Shields, 1997; Christoffels et al, 2004; McLysaght et al, 2002; Blanc et al., 2000; Dehal & Boore, 2005*). Some types of duplications have been identified, so far, only in certain lineages. For instance, large-scale segmental duplications have been detected in all primate genomes (*Bailey et al., 2002; Bailey & Eichler, 2006*). On the other hand, almost all eukaryotic lineages such as animals, fungi, protists, and especially plants have undergone one or more rounds of WGDs in their evolutionary past. For example, in animals, two successive rounds of WGDs occurred at the origin of vertebrates (the 2R event) (*Dehal and Boore, 2005; Panopoulou and Poustka, 2005*) and one in the bony fish lineage (the 3R event) (*Jaillon et al., 2004; Meyers & Van de Peer, 2005*). In the yeast lineage, a WGD occurred around 100 million years ago (*Wolfe and Shields, 1997*), whereas in the ciliate Paramecium, 3 or 4 WGDs have occurred (*Aury et al., 2006*). In plants, one or two genome duplications are shared between all flowering plants, whereas many of them have undergone additional rounds of polyploidization (*Blanc and Wolfe, 2004; Cui et al., 2006*).

2. Subsequent fates of duplicated genes

One benefit of the genomic era is that it has provided a complete and unbiased view of the landscape of duplicates in each genome. A landmark paper by *Lynch and Conery* (2000), which has demonstrated the relaxation of selective constraints on duplicated genes in eukaryotes, is one of the first genome-wide studies estimating the rates of birth and death of duplicated genes. Many duplicated genes have a short lifespan, as one of the two copies is either lost or degenerates and becomes nonfunctional (nonfunctionalization) (Figure 5). In fact, the two rounds of WGD in the vertebrate lineage have been followed by a period of excessive gene losses (*Lynch & Conery, 2000*). The same evolutionary scenario has been observed in the yeast clade as well (*Scannell et al, 2006*).

In the relatively rare cases in which both copies are retained in the genome, one copy can diverge and acquire a novel function that is completely different from the ancestral one (neofunctionalization), or the two duplicated genes partition the ancestral function (subfunctionalization) (*Force et al, 1999*) (Figure 5).

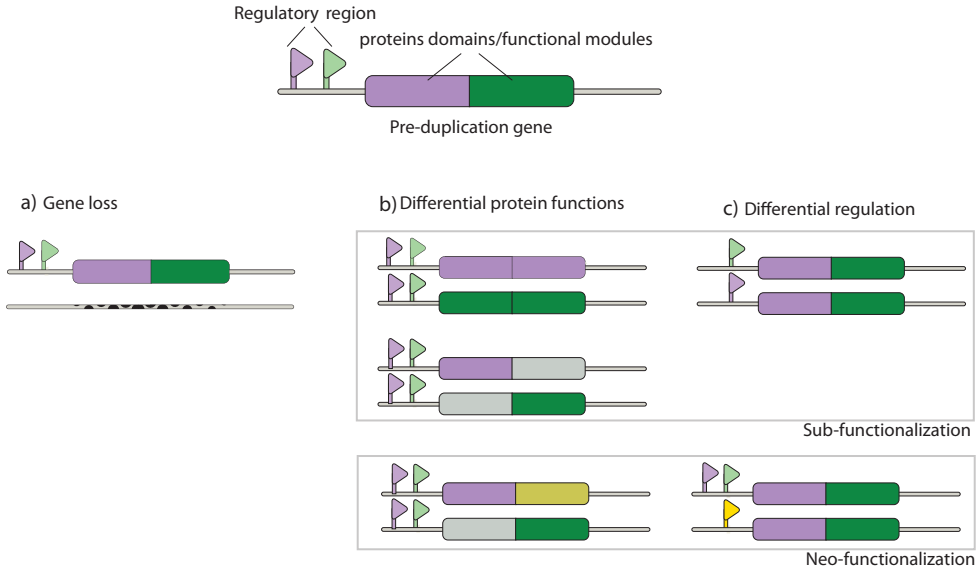


Figure 5: The subsequent fates of duplicated genes. a) A common outcome is loss of one of the two copies of the duplicated gene by deletion or degeneration (non-functionalization). b) Alternatively, the two copies of a gene can either diverge in sequence, resulting into complement or non-overlapping functions (sub-functionalization) or even expand their functionality (neo-functionalization) (e.g. through new domain acquisition, i.e. gene fusion). c) A Finally, the regulatory regions of the two copies can diverge; either to share the ancestral functionality or to expand it. Of course, all possible combinations of these scenarios are possible and usually the regulatory and structural modifications are accompanied (*Khaitovich et al, 2005*).

Sub- and neo-functionalization allow spatiotemporal specialization and expansion of functionality, respectively, but it is usually hard to draw a line between the two fates (*He & Zhang, 2005*). *Conant & Wolfe (2007)* propose that what we consider as a "new" function is a secondary, inferior function of the ancestral sequence, which under certain conditions is exapted. Generally, sub- and neo-functionalization can be achieved through changes in amino acid sequence (*Merritt & Quattro, 2003*) or through changes in genes expression patterns (*Bassham et al, 2008*). In any case, a multi-functional ancestral gene diverges either at the regulatory or sequence level to facilitate adaptation. Since genes with a larger number of cis-regulatory regions, expressed in many tissues (*Lynch et al. 2001*) or encoding multi-domain proteins (*Gibson and Spring 1998; Stoltzfus 1999*) are preferentially preserved, makes the hypothesis of co-option more plausible.

3. Which genes undergo duplication?

Understanding why certain duplicates are retained in the genome to generate multi-gene families, while others have degenerated and been lost, resulting in single copy orthologs, not only will provide insight into the mechanism of duplication, but, will ultimately help us to understand how phenotypic novelty is gained (*Seoighe and Wolfe 1999; Lynch and Conery 2000; Dermitzakis and Clark, 2001; Gu et al. 2002; Kitami and Nadeau, 2002; Wapinski et al, 2007 (a)*). It has been well established that there are functional biases on the molecular functions that have been retained in duplicated copies (*Kondrashov et al. 2002; Davis and Petrov, 2004; Paterson et al, 2006*). Studies from many different organisms (bacteria, yeasts, plants, and animals) have shown that transcription factors, kinases, enzymes and transporters are the most common classes of duplicated genes (*Taylor & Raes, 2005* and references therein). Many genomic studies have focused on how the evolutionary opportunities that gene duplication provides differ depending on whether the duplicate gene pair in question has been formed by whole-genome duplication or by single-gene duplication (reviewed extensively by *Conant & Wolfe, 2007(b)*). The functional categories of duplicate genes retained after WGD are similar across diverse lineages (*Blanc & Wolfe, 2004; Aury et al, 2006; Taylor & Raes, 2004; Paterson et al, 2006*), including ribosomal proteins and kinases. In fact, these two mechanisms (WGD and SSD) can produce different kinds of adaptations (*Wapinski, et al, 2007*); for instance, based on the "dosage balance hypothesis" (*Veitia et al, 2008*), WGD paralogs should not affect the stoichiometric balance of a complex or pathway, and thus are more likely to be retained (*Deluna et al, 2008*). Actually, in plant and yeast it has been found that genes belonging to a functional category are duplicated either via SSD or WGD (*Maere et al, 2005; Guan, Dunham, & Troyanskaya, 2007*).

On the other hand, WGD has been associated with speciation events through lineage-specific gene losses, suggesting that there is a balance between general biological trends (certain functional

classes can be retained after WGD events across different lineages) and environmental constraints (functional class A is preferably retained in environment X, but not class B). *Scanell* and co-authors (2006) have argued that the alternative loss of duplicated genes has led to the speciation between *S. cerevisiae, S. castellii* and *C. glabrata* through sexual isolation. Similarly, two and three rounds of WGD predate the radiation of vertebrates and teleosts, respectively (*Sémon & Wolfe, 2007*). The initial polyploid genome of the vertebrate ancestor through a series of chromosomal rearrangements, losses of genes, and expansions of gene families has diversified into different vertebrate lineages. The evolutionary history of plants has multiple examples of polyploid genomes; however, only recently, with the increasing number of plant genomes, has it been possible to date accurately two whole genome expansions; one in the ancestor of all seed plants and another one in the ancestor of angiosperms (*Jiao et al, 20110*).

There are multiple pieces of evidence that duplication facilitates adaptation to new environments (*Hittinger & Caroll, 2007; Ames et al, 2010; Colbourne et al, 2011; Gonzales-Vigil et al, 2011*). Plants synthesize a large number of compounds as defense mechanism against herbivory. Many of the enzymes used for the production of this chemical repertoire have evolved through gene duplication (*Gonzales-Vigil et al, 2011*). Recently, the genome of a cosmopolitan crustacean, *Daphnia pulex*, revealed a large number of lineage-specific duplicated genes. *Daphnia* is a model-organism for polyphenism, a trait through which multiple discrete phenotypes have emerged from a single genotype through environmental induction. The majority of duplicated genes of this phenotypically plastic animal have diverged their expression soon after birth to adapt to environmental conditions (*Colbourne et al, 2011*).

To summarize, a plethora of orthology prediction repositories provide orthology assignments for different sets of species and genome annotations using their own developed methodology. This leads to large inconsistencies between the predictions. Given the vital role of orthology in the modern biology era, this uncertainty is worrying. Therefore, the major aim of this thesis has been to establish a quality control dataset for orthology prediction (chapter D). At the same time, I have focused on the evolution of functional organization in Eukaryotes and, in particularly, the role of paralogy on their spatiotemporal patterning (Chapters E & F). Although a well-studied topic, there is still obscurity as to how those patterns have emerged, especially for the multicellular organisms, like plants and animals, with different levels of organization (i.e., cells and tissues. Since I had the opportunity to focus on different aspects of systems biology, I briefly introduce the aim of each project at the beginning of each chapter.

Results

D. A phylogeny-based test for metazoan orthology prediction

For an extended discussion of the results, please see Appendix A. It includes the manuscript and the supplementary materials by *Trachana et al, 2011*.

1. Aim of the project

I outlined in the Introduction the importance of orthology in functional and comparative "omics" and how robust orthology prediction can be hindered by several biological or technical factors. Therefore, the quality assessment of the predictions is indispensable. Thus far, the majority of quality assessment tests have been based on the functional conservation of predicted orthologs (*Hulsen et al, 2006; Chen et al, 2007; Altenhoff & Dessimoz, 2009*); however, there are certain biases on these approaches (Appendix A). At the same time, the orthology community has acknowledged that a phylogeny-based evaluation would be more appropriate (*Gabaldon et al, 2009; Boeckmann et al, 2011*). This project aims to generate a phylogeny-based test, which evaluates the accuracy of orthology predictions for single copy to complex large families.

2. Design, generation and application of a benchmark set for bilaterian orthology

There is extensive literature on factors that hinder the accuracy of orthology predictions (i.e. duplications (paralogy)/ losses, domain architecture) (extensively reviewed on *Koonin EV, 2005 and Kuzniar et al, 2008*). With a view to understanding their impact, we selected 70 protein families that range from single copy orthologs to OGs with one hundred members (Appendix A; Table S1). The phylogenetic analyses were performed for 16 species; 12 bilaterian spanning from nematodes to mammals and 4 outgroups (cnidaria, *Trichoplax* and *Monosiga*). The details about the phylogenetic analyses are shown in Appendix A (Box 2). The manually curated benchmarking set was used for two different comparisons: 1) with the automatically predicted OGs of five publicly available databases and 2) with different customized versions of the in-house database; eggNOG (*Muller et al, 2010*). The second analysis took place to quantify the effect of two confounding variables of the first comparison, namely, species representation/distribution and genome annotation quality.

3. Quantifying the impact of biological complexity on orthology prediction

After classifying the RefOGs based on their size, which is related to duplication events, we observed that the numbers of missing orthologs and RefOG fissions correlate significantly with the family size for all methods (Figure 3 in Appendix A). Additionally, the proportion of accurately predicted RefOGs decreases as the number of average domains per family increases (Figure 3 in Appendix A). Interestingly, the error source that significantly correlates with the complex domain architecture is the rate of erroneously assigned genes, suggesting that protein families with multiple protein domains "attract" non-orthologous proteins due to domain sharing.

The rate of evolution and the quality of MSA affect the number of missing orthologs for the graph-based methods (eggNOG, OrthoDB, OMA and OrthoMCL). Those approaches tend to accumulate a larger number of missing orthologs as the MSA quality drops or the rate of evolution increases. On the other hand, TreeFam, a tree-based method, is significantly more influenced by MSA quality rather than the rate of evolution. Again, this is not surprising as TreeFam uses MSA for tree-building and reconciliation steps to infer orthology, thus alignment quality is an essential standard step. Taken together, classification of the families from slow-evolving single copy to fast-evolving large families revealed method-specific limitations. In all cases, complex families failed to be predicted accurately.

4. Estimating the impact of other confounding factors.

Biological complexity is unlikely to be the primary source of errors in automatically predicted OGs, as there are single-copy, slow evolving or single-domain protein families in our dataset, which are not assigned correctly by several prediction methods. By investigating these families, we identified additional technical factors that influence orthology assignment including species range, species coverage and genome annotation. Species range is, actually, the most important confounding variable in public database comparisons. In Appendix A, we have interpreted some of the detected differences among the databases based on their species distribution. Additionally, to measure the impact of species coverage, we prepared new OGs using the 12 reference species with the eggNOG pipeline. Although, the phylogenetic range of the customized and public available eggNOG used in this study is the same (human to nematodes); the metazoan level of public eggNOG dataset contains double the number of species. We have identified that 30% of the missing genes in this dataset are due to the change in species coverage (Figure 6). It seems that sequences of the 34 species facilitate correct clustering, presumably, by breaking long branches so that faster evolving genes can be connected. Again more details for this comparison can be found in Appendix A. Finally, to directly test the effect of the genome annotation quality, we generated OGs for the 12 reference species based on the

Ensembl v60 gene annotations. We found 45% fewer erroneously assigned genes (149 vs. 271) in the 12-species-new-annotation-OGs compared to the 12-species-old-annotation-OGs (Figure 6). However, the number of missing genes is similar between the two datasets and higher compared to the 34-species-OGs, highlighting, once again, the impact of species coverage.

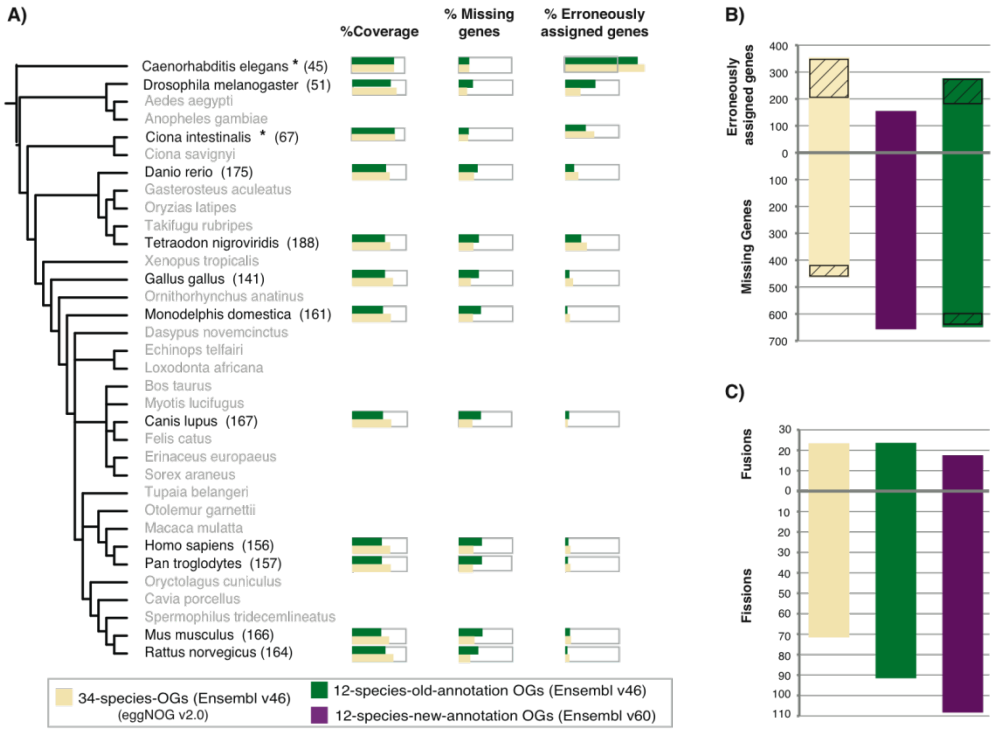


Figure 6: The impact of species coverage and genome annotation. (A) Comparison of the error rate for the 34-species and 12-species OGs using RefOGs. The genome annotation is the same for both datasets (Ensembl v46). We measured the percentage of the reference orthologs that were recovered (coverage), missing orthologs and erroneously assigned genes for each of the 12 reference species. The reference species are highlighted by black letters, while the extra species to complete the set of 34 species are written in grey letters. Numbers in brackets show the total amount of orthologs per species in the benchmarking set. The grey boxes enclosing the colored bars correspond to 100% coverage. Notice that the coverage is always higher for the 34-species-OGs compared to the 12-species-OGs except in the cases of *C.elegans* and *C.intestinalis* (marked by asterisk), which are separated by long branches in both datasets. (B,C) Comparison of 34-species (yellow bar) and 12-species (green bar) OGs with 12-species OGs using new annotation (purple bar) at the gene (B) and group (C) level. Shattered boxes label the fraction of mispredicted genes of 34-species and 12-species-old annotation datasets that do not exist in Ensembl v60 genome annotations, indicating the high number of errors due to old genome annotations. Notice that the 12-species datasets (either with old or new annotation) always introduce a larger number of fission events than the 34-species-OGs, highlighting again the importance of species coverage.

To conclude, in the course of this thesis, a phylogeny-based benchmark set of orthology prediction was generated. Despite the fact that it is focused on an animal clade, it has been proven a valuable tool to quantify the impact of biological and technical caveats on orthology prediction. To test its applicability, we compared five commonly used databases. At the end, method-specific errors, hidden correlations and confounding variables were revealed. All tested algorithms need to be improved to be able to handle the "complex" families (duplication/losses, complex domain architectures). For the eggNOG database, the in-house orthology pipeline, we were able to estimate the error rate related to species coverage and genome annotation. ~40% of the mispredicted genes in eggNOG OGs would have been avoided by using an updated version of genome annotations, highlighting the importance of frequent updates of orthology repositories. At the end, the ultimate outcome of this project is the choice of the most "trust worthy" repository. The interplay between specificity and sensitivity seems to be balanced for repositories with similar phylogenetic range as the user's interest; for instance, for studies on vertebrate species repositories as eggNOG and OrthoDB that provide OGs for this level are more suited. For a certain phylogenetic depth, the greater the number of species that are used to infer orthology, the higher accuracy overall is achieved. These important factors have been taken into consideration in the next functional analyses.

E. The role of paralogy in the temporal orchestration of the cell

For an extended discussion of the results, please see Appendix B. It includes the manuscript and the supplementary materials by *Trachana, Jensen & Bork, 2010*.

1. Aim of the project

There are many studies focusing on the evolution of duplicated genes in spatial scales (i.e. tissue-specific expression), but the role of duplicated genes in the temporal organization of the cell remains unclear (*Wagner A, 2002; Gu et al, 2002*). Clocks, rhythms and cycles are universal from unicellular to multi-cellular organisms and coordinate many biological pathways that respond to extracellular or intracellular signals to consequently adapt the organism to periodically changing environments. In multi-cellular organisms, as animals and plants, a 24hour diurnal rhythm (circadian clock) has been detected (*Doherty & Kay, 2010*), while in the unicellular budding yeast, a robust ~40min metabolic cycle has been reported (*Klevecz et al, 2004*). This project investigates how the cellular periodic processes are organized and how gene/genome duplication contributes to it.

2. Design of the analysis

Jensen et al (2006) previously identified 600, 400 and 600 cell cycle-regulated genes in budding yeast, Arabidopsis and human, respectively. To identify diurnal and ultradian regulated genes, we analyzed multiple time-series microarray experiments using the same algorithm as the cell cycle study to minimize technical noise (detailed experimental procedure is available in Supplementary Information in Appendix B). An important caveat of this study is the identification of diurnal regulated genes, as they are expressed in a tissue-specific manner (*Delaunay & Laudet, 2002*). To eliminate this effect, we combined the transcriptomes of several experiments that analyzed different tissues. At the end, we detected 600 ultradian-regulated budding yeast genes, 600 diurnal-regulated Arabidopsis genes and 491 diurnal-regulated human genes. The orthology/paralogy relationships of the temporally regulated genes were obtained using the eggNOG pipeline (*Jensen et al, 2008*).

3. Paralogs with distinct temporal regulation in eukaryotes

Mapping the genes to a set of eukaryotic orthologous groups revealed that there is a significant enrichment of cell cycle/diurnal regulated paralog pairs in human and Arabidopsis (Figure 1 in Appendix B). Although, cell cycle and diurnal regulated genes do not significantly overlap in the two multi-tissue species. Similarly, we can identify 58 paralogs that have diverged their regulation under cell cycle and ultradian rhythm in budding yeast (Figure 2 in Appendix B). Contrary, to human and Arabidopsis results, yeast has a significant number of shared genes between the two temporal processes. Still, the current data does not provide enough evidence to distinguish between sub- and neo-functionalization, as the three studied species spanning a large evolutionary distance and any inference of the ancestral state would have been more than inappropriate. For simplification below, I will refer to this functional divergence as subfunctionalization.

4. The temporal subfunctionalization of paralogs has evolved in parallel in the three eukaryotic lineages.

Despite the common trend of temporal subfunctionalization in all three species, the functional repertoires of cell cycle/circadian regulated paralogs are different in Arabidopsis, human and yeast (Appendix B). The functional analysis of the cell cycle – ultradian orchestration was more insightful due to the larger number of genes. Mapping the cell cycle/ultradian regulated proteins to the metabolic network of *S.cerevisiae* (Figure 7), revealed that cell cycle/ultradian sub-functionalization has frequently occurred in paralogs that regulate important metabolic substrates (e.g, glucose, pyruvate and sulfate). For example, glucose is transported by the HXT transporters; a subfamily of which composes a cell cycle/ultradian regulated paralogous group. Since the functional repertoires of subfunctionalized paralogs have been accommodated to species biology, the most parsimonious scenario is that this mode of regulation has evolved independently in these organisms, or in their lineages. This is further supported by the fact that the temporal subfunctionalized paralogs in budding yeast compose the 10% and 5% of the SSD and WGD pool, respectively. It seems that there is a stronger selection on SSDs, which enhances the idea of a lineage-specific functional repertoire of periodic divergent paralogs.

To conclude, all above suggest that the orchestration of cellular pathways under different periodic processes provides a selective advantage and that use of temporal regulation of newly emerging paralogs in different contexts (i.e. distinct cyclic processes) appears to be an efficient way to achieve it. As the functional repertoire of these duplicated genes in yeast, plant and animals are different, we

hypothesize that gene duplication and subsequent sub-functionalization have taken place independently during evolution.

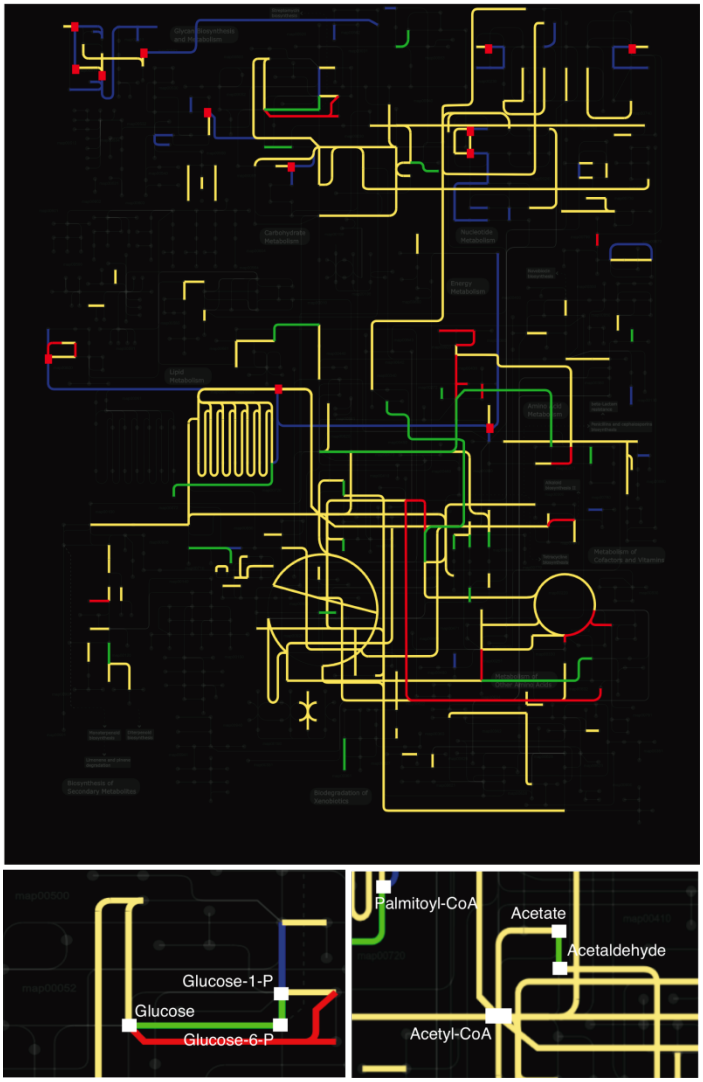


Figure 7: Core metabolic network of cell cycle and ultradian rhythm regulated genes in *S.cerevisiae*. The core metabolic network of *S.cerevisiae* is shown in yellow. Reactions that are catalyzed by cell cycle and ultradian regulated genes are highlighted with red and dark blue, respectively, while cell cycle/ultradian regulated paralogs are mapped with green lines. Metabolic substrates that are under cell cycle and ultradian regulation are indicated with red squares. A few of common cell- cycle- and ultradian-regulated substrates, like glucose-6-phosphate and acetate, are important for glycolysis and fatty acids biosynthesis, respectively, in budding yeast. The custom metabolic map shown here was generated using iPath (*Letunic et al., 2008*).

F. Studying the origin of tissue inventories and their functional divergence

This section outlines the results of the last year of my thesis, which have not yet been submitted to a peer-reviewed journal. It is still an on-going project and future experiments are discussed on the next chapter.

1. Aim of the project

Comparative analyses of the expression patterns of modern cell types ("molecular fingerprints") have suggested that ancestral cell types should be multifunctional (reviewed extensively in *Arendt D, 2008*). The specialized descendants of the latter may evolve through divergence of the ancestral functions or acquisition of a new function. This process can be parallelized to sub- or neo-functionalization of duplicated genes, respectively (Figure 5). In the same manner, the specialized tissues of vertebrates should have been evolved through functional divergence of multifunctional tissues/cell types in their ancestor; indeed, multifunctional structures exist in basal chordates (e.g. amphioxus or tunicates). For instance, the chordate endostyle, the homologous structure of thyroid in vertebrates (*Venkatesh et al, 1999; Hiruta et al, 2005*), is a pharyngeal structure. This implies that pharynx and thyroid have a common origin, even if they are functionally distinct in extant vertebrates, and raise the question if we can identify remnants of their common origin by comparing their transcriptomes. Furthermore, we want to understand the role of gene duplication in the divergence of these two structures. So far, comparisons of tissue-specific transcriptomes within species have reported that spatial (tissue) divergence between duplicate genes increases with evolutionary time (*Makova & Li, 2003; Blanc & Wolfe, 2004; Khaitovich et al, 2005*), suggesting that the age of duplications should be taken into consideration. Phylostratigraphy, a new methodology, which correlates the origin of genes (age) with macroevolutionary transitions (*Domazet-Lošo & Tautz, 2007*), has been proven a valuable tool to associate developmental transitions using transcriptomic data (*Domazet-Lošo & Tautz, 2010*). Herein, I present a meta-analysis of 31 human-tissues-transcriptomes using a similar approach to phylostratigraphy to associate the origin of the tissues with the origin of the expressed gene families and their duplication patterns.

2. Tissue transcriptomic datasets and biases

The studied tissues were selected based on their embryonic origin and the availability of transcriptomic data in public repositories, precisely in ArrayExpress (*Parkinson et al, 2007*). The ArrayExpress team has constructed a global gene expression map by integrating data from a large number of microarray experiments representing 369 different cell and tissues types, disease states and cell lines (*Lukk et al, 2010*). Currently, to simplify the task and be able to make direct associations between expression and functionality, we focused on the up-regulated genes. 31 out of the 75 human tissues, which are represented in the aforementioned dataset, are analyzed here. Tissues were excluded to avoid either statistical (i.e. tissue with a very small number of up-regulated genes; 44 up-regulated genes in rectum vs. >2000 up-regulated genes in esophagus) or biological (e.g. tissues tested for drug treatments or other conditions; smoker lung tissues) biases. The embryonic origin of the tissues is presented in Table 1 and highlights the two major tissue classes in our data. Each of them includes 10-13 tissues, thus any gene expressed in more than 13 tissues is considered as broadly expressed.

Brain tissues		**Foregut/gut related tissues**		**Other tissues**	
Name	Genes	Name	Genes	Name	Genes
Amygdala	2350	Tongue	433	Adrenal gland	1354
Caudate nucleus	2541	Hypopharynx	2192	Kidney	5437
Globus pallidus	3013	Oropharynx	1510	Bladder	2818
Hippocampus	2318	Esophagus	2237	Uterus	595
Frontal cortex	2404	Trachea	2823	Smooth muscles	2685
Olfactory bulb	694	Lung	5507	Heart	4993
Thalamus	887	Thymus	1991		
Hypothalamus	2461	Tonsil	1112		
Pituitary	552	Thyroid	3272		
Cerebellum	3808	Liver	3755		
Cerebellum penduncles	922	Pancreas	585		
Medulla oblongata	1282	Colon	1663		
Pons	1987				

Table 1: The transcriptomes of 31 studied tissues. Tissues have been separated based on their embryonic origin or body part they belong. The number of genes corresponds to up-regulated genes with orthology information. Tissues with relative small number of expressed genes are indicated with grey letters.

We inferred orthologous relationships for human genes at two different taxonomical levels: i) eumetazoa, including 48 species from the major phyla (placozoa, cnidaria, and bilateria) and ii)

vertebrate, including 25 species (Figure 8). The eumetazoan-specific OGs were used to classify human proteins based on their most ancient origin ortholog. On contrary to phylostratigraphic and other studies, which assigns the evolutionary origin based on the homology detection (*Domazet-Lošo& Tautz, 2007; Wolf YI et al, 2009*), we assigned the evolutionary age of each protein based on stringent orthologous relationships. This allows us to transfer functional information across species (*von Mering et al, 2005*; *Koonin EV, 2005*) and speculate about the function of the ancestral tissue. Additionally, the delineation of orthologous and paralogous relationships is more precise facilitating a robust analysis about paralogy and functional divergence. By comparing the phylogenetic age of all human genes with expression data, we observe that ancient genes are enriched in the expression dataset, while the mammalian-specific genes are underrepresented (Figure 8). This observation is consistent with previous studies (*Subramanian & Kumar, 2004; Freilich et al, 2005*), which also reported that ancient genes have higher expression levels, and thus are easier detectable.

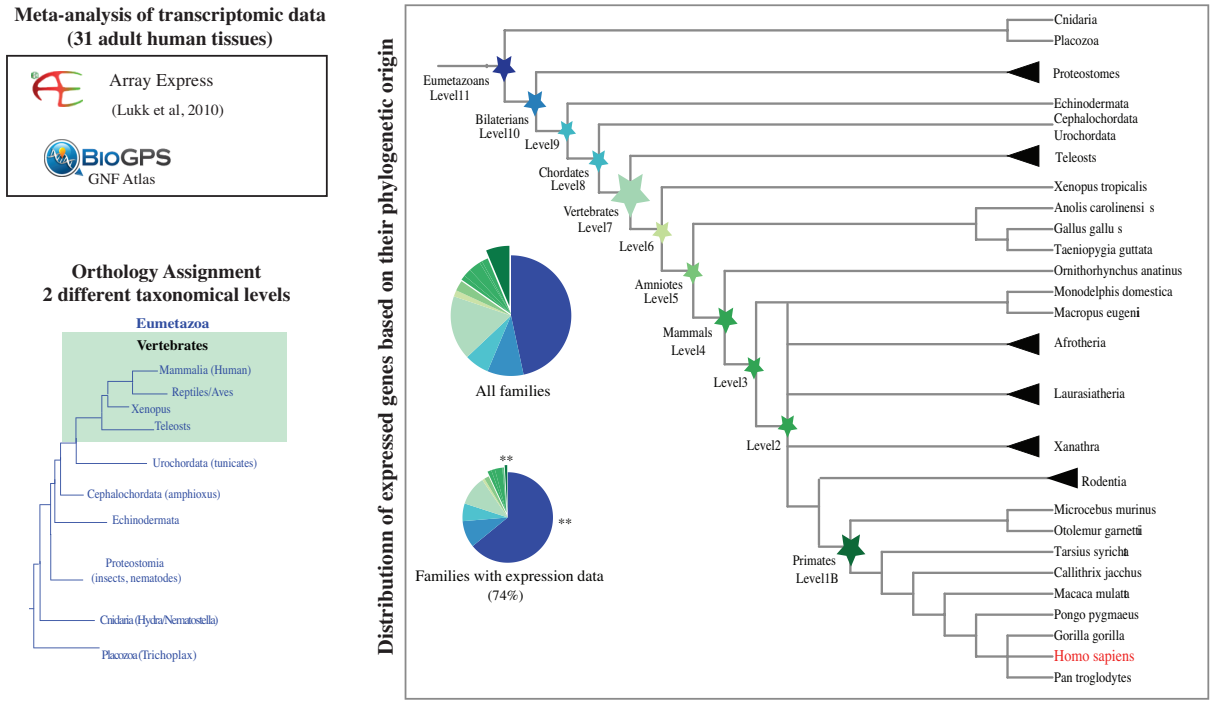


Figure 8: The flowchart of the study. We downloaded data for 31 tissue transcriptomes through ArrayExpress database. The majority of the tissues were samples for the GNF Atlas project (*Su et al, 2002*). We assigned OGs for two different levels (vertebrates in green box and metazoans in blue). We use the vertebrate OG to detect the human duplicated genes with respect the LCA of vertebrates. The more detailed tree on the right illustrates the phyla of the 48 animals used in this analysis. The pie charts show the biases exist on our dataset due to highly expressed genes with ancient origin. Mammalian-specific genes are under-represented; this affects the analysis of paralog genes with young origins.

In parallel, the OGs with respect the LCA of vertebrates (Level 7) were used to identify vertebrate-specific duplications; 2,093 OGs includes human paralogs. For 65% of them, we know the expression pattern of only one paralog, resulting in 717 OGs with expression data for multiple paralogs.

3. Phylogenetic origins of tissue inventories

In total, 10,597 human genes with expression and orthology information were analyzed. Since we are interested in the morphological transition from chordates to vertebrates, we focused on genes younger or contemporary to the vertebrate origin or genes that have undergone duplications in the vertebrate lineage. We classified the genes using both their phyletic age and their duplication status; class I, includes single copy human orthologs with vertebrate origin, while class II and class III, are duplicated genes (always with respect the vertebrate LCA), but their family origin predates or postdate the transition, respectively (Figure 9). A few studies have already reported the importance of duplication age in the divergence of paralogs (*Makova & Li, 2003; Blanc & Wolfe, 2004; Huminiecki & Wolfe, 2004; Freilich et al, 2006*). In this study, the age of duplication is constant (LCA of vertebrates), and the most important differentiation factor is the origin of the family. Class II genes – ancient phylogenetic origin - behave differently than class I and class III genes - young phylogenetic origin (Figure 9).

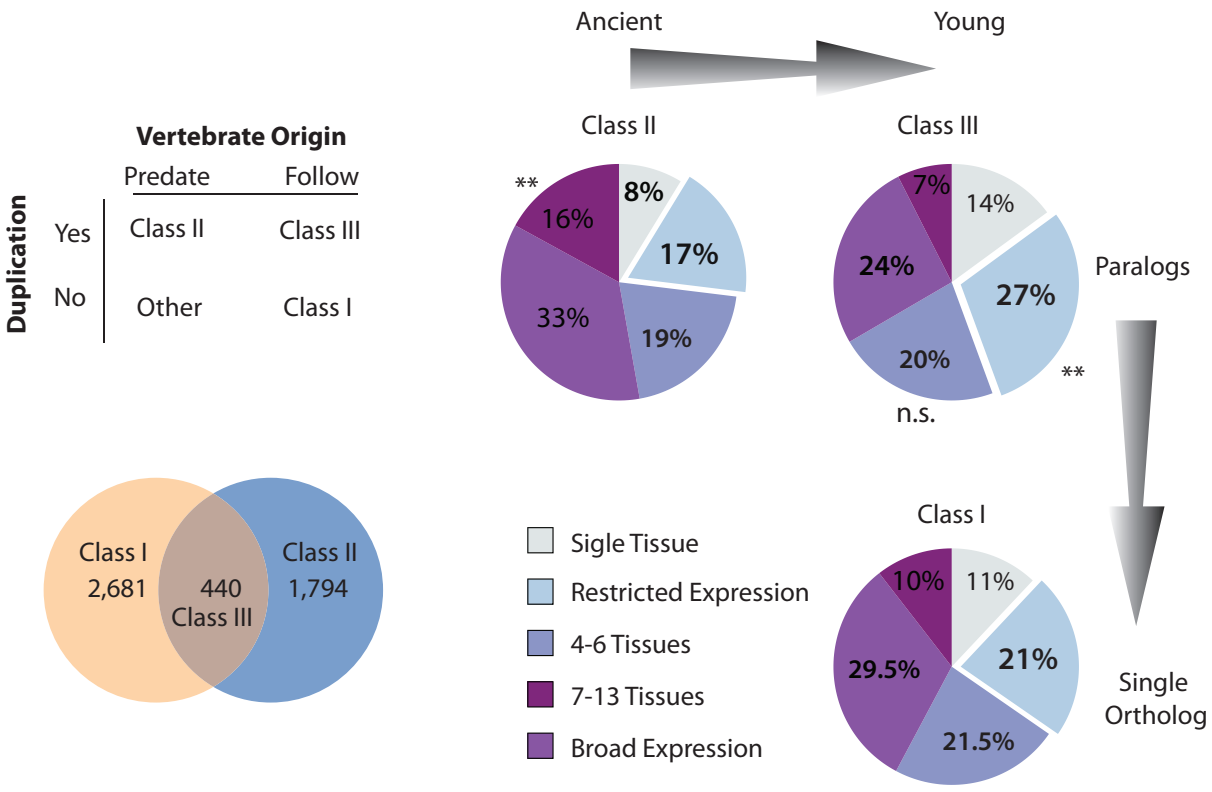


Figure 9: Genes and orthologous groups were classified into 4 classes based on their phylogenetic age and their duplication status. Then, we checked the broadness of expression for the different classes. As has been previously reported ancient origin genes (class II) are enriched for broadly expressed genes, while the class with the most restricted profiles are the young duplicates (class III).

In particular, they have broader expression patterns than duplicated or single copy genes of young origin. Indeed, only 17% of duplicated genes with ancient origin are expressed in less than 3 tissues, while this number increases to 27% for duplicated genes with young origin. Although, previous studies have reported similar findings (*Makova & Li, 2003; Blanc & Wolfe, 2004; Khaitovich et al, 2005*; *Huminiecki & Wolfe, 2004*), they didn't provide a working framework. Adopting the "division of labour" model, proposed for eye evolution (*Arendt et al, 2009*), our working hypothesis suggests that expression patterns of paralogs in human tissues can elucidate how tissue-specificity takes place from ancestral multifunctional tissues (Figure 10). Herein, we investigated the phylogenetic relationships of the tissues based on their expression profiles and quantified which is the contribution of these three classes of genes in tissue evolution and if certain classes are retained more frequently in certain tissues than others?

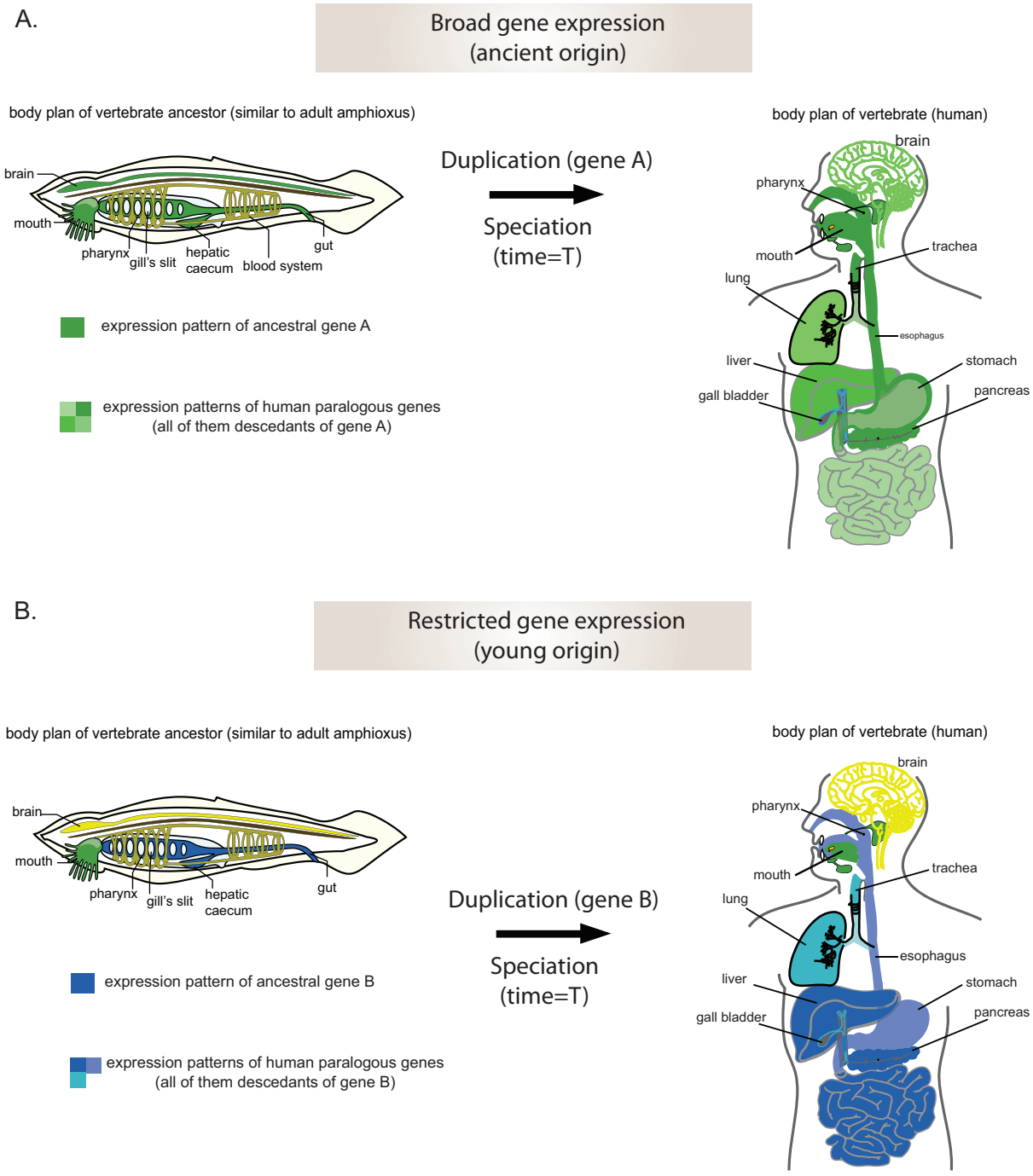


Figure 10: Studying the expression patterns of paralogous genes can elucidate the ancestry of tissues. A) Gene A, which predates the origin of vertebrates, is expressed in multiple tissues in the common ancestor of vertebrates. After gene duplication(s), the newly formed paralogs divide the ancestral activity of the multi-functional tissue into specialized human tissues. B) Gene B, which is vertebrate-specific, is expressed in a specific multi-functional tissue of the vertebrate ancestor. After gene duplication(s), the newly formed paralogs are detected in multiple human specialized tissues. Still, they present a more restricted pattern than paralogs of gene A.

4. Functional divergence of class II paralogs is more common between brain and other tissues.

Brain tissues present a similar pattern, distinct from the rest tissues; they are enriched for class II genes (Figure 11). There are many studies that provide evidence that the elaborated brain of vertebrates is a consequence of the 2R whole genome duplications. This hypothesis fits nicely to our data, as 80% of the class II OGs that expressed in brain tissues present a uniform duplication pattern across the 25 vertebrate species; however, till now, we haven't investigate in detail their syntenic relationships, which will help us to trace the mode of duplication (WGD vs. single gene).

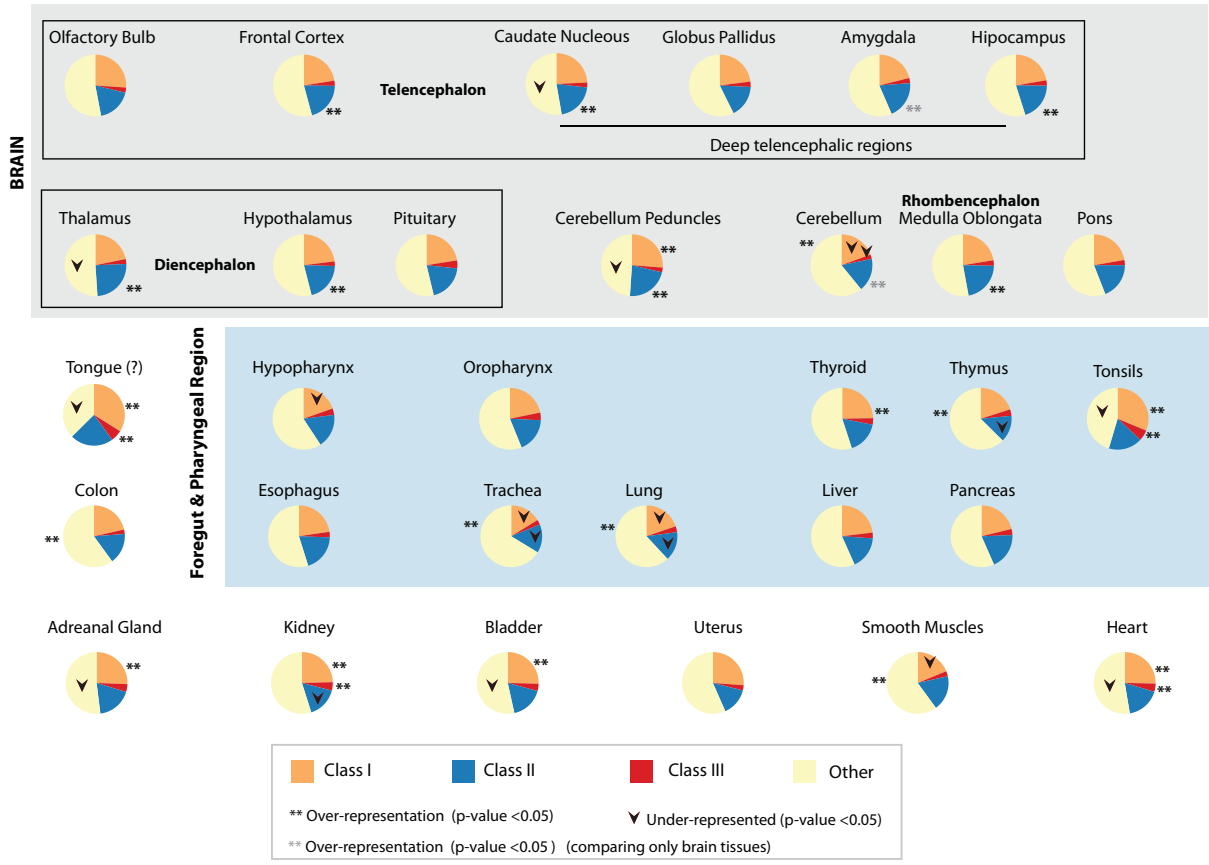


Figure 11: Representation of the 4 different classes of genes (orange=class I, red=class III, blue=II and yellow=ancient single copy orthologs) in the 31 transcriptomes. Brain tissues (grey area), despite their variety in the number of up-regulated genes (Table 1), present a common pattern; all of them are enriched in duplicated genes with ancient origin. On the other hand, foregut related tissues (blue area) are more variable. Newly acquired tissues, as tonsils that exist only in birds and mammals, are enriched for young origin genes, while old tissues, as trachea and colon are over-represented with ancient origin genes. Other tissues, unrelated to the previous clusters, like kidney and heart, are also enriched for genes of younger origin.

For the rest of the tissues, there is not any consensus (Figure 11). For instance, among the foregut derivatives tonsils are enriched for class I and class III genes (young origin), while class I genes are under-represented in the transcriptomes of hypopharynx, trachea and lung. Encouragingly,

tissues like colon and smooth muscles are enriched with genes of ancient origin and ontogeny-related tissues, like trachea and lung, present a parallel acquisition of genes. Nevertheless, there are surprising results as well, which can be explained by both technical and biological scenarios. Cerebellum, for example, although it is phylogenetically and ontogenetically newer structure than other brain tissues (e.g. pons) related to the regulation of highly skilled movements (*Purves et al, 2001*), expresses significantly the largest fraction of ancient genes. A part of cerebellum, however, the vestibulocerebellum (or archicerebellum), is primarily concerned with the regulation of movements underlying posture and equilibrium. Thus, by sampling this histological part, the transcriptome may reflect an ancient tissue inventory (*Dharani NE, 2005*). A similar scenario could explain the enrichment of ancient genes in the thymus transcriptome (*Bajoghli et al, 2009*). Bajoghli and co-authors suggested that the ancestral networks of the pharyngeal epithelium were expanded to evolve thymus. Indeed, we can identify 700 genes of vertebrate origin in thymus expression profile, including CCR9 and CCL15 that have been reported as primary cytokines for thymopoiesis (*Bajoghli et al, 2009*).

Figure 12 shows the intersection of shared class II OGs between any two tissues. Brain tissues, again, form a distinct cluster by sharing a statistical significant number of OGs between them. In a similar manner, gut related tissues tend to share class II OGs; this is, particularly, true for the 'tube' tissues (hypo- and oro-pharynx, esophagus and trachea). Additionally, there are a large number of shared OGs between kidney, heart and the rest of the tissues (more than 200 families), although it is not statistically significant.

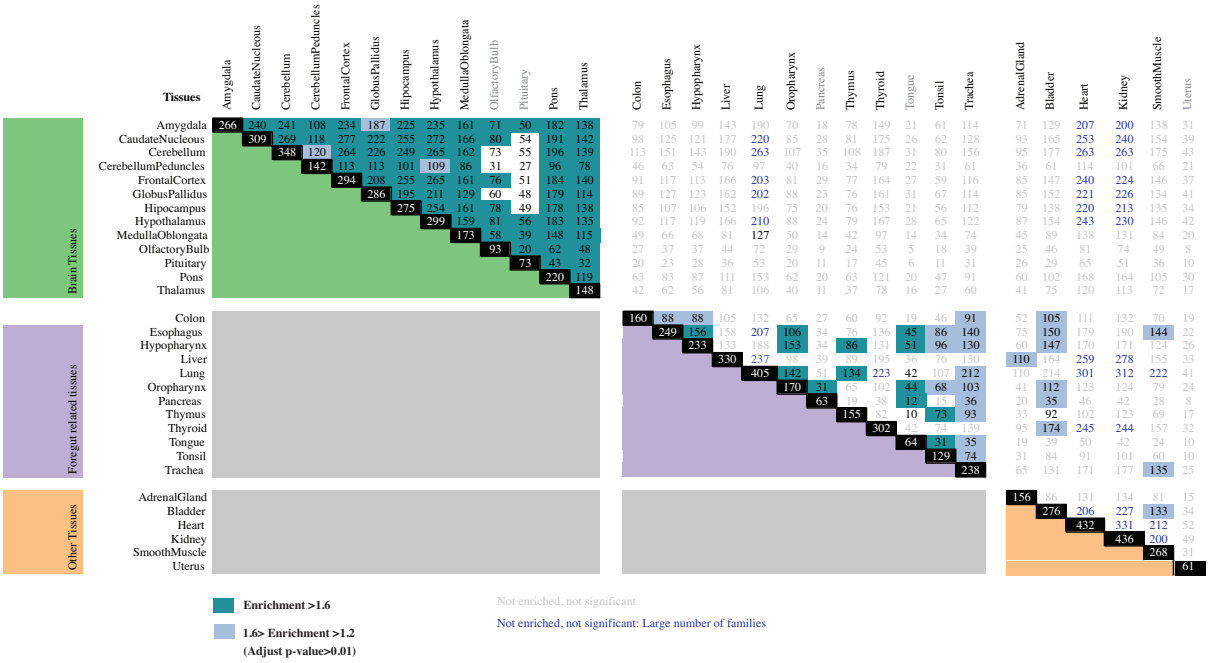


Figure 12: Across tissue comparison of class II families. Significant co-expressed pairs are indicated by dark green (2fold enrichment over the random expectation) and light blue (1.5 fold enrichment, p-values

for both categories is <0.01, after multi-test correction). Numbers in blue highlight large number of shared OGs that are not significant.

Investigating the pattern of functional divergent genes, we detect that the most common pattern of sub-functionalization is between brain and other tissues (Figure 13). We hypothesize that after the two rounds of WGD, the newly duplicated genes committed either to neuronal or non-neuronal fate. There are reported cases when house keeping genes, implying an ancestral origin, has been sub-functionalized between neurons and other tissues (*Plaitakis et al, 2003; Smith et al, 2006; Baldi et al, 2004; Serneels et al, 2005*). Despite the ongoing debate whether regulatory divergence is more important than changes in the biochemical function for functional innovation (*King & Wilson, 1975; Hoekstra & Coyne, 2007*); they seem to be parallel processes (*Khaitovich et al, 2005*). The case of glutamate dehydrogenase (GDH) illustrates nicely the parallel functional divergence in biochemical and regulatory level. Human GDH exists in GLUD1 (housekeeping) and GLUD2 (neural tissue-specific) isoforms, the protein sequences of which are 93% identical. However, the 15 different amino acids permit the neural enzyme to be recruited under conditions of low energy charge, similar to those that exist in synaptic astrocytes during intense glutamatergic transmission; leading to the adaptation of the GLUD2 to the unique metabolic needs of the nerve tissue (*Plaitakis et al, 2003*).

Many enzymes and cell cycle related molecules belong to the class II families, as it is expected. Due to the ancestry of the families, we can find multiple tissues sharing the same subfunctionalization events. For instance, pharynx, esophagus and trachea have a common intersection of 30 diverged paralogous families between themselves and brain tissues. Enzymes and receptors with multiple substrates – suggesting an ancient multifunctional molecule - are the best candidates; indeed, we can identify enzymes such as glycogen phosphorylase (PYGL, PYGLM) and GABA receptor channels.

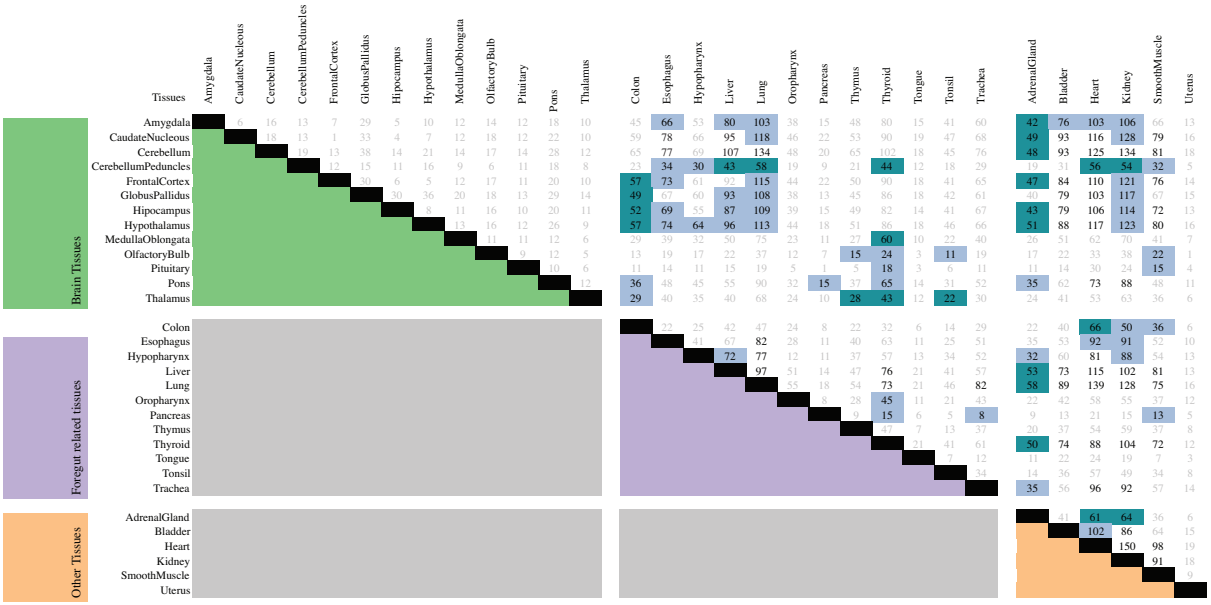


Figure 13: Across tissue comparison of families with sub-functionalized genes. Dark green indicates 2 fold enrichment over the random expectation, while blue indicates enrichment less than 2, but larger than 1.5 (p-values for both categories is <0.01, after multitest correction). Numbers in black highlight large number of subfunctionalized OGs even if not significant.

5. Young origin genes and their tissue-specificity

To investigate the impact of class III paralogs, we performed the same analysis as before. We focused on families with expression information for more than one member, qualifying 50% of the families. This is only a small number (141 OGs) impeding the statistical analysis. In the same manner as class II paralogs, brain- and gut- clusters share more OGs within their tissues rather than between them. Two pairs of tissues, namely hypopharynx-trachea and oropharynx-esophagus, present a number of sub-functionalized paralogs, which are statistically supported (Fisher test, adj. p-value<0.01). Interestingly, among the diverged families are cytokines (i.e. CCL8 expressed in the pharynx, CCL2 in trachea and pharynx, CCL11 in esophagus) and claudins (i.e. CLDN8 pharynx and thyroid and CLDN17 in esophagus, pharynx, tongue, salivary glands and thyroid). Foregut derivatives, as are thymus, pharynx and thyroid, evolved dramatically in the vertebrate lineage compared to their chordate homologous organs. As *Bajoghli* and co-authors (2009) reported the evolution of thymopoiesis is related to the evolution of cytokine families and the co-option of an ancestral network to a new function by introducing a few new genes. Again, the small number of OGs does not allow us to draw significant conclusions. Given that families that arose concomitantly with or after the vertebrate transition have a more restricted expression than the duplications of ancient families, we can hypothesize that, they are important to separate modern patterns of the vertebrate body plan. Thus,

the surface of ancient pharyngeal cells is transformed to distinct territories (hypo- and oro-pharynx, thyroid and thymus) by the functional divergence of paralogs.

There is a significant amount of class I OGs in the transcriptomes. Once again, this is a human-centric analysis, meaning that those genes might have been duplicated in some animal lineages –e.g. teleosts – or might have been lost in some others; the general tendency (60% of the families), however, is that are single copy orthologs. Similarly to the class II duplicated genes, tissues of the same cluster (brain vs. foregut) have a significant number of co-expressed genes compared to other tissue combinations (e.g. thyroid-pons). Interestingly, the tissue pairs with statistical supported co-expression patterns are almost identical as the class II genes (Figure 12). To be able to answer if those co-expression patterns are due to common ancestry or co-option (tissues with similar function recruit the same genes), we have to study more than one species and infer what might be the ancestral state.

To conclude, herein, we classified the expressed genes in 31 adult human tissues based on their phylogenetic age and duplication state and explored their distribution across tissues. As gene/genome duplication is a dominant aspect in the evolution of vertebrates, we mainly focused on how vertebrate-specific paralogs affect tissue specialization. We separated the paralogs based on their phylogenetic age; before (class II) and after (class III) chordate-vertebrate transition. We detected a significant number of class II co-expressed OGs within brain or gut-related tissues. However, functionally diverged class II paralogs were enriched between brain and the rest tissues. Multiple evo-devo studies associate the elaborated brain of vertebrates with multiple copies of developmental genes found in their genomes. In our list, despite important developmental genes and transcription factors, there many enzymes, transporters, channels and membrane associated proteins. We hypothesize that the commitment of basic (house-keeping) genes to neuronal or non-neuronal functional fate might associate with the brain evolution in the vertebrate lineage. *Khaitovich* and co-authors (2005), when compared human and chimpanzee brains, reported that there is an excess of gene expression and amino acid changes on the human brain compared to other tissues. They suggested that evolutionary changes at both the level of gene regulation and the level of protein sequence have played crucial roles in the evolution of certain organ systems, such as those involved in cognition (*Khaitovich et al, 2005*). Unfortunately, the second class of paralogs (class III) was limited to 141 families hindering statistically supported results; it could be, however, insightful to study their biological functions.

Discussion and future perspectives

Nowadays, several years into the era of systems biology, it has become clear that many previously unknown levels of biological function and organization exist (*Nurse & Hayles, 2011*). Biological systems – ranging from organelles to ecosystems – can be investigated with a holistic approach, by quantifying their overall molecular components and the interactions among them. As biology changes from a descriptive to a quantitative field, a big challenge is to evaluate the quality of the huge amounts of datasets that have been generated and quantify biological and technical perturbations (*Jensen & Bork, 2004; Aebersold R, 2011; Bork P, 2011*).

In the field of comparative and functional genomics, the quality control of orthology predictions is of the uttermost concern. A recent meta-analysis of the predictions based on several automated methods has revealed a large number of inconsistencies (*Pryszcz, Huerta-Cepas & Gabaldón, 2010*). This implies that the outcomes of the genomic and functional comparisons depend on the repository used for the analyses and thus, raises the question of how robust results are. Thus far, the assessment of orthology-prediction methods has been based on functional tests, which estimate the accuracy of methods based on gene order and functional conservation (*Hulsen et al, 2006; Chen et al, 2007; Altenhoff & Dessimoz, 2009)*. Yet, genomic rearrangements or functional divergence among orthologs influence the evaluation (*Gabaldon et al, 2009)*. A phylogeny-based assessment test is preferable compared to the aforementioned tests (*Altenhoff & Dessimoz, 2009*).

In the course of this thesis, a phylogeny-based benchmark dataset dedicated to animal orthology prediction was generated to address this issue (*Trachana et al, 2011*). Contrary to function-based tests, which are biased towards conserved, single-copy families, our aim was to generate a quality test set for complex families that exemplify known caveats in the analyses of eukaryotic genomes. This is reflected in our results; all studied repositories [TreeFam (*Ruan et al, 2008*), eggNOG (*Muller et al, 2010*), OrthoMCL (*Chen et al, 2006*), OrthoDB (*Waterhouse et al, 2011*) and OMA (*Altenhoff et al, 2011*)] predict only a fraction of the RefOGs accurately (Figure 2 in Appendix A). On average, 36% of the RefOGs were not predicted accurately by any tested databases, revealing general limitations of orthology predictions that are associated with biological complexity. For instance, it has been reported that protein families with multiple protein domains "attract" non-orthologous proteins due to domain sharing (reviewed extensively in *Koonin EV, 2005*). We observed that increasing domain complexity results in a significantly higher number of false assignments in all databases (*Trachana et al, 2011*). This indicates that there is considerable room for improvement for all orthology assignment methods. In an analogous study based on the phylogeny of only three and rather simple families, *Boeckmann et al* (2011) similarly concluded that none of the phylogenomic databases was in perfect agreement with the reference trees.

On the other hand, database-specific framework choices, i.e. species taxonomical range, or further confounding factors, such as genome annotation quality and species coverage, might influence

the final outcome. The necessity to adopt a standardized dataset to facilitate benchmarking has already been acknowledged by the orthology community (*Gabaldon et al, 2009; Altenhoff & Dessimoz, 2009*). By applying the in-house pipeline – eggNOG (*Muller et al, 2010*) - to certain datasets with distinct annotation and species information, we managed to quantify the impact of the aforementioned factors. Strikingly, 40% of the errors reported for the eggNOG database are associated with the old genome annotations. This suggests that frequent updates of the databases are necessary and the quality control of genomes, although a tedious, is still an essential task (*reviewed in Reeves et al, 2009*) that should be adopted by orthology prediction databases. *Milinkovitch et al* (2010) quantified the impact of low coverage genomes, mainly mammalian ones, on tree-based orthology methods. They reported that the inclusion of genomes with a large number of misannotated genes produces low quality MSAs, resulting in an inflated number of duplicated genes through tree reconciliation.

Another important point to be made, and perhaps the most considerable for any comparative and functional study, is that orthology depends on the phylogenetic context. The membership of orthologous groups depends on the evolutionary position of the LCA of the compared species. In practice, shifting the LCA deeper into the phylogeny will result in larger OGs (Figure 3). Therefore, many repositories provide orthologous groups for multiple taxonomical levels to increase the resolution of evolutionary relationships between proteins (*Muller et al, 2010; Waterhouse et al, 2011*). Apart from the species range, it seems that species coverage is equally important. We identified that orthology accuracy could be increased almost 20% by using more species, which break long branches and enable the detection of fast- evolving orthologs (*Trachana et al, 2011*). Many phylogenetic studies, which rely on orthology assignment, have reported that species tree topology depends on taxonomic sampling (*Philippe & Telford, 2006; Telford & Copley, 2011*). Although, there have been several studies evaluating the tree topologies generated by EST libraries and how missing data influences the final species tree topology (e.g. *de la Torre-Bárcena JE et al, 2009*), a similar study has not been performed for whole genome data; it would be interesting to quantify the exact impact of orthology accuracy in inferring phylogeny.

Despite the current uncertainty in the field of orthology mentioned above, the increasing number of genomes and other functional datasets greatly improves our understanding in the organization of biological systems. For instance, early functional and phylogenetic studies for the clade of eukaryotes have placed parasitic unicellular eukaryotes at the root of the eukaryotic evolutionary tree. However, the newly sequenced genomes of free-living, unicellular eukaryotes that have facilitated the reconstruction of the gene repertoire for a complex non-parasitic last common ancestor of eukaryotes (*Embley & Martin, 2006; Fritz-Laylin et al, 2010*). Additionally, integration of different types of large-scale datasets increases the signal-to-noise ratio (reviewed in *Jensen & Bork, 2004*) and results in more robust explanations. For example, we have recognized that the poor overlap between Y2H networks and co-expression networks is not primary due to technical constraints, but

rather to cellular regulation (post-translational modifications, mRNA and protein turnover) (*Schwanhausser et al, 2011; Maier et al, 2011*). Under the light of new findings and data, we have to re-examine previous findings and re-analyze publicly available resources. Therefore, in the course of this thesis, I mined publicly available data in another way, beyond the scope of the initial study, investigating an open question: how spatiotemporal patterns of biological systems are shaped through divergence of paralogs.

The first project is related to the temporal organization of the cell. We have reported, for the first time, the functional divergence of paralogs under cell cycle or diurnal (circadian)/ ultradian rhythm for three eukaryotic species (*A. thaliana, H. sapiens* and *S. cerevisiae*). Despite the reported links between periodic processes within a cell (*Hunt & Sassone-Corsi, 2007; Klevecz et al, 2004*), no systematic analysis had been previously performed till now (*Trachana, Jensen & Bork, 2010*). In higher complexity animals and plants, the circadian regulated genes are tissue-specific, comprising the most important caveat in this analysis. This explains the small, yet significant, number of functionally diverged paralogs that was detected in our study. One of the biggest challenges in systems biology is to design and interpret experiments that would explain complexity in larger systems (such as plants and animals) (*Walhout M, 2011*). To understand how building blocks of the larger systems have emerged and managed to function together, a list of their components and their interactions should be initially completed. As the differential temporal regulation of paralogs occurs in different lineages (*Trachana, Jensen & Bork, 2010*), it appears to be an efficient way to orchestrate cellular responses to extrinsic and intrinsic signals and thus, should be taken into consideration in the designs of future experiments. The analysis of the temporal organization of budding yeasts stresses the differences between "small" (unicellular) and "large" systems. In budding yeast, we were able to identify a larger number of paralogous pairs being subfunctionalized for cell cycle and ultradian rhythm regulation, as well as a significant number of genes being under the regulation of both processes. However, we cannot clarify if this is due to experimental perturbations or a sound biological finding, meaning that budding yeast exhibits a more tight control of the periodic processes within the cell. Further analysis of the functional repertoire of cell cycle/ultradian regulated paralogs indicates that they have arisen through both WGD and SSD and they are enriched in metabolic functions (Figure 8). So far, an ultradian transcriptome is available only for *S.cerevisiae,* while evidence for similar behavior in other post-WGD species (e.g. *Candida* genus) have been reported only recently. On the other hand, *S.pombe*, a pre-WGD species, has an ultradian oscillator, which, it is temperature-dependent, similar to the ultradian rhythm of other protists (all aforementioned ultradian rhythms are reviewed in *Lloyd & Murray, 2005*). All these rhythms are species-specific (*Lloyd D, 2008*) and related to the external and environmental stimuli of each species, suggesting that SSD duplication should be the major mode of duplication to orchestrate the temporal subfunctionalization. Indeed, our analysis has revealed that

despite the comparable number of WGD and SSD paralogs, the SSD paralogs are significantly over-represented in the list of temporally diverged paralogs.

Actually, for all three species, the functional repertoires of duplicated genes are different, comprising an additional evidence that the temporal sub(neo)functionalization has evolved in a species-dependent manner to distinguish cell cycle regulation from other periodic processes, perhaps even to coordinate them. However, as the three species are evolutionarily distant, with multiple taxonomical ranks in between, and the number of detected paralogs is small, we cannot infer the ancestral state and distinguish between sub- and neo-functionalization. Given the different functional repertoires of circadian regulated genes in mouse and Arabidopsis (*Harmer et al, 2000; Panda et al, 2002; Doherty & Kay, 2010*), or even between mouse and Drosophila (reviewed by *Doherty & Kay, 2010*), the temporal regulation of the cell and the role of paralogs in cellular periodicity should be first understood in closely related organisms. Primates would be an ideal group, as genomic information of multiple species is available and they have different daily cycles (e.g. nocturnal lemur vs. diurnal human). As I have already mentioned above, it should also be addressed in a systematic way by investigating transcriptomic profiles of different tissues and thus, completing the biology of a "large system".

Apart from deciphering the tempo of a tissue, in the case of multi-cellular animals and plants, it is necessary to understand how tissue-specificity itself has evolved. The spatial organization of animal body parts and tissues has been studied more thoroughly; yet, integrating all the available data and formulating a conceptual framework remains a challenging task. In the course of this thesis, I had the chance to be involved in two different projects investigating two different aspects of tissues evolution; the first one was focused on the role of miRNAs in tissue evolution and in the second one investigated whether ontogenic relationships of tissues can be traced through the tissue-specific patterns of duplicated genes.

Regarding the second project particularly, I focused on vertebrate evolution, since genomic and transcriptomic data are available for this group, and most importantly, because their morphological complexity, as compared to other chordate phyla (e.g. *Ciona* and *Amphioxus*) has been linked to multiple rounds of genome duplications in the ancestral population (*Meyer & Schartl, 1999; Hokamp et al, 2003; Panopoulou & Poustka, 2005*). To identify paralogs that are related to this morphological transition and study their spatial (tissue) divergence, we combined orthology information of 43 metazoans and 25 vertebrtates with transcriptomic data of 31 human tissues. The metazoan-and vertebrate- specific groups were used to define the phyletic age of each human gene and to determine human duplicated genes based on the LCA of vertebrates, respectively. At the end, human genes were classified in four classes based on their phyletic age and duplication status (Figure 10). Many of our observations are consistent with previously published results; for instance, we verified that i) expression divergence between duplicate genes increases with evolutionary time and ii)

mammalian-specific paralogs tend to be expressed in a tissue-specific pattern (*Makova & Li, 2003; Blanc & Wolfe, 2004; Khaitovich et al, 2005*; *Huminiecki & Wolfe, 2004; Freilich et al, 2005; Freilich et al, 2006*)). Nevertheless, to our best knowledge, no study before had quantified the representation of the different gene classes on tissue expression. Interestingly, different classes of genes accumulate in different fractions in the 31 transcriptomes (Figure 11). Brain tissues have presented a common pattern; the tissues are enriched for paralogs that belong to families with ancient origin. On the other hand, the tissues with the largest representation of paralogs, which belong to young origin families, are heart, kidney, tonsils and tongue. Although the heart is an ancient structure, the development of which is controlled by a conserved network of transcription factors, its morphological complexity has increased dramatically in the lineage of vertebrates through gene duplication and co-option of additional networks (*Olson EN, 2006*). Our analysis has revealed that the age of the transcriptome can reflect the phylogenetic time in which morphological modifications have taken place. *Kalinka et al* (2010) and *Domazet-Loso & Tautz* (2010) observed the same trend by studying expression profiles on different developmental stages in the fruit fly, nematode and zebrafish.

Previous studies in comparative transcriptomics have suggested that variation in tissue-specific patterns within species can be substantial (*Whitehead & Crawford, 2005; Khaitovich et al, 2006*). Understanding why different tissues retain other classes of genes may give insights to its functional constraints and why tissues diverge with different rates. Given the fact that genes with ancient origins are expressed more broadly and genes that present tissue-specific (or restricted expression) patterns are fast-evolving, we would expect that tissues enriched in genes of pre- and post-vertebrate origin are slow- and fast-evolving, respectively. If differences in expression between tissues increase with time since the last common ancestor in which the tissues have not diverged, then we hypothesize that tissues like tonsils, bladder, adrenal glands (Figure 12) evolve faster than lungs and brain. This would explain why, contrary to anthropologists' expectations, gene expression in the brain has diverged less than other tissues (like heart and testes) between human and primates.

Furthermore, the investigation of the expression patterns of paralogs revealed that paralogs that belong to families with ancient origins tend to be subfunctionalized between brain and non-brain tissues (Figure 13). Once again, the large-scale comparisons between human and chimpanzees have elucidated that evolutionary changes at both the level of gene regulation and the level of protein sequence have played crucial roles in the evolution of certain organ systems, such as those involved in cognition or male reproduction (*Khaitovich et al, 2005*). These authors proposed that both types of changes are likely to have acted in concert. Considering all the above, we assume that during the evolution of human, many of the retained paralogs, after WGD in the LCA of vertebrates, have functionally diverged between neuronal and non-neuronal tissues. The brain-specific paralogs escape the functional constraints of other tissues enabling their adaptation to neuronal fate, which may have

facilitated the development of a more elaborate brain. Of course, this is only a speculation, as the teleost lineage, although it has undergone an extra round of WGD, does not present a more elaborated brain than land vertebrates.

In the future, my intension is to increase the evolutionary distance and expand the cross-tissues comparison with more vertebrate species. Unfortunately, tissue transcriptomics of other vertebrate species, apart of human and mouse, suffers from data discrepancies due to incomplete annotation similar to many other microarray experiments (*Ioannidis et al, 2009*). However, there is a deluge of human and mouse datasets, which can be analyzed to provide new insights to the tissue evolution of mammalian species. Despite the further improvements that might be gained by integrating more data, this analysis will always suffer from the low resolution of the studied organizational level (tissue). If we scale the biological systems based on their level of organization and complexity, then cells is the simplest unit, which might be the most plausible one to understand. Therefore, large-scale datasets at single cell resolution would be ideal for understanding how multifunctional systems (cell types) evolve to more specialized structures. The conceptual framework for cell type evolution exists (*Arendt D, 2008*) and the study of many genes through the candidate-approach has proven fruitful (*Tomer et al, 2010*); however, a series of high-throughput experiment would allow for a comprehensive view of the system and drive unbiased conclusions. Then, we can apply this knowledge to understand the evolution and organization of more complex systems (such as tissues and organs).

Finally, all three projects highlight the importance of one factor: the impact of species selection on the evolutionary signal we analyze. For instance, in the last project, the absence of slow-evolving animals of the proteostome clade (e.g. annelids) and the presence of the fast-evolving clade of ecdysozoans has probably influnced our analysis; families that have been characterized as chordate- or vertebrate- specific may have a bilaterian origin if a counterpart ortholog exists in slow-evolving proteostomes. Our concern is valid given the results of the orthology study. Similar to any phylogenetic study, fast-evolving species introduce more mispredictions than species belonging to a well-represented clade. This has been a common problem for many genome content analyses (presence/absence studies of homologous characters, e.g. protein domain, family expansions, chromosomal rearrangements) in which the species distribution does not mirror properly the situation of the ancestor (*Copley et al, 2004; Telford & Copley, 2011*). Indeed, the analysis of only three eukaryotic species (yeast, plant and human) separated by large evolutionary distance didn't allow us to distinguish between sub- or neo-functionalization of the studied paralogs. Except of the expression data that have been proven valuable to understand the functional evolution of eukaryotes, we have to collect other types of functional data, such as biochemical specificity of molecules, their physical interactions and their location across species. Biologically evident, as moonlighting (gene sharing) and mRNA- protein stability raises the question how well equipped we are for the quest of cellular

organization. It is crucial to consider all the properties of a protein to understand functional innovation. *Raes et al* (2007) asked how limited our view of the protein space is and how sampling new (eco)systems yields an ever-increasing number of novel families.

To conclude, in the course of this thesis, I studied the evolution of gene function at different levels of biological organization: cells, tissues and species supporting the view of *Nurse and Hayles* (2011) that to understand biological function at the level of the cell, the simplest unit of biological organization, means to decipher - among the others - the spatiotemporal organization and the communication between and within the units of organization. To their view, I add Darwin's words that "If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down" to propose that we have to sample the evolutionary space more elegantly to be able to understand the cellular tempo (dynamic behavior).

References

Aebersold R (2011) Systems Biology: What's the Next Challenge? Cell. 144(6), 837 – 838.

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2,375 human brain genes. *Nature*. 355(6361): 632-4.

Akerborg O, Sennblad B, Arvestad L, Lagergren J (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci.* 106(14): 5714-9.

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol.* 5: e1000262.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011. orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39(Database issue): D289-294.

Ames RM, Rash BM, Hentges KE, Robertson DL, Delneri D, Lovell SC. (2010) Gene duplication and environmental adaptation within yeast populations. *Genome Biol Evol.* 2: 591-601.

Arendt D, Tessmar-Raible K, Snyman H, Dorresteijn AW, Wittbrodt J. (2004) Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science.* 306: 869–871.

Arendt D. (2008) The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet.* 9(11): 868-82.

Arendt D, Hausen H, Purschke G. (2009) The 'division of labour' model of eye evolution. *Philos Trans R Soc Lond B Biol Sci.* 364(1531): 2809-17.

Aury JM, Jaillon O, Duret L, et al (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature.* 444(7116): 171-8.

Axelsson E, Hultin-Rosenberg L, Brandström M, Zwahlén M, Clayton DF, Ellegren H (2008) Natural selection in avian protein-coding genes expressed in brain. *Mol Ecol*. 17(12):3008-17.

Averof M, Patel NH. (1997) Crustacean appendage evolution associated with changes in Hox gene expression. *Nature.* 388(6643): 682-6.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. *Science.* 297(5583): 1003-7.

Bailey JA, Eichler EE. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet.* 7(7):552-64.

Bassham S, Canestro C, Postlethwait JH (2008) Evolution of developmental roles of Pax2/5/8 paralogs after independent duplication in urochordate and vertebrate lineages. *BMC Biol.* 6: 35-41.

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. (2000) Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell*. 2(7): 1093-101.

Blanc G, Wolfe, KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679–1691.

Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform*. [Epub ahead of print]

Bork P (1989) Recognition of functional regions in primary structures using a set of property patterns. *FEBS Lett.* 257(1): 191-5.

Bork P (1991) Shuffled domains in extracellular proteins. *FEBS Lett.* 286(1-2): 47-54.

Bork P (2011) Systems Biology: What's the Next Challenge? Cell. 144(6), 837 – 838**.**

Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15(10): 1456-61.

Campillos M, Doerks T, Shah PK, Bork P (2006) Computational characterization of multiple Gag-like human proteins. *Trends Genet.* 22(11): 585-9.

Carroll SB (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell.* 101(6):577-80.

Carroll SB, Grenier JK, Weatherbee SD (2005) From DNA to diversity: Molecular genetics and the evolution of animal design - 2nd edition (Blackwell publishing, UK).

Catchen JM, Conery JS, Postlethwait JH. (2009) Automated identification of conserved synteny after whole-genome duplication. *Genome Res.* 19(8): 1497-505.

Chapman JA, Kirkness EF, Simakov O et al (2010) The dynamic genome of Hydra. *Nature.* 464(7288): 592-6.

Chan ET, Quon GT, Chua G, et al (2009) Conservation of core gene expression in vertebrate tissues. *J Biol.* 8(3):33.

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*. 34(Database issue): D363-368.

Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2: e383.

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* 21, 1146-51.

Ciccarelli FD, von Mering C, Suyama M, et al (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* 15(3): 343-351.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science.* 311(5765): 1283-7.

Colbourne JK, Pfrender ME, Gilbert D et al. (2011) The ecoresponsive genome of Daphnia pulex. *Science*. 331(6017): 555-61.

a. Conant GC, Wolfe KH (2007) Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol* 3: 129-141.

b. Conant GC, Wolfe KH (2007) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9(12): 938-50.

Copley RR (2008) The animal in the genome: comparative genomics and evolution. *Philos Trans R Soc Lond B Biol Sci.* 363(1496): 1453-61.

Copley RR, Letunic I, Bork P (2002) Genome and protein evolution in eukaryotes. *Curr Opin Chem Biol.* 6(1): 39-45.

Cui L, Wall PK, Leebens-Mack JH, et al (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16(6): 738-49.

Cusack BP, Wolfe KH (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol.* 24(3): 679-86.

Danchin É, Charmantier A, Champagne FA, Mesoudi A, Pujol B, Blanchet S. (2011) Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nat Rev Genet*. 12(7):475-86.

Darwin CR (1859) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. (John Murray, London).

Datta RS, Meacham C, Samad B, Neyer C, Sjölander K (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res.* 37: W84-9.

Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science.* 311(5762): 796-800.

Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* 2, E55.

Degnan BM, Vervoort M, Larroux C, Richards GS (2009) Early evolution of metazoan transcription factors. *Curr Opin Genet Dev* 19(6): 591-9.

Dehal P, Satou Y, Campbell RK, et al. (2002) The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins. *Science*. 298:2157–2167.

Dehal P, Boore, JL (2005) Two rounds of genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:e314

Delaunay F, Laudet V (2002) Circadian clock and microarrays: mammalian genome gets rhythm. *Trends Genet.* 18(12):595-7.

Delsuc F, Brinkmann H, Philippe H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6(5):361-75.

Deluca TF, Wu IH, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics.* 22(16):2044-6.

Denes AS, Jékely G, Steinmetz PR, Raible F, Snyman H, Prud'homme B, Ferrier DE, Balavoine G, Arendt D. (2007) Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell.* 129(2): 277-88.

Denoeud F, Henriet S, Mungpakdee S, et al (2011) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science.* 330(6009): 1381-5.

Dermitzakis ET, Clark AG (2001) Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol.* 18(4): 557-62.

de la Torre-Bárcena JE, Kolokotronis SO, Lee EK, Stevenson DW, Brenner ED, Katari MS, Coruzzi GM, DeSalle R (2009) The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. *PLoS One.* 4(6): e5764.

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst.* 5(12):1512-26.

Doherty CJ, Kay SA (2010) Circadian control of global gene expression patterns. *Annu Rev Genet.* 44: 419-44.

Domazet-Loso T, Brajković J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*. 23(11):533-9.

Domazet-Lošo T, Tautz D. (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature.* 468(7325):815-8.

Doolittle RF (1995) The origins and evolution of eukaryotic proteins. *Philos Trans R Soc Lond B Biol Sci.* 349(1329): 235-240.

Doolittle WF (1999) Lateral genomics. *Trends Cell Biol.* 9(12):M5-8.

Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. *Trends Genet.* 27(4):157-63.

Duret L (2000) tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16(7):287-9.

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol.* 17(1):68-74.

Edlund T, Normark S (1981) Recombination between short DNA homologies causes tandem duplication. *Nature.* 292(5820): 269-71.

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8(3): 163-7.

Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature.* 440(7084): 623-30.

Fang H, Yang Y, Li C, Fu S, Yang Z, Jin G, Wang K, Zhang J, Jin Y. (2010) Transcriptome analysis of early organogenesis in human embryos. *Dev Cell.* 19(1): 174-84.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99-113.

Force A, Lynch M, Pickett FB, Amores A, Yi-lin Y, Postlethwait JH (1999) Preservation of duplicated genes by complementary, degenerative mutations. *Genetics.* 151:1531-1545.

Freilich S, Massingham T, Bhattacharyya S, Ponsting H, Lyons PA, Freeman TC, Thornton JM (2005) Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol.* 6(7): R56.

Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM (2006) Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol.* 7(10): R89.

Fritz-Laylin LK, Prochnik SE, Ginger ML, et al (2010) The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell.* 140(5): 631-42.

Gabaldon T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9(10): 235.

Gabaldón T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. (2009) Joining forces in the quest for orthologs. *Genome Biol.* 10(9): 403.

Garcia-Fernández J, Holland PW (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature.* 370(6490): 563-6.

Gaudieri S, Kulski JK, Balmer L, Giles KM, Inoko H, Dawkins RL (1997) Retroelements and segmental duplications in the generation of diversity within the MHC. *DNA Seq.* 8(3): 137-41.

Gonzales-Vigil E, Bianchetti CM, Phillips GN Jr, Howe GA (2011) Adaptive evolution of threonine deaminase in plant defense against insect herbivores. *Proc Natl Acad Sci.* 108(14): 5897-902.

Goodstadt L, Ponting CP (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol.* 2(9): e133.

Gravel S, Henn BM, Gutenkunst RN, et al (2011) Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*. 108(29): 11983-8.

Gu Z, Nicolae D, Lu HH, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*. 12: 609-13.

Guan Y, Dunham MJ, Troyanskaya OG (2007) Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics* 175, 933–943.

Güell M, van Noort V, Yus E, et al (2009) Transcriptome complexity in a genome-reduced bacterium. *Science.* 326(5957): 1268-71.

Harmer SL, Hogenesch JB, Straume M, et al (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science.* 290: 2110-2113.

He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*. 169, 1157–1164.

Hittinger CT, Caroll SB (2007) Gene duplication and the adaptive evolution of a classic switch. *Nature.* 440: 677-681.

Hokamp K, McLysaght A, Wolfe KH (2003) The 2R hypothesis and the human genome sequence. *J Struct Funct Genomics*. 3(1-4): 95-110.

Holland PW, Garcia-Fernàndez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. *Dev Suppl.* 125-33.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, et al. (2011) PhylomeDB v3.0. an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* 39(Database issue): D556-560.

Hulsen T, Huynen MA, de Vlieg J, Groenen PM (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 7: R31.

Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.*14(10A): 1870-9.

Hunt T, Sassone-Corsi P (2007) Riding tandem: circadian clocks and the cell cycle. *Cell.* 129(3): 461-4.

Huynen MA, Snel B, von Mering C, Bork P (2003) Function prediction and protein networks. *Curr Opin Cell Biol.* 15(2):191-8.

Ioannidis JP, Allison DB, Ball CA, et al (2009) Repeatability of published microarray gene expression analyses. *Nat Genet.* 41(2): 149-55.

Jaillon O, Aury JM, Brunet F, et al. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*. 431: 946–957.

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and posttranslational cell cycle regulation. *Nature*. 443: 594–597.

Jensen LJ, Julien P, Kuhn M, et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* 36(Database issue): D250-254.

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. (2008) The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell.* 132(5):783-93.

Jiao Y, Wickett NJ, Ayyampalayam S, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature. 473(7345): 97-100

Jordan IK, Wolf YI & Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* 4, 22.

Joyce AR, Palsson BØ. (2006) The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol.* 7(3):198-210.

Jun J, Ryvkin P, Hemphill E, Mandoiu I, Nelson C. (2009) The birth of new genes by RNA- and DNA-mediated duplication during mammalian evolution. *J Comput Biol.* 16(10): 1429-44.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P (2010) Gene expression divergence recapitulates the developmental hourglass model. *Nature.* 468(7325): 811-4.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science.* 309(5742):1850-4.

Khaitovich P, Enard W, Lachmann M, Pääbo S. (2006) Evolution of primate gene expression. *Nat Rev Genet.* 7(9):693-702.

King N, Westbrook MJ, Young SL, et al (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*. 451(7180): 783-8.

Kitami T, Nadeau JH (2002) Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nat Genet.* 32(1): 191-4.

Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. *PNAS.* 1200-1205.

Koonin EV (2001) An apology for orthologs - or brave new memes. *Genome Biol.* 2(4):COMMENT1005.

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338.

Koonin EV (2010) The incredible expanding ancestor of eukaryotes. *Cell* 140(5):606-8.

Koonin EV, Aravind L & Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101, 573–576.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. *Brief Bioinform*. [PMID: 21690100]

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* 36(Database issue): D271-5.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13(10):2229-35.

Kuzniar A, van Ham RC, Pongor S, Leunissen JA (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* 24(11): 539-51.

Kühner S, van Noort V, Betts MJ, et al (2009) Proteome organization in a genome-reduced bacterium. *Science.* 326(5957): 1235-40.

Lang T, Alexandersson M, Hansson GC, Samuelsson T (2007) Gel-forming mucins appeared early in metazoan evolution. *Proc. Natl. Acad. Sci.* 104(41): 16209-16214.

Lehner B, Fraser AG (2004) Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet.* 20(10):468-72.

Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci*. 33(3): 101-3.

Letunic I, Doerks T, Bork P (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.* 37(Database issue): D229-232.

Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC (2009) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 38(Database issue): D346-54.

Lloyd D (2008) Respiratory oscillations in yeasts. *Adv Exp Med Biol.* 641: 118-40.

Lloyd D, Murray DB (2005) Ultradian metronome: timekeeper for orchestration of cellular coherence. *Trends Biochem Sci.* 30(7): 373-7.

Lloyd D, Murray DB (2007) Redox rhythmicity: clocks at the core of temporal coherence. *BioEssays.* 29: 465-473.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155.

Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A. (2010) A global map of human gene expression. *Nat Biotechnol.* 28(4): 322-4.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. *PNAS*. 102:5454-9.

Maier T, Schmidt A, Güell M, Kühner S, Gavin AC, Aebersold R, Serrano L. (2011) Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol Syst Biol.* 7:511.

Makova KD, Li WH (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13(7): 1638-45.

Merrit TJS, Quattro JM (2002) Negative charge correlates with neural expression in vertebrate aldolase isozymes. *J Mol Evol* 55: 674-683.

Meyer A, Schartl M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol.* 11(6): 699-704.

Meyer A, Van de Peer Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays.* 27(9): 937-45.

Muller J, Szklarczyk D, Julien P, et al. (2010) eggNOG v2.0. extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* 38(Database issue): D190-195.

Nakatani Y, Takeda H, Kohara Y, Morishita S. (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17(9): 1254-65.

Nurse P, Hayles J. (2011) The cell in an era of systems biology. *Cell.* 144(6):850-4.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet.* 39(6):730-2.

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York.

Olson EN (2006) Gene regulatory networks in the evolution and development of the heart. *Science.* 313(5795): 1922-7.

Ostlund G, Schmitt T, Forslund K, et al (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38(Database issue): D196-203.

Pal C, Papp B & Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics.* 158, 927–931.

Panda S, Antoch MP, Miller BH, et al (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell.* 109: 307–320.

Panopoulou G, Poustka AJ (2005) Timing and mechanism of ancient vertebrate genome duplications -- the adventure of a hypothesis. *Trends Genet.* 21(10): 559-67.

Pareek CS, Smoczynski R, Tretyn A. (2011) Sequencing technologies and genome sequencing. *J Appl Genet*. PMID: 21698376.

Parkinson H, Kapushesky M, Shojatalab M, et al (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 35(Database issue): D747-50.

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC (2006) Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends Genet.* 22(11): 597-602.

Paterson AH, Bowers JE, Bruggmann R, et al (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature.* 457(7229): 551-6.

Paxton CN, Bleyl SB, Chapman SC, Schoenwolf GC (2010) Identification of differentially expressed genes in early inner ear development. *Gene Expr Patterns*. 10(1):31-43.

Philippe H, Telford MJ (2006) Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol.* 21(11): 614-20.

Piatigorsky J, O'Brien WE, Norman BL, Kalumuck K, Wistow GJ, Borras T, Nickerson JM, Wawrousek EF (1988) Gene sharing by delta-crystallin and argininosuccinate lyase. *Proc Natl Acad Sci.* 85(10): 3479-83.

Pigliucci M (2009) An extended synthesis for evolutionary biology. *Ann. NY Acad. Sci*. 1168:218-28.

Pigliucci M (2008) Is evolvability evolvable? *Nat Rev Genet.* 9(1): 75-82.

Plaitakis A, Spanaki C, Mastorodemos V, Zaganas I (2003) Study of structure–function relationships in human glutamate dehydrogenases reveals novel molecular mechanisms for the regulation of the nerve tissue-specific (GLUD2) isoenzyme. *Neurochem. Int.* 43, 401–410.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS (2000) Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* 10(12): 1890-902.

Prochnik SE, Umen J, Nedelcu AM, et al. (2010) Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science*. 329(5988): 223-6.

Prud'homme B, de Rosa R, Arendt D, Julien JF, Pajaziti R, Dorresteijn AW, Adoutte A, Wittbrodt J, Balavoine G. (2003) Arthropod-like expression patterns of engrailed and wingless in the annelid Platynereis dumerilii suggest a role in segment formation. *Curr Biol.* 13(21):1876-81.

Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci* 104 Suppl 1:8605-12.

Pryszcz LP, Huerta-Cepas J, Gabaldon T (2010) MetaPhOrs. orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.* 39(5):e32.

Putnam NH, Srivastava M, Hellsten U, et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science.* 317(5834): 86-94.

Putnam NH, Butts T, Ferrier DE, et al (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature.* 453(7198): 1064-71.

Raes J, Harrington ED, Singh AH, Bork P (2007) Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol.* 17(3): 362-9.

Ravasi T, Suzuki H, Cannistraci CV, et al (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell.* 140(5): 744-52.

Reeves GA, Talavera D, Thornton JM. (2009) Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface.* 6(31): 129-47.

Ruan J, Li H, Chen Z, et al. (2008) TreeFam. 2008 Update. Nucleic Acids Res. 36 (Database issue): D735-40.

Roux J, Robinson-Rechavi M (2008) Developmental constraints on vertebrate genome evolution. *PLoS Genet.* 4(12):e1000311.

Sodergren E, Weinstock GM, Davidson EH, et al. (2006) The genome of the sea urchin Strongylocentrotus purpuratus. *Science.* 314: 941–952.

Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics.* 168(1): 373-81.

Salichos L, Rokas A (2011) Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One*. 6(4): e18755.

Scanell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature.* 440: 341-345.

Schmidt TR, Doan JW, Goodman M, Grossman LI (2003) Retention of a duplicate gene through changes in subcellular targeting: an electron transport protein homologue localizes to the golgi. *JMol Evol* 57:222-228.

Schmidt D, Wilson MD, Ballester B, et al (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*. 328(5981): 1036-40.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. (2011) Global quantification of mammalian gene expression control. *Nature.* 473(7347): 337-42.

Sémon M, Wolfe KH (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet*. 23(3): 108-12.

Seoighe C, Wolfe KH (1999) Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* 2, 548–554.

She X, Cheng Z, Zöllner S, Church DM, Eichler EE. (2008) Mouse segmental duplication and copy number variation. *Nat Genet.* 40(7): 909-14.

Singh AH, Doerks T, Letunic I, et al (2009) Discovering functional novelty in metagenomes: examples from light-mediated processes. *J Bacteriol.* 191(1): 32-41.

Sonnhammer EL & Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18:619–20.

Srivastava M, Begovic E, Chapman J, et al (2008) The Trichoplax genome and the nature of placozoans. *Nature.* 454(7207): 955-60.

Stoltzfus A (1999) On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49, 169–181.

Su AI, Cooke MP, Ching KA, et al (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci.* 99(7):4465-70.

Sudmant PH, Kitzman JO, Antonacci F, et al (2010) Diversity of human copy number variation and multicopy genes. *Science*. 330(6004):641-6.

Tatusov RL, Koonin EV & Lipman DJ (1997) A genomic perspective on protein families. *Science* 278: 631–637.

Tatusov RL, Fedorova ND, Jackson JD, et al (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 4-41.

Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* 38, 615–643.

Telford MJ, Copley RR (2011) Improving animal phylogenies with genomic data. *Trends Genet.* 27(5): 186-95.

Tessmar-Raible K, Raible F, Christodoulou F, Guy K, Rembold M, Hausen H, Arendt D. (2007) Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell*. 129(7):1389-400.

Tomer R, Denes AS, Tessmar-Raible K, Arendt D. (2010) Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell.* 142(5):800-9.

Tordai H, Nagy A, Farkas K, et al (2005) Modules, multidomain proteins and organismic complexity. *FEBS J.* 272(19): 5064-5078.

Trachana K, Jensen LJ, Bork P. (2010) Evolution and regulation of cellular periodic processes: a role for paralogues. *EMBO Rep.* 11(3):233-8.

Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. (2011) Orthology prediction methods: A quality assessment using curated protein families. *Bioessays.* doi:10.1002/bies.201100062.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. (2009) A census of human transcription factors: function, expression and evolution. *Nat Rev Genet.* 10(4):252-63.

van der Heijden RT, Snel B, van Noort V, Huynen MA (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*. 8:83.

Venkatesh TV, Holland ND, Holland LZ, Su MT, Bodmer R (1999) Sequence and developmental expression of amphioxus AmphiNk2-1: insights into the evolutionary origin of the vertebrate thyroid gland and forebrain. *Dev Genes Evol.* 209(4): 254-9.

Viettia

Vilella AJ, Severin J, Ureta-Vidal A, et al. (2009) EnsemblCompara GeneTrees. Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19(2): 327-335.

Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol*. 6:400.

von Mering C, Jensen LJ, Snel B, et al (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*.33: D433-7.

Wardle FC, Odom DT, Bell GW, Yuan B, Danford TW, Wiellette EL, Herbolsheimer E, Sive HL, Young RA, Smith JC. (2006) Zebrafish promoter microarrays identify actively transcribed embryonic genes. *Genome Biol.* 7(8):R71.

Wagner A (2002) Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol.* 10: 1760-8.

Walhout M (2011) Systems Biology: What's the Next Challenge? Cell. 144(6), 837 – 838.

a. **Wapinski I, Pfeffer A, Friedman N, Regev A (2007)** Natural history and evolutionary principles of gene duplication in fungi. *Nature*. 449, 54–61.

b. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics.* 23(13): i549-58.

Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.* 39(Database issue): D283-288.

Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. *Mol Ecol.* 15(5): 1197-211.

Wolf, YI, PS Novichkov, GP Karev, EV Koonin, DJ Lipman (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci*. 106:7273-7280.

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature.* 387: 708-713.

Wong A, Vallender EJ, Heretis K, Ilkin Y, Lahn BT, Martin CL, Ledbetter DH (2004) Diverse fates of paralogs following segmental duplication of telomeric genes. *Genomics.* 84(2): 239-47.

Wray GA. (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet.* 8(3):206-16. Review.

Yus E, Maier T, Michalodimitrakis K, et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science.* 326(5957):1263-8.

Venter JC, Adams MD, Myers EW, et al. (2001) The sequence of the human genome. *Science* 291, 1304-51.

Zhang J (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292-298.

Zheng-Bradley X, Rung J, Parkinson H, Brazma A. (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*. 11(12):R124.

Contribution in publications

The next appendixes contain reprints of the published articles that support this cumulative thesis. In the case of the article "Ancient animal microRNAs and the evolution of tissue identity" by Christodoulou et al (2010), this is not possible due to copyright restrictions. This article and the complete supplementary materials can be accessed through:
http://www.nature.com/nature/journal/v463/n7284/full/nature08744.html
Below, I list my contribution to the articles used for this thesis. The next information have been approved by Prof. Detlev Arendt.

Accepted publications:

Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. (2011) Orthology predictions quantified: An assessment of current methods using curated protein families. *Bioessays.* 33(10): 769-80.

- (My contribution to the publication: 60%) I assembled input data, performed data analysis and wrote the paper.
- This paper contributes to 45% of my thesis.

Trachana K, Jensen LJ, Bork P. (2010) Evolution and regulation of cellular periodic processes: a role for paralogues. *EMBO Rep.* 11(3): 233-8.

- (My contribution to the publication: 80%) I performed data analysis and wrote the paper.
- This paper includes results for the 25% of my thesis.

Christodoulou F, Raible F, Tomer R, Simakov O, **Trachana K**, Klaus S, Snyman H, Hannon GJ, Bork P, Arendt D. (2010 Ancient animal microRNAs and the evolution of tissue identity. *Nature*. 463(7284): 1084-8.

- (My contribution to the publication: 5%) I contributed to the bioinformatic analysis of this paper.

Appendix A

Orthology prediction methods: A quality assessment using curated protein families.

Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P.

Bioessays (2011) 33(10): 769-80.

Access in all supplementary material through:

http://onlinelibrary.wiley.com/doi/10.1002/bies.201100062/suppinfo

Orthology prediction methods: A quality assessment using curated protein families

Kalliopi Trachana[1)], *Tomas A. Larsson*[1)2)], *Sean Powell*[1)], *Wei-Hua Chen*[1)], *Tobias Doerks*[1)], *Jean Muller*[3)4)] *and Peer Bork*[1)5)*]

The increasing number of sequenced genomes has prompted the development of several automated orthology prediction methods. Tests to evaluate the accuracy of predictions and to explore biases caused by biological and technical factors are therefore required. We used 70 manually curated families to analyze the performance of five public methods in Metazoa. We analyzed the strengths and weaknesses of the methods and quantified the impact of biological and technical challenges. From the latter part of the analysis, genome annotation emerged as the largest single influencer, affecting up to 30% of the performance. Generally, most methods did well in assigning orthologous group but they failed to assign the exact number of genes for half of the groups. The publicly available benchmark set (http://eggnog.embl.de/orthobench/) should facilitate the improvement of current orthology assignment protocols, which is of utmost importance for many fields of biology and should be tackled by a broad scientific community.

Keywords:
metazoan; orthology; quality assessment

Introduction

The analysis of fully sequenced genomes offers valuable insights into the function and evolution of biological systems [1]. The annotation of newly sequenced genomes, comparative and functional genomics, and phylogenomics depend on reliable descriptions of the evolutionary relationships of protein families. All the members within a protein family are homologous and can be further separated into orthologs, which are genes derived through speciation from a single ancestral sequence, and paralogs, which are genes resulting from duplication events before and after speciation (out- and in-paralogy, respectively) [2, 3]. The large number of fully sequenced genomes and the fundamental role of orthology in modern biology have led to the development of a plethora of methods (e.g. [4–11]) that automatically predict orthologs among organisms. Current approaches of orthology assignment can be classified into (i) graph-based methods, which cluster orthologs based on sequence similarity of proteins, and (ii) tree-based methods, which not only cluster, but also reconcile the protein family tree with a species tree (Box 1). Despite the fact that orthology and paralogy are ideally illustrated through a phylogenetic tree, where all pairwise relationships are evident, tree-based methods are computationally

DOI 10.1002/bies.201100062

[1)] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[2)] Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[3)] Institute of Genetics and Molecular and Cellular Biology, CNRS, INSERM, University of Strasbourg, Strasbourg, France
[4)] Genetic Diagnostics Laboratory, CHU Strasbourg Nouvel Hôpital Civil, Strasbourg, France
[5)] Max-Delbruck-Centre for Molecular Medicine, Berlin-Buch, Germany

***Corresponding author:**
Peer Bork
E-mail: bork@embl.de

Abbreviations:
COG, clusters of orthologous group; **EGF,** epidermal growth factor; **HMM,** hidden Markov models; **LCA,** last common ancestor; **MSA,** multiple sequence alignment; **OG,** orthologous group; **RefOG,** reference orthologous groups; **VWC,** von Willebrand factor type C; **VWD,** von Willebrand factor type D.

Authors' contributions:
T. A. L., K. T., T. D., W. H. C., and S. P. did the phylogenetic analysis of the families. K. T., T. A. L., and W. H. C. performed the analysis. S. P. created the website. T. D., J. M., and P. B. designed the analysis. All the authors contributed to write the paper.

Supporting information online

Box 1 Comparison of orthology prediction methods

Orthology prediction methods can be classified based on the methodology they use to infer orthology into (i) graph-based and (ii) tree-based methods [12, 16, 17]. Different graph-based methods are designed to assign orthology for two (pairwise) or more (multiple) species. Graph-based methods assign proteins into OGs based on their similarity scores, while tree-based methods infer orthology through tree reconciliation.

Pairwise species methods (e.g. BHR, InParanoid, RoundUp):

Based on these methods, orthologs are best bi-directional hits (BBH) between a pair of species. BRH [46] is the first automated method and does not detect paralogs. InParanoid [47] implements an additional step for the detection of paralogs. RoundUp [48] uses evolutionary distances instead of BBH. In addition to the restriction of only two-species at a time, these methods are disadvantageous for long evolutionary distances.

Multi-species graph-based methods (e.g. COG, eggNOG, OrthoDB, OrthoMCL, OMA):

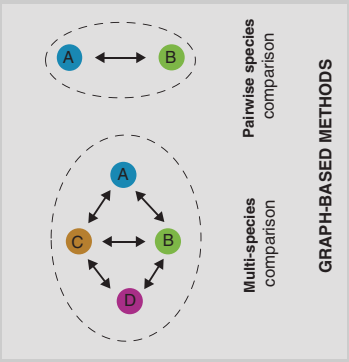
Due to the fast implementation and high scalability, there are many graph-based methods for multi-species comparisons. So far, all of them use either BLAST or Smith-Waterman (e.g. PARALIGN, SIMG) as sequence-similarity search algorithms. However, they are quite diverse regarding the clustering algorithms. COG, eggNOG, and OrthoDB share the same methodology: they identify three-way BBHs in three different species and then merge triangles that share a common side. OrthoMCL is a probabilistic method that uses a Markov clustering procedure to cluster BBH into OGs. OMA removes from the initial graph BBHs characterized by high evolutionary distance; a concept similar to RoundUp. After that, it performs clustering based on maximum weight cliques. Unique database characteristics are the hierarchical groups (OGs in different taxonomic levels) and ''pure orthologs'' (generate groups of one-to-one orthologs without paralogs), which has been introduced only by OMA (indicated as ** in the figure). Hierarchical groups can substitute the view of phylogenetic trees.

Multi-species tree-based methods (e.g. TreeFam, Ensembl Compara, PhylomeDB, LOFT):

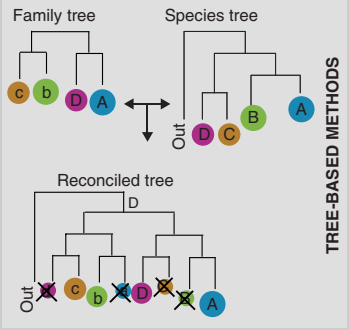
Tree-based prediction methods can be separated into approaches that do (like EnsemblCompara, TreeFam, and PhylomeDB) and do not, e.g. LOFT [49], use tree-reconciliation. Tree-based methods also initially use homology searches; however, their criteria are more relaxed, as the orthology is resolved through tree topology. Although a reconciled phylogenetic tree is the most appropriate illustration of orthology/paralogy assignment, there are a few caveats to such an approach, namely their scalability and sensitivity to data quality.

For a more detailed and extensive discussion of the differences among orthology methodology, we recommend refs. [12, 16, 17].

Phylogenetic distribution describes the species range of each database. Homology search shows a few technical differences for recruiting orthologs. §: Supplies OGs whose members share only orthologous relationships. *: The user can compare any two genomes spanning a phylogenetic distance from bacteria to animals.



	Phylogenetic distribution	Paralogs	Homology search	Clustering strategy	Hierarchical groups
BRH	ALL*	NO	BLAST	None	-
InParanoid	ALL*	YES	BLAST	None	-
RoundUp	ALL*	NO	Evol. Distance	None	-
COG	ALL	YES	BLAST	Triangles	NO
eggNOG	ALL	YES	BLAST	Triangles	YES
OrthoDB	Eukaryotes	YES	PARALIGN	Triangles	YES
OMA	ALL	YES**	SIMD & Evol. Distance	Maximum weight cliques	YES
OrthoMCL	ALL	YES	BLAST	Markov Clustering	NO



TreeFam	Metazoa	YES	BLAST & HMM	Hierarchical clustering	-
Ensembl Compara	Metazoa	YES	BLAST	Hierarchical clustering	-
PhylomeDB	ALL	YES	BLAST§	None	-

expensive and at times fail due to the complexity of the family or to the substantial number of species in the comparison [12]. As a trade-off between speed and accuracy, the evolutionary relationships among proteins in comparisons that include a large number of species are better explored using graph-based methods. During the first large-scale orthology assignment project of multiple species, the concept of clusters of orthologous groups (COGs) was introduced [4]. A COG consists of proteins that have evolved from a single ancestral sequence existing in the last common ancestor (LCA) of the species that are being compared, through a series of speciation and duplication events [4]. The orthologous/paralogous relationships among proteins of multiple species are better resolved through orthologous groups (OGs) rather than pairs of orthologs. This is particularly evident in the instances of complex protein family histories (e.g. tubulins) or families over significant phylogenetic distances (e.g. proteins conserved across all domains of life) [13].

Despite the clear definition of OGs, their automated prediction is challenged by a number of biological and technical factors exemplified by the evolution of mucins (see Fig. 1), a family with a complex evolutionary history [14]. The phylogenetic tree of mucins resolves the orthologous relationships among the members of the family at every evolutionary level (Fig. 1). Still, how they are grouped into OGs depends on the phylogenetic range of the species compared. For instance, a vertebrate-specific OG will include otogelin and VWF mucins, but not the additional gel-forming mucins (MUC5, MUC2, and MUC6). Conversely, all gel-forming mucins encompass a large OG when considering bilaterians (an animal clade that includes vertebrates, insects, and nematodes among others) as the level of comparison. Analyzing the OGs at different taxonomic levels (e.g. vertebrates vs. bilaterians) sheds light on the evolutionary history of the family; however, big protein families, which have expanded and contracted many times in the history of a lineage, require an increased resolution of orthologous-paralogous relationships within the same taxonomic level. The inclusion of outgroup species of a taxonomic level delineates the aforementioned relationships. For instance, Hydra sequences revealed the existence of two paralogous sequences in the LCA of bilaterians (marked by an asterisk in Fig. 1); thus, according to the OG definition, membrane-bound and gel-forming mucins should be clustered into two different OGs. Despite the lineage-specific duplications and losses of domains [14], many complex domain architectures are found across the family but not always conserved, which contributes to erroneous assignments of orthologs. Repeated domains and fast-evolving mucin domains also hamper the automatic sequence alignment of the family [15]. All these factors and more (see Fig. 1) can influence the accuracy of the many emerging resources for orthology assignment [13, 16, 17]. To understand the impact on individual resources, one needs to understand the design of different orthology prediction methods (briefly introduced in Box 1). However, an appropriate comparison is extremely difficult for two major reasons, both of which contribute to conflicting orthology assignments: (i) each method differs in technical (e.g. species distribution, similarity cut-offs) and conceptual (e.g. OG definition) aspects, and (ii) the lack of a common set of species obtained from the same release of genome repositories and tested across all methods [16].

Benchmarking orthology prediction methods using a phylogeny approach

Despite the acknowledged necessity of a phylogeny-based evaluation of orthology, thus far the majority of quality assessment tests are based on the functional conservation of predicted orthologs [18–21]. However, orthology is an evolutionary term defined by the relationships among the sequences under study, and functional equivalences are not always inferable [13]. Moreover, the functional divergence between orthologs and paralogs (sub-/neo-functionalization of paralogs) or alteration of function during long evolutionary distances [13] suggests that those tests are biased toward single copy genes or conserved families and less suited for large diversified families. It has been proposed that the inclusion of synteny information limits the errors arisen due to low sequence similarity and increases orthology accuracy [22]. However, this requires a certain level of synteny conservation among the compared species. It has been illustrated that synteny information combined with sequence similarity identifies accurately the paralogs that have arisen through WGD in six closely related yeast species [23]. Further refinement of this dataset using tree reconciliation [24, 25] ends up with a phylogeny-based dataset. However, it is still biased toward simple evolutionary scenarios, highlighting mostly the impact of lineage-specific losses in orthology prediction [26]. For a much more fine-grained analysis that also involves complex OGs, we developed a phylogeny-based benchmark set and applied it to a much more diverse taxonomic clade, namely metazoans. The set involved the manual curation of the phylogeny of 70 protein families that range from single copy orthologs to OGs with 100 members (Table S1 of Supporting Information). The phylogenetic analysis of each protein family for 12 reference bilaterian species and 4 basal metazoans as outgroups (Box 2) resulted in the reference orthologous groups (RefOGs), including in total 1,638 proteins.

The manually curated benchmarking set was used for two different analyses: (i) comparison of RefOGs to the automatically predicted OGs of five publicly available databases, and (ii) comparison of RefOGs to different customized versions of the eggNOG database. The first comparison aimed at demonstrating the power of this dataset to guide the improvement of current methods. We selected five databases, namely TreeFam [5], eggNOG [6], OrthoDB [7], OrthoMCL [8], and OMA [9], since each is designed for multiple-species comparison, but with unique database features (Box 1). Although the comparisons are against the same benchmarking set, we are aware of several other confounding variables, such as algorithmic differences, species representation/distribution or genome annotation, that can all affect the results. Yet, it quantifies the status of the compared databases in an objective way. To quantify the impact of some specific biological and technical factors, we additionally generated different versions of the eggNOG database to monitor several influencing factors one by one.

We assessed the quality of the OGs at two different levels of resolution: (i) gene count, measuring mispredicted genes, and

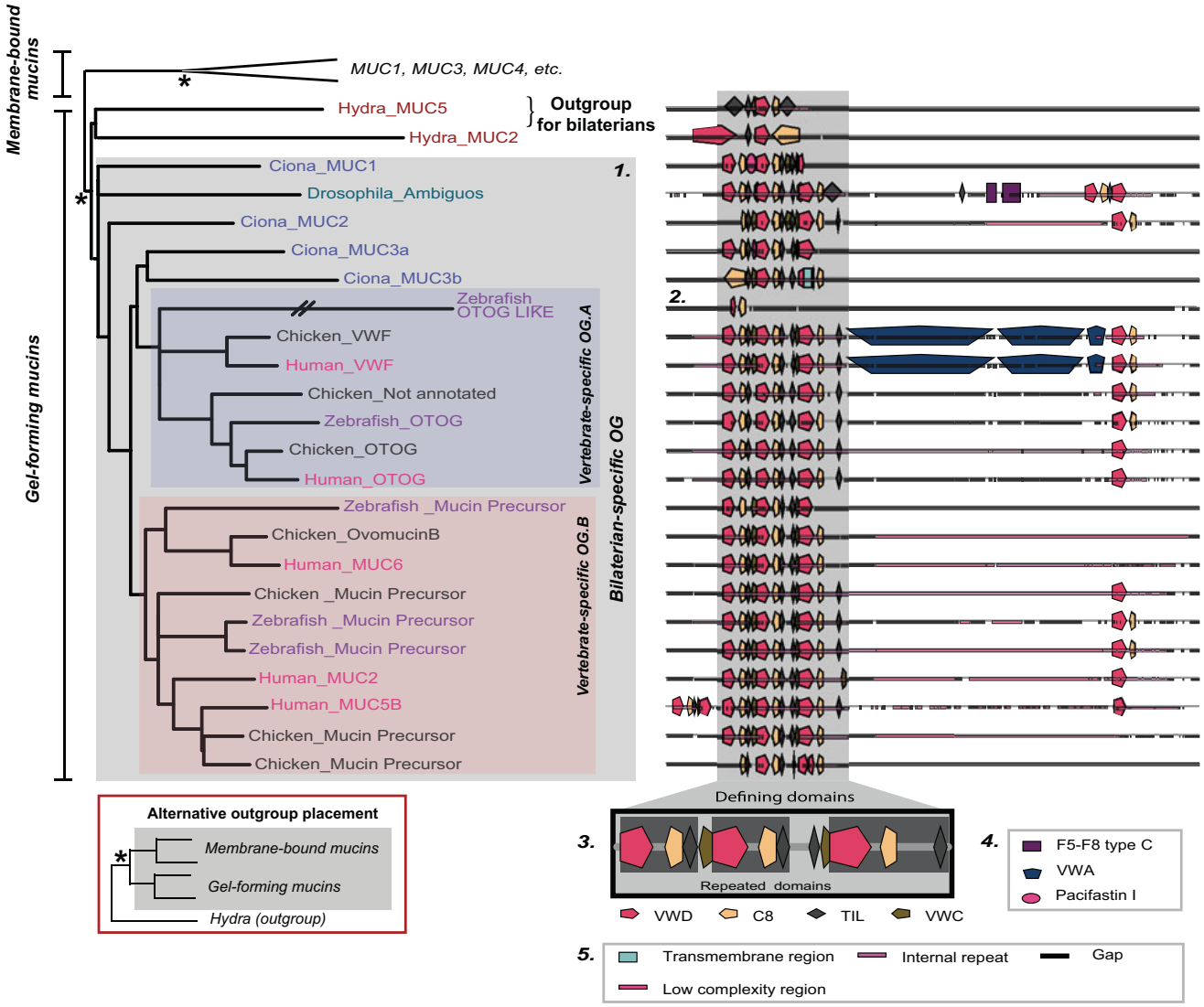


Figure 1. Mucins: a challenging family for orthology prediction. This figure shows the phylogenetic tree and domain architecture of aligned mucins. The identification of cnidarian (an outgroup for bilaterians) mucin2/5 orthologs separates the gel-forming mucins from other mucins, defining a bilaterian-specific OG (gray box). An alternative topology of Hydra in respect to the LCA of bilaterian species (shown schematically in the red box) would propose that those two different classes of mucins should be clustered together at the bilaterian level. The bilaterian OG can be further resolved at the vertebrate level into OG.A (blue) and OG.B (red), illustrating the hierarchical nature of OGs. This family, besides its large size due to vertebrate-specific duplications, exemplify five additional problems that often lead to orthology misassignment: (1) uneven evolutionary rate illustrated as branch lengths, lowering the sequence similarity among members of the family; (2) quality of genome annotation: the particular zebrafish protein can be either a derived member of the mucin family or a erroneous gene prediction; (3) repeated domains: the domain combination VWD-C8-VWC, which is the core of the family, is repeated multiple times within the protein; (4) complexity of domain architectures: there are multiple unique domain combinations (e.g. the VWD domain is combined with the F5-F8 type C domain only in the *Drosophila* ortholog); and (5) low complexity regions: internal repeats within the amino acid sequences and other low complexity features impede the correct sequence alignment of the mucins. *Possible orthologous sequence at the LCA of cnidarians bilaterians.

(ii) group count, reflecting errors at the level of OG (Fig. 2). Additionally, for each of the two resolution levels, we used three counting schemes allowing us to distinguish database-specific trends. At a strict requirement of all genes being correctly assigned (gene count level) only as little as 3–22% of the RefOGs were recovered, while a more relaxed requirement that curated orthologs are not clustered in multiple OGs or with other homologous proteins that are not part of the RefOG (group count level) results in 10–48% correctly predicted RefOGs. Limiting our analysis to the 35 most challenging families decreases this percentage even more (Fig. S1 of Supporting Information), reflecting our initial aim to select families that hamper accurate orthology prediction; we aimed at a benchmark set that points out shortcomings of each method and leads to its improvement. All above indicated that there is room

for improvement for all methods, but most importantly, we have to understand which factors contributed to this result.

The phylogenetic range of the compared species affects the accuracy of prediction

The phylogenetic distribution of the compared species influences the orthology/paralogy assignment, as we exemplified with the mucin family (vertebrate- vs. bilaterian-specific groups). The broader the phylogenetic range of the compared species the larger the OGs, as the single ancestral sequence from which all the orthologs and paralogs are derived is placed deeper in the tree. This is reflected in the ranking of the five databases that varies considerably in the six different scoring schemes used (Fig. 2). For instance, although OrthoMCL contains the highest number of erroneously assigned genes (Fig. 2C), the number of RefOGs that are affected by erroneously assigned genes is higher for eggNOG than OrthoMCL (Fig. 2D). On closer examination, OrthoMCL overpredicts many

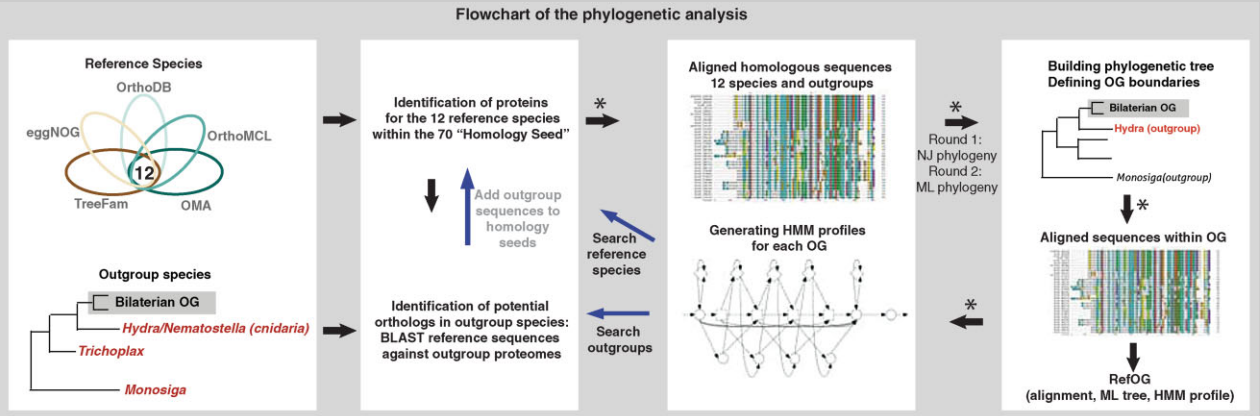
Box 2

Phylogenetic analysis of the 70 protein families

Selecting families for exploring caveats of orthology prediction: we focused on five major affecting factors of orthology prediction, mostly related with metazoan (eukaryotic) biology: rate of evolution (fast- vs. slow-evolving families), domain architecture (single domain vs. multiple repeated domains), low complexity/repeats, lineage-specific loss/duplication (single copy families vs. multiple duplication events), and alignment quality (high- vs. low-quality alignment). We used the eggNOG database to select 70 families (Supporting Information) that we refer to as ''homology seeds.'' Of the selected families, 35 exemplify known biological and technical challenges. Five additional slow-evolving, well-aligned families were chosen as counterbalance, while the remaining 30 families were chosen randomly to avoid prior biases (Table S1 of Supporting Information).

Defining of reference species: for an applicable comparison of the five databases studied, we had to confine the analysis to 12 reference species that are shared by all resources: *Caenorhabditis elegans*, *Drosophila melanogaster*, *Ciona intestinalis*, *Danio rerio*, *Tetraodon nigroviridis*, *Gallus gallus*, *Monodelphis domestica*, *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*, *Pan troglodytes*, *Homo sapiens*. All 12 species belong to the bilaterians, a metazoan subgroup simplifies the objective of this study since (i) the phylogeny of bilaterians is reasonably defined, and (ii) a few fully sequenced basal metazoan genomes (like cnidarians) can be used as outgroups of bilaterians [29, 50–52].

The phylogenetic analysis: briefly, we selected 70 aforementioned COG/KOGs, as they exist in eggNOG v2 [6], which we refer to as ''homology seeds.'' To exclude errors due to old genome annotation (eggNOG v2 is based on Ensembl v46), we mapped the ''homology seed'' identifiers to Ensembl v60. The following steps were performed uniformly to all families certifying that RefOGs are not biased toward their initial ''homology seeds.'' BLAST [53] searches were performed in the 16 animals using query sequences from well-annotated genomes (e.g. human, zebrafish, and fly). The homologous sequences were aligned with MUSCLE [54] and the alignments were used to build initial NJ trees with Clustal X [55] (indicated as Round 1 in the illustration below). Large groups were thereafter divided based on the positions of orthologs in the outgroups, as exemplified by the family of mucins (Fig. 1). In several cases where no clear outgroup was found, RefOGs were defined based on (i) the domain content, (ii) manual inspection of the alignments, and (iii) previous published descriptions of the families. After the initial curation of the families, all sequences determined to be members of the bilaterian RefOGs were aligned using MUSCLE [54]. Alignments were refined [56] and hidden Markov models (HMM) were built using the HMMER3 package [57]. In a second refinement step (indicated as Round 2), the HMM models were used to identify related sequences that were left out from the 16 aforementioned genomes. As a last step, all qualified members of each RefOG were realigned, using the same procedure as before, final HMM models were generated and phylogenetic trees were calculated using PhyML version 3.0 [58]. The detailed analysis is described in the supplementary file. Black arrows indicate the flow of the analysis. * Steps that are repeated after HMM profile searches resulting in RefOGs after Round 2 (red arrow).



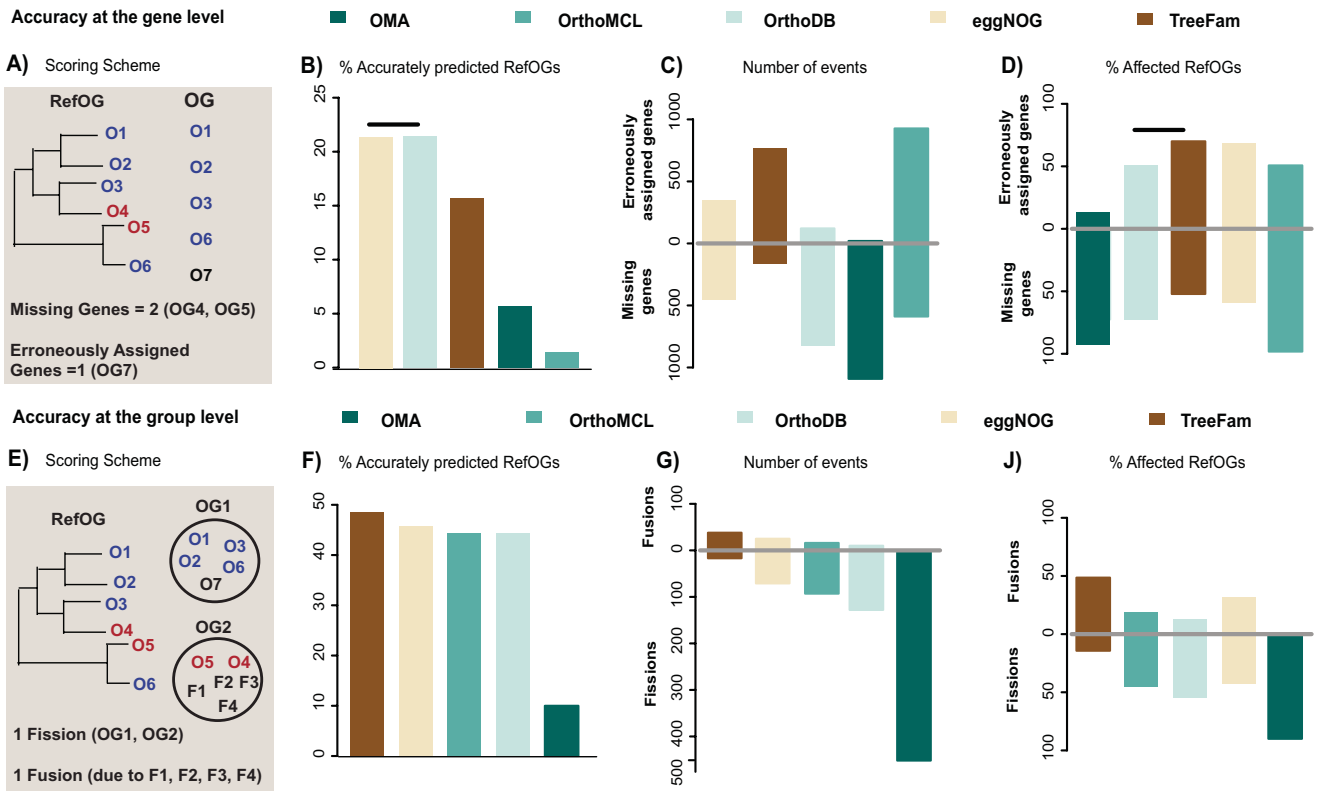


Figure 2. The 70 manually curated RefOGs as a quality assessment tool. Five databases were used to illustrate the validating power of the benchmark set. The performance of each database was evaluated at two levels: gene (focus on mispredicted genes; upper panel) and Group (focus on fusions/fissions; lower panel) level. **A:** Gene count – for each database we identified the OG with the largest overlap with each RefOG and calculated how many genes were not predicted in the OG (missing genes) and how many genes were over-predicted in the OG (erroneously assigned genes) and **E:** group count – for each method we counted the number of OGs that members of the same RefOG have been separated (RefOG fission) and how many of those OGs include more than three erroneously assigned genes (RefOG fusion). To increase the resolution of our comparison, three different measurements for each level were provided, resulting in six different scoring schemes. **B:** Percentage of accurately predicted RefOGs in gene level (RefOGs with no mispredicted genes); **C:** number of erroneously assigned and missing genes; **D:** percentage of affected RefOGs by erroneously assigned and missing genes; **F:** percentage of accurately predicted RefOGs in grouplevel (all RefOG members belong to one OG and are not fused with any proteins); **G:** number of fusions and fissions; and **J:** percentage of affected RefOGs by fusion and fission events. Databases are aligned from the more to the less accurate, taking into account the total number of errors (length of the bar in total). Black bars indicate identical scores.

orthologs in only a few families, while eggNOG overpredicts a few proteins in many families (Table S2 of Supporting Information), partially due to mispredicted genes (later characterized as pseudogenes or wrong gene models) inherited from an old genome annotation (see below). We assume this observation is partly due to the diverse species ranges of the studied repositories (Box 1). EggNOG, although it provides a broad species coverage (630 prokaryotes and 55 eukaryotes), supplies OGs for several taxonomic levels, such as metazoans (meNOGs) that are used in this study and are build from 34 bilaterians in the eggNOG version studied here. On the other hand, OrthoMCL builds its OGs from all 138 eukaryotic and prokaryotic species in the database. In other words, ancient families, e.g. ABC transporters, which expanded before the bilaterian radiation, form huge OGs in OrthoMCL, but not in the meNOG subset of eggNOG. As different scientific questions require a different species range, hierarchical groups as provided by eggNOG [27], OrthoDB [28], and OMA [9] appear to be a balanced solution to serve many different questions, compared to databases that are only dedicated to a particular phylogenetic range [be they narrow (TreeFam) or broad (OrthMCL)].

Despite being specifically designed for metazoans, TreeFam has the second largest number of erroneously assigned genes after OrthoMCL (Fig. 2C), which is accompanied by the largest number of fusion events (Fig. 2G). This can be attributed to the choice of outgroups used by Treefam. TreeFam families are phylogenetically separated by a non-animal outgroup (yeast or plant), while, for example, *Monosiga brevicollis* [29] or other proposed species [30] would be much better suited. The choice of a phylogenetically closer species would presumably split artificially large families. Furthermore, delineating orthology through tree reconciliation benefits TreeFam in the category of missing genes (Fig. 2C), since the lack of a closer outgroup prevents the bilaterian OGs from splitting, as illustrated in Fig. 1. In contrast, the database with the largest number of missing genes and fission events is OMA (Figs. 2C and G) due to an alternative operational definition of an OG [31]; only proteins with one-to-one orthologous relationships are included in an OG, so that

large families with multiple paralogs are split artificially into multiple smaller OGs. The latest release of the OMA database, publicly available after the completion of our analysis, has been redesigned and now provides OGs based on both OMA and COG formulations [9].

In summary, the initial design of an orthology resource, e.g. phylogenetic range of species, ''hierarchical groups'', or formulation of OG, is crucial for its performance. In any case, all methods only predict a fraction of RefOGs accurately and mispredict a large number of genes (Fig. 2). It is noteworthy that there are RefOGs that none of the methods infer accurately, indicating that there are biological and technical factors that affect the performance of orthology assignment more generally. We have thus tried to relate a few of them with the outcome of this comparison.

The impact of family complexity on orthology prediction

Due to the central role of orthology in comparative and functional genomics, there is an extensive literature on accuracy-restricting factors of its assignment [13, 16, 17]. We have already mentioned several caveats of orthology prediction using the mucin family, the majority of which are exemplified by the 70 RefOGs. The families were selected under certain criteria (Box 2), mostly with a view to understanding the impact of a few biological and technical factors, namely duplications (paralogy)/losses, rate of evolution, domain architecture, and alignment quality. All these factors have been reported to affect the quality of orthology prediction [17]. Paralogy as manifested in multi-gene families hamper the accurate orthology prediction [4, 13]. Multiple lineage-specific gene losses and duplications result in complex evolutionary scenarios, which are hard to interpret. Classifying the RefOGs based on their size, we observed that the larger the RefOG, the more mispredictions are introduced by the methods (Fig. 3A). For all methods, the numbers of missing genes (Fig. 3A) and OG fissions (Fig. S2 in Supporting Information) increases significantly with the RefOG size (Table S5 of Supporting Information). Additionally, families with more than 40 members accumulate both fusion and fission events. For instance, GH18-chitinases, a RefOG that consists of 45 members, is characterized by multiple vertebrate-specific duplication events. All graph-based methods split the vertebrate subfamilies of the GH18-chitinases into distinct groups (Table S2 of Supporting Information), and TreeFam lumps the RefOG with insect-specific homologs due to the presence of the glyco-hydro-18 domain, although phylogenetic analysis of the family indicates a general lack of orthology among those groups [32].

Some large-size RefOGs, like ribosomal proteins or SAM-synthetases are, however, predicted accurately by several methods. Since these two well-predicted large families are well conserved, we decided to investigate the impact of the rate of evolution on orthology prediction. We categorized our benchmarking families into fast-, medium-, and slow-evolving based on their MeanID score (described as the ''FamID'' in [33]), which indicates the rate of evolution (Supporting Information). Fast-evolving families tend to accumulate a larger number of errors (Fig. 3B). All graph-based methods miss a larger number of genes and introduce more fission events (Fig. S2 in Supporting Information) in fast-evolving RefOGs compared to the more slowly evolving groups. Since the MeanID score is calculated based on the multiple sequence alignment (MSA), we investigated the impact of MSA quality by calculating the norMD score [34], an alignment score that depends on the number and the length of aligned sequences as well as their estimated similarity (Supporting Information). We expected TreeFam to be more sensitive to low-quality MSAs compared to graph-based methods, since it uses MSA for tree-building and reconciliation steps to infer orthology. Indeed, it presents the highest deviation for all sources of errors (Table S5 of Supporting Information). We also found that the number of missing genes is also affected by the alignment quality in graph-based methods (Fig. 3C). Because MeanID and norMD scores are correlated, many of the fast-evolving families are also poorly aligned. Still, we can see that TreeFam is significantly more affected by MSA quality rather than rate of evolution.

The vast majority of proteins contain only one domain, and the most common multi-domain proteins tend to have few (two or three) domains [35, 36]. Due to a variety of genetic processes (duplication, inversion, recombination, retrotransposition, etc.) proteins consisting of multiple domains with independent evolutionary origin can arise [37–40]. This leads to conceptual but also practical challenges (e.g. alignment) in orthology prediction, as the domains have followed distinct evolutionary trajectories [16]. We identified the domains of each protein in each RefOG through the SMART database [41]. Out of the 70 RefOGs, 75% contain multi-domain (more than two domains) proteins, compared to 62% in the random subset and a report of 40% multi-domain occurrence in metazoans [36], which illustrates the tendency of the benchmark set toward more challenging families. As expected, the proportion of accurately predicted RefOGs decreases as the number of average domains per family increases (Fig. 3D). Interestingly, the rate of erroneously assigned genes presents the most significant correlation with domain complexity, suggesting that protein families with multiple protein domains ''attract'' non-orthologous proteins due to domain sharing. Repeated domains within proteins, as the Von Willebrand factor (VW) D-C8-VWC repeat in mucins (Fig. 1) or the epidermal growth factor (EGF) domains in collagen, also lead to lower quality of OGs. All of the 27 RefOGs containing repeated domains are more error prone than RefOGs without repeated domains (Fig. S3 of Supporting Information).

Taken together, classification of the families from slow-evolving single copy to fast-evolving large families revealed method-specific limitations, but also that all pipelines fail to predict complex families accurately. The rates of missing genes and fissions significantly correlate with the family size and rate of evolution, as expected, whereas the domain complexity seems to affect the recruitment of non-orthologous genes (Fig. 3, Figs. S2 and S4 of Supporting Information).

Species coverage affects orthology prediction

Biological complexity is unlikely to be the primary source of errors in automated predicted OGs, as there are single-copy,

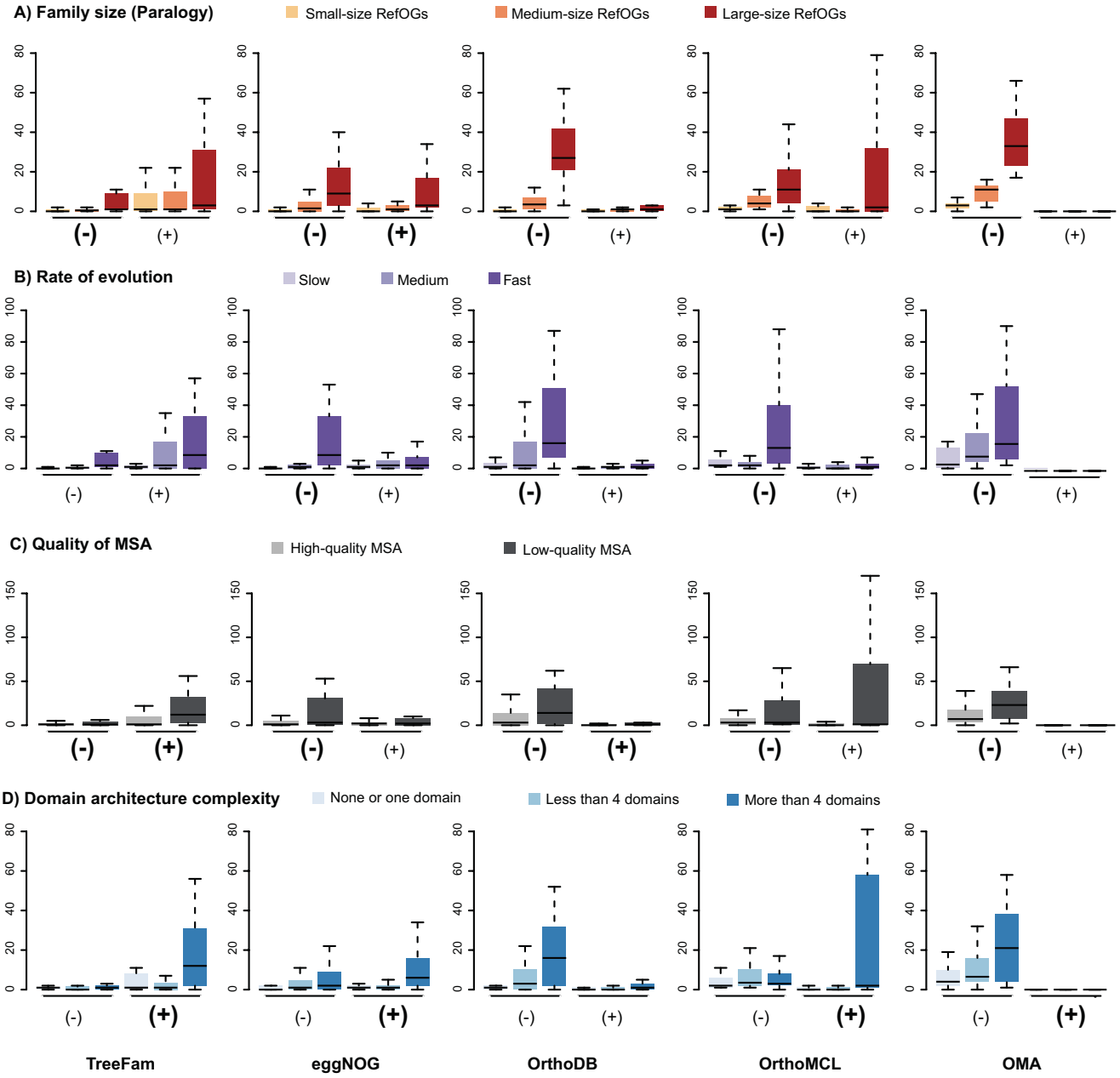


Figure 3. The impact of biological complexity in orthology assignment. To evaluate the impact of important caveats in orthology prediction, the RefOGs were classified based on their family size, rate of evolution, alignment quality and domain complexity. **A:** Family size (reveals the impact of paralogy): the RefOGs were separated into (i) small (less than 14 members), (ii) medium (more than 14 members, but less than 40), and (iii) large (more than 40 genes). **B:** Rate of evolution: the RefOGs were classified based on the MeanID score (described as the ''FamID'' in [33]), an evolutionary rate score derived from the MSA of each family. There are: (i) slow-evolving (MeanID >0.7), (ii) medium-evolving (MeanID <0.7, but >0.5), and (iii) fast-evolving (MeanID <0.5) RefOGs. **C:** Quality of alignment: we classified the families based on their norMD score [34] into: (i) high-quality alignment (norMD >0.6), and (ii) low-quality alignment [44, 45]. We can observe that high amino acid divergence correlates with an increasing number of mispredicted genes. **D:** Domain architecture complexity; each RefOG is associated with the average number of domains, which is equal to the sum of predicted domains of the members of one RefOG divided by the family size. There are three levels of complexity, starting from (i) none or one domain on average, to (ii) two to four, to (iii) more than four. We observe that the performance of the five databases correlates with the biological complexity of RefOGs; as families increasing their complexity (more members, fast-evolving or multiple domains), the accuracy of predictions drops. (+) and (−) symbolize erroneously assigned and missing genes, respectively. Significant correlations (Table S5 of Supporting Information) between the distribution of missing/erroneously assigned genes and the tested factor are indicated in bold [**(+)**, **(−)**]. Figures S2 and S4 of Supporting Information show similar observations at the group level (fusions/fissions of RefOGs).

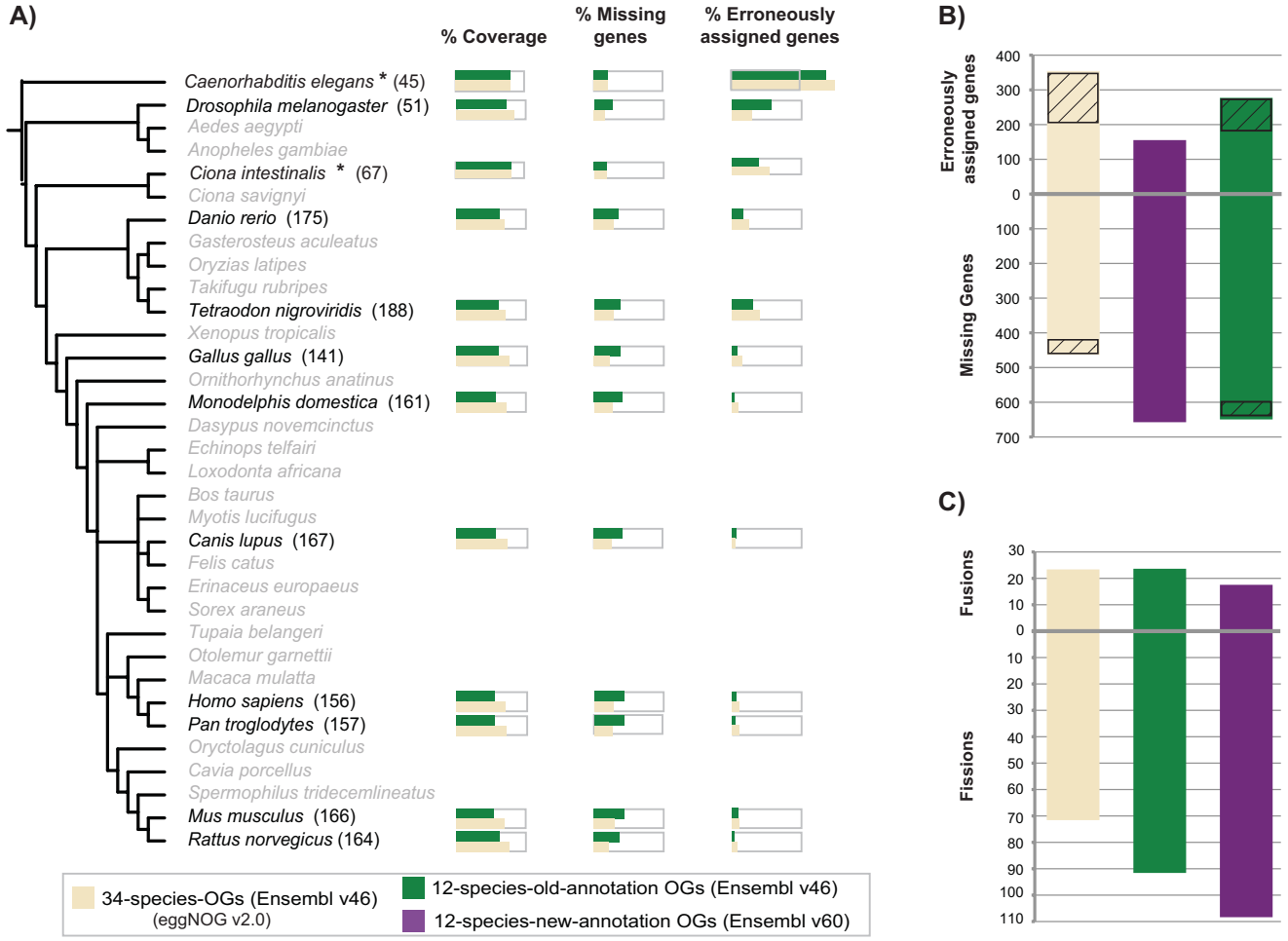


Figure 4. The impact of species coverage and genome annotation. **A:** Comparison of the performance of 34-species and 12-species OGs using RefOGs. We measure the percentage of orthologs recovered (coverage), missing genes and erroneously assigned genes for each reference species for those datasets [yellow bar: publicly available OGs in eggNOG (same measurements as Fig. 2) and green bar: customized OGs of the 12 selected species using same genome annotations as the public eggNOG]. The reference species are highlighted by black letters, while the unconsidered species that complete the set of 34 eggNOG species are written in gray letters. Numbers in parentheses show the total number of orthologs per species in the benchmarking set. The gray boxes enclosing the colored bars correspond to 100% coverage. Notice that the coverage is always higher for the 34-species OGs compared to the 12-species OGs except in the cases of *C. elegans* and *Ciona* (marked by asterisk), which are separated by long branches in both datasets. **B:** Comparison of the public eggNOG (yellow bar), 12-species-old-annotation OGs (green bar) and 12-species-new-annotation OGs (purple bar) at the gene level. Hatched boxes label the fraction of mispredicted genes of 34-species- and 12-species-old-annotation datasets that do not exist in Ensembl v60 genome annotations, indicating the high number of errors due to old genome annotations. **C:** Comparison of public eggNOG (yellow bar), 12-species-old-annotation OGs (green bar) and 12-species-new-annotation OGs (purple bar) at the group level. Notice that the 12-species datasets (either with old or new annotation) always introduce a larger number of fission events than the 34-species OGs, highlighting again the importance of species coverage.

slow-evolving, or single-domain protein families, which are not assigned correctly by several prediction methods. By investigating these families, we identified two additional technical factors that influence orthology assignment: genome annotation and species coverage. To quantify the impact of these, we used the method in our own hands, eggNOG, as we could apply it to different species sets (Fig. 4, Table S3 of Supporting Information) and genome annotation versions (Fig. 4, Table S4 of Supporting Information).

To measure the impact of species coverage, we prepared new OGs from only the 12 reference species, but kept the same genome annotation version (Ensembl v46) that the public eggNOG v2 uses. The 12-species-Ensembl46 OGs were compared to the RefOGs as well as the 34-species-Ensembl46 OGs (referred to as eggNOG in Fig. 2). In the 12-species-Ensembl46 OGs, a larger number of genes are missing compared to the 34-species OGs (eggNOG_v2) (Fig. 4B), implying that 30% of the missing genes in this dataset are due to the change in species coverage. It seems that sequences of the 34 species facilitate correct clustering, presumably, by breaking long branches so that faster evolving genes can be connected (Fig. 4A). For mammals, fish and insects, which contain more representatives in 34-species OGs, we identified fewer missing genes in the 34-species

OGs than the 12-species OGs. On the other hand, *C. elegans* and *C. intestinalis*, which are separated by long branches from their nearest phylogenetic neighbors in both datasets, are not influenced as the sequence similarity for ortholog detection remains limited (Fig. 4A). While 34-species perform better than 12-species in terms of missing genes, they contain more erroneously assigned genes. A large fraction of erroneously assigned genes is due to inclusion of low-quality genomes, i.e. Tetraodon in Ensembl v60 contains almost 5,000 gene predictions less than the same genome in Ensembl v46. In summary, the total number of mispredicted genes is higher for the 12-species OG (Figs. 4A and C), indicating that the more genomes and in particular those at the right evolutionary distance, increase the quality of the OGs.

Number of errors inflates because of inaccuracies in genome annotation

The quality of the genome annotation of a species included in a genomic or phylogenetic study has been reported to affect the results of the study [42]. All resources in this study rely on Ensembl [43] genome annotations for all 12 species, but the annotation status is considerably different from version to version. While eggNOG uses Ensembl v46 (the oldest among the compared resources) OrthoDB uses Ensembl v59, thus it is the most updated and closest to the RefOG annotation, for which Ensembl v60 was used. By tracing the identifiers of the mispredicted genes through Ensembl history, we discovered that 7% of the missing genes of eggNOG only exist in the latest versions of Ensembl (v54 to v60) (Fig. 4B). Genomes like human, zebrafish and puffer fish, which were updated after Ensembl v46, contribute significantly to the pool of missing genes. Likewise, only 58% of the erroneously assigned genes of eggNOG map to Ensembl v60, while 40% of them have been retracted and 2% have been characterized as pseudogenes. Taken together, almost half of all errors result from genome annotation artifacts, which is thus a major factor to consider. To directly test the effect of the genome annotation and separate the impact of species coverage from this analysis, we clustered the proteins of the 12 reference species based on the Ensembl v60 gene annotations. The impact of genome annotation is elucidated by comparing the number of errors between the 12-species-Ensembl60 OGs with the 12-species-Ensembl46 OGs. Comparing the overall number of mispredicted genes, at the gene level, the 12-species-Ensembl60 OGs perform better than the 12-species-Ensembl46 OGs (Fig. 4B). We found 45% fewer erroneously assigned genes (149 vs. 271) in the 12-species-new-annotation OGs compared to the 12-species-old-annotation OGs. Again, a large fraction or erroneously assigned genes of the latter dataset (33%) do not exist in Ensembl v60 (Table S4 of Supporting Information). However, the number of missing genes is similar between the two datasets and higher compared to the 34-species OGs, indicating, once again, the impact of species coverage. The fact that ~40% of the mispredicted genes in eggNOG OGs would have been avoided by using an updated version of genome annotations, highlights the importance of frequent updates and points to the sensitivity of genome annotations.

A transparent benchmark set made publicly available

To facilitate the access to the curated benchmark families, we have created a web interface through which details on the 70 RefOGs can be retrieved. In addition, alignments, protein sequences, phylogenetic trees and HMM of each RefOG can be downloaded and used for future analyses of the 70 bilaterian OGs. The data are available under the Creative Commons Attribution 3.0 License at: http://eggnog.embl.de/orthobench.

Conclusions

The quality assessment introduced here is independent of functional associations and, instead, directly approaches the phylogenetic foundations of OGs. The benchmark set was applied to five commonly used databases and revealed the impact of several biological and technical factors that challenge orthology prediction. All studied repositories predict only a fraction of RefOGs accurately and thus indicate that there is considerable room for improvement for all orthology assignment methods. Although it is impossible to completely quantify the individual factors that contribute to the errors of each method due to the diversity of the methodologies, hidden correlations, and confounding variables, the 70 RefOGs reveal biological and technical limitations that affect each method significantly. For example, domain complexity is significantly correlated with an increased accumulation of erroneously assigned genes in all databases. Our results also illustrate that all the tested algorithms need to be improved to be able to handle the ''complex'' families (duplication/losses, complex domain architectures). Of the RefOGs, 36% are not accurately predicted by any tested databases, revealing ''global'' limitations of orthology predictions that are associated with the factors we outlined here. There are also RefOGs that only some of the databases mispredict, and, thus, hint at database-specific improvements, i.e. several operational differences, such as the delineation of hierarchical groups, and the usage of (as close as possible) outgroups affect the accuracy of predicted OGs.

However, the most striking outcome of this study is that technical factors, such as genome quality followed by the phylogenetic coverage of the compared species seem to be the most limiting factors, causing up to 40% of the errors observed. The last observation suggests that frequent updates of the databases are necessary. Although we only tested bilaterian OGs in this study, we realize the importance of the expansion to other taxonomic groups, and have therefore provided sequences, alignments, HMM profiles, and trees of the RefOGs publicly at http://eggnog.embl.de/orthobench for further curation in other species. As this benchmark set proved valuable for assessing the quality of predicted OGs in metazoans, we believe that an analogous dataset covering the entire tree of life and capturing additional challenges more prominent in prokaryotes, such as horizontal gene transfer, should be the next step in guiding orthology prediction.

Acknowledgments

We would like to thank the OrthoDB team for providing us with a customized dataset. We are grateful to Lars Juhl Jensen, Martijn Huynen, Evgeny Zdobnov, and Michael Kuhn for their insightful comments, Venkata Satagopam for his help with EnsEMBL History, and all members of the Bork group for the fruitful discussions, and in particular Ivica Letunic for his help with the multidomain analysis. We would like also to thank the three anonymous reviewers for their constructive comments.

References

1. **Koonin EV**, **Galperin MY.** 2003. *Sequence - Evolution - Function. Computational Approaches in Comparative Genomics*. Kluwer Academic: Boston.
2. **Fitch WM.** 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99–113.
3. **Sonnhammer EL**, **Koonin EV.** 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**: 619–20.
4. **Tatusov RL**, **Koonin EV**, **Lipman DJ.** 1997. A genomic perspective on protein families. *Science* **278**: 631–7.
5. **Ruan J**, **Li H**, **Chen Z**, **Coghlan A**, et al. 2008. TreeFam. 2008. Update. *Nucleic Acids Res* **36**: D735–40.
6. **Muller J**, **Szklarczyk D**, **Julien P**, **Letunic I**, et al. 2010. eggNOG v2.0. extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **38**: D190–195.
7. **Waterhouse RM**, **Zdobnov EM**, **Tegenfeldt F**, **Li J**, et al. 2011. OrthoDB, the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res* **39**: D283–288.
8. **Chen F**, **Mackey AJ**, **Stoeckert, CJ Jr.**, **Roos DS.** 2006. OrthoMCL-DB. Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**: D363–368.
9. **Altenhoff AM**, **Schneider A**, **Gonnet GH**, **Dessimoz C.** 2011. OMA 2011. Orthology inference among 1000 complete genomes. *Nucleic Acids Res* **39**: D289–294.
10. **Vilella AJ**, **Severin J**, **Ureta-Vidal A**, **Heng L**, et al. 2009. EnsemblCompara GeneTrees. Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**: 327–35.
11. **Huerta-Cepas J**, **Capella-Gutierrez S**, **Pryszcz LP**, **Denisov I**, et al. 2011. PhylomeDB v3.0. An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* **39**: D556–560.
12. **Kristensen DM**, **Wolf YI**, **Mushegian AR**, **Koonin EV.** 2011. Computational methods for Gene Orthology inference. *Brief Bioinform*, in press, DOI: 10.1093/bib/bbr030.
13. **Koonin EV.** 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309–38.
14. **Lang T**, **Alexandersson M**, **Hansson GC**, **Samuelsson T.** 2007. Gel-forming mucins appeared early in metazoan evolution. *Proc Natl Acad Sci USA* **104**: 16209–14.
15. **Lang T**, **Alexandersson M**, **Hansson GC**, **Samuelsson T.** 2004. Bioinformatic identification of polymerizing and transmembrane mucins in the puffer fish Fugu rubripes. *Glycobiology* **14**: 521–7.
16. **Gabaldon T.** 2008. Large-scale assignment of orthology. Back to phylogenetics? *Genome Biol* **9**: 235.
17. **Kuzniar A**, **van Ham RC**, **Pongor S**, **Leunissen JA.** 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* **24**: 539–51.
18. **Pryszcz LP**, **Huerta-Cepas J**, **Gabaldon T.** 2010. MetaPhOrs. Orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res* **39**: e32.
19. **Hulsen T**, **Huynen MA**, **de Vlieg J**, **Groenen PM.** 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7**: R31.
20. **Chen F**, **Mackey AJ**, **Vermunt JK**, **Roos DS.** 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* **2**: e383.
21. **Altenhoff AM**, **Dessimoz C.** 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5**: e1000262.
22. **Goodstadt L**, **Ponting CP.** 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* **2**: e133.
23. **Byrne KP**, **Wolfe KH.** 2005. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**: 1456–61.
24. **Wapinski I**, **Pfeffer A**, **Friedman N**, **Regev A.** 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* **23**: i549–58.
25. **Akerborg O**, **Sennblad B**, **Arvestad L**, **Lagergren J.** 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA* **106**: 5714–9.
26. **Salichos L**, **Rokas A.** 2011. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One* **6**: e18755.
27. **Jensen LJ**, **Julien P**, **Kuhn M**, **von Mering C**, et al. 2008. eggNOG. automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **36**: D250–254.
28. **Kriventseva EV**, **Rahman N**, **Espinosa O**, **Zdobnov EM.** 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* **36**: D271–5.
29. **King N**, **Westbrook MJ**, **Young SL**, **Kuo A**, et al. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451**: 783–8.
30. **Rokas A.** 2008. The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet* **42**: 235–51.
31. **Schneider A**, **Dessimoz C**, **Gonnet GH.** 2007. OMA Browser-exploring orthologous relations across 352 complete genomes. *Bioinformatics* **23**: 2180–2.
32. **Funkhouser JD**, **Aronson, NN. Jr.**, 2007. Chitinase family GH18. Evolutionary insights from the genomic history of a diverse protein family. *BMC Evol Biol* **7**: 96.
33. **Muller J**, **Oma Y**, **Vallar L**, **Friederich E**, et al. 2005. Sequence and comparative genomic analysis of actin-related proteins. *Mol Biol Cell* **16**: 5736–48.
34. **Thompson JD**, **Plewniak F**, **Ripp R**, **Thierry JC**, et al. 2001. Towards a reliable objective function for multiple sequence alignments. *J Mol Biol* **314**: 937–51.
35. **Doolittle RF.** 1995. The origins and evolution of eukaryotic proteins. *Philos Trans R Soc Lond B Biol Sci* **349**: 235–40.
36. **Tordai H**, **Nagy A**, **Farkas K**, **Banyai L**, et al. 2005. Modules, multidomain proteins and organismic complexity. *FEBS J* **272**: 5064–78.
37. **Copley RR**, **Letunic I**, **Bork P.** 2002. Genome and protein evolution in eukaryotes. *Curr Opin Chem Biol* **6**: 39–45.
38. **Ciccarelli FD**, **von Mering C**, **Suyama M**, **Harrington ED**, et al. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Res* **15**: 343–51.
39. **Forslund K**, **Henricson A**, **Hollich V**, **Sonnhammer EL.** 2008. Domain tree-based analysis of protein architecture evolution. *Mol Biol Evol* **25**: 254–64.
40. **Singh AH**, **Doerks T**, **Letunic I**, **Raes J**, et al. 2009. Discovering functional novelty in metagenomes: examples from light-mediated processes. *J Bacteriol* **191**: 32–41.
41. **Letunic I**, **Doerks T**, **Bork P.** 2009. SMART 6. Recent updates and new developments. Nucleic Acids Res 37: D229–232.
42. **Milinkovitch MC**, **Helaers R**, **Depiereux E**, **Tzika AC**, et al. 2010. 2x genomes–depth does matter. *Genome Biol* **11**: R16.
43. **Flicek P**, **Amode MR**, **Barrell D**, **Beal K**, et al. 2011. Ensembl 2011. *Nucleic Acids Res* **39**: D800–806.
44. **Thompson JD**, **Thierry JC**, **Poch O.** 2003. RASCAL. Rapid scanning and correction of multiple sequence alignments. *Bioinformatics* **19**: 1155–61.
45. **Muller J**, **Creevey CJ**, **Thompson JD**, **Arendt D**, et al. 2009. AQUA. Automated quality improvement for multiple sequence alignments. *Bioinformatics* **26**: 263–5.
46. **Huynen MA**, **Bork P.** 1998. Measuring genome evolution. *Proc Natl Acad Sci USA* **95**: 5849–56.
47. **Ostlund G**, **Schmitt T**, **Forslund K**, **Kostler T**, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**: D196–203.
48. **Deluca TF**, **Wu IH**, **Pu J**, **Monaghan T**, et al. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* **22**: 2044–6.
49. **van der Heijden RT**, **Snel B**, **van Noort V**, **Huynen MA.** 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* **8**: 83.
50. **Putnam NH**, **Srivastava M**, **Hellsten U**, **Dirks B**, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**: 86–94.

51. **Chapman JA**, **Kirkness EF**, **Simakov O**, **Hampson SE**, et al. 2010. The dynamic genome of Hydra. *Nature* **464**: 592–6.
52. **Srivastava M**, **Begovic E**, **Chapman J**, **Putnam NH**, et al. 2008. The Trichoplax genome and the nature of placozoans. *Nature* **454**: 955–60.
53. **Altschul SF**, **Gish W**, **Miller W**, **Myers EW**, et al. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–10.
54. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–7.
55. **Larkin MA**, **Blackshields G**, **Brown NP**, **Chenna R**, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–8.
56. **Castresana J.** 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–52.
57. **Eddy SR.** 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**: 205–11.
58. **Guindon S**, **Dufayard JF**, **Lefort V**, **Anisimova M**, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies assessing the performance of PhyML 3.0. *Syst Biol.* **59**: 307–21.

Table S1: Benchmark Set of Orthology Prediction

This table presents the 70 selected "homology seeds" from eggNOG, as well as the corresponding RefOG_ID after curation. The biological and technical challenges that this benchmark set tries to reveal are highligthed in bold letters. The phylogenetic range of its "homology seed" is shown (column D). Further details as the MeanID (measure the speed of evolution), norMD (quality of sequence alignment) and family size are given for uncurated and curated orthologous groups. The 35 challenging families are highlighted in grey (used Figure S1).

Category I: Low Complexity/Small Repeats

Cases	eggNOG_OG	Description	Occurance	MeanID	NorMD	Reference_OG	Description	MeanID	NorMD	RefOG_Size
1	COG2319	FOG: WD40 repeat	A, E	NA	0.362	RefOG001	Sec13	0.853	1	11
2	KOG1181	FOG: Low-complex	E	0.126	0.368	RefOG002	Retinitis pigmentc	0.318	1.412	18
3	COG0666	FOG: Ankyrin repe	B,A,E	NA	0.348	RefOG003	Ankyrin repeat an	0.328	0.72	99
4	euNOG05920	Osteoclast protein	E	0.704	1.333	RefOG004	Osteoclast proteir	0.348	0.72	13
5	COG5022	Myosin heavy chai	E	0.369	0.382	RefOG005	Dilute myosin hea	0.586	0.554	32
6	KOG0161	Myosin class II hea	E	0.456	0.39	RefOG006	Myosin heavy cha	0.697	0.859	32
7	KOG4193	G protein-coupled	E	0.21	0.468	RefOG007	GPS domain-conta	0.35	0.905	50
8	COG4886	Leucine-rich repea	B,A,E	0.223	0.415	RefOG008	Leucine-rich repea	0.652	0.569	12
9	KOG1836	Extracellular matri	E	0.205	0.926	RefOG009	Laminin alpha	0.527	0.566	18

Category II: Speed of Evolution

Cases	eggNOG_OG	Description	Occurance	MeanID	NorMD	Reference_OG	Description	MeanID	NorMD	RefOG_Size
Fast 1	NOG139072	Possible pheromor	E	0.273	0.64	RefOG010	Vomeromodulin	0.532	1.043	5
Fast 2	KOG4160	BPI/LBP/CETP fami	E	0.215	0.544	RefOG011	PLUNC proteins	0.232	0.369	60
Fast 3	KOG3614	Ca2+/Mg2+-perme	E	0.375	0.699	RefOG012	Transient recepto	0.536	1.16	41
Fast 4	KOG1052	Glutamate-gated k	E	0.256	0.433	RefOG013	Ionotropic glutam	0.213	0.829	7
Fast 5	NOG46262	Phophodiesterase	E	0.276	1.347	RefOG014	Phophodiesterase	0.482	1.554	11
Fast 6	KOG4338	Predicted lipoprot	E	0.191	0.989	RefOG015	Vitellogenin	0.373	0.84	9

Fast 7	euNOG04722	LIM domain kinase	E		0.418	1.31	RefOG016	LIM domain kinas	0.608	1.004	23
Fast 8	KOG1366	Membrane-associa	E		0.197	0.803	RefOG017	Otoferlin	0.558	1.061	26
Slow 9	KOG4764	26 proteasome co	E		0.819	1.199	RefOG018	Split hand/Split fo	0.936	1	8
Slow 10	KOG1705	Uncharacterized c	E		0.765	0.907	RefOG019	PHd Finger family	0.971	1	12

Category III: MSA Quality

Cases	eggNOG_OG	Description	Occurance	MeanID	NorMD	Reference_OG	Description	MeanID	NorMD	RefOG_Size
Low Quality 1	KOG3544	Collagen	B,E	NA	0.542	RefOG020	Collagen type IV a	0.517	0.971	49
Low Quality 2	NOG40200	Protein involved in	B,E	0.231	0.543	RefOG021	Fillagrin	0.39	0.952	6
Low Quality 3	KOG1601	GATA Zn-finger-cor	E	0.24	0.314	RefOG022	GATA 1/2/3	0.557	0.816	32
Low Quality 4	KOG1216	von Willebrand fac	E	0.126	0.483	RefOG023	Mucins	0.319	0.854	60
Low Quality 5	COG3325	Chitinase	B, A, E	0.279	0.511	RefOG024	Chitinase	0.498	0.67	45
Low Quality 6	KOG0297	TNF receptor-asso	E	0.185	0.273	RefOG025	TRAF4	0.592	1.376	13
Low Quality 7	KOG0200	Fibroblast/platelet	E	0.231	0.498	RefOG026	FGFR1/2/3/4	0.612	1.024	41
Low Quality 8	KOG0032	Ca2+/calmodulin-d	E	0.286	0.504	RefOG027	Serine/threonine-	0.557	0.864	21
High Quality 9	NOG77505	Ski oncogene prote	E	0.591	46.684	RefOG028	Ski oncogene prot	0.519	0.834	23
High Quality 10	KOG2612	Predicted integral	E	0.521	1.028	RefOG029	Ataxin-7-like prote	0.678	1.25	11
High Quality 11	COG0048	Ribosomal protein	B, A, E	0.593	0.919	RefOG030	Ribosomal protein	0.642	0.756	13

Category IV: Multigene family (size of groups) and Paralogy

Cases	eggNOG_OG	Description	Occurance	MeanID	NorMD	Reference_OG	Description	MeanID	NorMD	RefOG_Size
1	KOG2087	Glycoprotein horm	E	0.305	0.495	RefOG031	GPCR hormone re	0.565	0.794	28
2	COG0515	Serine/threonine p	B, A, E	NA	NA	RefOG032	Tyrosine-protein k	0.305	0.935	67
3	COG2124	Cytochrome P450	B, A, E	0.217	0.344	RefOG033	Cytochrome P450	0.468	0.724	27
4	COG1028	Dehydrogenases	B, A, E	NA	0.28	RefOG034	Carbonyl reductas	0.704	0.891	24
5	COG1131	ABC transporter	B, A, E	NA	0.595	RefOG035	ATP-binding casse	0.562	0.576	31
6	COG0642	Signal transductior	B, A, E	NA	NA	RefOG036	Pyruvate dehydro	0.676	0.909	41

Category V: Domain shuffling and other inconsistancies (domain evolution)

Cases	eggNOG_OG	Description	Occurance	MeanID	NorMD	Reference_OG	Description	MeanID	NorMD	RefOG_Size
1	NOG149771	Thrombospondin p	B, E	0.666	0.822	RefOG037	Thrombospondin	0.528	0.671	55
2	KOG1215	Low-density lipopr	E	0.172	0.54	RefOG038	Low-density lipop	0.595	1.897	11
3	KOG4475	FOG- Immunoglob	E	0.152	0.342	RefOG039	FOG- Immunoglol	0.541	2.031	16
4	KOG0951	RNA helicase BRR2	E	0.491	1.351	RefOG040	Subunit of activa	0.737	1.112	11

Category VII:Randomly Selected

Cases	eggNOG_OG	Description	Occurance	MeanID	NorMD	Reference_OG	Description	MeanID	NorMD	RefOG_Size
1	COG0094	Ribosomal protein	B, A, E	0.482	0.553	RefOG041	Ribosomal protein	0.79	0.972	22
2	COG0229	annotated cluster	B, A, E	NA	0.655	RefOG042	Methionine-R-sulf	0.53	0.928	18
3	COG5252	Zn-finger protein	E	0.617	1.204	RefOG043	Erythropoietin 4	0.716	1.499	15
4	COG0506	Proline dehydroge	B, A, E	0.317	0.759	RefOG044	Proline oxidase	0.501	1.209	18
5	COG2030	Acyl dehydratase	B, A, E	0.225	0.358	RefOG045	Peroxisomal multi	0.603	0.33	13
6	COG1997	Ribosomal protein	A, E	0.494	0.518	RefOG046	Ribosomal protein	0.767	1.158	24
7	COG0192	S-adenosylmethior	B, A, E	0.568	0.645	RefOG047	S-adenosylmethic	0.8	1	26
8	COG2051	Ribosomal protein	A, E	0.568	0.705	RefOG048	Ribosomal protein	0.836	1.581	22
9	COG3643	Glutamate formim	B, A, E	0.446	0.749	RefOG049	Glutamate formin	0.659	0.948	9
10	COG1454	Alcohol dehydroge	B, A, E	0.321	0.519	RefOG050	Alcohol dehydrog	0.701	0.945	12
11	KOG1712	Adenine phosphor	E	0.45	0.64	RefOG051	Adenine phospho	0.559	1.346	12
12	KOG3511	Sortilin and relatec	E	0.307	0.837	RefOG052	Neurotensin recep	0.721	0.706	10
13	KOG3781	Dystroglycan	E	0.528	0.663	RefOG053	Dystroglycan 1	0.559	1.592	11
14	NOG08919	Tyrosinase	B, E	0.328	0.654	RefOG054	Tyrosinase, TYR	0.644	1.102	11
15	NOG25268	Annotation not ava	E	0.547	2.145	RefOG055	C8orf13	NA	1.085	21
16	NOG73730	Centrosomal prote	E	0.377	0.692	RefOG056	Centrosomal prot	0.523	1.237	10

17	NOG85443	Annotation not ava	E		0.637	1.3	RefOG057	C1orf43	0.542	1.192	13
18	NOG39168	Tumor necrosis fac	E		0.442	1.017	RefOG058	Tumor necrosis fa	0.445	1.156	9
19	NOG74284	Protein involved in	E		0.554	1.313	RefOG059	C6orf170	0.636	1.062	10
20	NOG43394	Calcium channel	E		0.685	1	RefOG060	Calcium channel g	0.406	0.428	74
21	NOG42950	Annotation not ava	E		0.584	0.614	RefOG061	Thyroid hormone	0.403	1.273	19
22	COG0500	SAM-dependent m	B,A,E	NA		0.306	RefOG062	Glutathione S-trar	0.582	1.129	11
23	NOG45942	Prokineticin 1	E		0.724	1.129	RefOG063	Prokineticin 1 (PR	0.743	0.948	8
24	NOG78005	Igg binding protein	B, E		0.392	1.011	RefOG064	Serum response fa	0.386	1.077	13
25	KOG1811	Predicted Zn2+-bir	E		0.574	1.81	RefOG065	Zinc finger, FYVE d	0.537	3.399	10
26	COG3332	Annotation not ava	B,A,E		0.358	1.036	RefOG066	C22orf25	0.532	1.21	12
27	KOG4107	MP1 adaptor inter	E		0.795	1.465	RefOG067	Mitogen-activated	0.685	0.872	12
28	KOG3604	Pecanex	E		0.332	0.787	RefOG068	Pecanex	0.349	1.124	46
29	NOG38852	Corticotropin prote	E		0.65	0.782	RefOG069	Corticotropin and	0.485	0.934	15
30	COG0123	Deacetylases, inclu	B, A, E		0.292	0.662	RefOG070	Histone deacetyla	0.857	0.942	11

Table S2: Evaluation of databases using RefOGs.

Table S2 presents the actual number of errors at the gene- and group- level foreach database. RefOGs that have been predicted accurately at all levels are highligthed with dark blue, while RefOGs that have been mispredicted upon 2 genes are highlighted with grey. In red and blue bold letters are indicated the fission and fusion events, respectively. You can see that OMA splits almost all RefOGs and TreeFam orthologous groups are more inclusive (indicating by fusions). RefOGs that map to orthologous groups with many erroneously assigned genes (above 80) are highlighted in black; OrthoMCL has 6 families like this. Summarizing numbers are given at the end of the table; Figure 3 numbers are indicated by red boxes.

		Tree-Based Method				Graph-Based Methods											
		TreeFam				**eggNOG**				**OrthoDB**				**OrthoMCL**				**OMA**			
	Level	Gene		Group		Gene		Group		Gene		Group		Gene		Group		Gene		Group	
#RefOG	RefOG_Size	**Missing Genes**	**Erroneously assigned Genes**	**Fission**	**Fusion**	**Missing Genes**	**Erroneously assigned Genes**	**Fission**	**Fusion**	**Missing Genes**	**Erroneously assigned Genes**	**Fission**	**Fusion**	**Missing Genes**	**Erroneously assigned Genes**	**Fission**	**Fusion**	**Missing Genes**	**Erroneously assigned Genes**	**Fission**	**Fusion**
RefOG010	5	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	2	0	**1**	0
RefOG021	6	2	17	**1**	**2**	2	7	0	**1**	5	5	**2**	**1**	3	8	**1**	**1**	3	0	**3**	0
RefOG013	7	7	0	0	0	2	1	0	0	7	0	0	0	3	0	**2**	0	6	0	**3**	0
RefOG018	8	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	1	0	0
RefOG063	8	0	8	0	**1**	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0
RefOG049	9	0	4	0	**1**	0	0	0	0	0	0	0	0	1	4	0	**1**	0	0	0	0
RefOG015	9	0	7	0	**1**	0	7	**1**	**1**	8	17	**2**	**2**	3	7	**1**	0	6	0	**6**	0
RefOG058	9	0	34	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	2	0	**1**	0
RefOG065	10	1	1	0	0	1	3	0	0	1	1	**1**	0	1	2	0	0	3	0	**3**	0
RefOG052	10	0	28	0	**1**	0	0	0	0	1	0	0	0	1	1	0	0	3	0	**2**	0
RefOG056	10	2	0	0	0	2	0	**1**	0	1	0	0	0	4	0	**2**	0	4	0	**3**	0
RefOG059	10	1	0	0	0	0	1	0	0	0	0	0	0	1	0	**1**	0	4	0	**4**	0
RefOG070	11	5	0	**1**	0	0	0	0	0	0	0	0	0	1	21	0	**1**	0	0	0	0
RefOG062	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	**1**	0
RefOG001	11	0	2	0	0	0	2	0	0	0	0	0	0	2	2	0	0	1	2	**1**	0
RefOG040	11	0	0	0	0	1	23	**1**	**1**	0	1	0	0	2	14	0	**1**	2	0	**1**	0
RefOG038	11	0	28	0	**1**	0	0	0	0	2	0	0	0	2	1	**1**	0	2	0	**1**	0
RefOG029	11	0	0	0	0	2	0	**1**	0	0	0	0	0	1	0	0	0	4	0	**3**	0
RefOG053	11	2	0	0	0	1	4	0	**1**	0	4	0	**1**	2	3	**1**	0	3	0	**3**	0
RefOG054	11	0	22	0	**1**	0	2	0	0	1	5	**1**	**1**	2	0	0	0	4	0	**2**	0
RefOG014	11	1	10	0	**1**	1	8	0	**1**	0	1	0	0	1	0	0	0	5	0	**4**	0
RefOG050	12	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RefOG066	12	0	0	0	0	0	0	0	0	0	0	0	0	2	0	**1**	0	5	0	**3**	0
RefOG008	12	0	5	0	**1**	0	0	0	0	0	0	0	0	1	0	0	0	4	0	**2**	0
RefOG019	12	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RefOG051	12	1	0	0	0	0	1	0	0	2	0	**1**	0	3	0	**1**	0	7	0	**4**	0
RefOG067	12	1	11	0	**1**	0	0	0	0	0	0	0	0	1	0	0	0	2	0	**2**	0
RefOG030	13	2	1	0	0	2	1	0	0	2	0	0	0	3	0	0	0	4	0	**2**	0
RefOG045	13	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	**2**	0
RefOG025	13	1	54	0	**1**	0	2	0	0	1	3	0	**1**	1	2	0	0	4	0	**3**	0
RefOG004	13	1	1	0	0	2	1	0	0	3	1	0	0	3	0	0	0	2	0	0	0
RefOG064	13	2	0	0	0	4	1	**1**	0	3	1	0	0	4	1	**1**	0	5	0	**2**	0
RefOG057	13	0	1	0	0	0	3	**1**	**1**	0	1	**1**	**1**	2	0	**2**	0	5	0	**3**	0
RefOG043	15	0	0	0	0	0	1	0	0	1	0	0	0	2	0	0	0	7	0	**4**	0
RefOG069	15	0	0	0	0	8	0	**1**	0	7	1	**1**	0	7	0	**1**	0	8	0	**3**	0
RefOG039	16	2	80	0	**1**	2	16	**2**	**1**	6	4	**2**	**1**	2	73	0	**1**	8	1	**5**	0
RefOG042	18	0	10	0	**1**	9	0	**1**	0	1	1	0	0	2	0	0	0	10	0	**5**	0

RefOG044	18	1	0	0	0	1	2	0	0	12	1	2	1	4	0	0	0	12	0	4	0
RefOG002	18	2	55	0	1	8	0	1	0	11	2	2	0	8	0	1	0	11	0	5	0
RefOG009	18	0	32	0	1	1	10	1	1	3	3	1	0	3	58	1	2	11	0	4	0
RefOG061	19	1	10	0	1	6	2	1	0	10	1	2	0	11	1	3	0	12	1	6	0
RefOG027	21	1	3	0	1	1	1	0	0	1	0	0	0	11	0	1	0	14	0	7	0
RefOG055	21	1	1	0	0	5	1	1	0	6	1	1	0	10	0	1	0	13	0	5	0
RefOG041	22	0	1	0	0	0	5	0	1	4	0	0	0	8	0	0	0	11	0	2	0
RefOG048	22	2	2	0	0	2	8	0	1	7	0	1	0	6	1	0	0	14	1	4	0
RefOG016	23	0	22	0	1	0	2	0	0	10	1	1	0	2	1	0	0	14	0	6	0
RefOG028	23	0	2	0	0	11	0	1	0	9	1	1	0	11	1	1	0	16	0	6	0
RefOG034	24	0	0	0	0	0	0	0	0	5	0	1	0	5	1	0	0	16	0	7	0
RefOG046	24	0	1	0	0	0	3	0	0	1	0	0	0	11	0	1	0	13	0	7	0
RefOG047	26	0	3	0	1	0	1	0	0	3	0	0	0	3	0	0	0	17	0	10	0
RefOG017	26	1	35	0	1	8	34	3	1	16	0	3	0	2	32	0	1	18	0	9	0
RefOG033	27	1	57	0	1	14	3	1	1	18	0	3	1	16	22	4	2	19	0	8	0
RefOG031	28	1	26	0	1	0	7	0	1	10	2	2	0	4	1	0	0	21	1	8	0
RefOG035	31	3	56	2	2	9	32	2	2	14	3	2	0	3	81	0	1	24	0	13	0
RefOG005	32	1	12	0	1	3	6	1	1	20	5	4	2	1	170	1	2	23	0	12	0
RefOG022	32	0	0	0	0	9	3	1	0	22	0	4	0	21	0	5	0	23	0	7	0
RefOG036	41	0	1	0	0	0	2	0	0	18	1	2	0	11	0	1	0	32	0	8	0
RefOG006	41	0	0	0	0	3	17	1	1	27	1	3	0	0	170	1	1	32	0	15	0
RefOG026	41	0	4	0	1	7	3	2	1	28	0	5	0	5	0	0	0	33	0	14	0
RefOG012	41	2	2	0	1	0	9	0	1	24	2	3	0	5	2	0	0	33	0	11	0
RefOG024	45	1	42	1	2	9	24	2	1	21	3	6	0	19	2	5	0	39	0	25	0
RefOG068	46	11	0	2	1	16	1	5	0	32	0	6	0	17	0	4	0	38	0	13	0
RefOG020	49	1	9	0	1	2	57	0	1	35	3	5	0	8	155	0	1	42	1	16	0
RefOG007	50	10	1	1	1	33	5	6	1	40	3	6	0	40	3	7	0	41	0	13	0
RefOG037	55	2	0	0	0	22	2	1	0	42	1	4	0	8	2	0	0	47	0	13	0
RefOG023	60	9	31	4	2	18	17	3	1	51	3	10	1	15	79	1	1	52	1	31	0
RefOG011	60	24	33	2	1	52	0	8	0	49	23	10	1	54	1	9	0	53	0	14	0
RefOG032	67	10	1	2	0	40	2	3	0	52	0	5	0	44	0	5	0	58	0	14	0
RefOG060	74	6	0	0	0	53	2	5	0	62	0	9	0	65	1	12	0	66	0	24	0
RefOG003	99	41	2	1	1	86	0	11	0	87	0	12	0	88	0	14	0	90	0	27	0
Total	1638	**163**	**771**	**17**	**38**	**459**	**347**	**71**	**24**	**808**	**121**	**128**	**12**	**589**	**927**	**93**	**16**	**###**	**10**	**450**	**0**
		9				**15**				**15**				**1**				**4**			
		17				**10**				**11**				**13**				**3**			

Accuracy at the gene level	Counts	%	Counts	%	Counts	%	Counts	%	Counts	%
Accurately predicted RefOGs	11	16	15	22	15	22	1	2	4	6
Erroneously assigned genes	771		347		121		927		10	
Missing genes	163		459		808		589		###	
RefOGs affected by erroneously affected	48	64	47	67	34	49	35	50	9	13
RefOGs affected by missing genes	37	53	39	56	50	71	68	97	65	93

Accuracy at the group level	Counts	%	Counts	%	Counts	%	Counts	%	Counts	%
Accurately predicted RefOGs	33	47	32	46	31	44	31	44	7	10
Fusion	38		24		12		16		0	
Fissions	17		71		128		93		450	
RefOGs affected by fusions	34	49	22	31	12	17	13	19	0	0
RefOGs affected by fissions	10	14	30	43	37	53	32	46	63	90

Table S3: Impact of species coverage

Table S3 presents the evaluation of two different eggNOG datasets using RefOGs. Public available eggNOG (v2.0) provide orthologous groups for 34 metazoan species (called meNOGs). Based on the same genome annotations, we built orthologous groups for the 12 reference species (12-species-OGs). RefOGs that have the same performance in both datasets are highlighted with grey. Black boxes indicate larger number of fissions for the corresponding dataset.

		34-species-OG (eggNOG_v2)					12-species-OG				
			Gene Level		Group-Level			Gene-Level		Group-Level	
#RefOG	Ref Fam Size	Coverage	Missing Genes	Erroneously assigned Genes	Fission	Fusion	Coverage	Missing Genes	Erroneously assigned Genes	Fission	Fusion
RefOG001	11	11	0	2	0	0	11	0	2	0	0
RefOG002	18	10	8	0	1	0	8	10	0	2*	0
RefOG003	99	13	86	0	11*	0	13	86	0	10	0
RefOG004	13	11	2	1	0	0	11	2	1	0	0
RefOG005	32	29	3	6	1	1	21	11	6	2*	1
RefOG006	41	38	3	17	1	1	38	3	18	1	1
RefOG007	50	17	33	5	6	2	9	41	1	6	1
RefOG008	12	12	0	0	0	0	12	0	0	0	0
RefOG009	18	17	1	10	2*	2	18	0	13	1	2
RefOG010	5	5	0	2	0	0	5	0	1	0	0
RefOG011	60	8	52	0	8	0	9	51	2	8	0
RefOG012	41	41	0	9	0	1	13	28	4	4*	1
RefOG013	7	5	2	1	0	0	NA	NA	NA	NA	NA
RefOG014	11	10	1	8	0	1	10	1	12	0	1
RefOG015	9	9	0	7	1*	1	7	2	8	0	1
RefOG016	23	23	0	2	0	0	15	8	1	1*	0
RefOG017	26	18	8	34	3	1	12	14	1	5*	1
RefOG018	8	8	0	0	0	0	8	0	0	0	0
RefOG019	12	12	0	0	0	0	12	0	0	0	0
RefOG020	49	47	2	57	0	1	15	34	38	4*	1
RefOG021	6	4	2	7	0	1	4	2	3	0	0
RefOG022	32	23	9	3	1	0	13	19	2	2*	0
RefOG023	60	42	18	17	3	1	30	30	14	5*	2
RefOG024	45	36	9	24	2	1	30	15	41	3*	1
RefOG025	13	13	0	2	0	0	13	0	2	0	0
RefOG026	41	34	7	3	2*	1	35	6	7	1	1
RefOG027	21	20	1	1	0	0	20	1	1	0	0
RefOG028	23	12	11	0	1*	0	23	0	1	0	0
RefOG029	11	9	2	0	1*	0	11	0	0	0	0
RefOG030	13	11	2	1	0	0	11	2	1	0	0
RefOG031	28	28	0	7	0	1	19	9	6	1	1
RefOG032	67	27	40	2	3	0	24	43	3	4*	0
RefOG033	27	13	14	3	1	1	13	14	1	2*	1
RefOG034	24	24	0	0	0	0	24	0	0	0	0
RefOG035	31	22	9	32	2	2	12	19	3	3*	1

RefOG036	41	41	0	2	0	0	34	7	2	1*	0
RefOG037	55	33	22	2	1	0	11	44	1	4*	0
RefOG038	11	11	0	0	0	0	11	0	1	0	0
RefOG039	16	14	2	16	2	1	14	2	26	2	1
RefOG040	11	10	1	23	1*	1	11	0	1	0	0
RefOG041	22	22	0	5	0	1	22	0	5	0	1
RefOG042	18	9	9	0	1	0	9	9	0	1	0
RefOG043	15	15	0	1	0	0	15	0	1	0	0
RefOG044	18	17	1	2	0	0	9	9	1	1*	0
RefOG045	13	13	0	0	0	0	13	0	1	0	0
RefOG046	24	24	0	3	0	0	24	0	3	0	0
RefOG047	26	26	0	1	0	0	26	0	1	0	0
RefOG048	22	20	2	8	0	1	17	5	7	1*	1
RefOG049	9	9	0	0	0	0	9	0	0	0	0
RefOG050	12	12	0	0	0	0	12	0	0	0	0
RefOG051	12	12	0	1	0	0	12	0	1	0	0
RefOG052	10	10	0	0	0	0	10	0	0	0	0
RefOG053	11	10	1	4	0	1	10	1	4	0	1
RefOG054	11	11	0	2	0	0	11	0	7	0	1
RefOG055	21	16	5	1	1	0	16	5	1	1	0
RefOG056	10	8	2	0	1*	0	9	1	1	0	0
RefOG057	13	13	0	3	1*	1	12	1	3	0	0
RefOG058	9	9	0	0	0	0	9	0	0	0	0
RefOG059	10	10	0	1	0	0	10	0	2	0	0
RefOG060	74	21	53	2	5	0	11	63	2	8*	0
RefOG061	19	13	6	2	1	0	13	6	2	1	0
RefOG062	11	11	0	0	0	0	11	0	0	0	0
RefOG063	8	8	0	0	0	0	8	0	0	0	0
RefOG064	13	9	4	1	1*	0	9	4	1	0	0
RefOG065	10	9	1	3	0	0	9	1	3	0	0
RefOG066	12	12	0	0	0	0	12	0	0	0	0
RefOG067	12	12	0	0	0	0	12	0	0	0	0
RefOG068	46	30	16	1	5	0	18	28	0	5	0
RefOG069	15	7	8	0	1	0	7	8	0	1	0
RefOG070	11	11	0	0	0	0	11	0	0	0	0
Total	1638	**1181**	**457**	**347**	**71**	**25**	**986**	**645**	**270**	**91**	**22**

Supplementary Information

Quality assessment of orthology prediction methods using curated protein families

Kalliopi Trachana, Tomas Larsson, Sean Powell, Wei-Hua Chen, Tobias Doerks, Jean Muller, Peer Bork

Material and Methods

- Selection of reference families
- Building the Reference Orthologous Groups (RefOGs)
- Mapping of RefOGs to the five databases

Figure S1: The effect of the 35 challenging families on the performance of the methods

Figure S2: The impact of biological complexity in orthology assignment at the group-level (fusions/fissions).

Figure S3: The effect of repeated domains on orthology assignment

Figure S4: The quality of MSA as a proxy for accurate orthology prediction

Table S1: Presentation of the 70 families that consist this benchmark set (separate xls file)

Table S2: Evaluation of the 5 databases using RefOGs (separate xls file)

Table S3: Effect of the species coverage (separate xls file)

Table S4: Effect of the genome annotation (separate xls file)

Table S5: Correlation and p-values of error distribution and error sources

Materials and Methods

Selection of reference families

70 Orthologous Groups (OGs) were selected from the second version of the eggNOG database (*Muller et al, 2010*), the majority of which were originally build for the COG database (*Tatusov et al, 1997*), to form a benchmark set of orthology prediction. Previous studies have reported certain biological (e.g. multi-gene families) or technical aspects (e.g. quality of MSA) that cause problems in assignment of orthologous groups (*Tatusov et al, 1997; Koonin EV, 2005; Gabaldon T, 2008).* 40 out of the 70 selected families were classified under a specific category of biological or technical challenge, while 30 OGs were selected randomly (Table S1). The trouble-making categories that are covered by our benchmark set are the following:

1. Multiple Sequence Alignment Quality: To select OGs with different alignment quality, we built multiple sequence alignments (MSA) of every OGs either at the universal (similar to COG) or eukaryotic-specific (similar to KOG) level on eggNOG database. The MSA were computed using the AQUA protocol (*Muller et al, 2009*) setup to use MUSCLE (v3.7) (*Edgar RC, 2004*) and RASCAL (v1.34) (*Thompson et al 2003*). AQUA makes use of the NORMD program (*Thompson et al. 2001*) to assess the quality of each individual MSA by comparing norMD scores and selecting the one with the highest score. The norMD score gives information about the general quality of the alignment, a norMD >0.6 indicates a reliable MSA, (*Thompson et al., 2003*). Looking at the distribution of the norMD score for all OGs, one can observe first, that the OGs dataset does contain the full spectrum from fast to slow evolving gene families and second, that the vast majority of the gene families have a reliably aligned MSA (i.e. norMD>0.6). 10 families were selected under this category; 8 of them represent families with low quality MSA (norMD<0.6), while 3 of them score a high quality MSA (norMD>2).

2. Speed of evolution: The multiple sequence alignments were also used to define the speed of evolution. To classify eggNOG OGs based on their evolutionary pace, we computed the mean percent identity for each of them. The mean percent identity (described as the "FamID" in *Muller et al, 2006*) is calculated as the mean pairwise percent identity of each sequence against each other within a given MSA. Positions in the alignment corresponding to gaps within the MSA were excluded from the calculation.

$$FamID = 2\frac{\sum_{1 \le i < j \le n} ID_{S_i,S_j}}{n(n-1)}$$

where:

n = total number of sequence tested, S_i and S_j are the *i*th and *j*th sequence,

$IDs_{i,}s_j$ = pairwise percent identity between the *i*th and *j*th sequence, excluding gap regions.

Only 2 of the 10 selected families for this category are slow evolving families, while the rest eight are characterized as fast-evolving families (MeanID<0.45) (Table S1).

3. Low complexity/repeats: Intrinsic features like low complexity, coiled coil and other variable repeated elements or repeated modules can affect the building of orthologous groups. 10 families were taken from this category.

4. Domain complexity/Domain shuffling: The complexity of the domain architecture of different protein families can lead to miss-assignment of orthologs. The vast majority of proteins consist of single or a few (2-3) domains; however, we collected 4 OGs that have been previously reported to hamper orthology prediction either due to the complex architecture within the protein (contain more than 4 domains) or due to the high variety of domain architecture among the members of an OG.

5. Multigene families: Of all above problem, the most-acknowledged one that affects all three domains of life is the multi-gene families and the detection of paralogy. 6 large OGs that contain several paralogs and orthologs were chosen to address this issue.

Building the Reference Orthologous Groups (RefOGs)

Starting with COG/KOGs (Table S1) as "homology seeds" we manually recovered orthologous groups for 12 bilaterian species that are referred to as Reference Orthologous Groups (RefOGs). Initially, we mapped the "homology seed" identifiers (Ensembl v46) to Ensembl v60 via Ensembl History. BLAST (*Altschul et al, 1990*) searches were performed in the 12 reference genomes and four outgroup species (*Monosiga brevicollis*, *Trichoplax adherens*, *Nematostella vectensis* and *Hydra magnipapillata*) using query sequences from well-annotated genomes (e.g. human, zebrafish and fly). The homologous sequences were aligned by MUSCLE (*Edgar, 2004*) and the alignments were used to build NJ trees with Clustal X (Larkin et al, 2007). Large groups were resolved by the presence of ortholog(s) in the outgroup(s) (Figure 2). However, in several cases no clear outgroup was found hampering the resolution on the bilaterian level. For these families, RefOGs were defined based on i) the domain content using SMART database (*Letunic et al., 2009*), ii) manual inspection of the alignments and iii) previous published descriptions of the families. After the initial curation of the families, all sequences determined to be members of the bilaterian RefOGs were aligned using MUSCLE (*Edgar, 2004*). Alignments were manually cut based on the first and last well aligned columns according to GBLOCKS (*Castresana J, 2000*) with the following parameters: (Minimum Length Of A Block: 10, Allowed Gap Positions: With Half, Use Similarity Matrices: Yes). This was made in order to remove highly divergent N- and C-terminal parts of each alignment where misalignment is assumed to be common. Alignments were further manually cleaned to remove large parts where all sequences but one had gaps or short sequences that did not align within a conserved "block". Based on the refined alignments, Hidden Markov Models (HMM) were built using the HMMER3 package (*Eddy SR, 2009*). At a second refinement step, the HMM models were used to identify related sequences that were left out from the 16 aforementioned genome. We did not define a global cut-off for sequence recruitment instead we treat each family uniquely by adding sequences with bit score within the range of bitscores of already known members. After the addition of those sequences phylogenetic trees were calculated using PhyML version 3.0 (*Guindon et al, 2010*) with the following settings: 100 bootstrap replicates, optimization of tree topology, branch lengths and rate parameters, 4 substitution rate categories and the NNI topology search option. RefOG identifiers, alignments, HMM models and trees are available on www.eggnog.embl.de/orthobench.

Mapping of RefOGs to the five databases

Five orthology prediction methods were benchmarked against the RefOGs: TreeFam (release 7.0), eggNOG (v2.0), orthoDB (customized orthologous groups for the 12 reference species), orthoMCL (v4.0) and OMA (release 3.0). Generally, we downloaded and benchmarked the latest version of each database (October 2010). TreeFam resolve the evolutionary relationships of big homologous families through tree reconciliation, thus we had to score each RefOG against the reconciled tree of the respective homologous family on the bilaterian level. TreeFam provides both curated and automatically predicted orthologous groups, we used the second category for our analysis. eggNOG generates OGs for different taxonomic levels, thus, in the current comparison we used OGs generated by bilaterian species only (called meNOGs). eggNOG and OrthoDB use a similar clustering procedure based on triangulars of best hits; OrthoMCL identifies OGs using Markov clustering and OMA applies its unique algorithm, which does not allow paralogs within OGs.

The RefOGs are built using the genome annotations of Ensembl_v60. However, all five repositories predicted OGs based on older Ensembl versions. For each RefOG sequence we track its identifiers to older Ensembl versions via Ensembl History (i.e. ENSTNIG00000002616 (annotated as RPL11 Ensembl_v60) mapped to GSTENG00003639001 in to Ensembl_v46). There are certain cases, where this automated procedure doesn't work, i.e. one protein of Ensembl_v60 maps to multiple identifiers in older Ensembl releases or genome assemblies predict a new gene locus (e.g. ENSDARP00000103772 (prok1) - a predicted locus after Ensembl v54- is identified as a missing ortholog in eggNOG, orthoMCL and TreeFam databases that use older releases of Ensembl).

References

1. **Tatusov RL, Koonin EV, Lipman DJ.** 1997, A genomic perspective on protein families. *Science* **278:** 631–637.

2. **Ruan J, Li H, Chen Z, Coghlan A,** *et al.* 2008. TreeFam. 2008 Update. *Nucleic Acids Res.* **36**(Database issue): D735-40.

3. **Muller J, Szklarczyk D, Julien P, Letunic I,** et al. 2010. eggNOG v2.0. extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research* **38**(Database issue): D190-195.

4. **Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV.** 2011. OrthoDB. the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.* **39**(Database issue): D283-288.

5. **Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS**. 2006. OrthoMCL-DB. querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**(Database issue): D363-368.

6. **Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C**. 2011. OMA 2011. orthology inference among 1000 complete genomes. *Nucleic Acids Res.* **39**(Database issue): D289-294.

7. **Muller J, Oma Y, Vallar L, Friederich E**, *et al.* 2005. Sequence and Comparative Genomic Analysis of Actin-related Proteins. *Molecular Biology Cell* **16**(12): 5736-48.

8. **Letunic I, Doerks T, Bork P**. 2009. SMART 6. recent updates and new developments. *Nucleic Acids Res.* **37**(Database issue): D229-232.

9. **Thompson JD, Plewniak F, Ripp R, Thierry JC**, et al. 2001. Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.* **314**: 937-951.

10. **Thompson JD, Thierry JC, Poch O.** 2003. RASCAL. rapid scanning and correction of multiple sequence alignments. *Bioinformatics* **19**.1155-1161.

11. **Muller J, Creevey CJ, Thompson JD, Arendt D**, *et al.* 2009. AQUA. Automated quality improvement for multiple sequence alignments. *Bioinformatics* **26**(2): 263-5.

12. **Flicek P, Amode MR, Barrell D, Beal K**, *et al.* 2011. Ensembl 2011. *Nucleic Acids Res.* **39**(Database issue): D800-806.

13. **Edgar RC.** 2004. MUSCLE. multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**(5): 1792-7

14. **Sorek R, Zhu Y, Creevey CJ, Francino MP,** *et al.* 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science.* **318**(5855): 1449-52

15. **Garcia-Vallve S, Guzman E, Montero MA, Romeu A**. 2003. HGT-DB. a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Research* **31**: *187-189.*

16. **Altschul SF, Gish W, Miller W, Myers EW,** et al. 1990. Basic local alignment search tool. *J Mol Biol.* **215**(3): 403-10.

17. **Larkin MA, Blackshields G, Brown NP, Chenna R,** *et al.* 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21): 2947-8.

18. **Castresana J**. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**: 540-552.

19. **Eddy SR**. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**(1): 205-11.

20. **Guindon S, Dufayard JF, Lefort V, Anisimova M**, *et al.* 2010. New algorithms and methods to estimate maximum-likelihood phylogenies. assessing the performance of PhyML 3.0. *Syst Biol*. **59**(3): 307-21.

Figure Legends

Figure S1: A benchmark set that highlights the challenges of orthology assignment. Using the manually curated RefOGs, we evaluated the performance of five databases. Among the 70 families, there are 35 families that illustrate challenges of orthology prediction. Green bar illustrates the performance of the databases for those 35 families, while the purple bar shows the performance of the databases for all the benchmark set. The upper panels illustrate the accurately predicted RefOGs at two different levels (gene- and group-level). The low panels show the % effected RefOGs by 4 different errors: erroneously assigned genes - missing genes (left) and fusion - fissions (right). The larger green bars on the lower panels illustrate the higher number of errors that accumulate the 35 complicated families.

Figure S2: The impact of biological complexity in orthology assignment at the group-level (fusions/fissions). *(A) The impact of family size (paralogy);* The RefOGs were separated into (i) small (contain less than 14 members), (ii) medium (contain more than 14 members, but less than 40) and (iii) large (contain more than 40 genes). For the graph-based methods (eggNOG, OrthoDB, OrthoMCL and OMA), we observe that they split larger RefOGs into more orthologous groups than the smaller ones. *(B) Speed of evolution;* The RefOGs were classified based on the MeanID score (described as the "FamID" in *Muller et al, 2006*), an evolutionary rate score derived from the multiple sequence alignment of each family. There are (i) slow-evolving (MeanID>0.7), (ii) medium-evolving (MeanID lower than 0.7, but larger than 0.5) and (iii) fast-evolving (MeanID<0.5) RefOGs. Similarly to the previous observation, as biological complexity increases (slow to fast-evolving families), we count more fission events for the graph-based methods. *(C) Domain architecture complexity;* each RefOG is associated with the average number of domains, which is equal to the sum of predicted domains of the members of one RefOG divided by the family size. Again, there are 3 levels of complexity, starting from (i) none or 1 domain on average to (ii) 2-4 to (iii) more than 4. By classifying RefOGs based on their domain complexity we can see a more diverse pattern; TreeFam seems to have a large number of fusion events on the most difficult category, while OrthoMCL seems to have a uniform distribution of fissions across all three categories. Significant correlations (Table S5) between the distribution of missing/erroneously assigned genes and the tested factor is indicated with an asteric.

Figure S3: Repeated domains affect the orthology assignment. We used SMART database to identify the number of domains for each protein of our benchmark dataset. 24 out of the 70 RefOGs contain proteins with repeated domains (Table S5). We observed that the percentage of RefOGs that failed to be predicted accurately is higher for these 24 families than the rest indicating that repeated domains have an impact in orthology assignment.

Figure S4: The quality of MSA as a proxy for accurate orthology prediction. (A) We classified the families based on their norMD score (*Thompson et al, 2001*) into (i) high quality alignment (norMD>0.6) and (ii) low quality alignment. We observed that all graph-based methods tend to have more fissions when the alignment quality is low. For TreeFam, on the other hand, low quality of alignment was correlated with fusions. (B) Effect of sequence length variation; the RefOGs were divided into three different categories (low, medium and high deviation) based on the sequence length variability of included orthologs. We can see that RefOGs with variable-size members accumulate the higher fraction of fusion and fission events. Significant correlations (Table S5) between the distribution of missing/erroneously assigned genes and the tested factor is indicated with bold letters. Significant correlations (Table S5) between the distribution of missing/erroneously assigned genes and the tested factor is indicated with an asteric.

Figure S1.

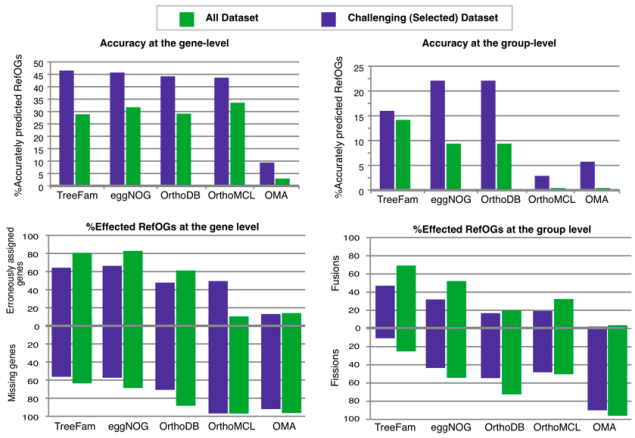


Figure S2.

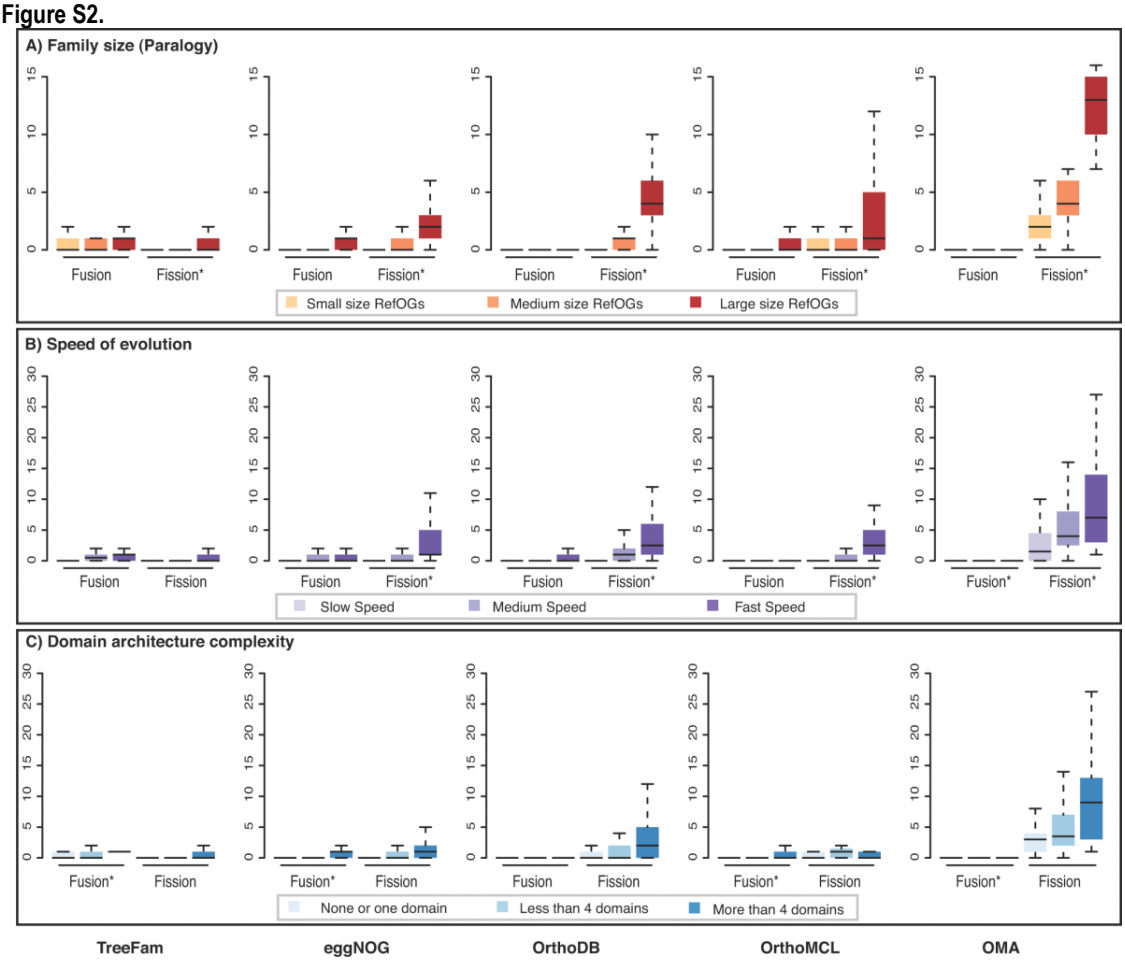


Figure S3.

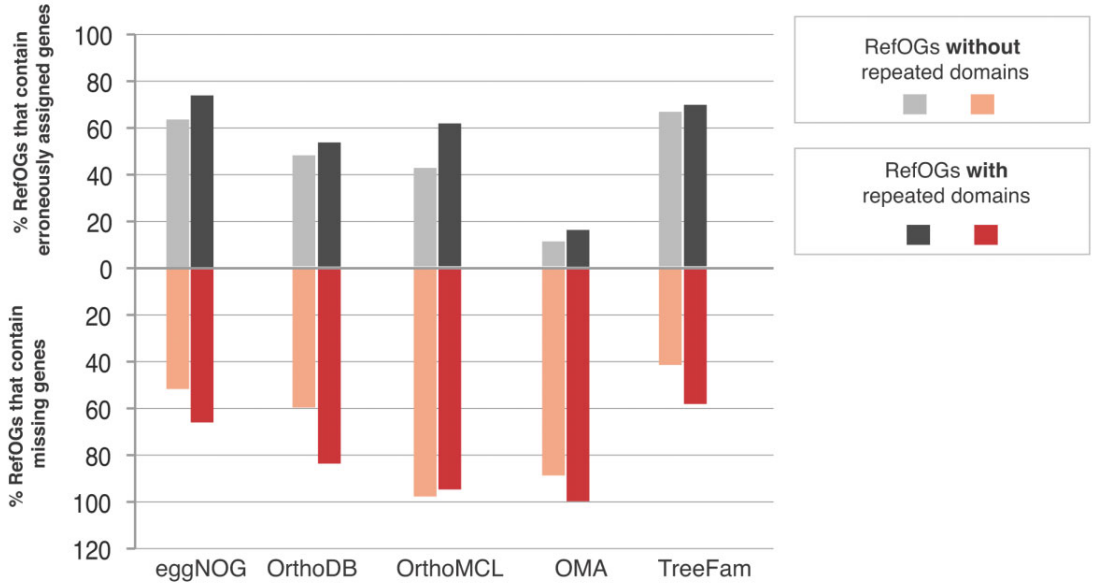
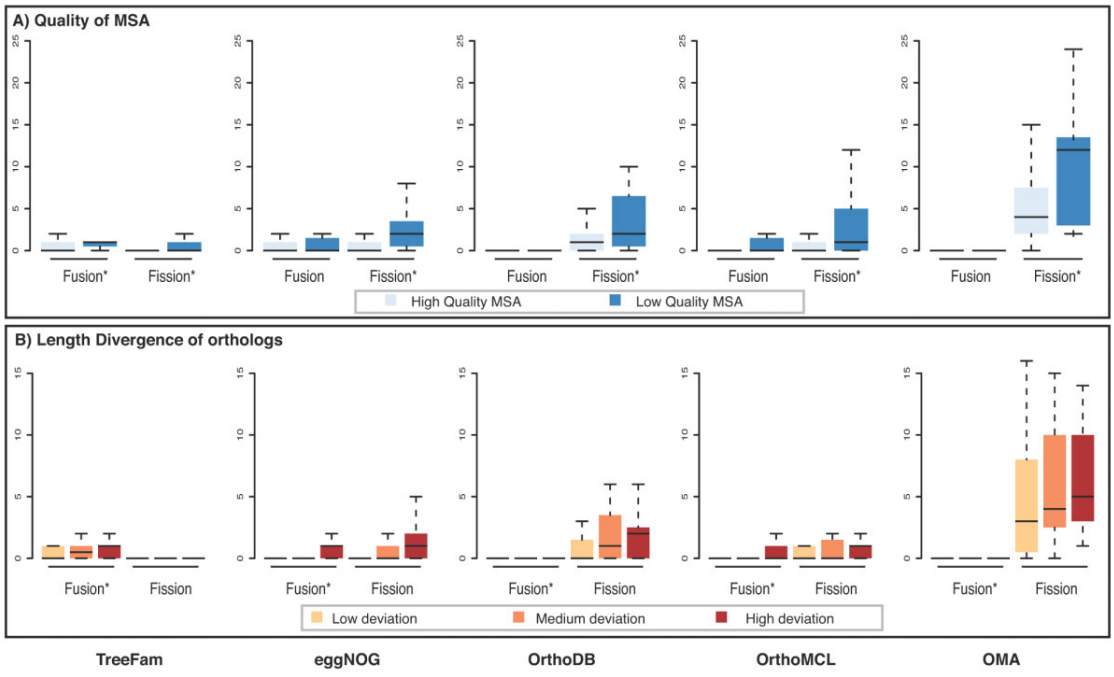


Figure S4.



Appendix B

Evolution and regulation of cellular periodic processes: a role for paralogues.

Trachana K, Jensen LJ, Bork P. *EMBO Rep.* (2010) 11(3):233-8.

EMBO reports

EMBO open

scientific report

Evolution and regulation of cellular periodic processes: a role for paralogues

Kalliopi Trachana[1], *Lars Juhl Jensen*[1,2+] & *Peer Bork*[1,3++]

[1]European Molecular Biology Laboratory Heidelberg, Heidelberg, Germany, [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, Copenhagen, Denmark, and [3]Max-Delbrueck-Center for Molecular Medicine, Berlin-Buch, Berlin, Germany

Several cyclic processes take place within a single organism. For example, the cell cycle is coordinated with the 24 h diurnal rhythm in animals and plants, and with the 40 min ultradian rhythm in budding yeast. To examine the evolution of periodic gene expression during these processes, we performed the first systematic comparison in three organisms (*Homo sapiens*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*) by using public microarray data. We observed that although diurnal-regulated and ultradian-regulated genes are not generally cell-cycle-regulated, they tend to have cell-cycle-regulated paralogues. Thus, diverged temporal expression of paralogues seems to facilitate cellular orchestration under different periodic stimuli. Lineage-specific functional repertoires of periodic-associated paralogues imply that this mode of regulation might have evolved independently in several organisms.

Keywords: cell cycle; diurnal rhythm; ultradian rhythm; metabolism; sub/neo-functionalization

EMBO *reports* (2010) **11**, 233–238. doi:10.1038/embor.2010.9

INTRODUCTION

Gene duplication is a major evolutionary force (Ohno, 1970) facilitating development of morphological novelties (Bassham *et al*, 2008), adaptation to new environments (Hittinger & Caroll, 2007) and speciation (Scanell *et al*, 2006). Small scale (SSD) and whole genome (WGD) duplications provide the raw genetic material on which mutation and selection act to evolve new functionalities. Many duplicated genes have a short lifespan, as one of the two copies is either lost or degenerates and becomes non-functional (non-functionalization). In the relatively rare case in which both copies are retained in the genome, one copy can diverge and acquire a new function that is completely different from the ancestral one (neo-functionalization), or the two duplicated genes partition the ancestral function (sub-functionalization; Force *et al*, 1999). The last two scenarios can be achieved through changes in amino acid sequence (Merritt & Quattro, 2002) or through changes in gene expression patterns (Bassham *et al*, 2008). Sub-functionalization and neo-functionalization allow spatio-temporal specialization and expansion of functionality, respectively, but it is usually difficult to draw a line between the two fates (He & Zhang, 2005). There are several studies of spatial sub/neo-functionalization revealing the function of paralogues as either tissue-specific (for example, in zebrafish, *pax6a* and *pax6b* paralogues are expressed in different tissues and both of them fulfil the functional role of mammalian pax6; Kleinjan *et al*, 2008) or even within a single cell compartment (for example, the paralogues of the mammalian COX7A family are expressed in either the mitochondrion or the Golgi; Schmidt *et al*, 2003). Although a few studies observe distinct expression profiles of duplicates during the developmental time scale (Bassham *et al*, 2008; Kleinjan *et al*, 2008), the role of duplicated genes in the temporal organization of the cell remains unclear (Gu *et al*, 2002; Wagner, 2002).

The temporal organization of a biological system—be that a single cell or an entire organism—is at least as intricate as its spatial organization. Clocks, rhythms and cycles are universal from unicellular to multicellular organisms and coordinate many intertwined biological pathways that respond to extracellular or intracellular signals, adapting the organism to periodically changing environments. The 24 h diurnal rhythm (circadian clock) controls many biological responses in animals and plants (Harmer *et al*, 2000; Panda *et al*, 2002). Similarly, in *Saccharomyces cerevisiae*, the cell cycle is coordinated with the ultradian rhythm, which is a robust 40 min (approximately) metabolic cycle that persists indefinitely when cultures are supplemented continuously with glucose (Klevecz *et al*, 2004). During this metabolic cycle,

[1]EMBL Heidelberg, Meyerhofstrasse 1, Heidelberg 69117, Germany
[2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, Blegdamsvej 3b, Copenhagen N 2200, Denmark
[3]Max-Delbrueck-Center for Molecular Medicine, Berlin-Buch, Robert-Rossle-Strasse 10, Berlin 13092, Germany
[+]Corresponding author. Tel: +45 35 325 025; Fax: +45 35 325 001;
E-mail: lars.juhl.jensen@cpr.ku.dk
[++]Corresponding author. Tel: +49 6221 387 526; Fax: +49 6221 387 517;
E-mail: bork@embl.de

Received 16 July 2009; revised 19 November 2009; accepted 17 December 2009; Published online 19 February 2010

transcription is organized into redox-state superclusters; for example, genes that are involved in DNA replication are transcribed in the reductive phase, suggesting a mechanism for reducing oxidative damage to DNA during replication. Apart from the ultradian rhythm, a 4–5 h yeast metabolic cycle that takes place under glucose-limited conditions in budding yeast has also been reported (Tu *et al*, 2005). However, our analysis is focused on the ultradian rhythm, as yeast metabolic cycle-synchronized culture is also synchronized inherently with the cell cycle (Rowicka *et al*, 2007), making it impossible to separate the two processes.

Here, we present the first systematic comparison of the genes that are transcribed periodically during the cell cycle, the diurnal rhythm and the ultradian rhythm. We observe that diurnal- and ultradian-regulated genes are more likely to have cell-cycle-regulated paralogues than would be expected by random chance. As the respective functional repertoires of these duplicated genes in yeast, plants and animals are different, we conclude that gene duplication and subsequent sub/neo-functionalization took place independently during evolution. This suggests that orchestration of cellular pathways under different periodic processes provides a selective advantage, and that use of temporal regulation of newly emerging paralogues in different contexts—that is, distinct cyclic processes—seems to be an efficient way in which to achieve this.

RESULTS AND DISCUSSION

Identification of periodically regulated genes

Recently, there have been numerous efforts aimed at capturing the temporal profiles of various periodic cellular processes. Time-course microarray experiments have provided much data on the global transcriptome of the cell cycle, and on diurnal and ultradian rhythms in plants, mammals and yeast (supplementary Table S1 online). We have previously identified 600, 400 and 600 cell-cycle-regulated genes in budding yeast, *Arabidopsis* and humans, respectively (Jensen *et al*, 2006). To maximize the comparability between data sets, we reanalysed the microarray experiments and identified diurnal-regulated and ultradian-regulated genes by using the same algorithm as the aforementioned cell cycle study (see Methods).

The identification of diurnal-regulated genes is particularly complicated, as there is a high biological variance that should be taken into account. The genes that have been identified as diurnal in different tissues overlap only in part (Delaunay & Laudet, 2002), indicating a tissue-specific regulation of diurnal genes that depends on the physiology of the tissue (Harmer *et al*, 2000; Panda *et al*, 2002). Unfortunately, it is not only biological variability that has to be considered. Only about 90 common genes (out of hundreds) were identified to cycle diurnally in the liver in two separate microarray experiments (Delaunay & Laudet, 2002), pointing to problems associated with microarray reproducibility. To eliminate the biological variance, we decided to average over many different tissues and experiments (supplementary Table S1 online). Benchmarks of the resulting lists against experimentally verified diurnal genes show that we obtained the best list by combining all available expression data across studies and tissues (supplementary Fig S2 online). We produced three further lists consisting of 600 ultradian-regulated budding yeast genes, 600 diurnal-regulated *Arabidopsis* genes and 600 diurnal-regulated mouse genes.

Cell cycle and diurnal rhythm regulation of paralogues

Comparison of the *Arabidopsis* regulated genes under diurnal rhythm and the cell cycle reveals that only seven genes (supplementary Table S3 online) are expressed periodically in both processes, which is no more than what would be expected by chance alone. However, mapping the genes to a set of eukaryotic paralogous groups (see Methods) reveals that 18 diurnal-regulated genes belong to paralogous groups with cell-cycle-regulated members, which corresponds to 3.4 times more genes ($P<10^{-5}$; Fisher's exact test) than obtained by random expectation, after taking into account the total number of genes, the number of periodic genes and the number of paralogues of periodic genes (supplementary information online). Similarly, 26 cell-cycle-regulated genes have diurnal-rhythm-regulated paralogues (3.8-fold enrichment; $P<10^{-8}$; Fisher's exact test; supplementary Table S4 online). The diurnal-regulated genes and the cell-cycle-regulated genes tend to be paralogues of each other (Fig 1A).

We observed the same trend when comparing the diurnal rhythm and cell cycle in humans. The cell cycle and diurnal rhythm analyses were based on human and mouse data, respectively. Assuming that at least one of the two processes is comparable between human and mouse, which should be the case for the cell cycle, we mapped diurnal-regulated genes to their 491 one-to-one orthologues in human and mouse (supplementary information online). Indeed, 15 paralogue pairs have been detected that consist of cell-cycle and diurnal-rhythm-regulated genes (2.5-fold more than that by random expectation; $P<4\times10^{-4}$; Fisher's exact test; supplementary Table S4 online). We thus get a statistically significant result despite there being interspecies differences due to the rapid evolution of transcriptional regulation in mice and humans (Odom *et al*, 2007) and intraspecies differences between tissues, both of which weaken the signal. Besides the paralogous pairs, 22 genes are regulated during both the cell cycle and the diurnal rhythm in humans (supplementary Table S3 online). As was the case for *Arabidopsis*, this is not significantly more than that expected by chance (Fig 1B).

Currently, we cannot distinguish between sub-functionalization and neo-functionalization as we can posit two different scenarios: (i) a gene is regulated periodically in the phylogenetically older cell cycle and after duplication its functional properties can be extended to another cyclic process (neo-functionalization) and (ii) a gene is regulated periodically under two periodic processes and its duplication enables two specialized temporal regulation profiles (sub-functionalization). In either case, we propose that there was only one ancestral response to extrinsic and intrinsic periodic signals. After gene/genome duplication, the ancestral response could be expanded or could become specialized in time.

Parallel evolution in *Arabidopsis* and human

When analysing the paralogue groups that are regulated in more than one cycle, we observed that their functional repertoires are different in *Arabidopsis* (supplementary Table S5 online) and humans (supplementary Table S6 online) and are in accordance with their specialized biology. To exemplify this, we focus on two pairs of paralogues and how their temporal regulation is related to plant and animal physiology, respectively.

In *Arabidopsis*, for example, we find periodic regulation of alpha-amylases during diurnal rhythm (AMY3) and cell cycle (AMY1) that does not occur in humans. Starch is prepared in

chroloplasts during day-time photosynthesis and is degraded during the night, providing sugars for leaf metabolism and exporting them to other organs such as seeds and the root (Smith *et al*, 2005). The diurnal regulation of AMY3 accompanies the diurnal (day/night) regulation of starch metabolism in leaves. The enzyme is targeted to the chloroplasts and participates in transitory starch degradation, although its exact role remains unclear (Zeeman *et al*, 2007). Cell-cycle-regulated AMY1, however, contributes to seed germination (Borisjuk *et al*, 2004). During germination, the cell is in rest in the G1 phase. Gibberellin is necessary to enter the S phase and complete cell division (Ogawa *et al*, 2003). Concurrently, gibberellin-induced alpha-amylase (AMY1) promotes degradation and mobilization of the starch accumulated in endosperm to fuel cell division (Fincher, 1989).

Among the 16 diverged regulated paralogues in humans, we identified, for instance, cell-cycle-regulated and diurnal-regulated ribonucleotide (nucleoside 5′-triphosphate; NTP) reductase subunits named RRM2 and RRM2B, respectively. These enzymes exemplify differential temporal regulation of isoenzymes. NTP reductase in mammals catalyses the reduction of ribonucleotides to deoxyribonucleotides, the balanced supply of which is essential for both accurate DNA replication and repair. NTP reductase consists of two non-identical subunits (R1 and R2), and its enzymatic activity is regulated by R2 expression—that is, by RRM2 or RRM2B. RRM2 peaks during the S phase and is blocked during G1 phase, pointing to a mechanism protecting the cell against unscheduled DNA synthesis (Chabes *et al*, 2003). However, RRM2B (the diurnal-regulated gene) is hardly expressed in proliferating cells. Recently, its role in DNA repair and mitochondrial DNA synthesis has been elucidated (Bourdon *et al*, 2007) in non-proliferating cells. Both the above-mentioned processes take place independently of the cell cycle, it has been reported that mitochondrial DNA synthesis cycles in the rat liver (Dallman *et al*, 1974).

As the functional repertoires of paralogues that have been sub/neo-functionalized under the regulation of the cell cycle and the diurnal rhythm in *Arabidopsis* and humans are different, the most parsimonious scenario is that this mode of regulation has evolved independently in both organisms.

Periodic regulation of metabolism in yeast

The cell cycle in budding yeast is orchestrated with the ultradian rhythm. Recent studies have shown that the latter gates cells into the S phase of the cell cycle, organizes the energetic (redox) status of the cell, and coordinates mitochondrial and metabolic functions (Klevecz *et al*, 2004). Basic redox molecules such as NAD(P)H and glutathione are under the temporal control of the cell cycle and ultradian rhythm (Lloyd & Murray, 2007). The cellular redox balance is also vital for organization of the cell cycle and the diurnal rhythm (Matés *et al*, 2008; Lepisto *et al*, 2009). Similar to the diurnal/cell cycle results presented above, we can identify 58 paralogues that have diverged their regulation under the cell cycle and the ultradian rhythm (twofold enrichment, $P<10^{-5}$; Fisher's exact test; supplementary Table S4 online). Besides paralogue pairs with divergent regulation, there are 64 genes that are expressed periodically during both the cell cycle and the ultradian rhythm (1.25-fold enrichment, $P<0.02$; Fig 2).

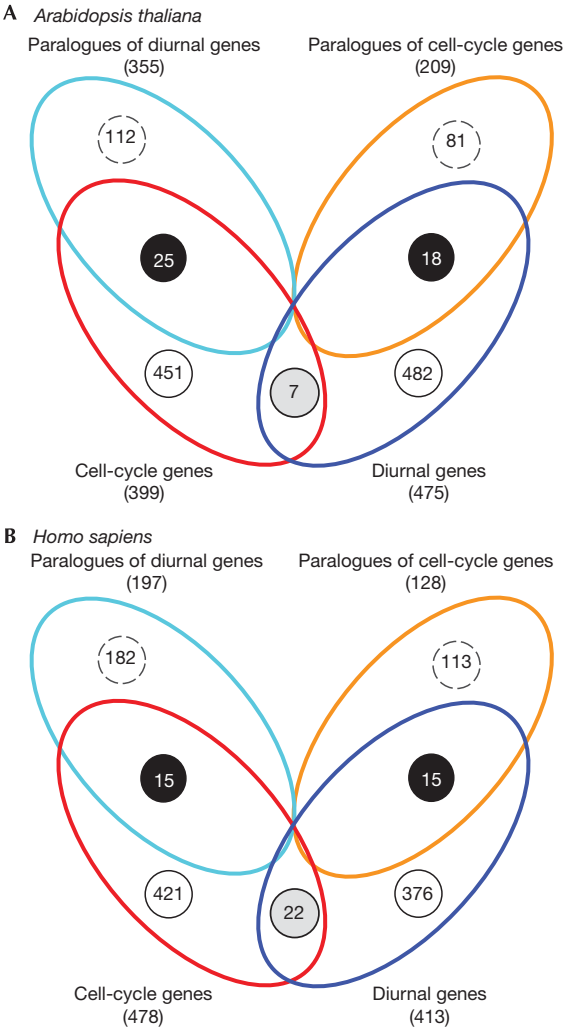


Fig 1 | Four-way Venn diagrams of cell-cycle-regulated genes, diurnal-regulated genes and their paralogues in *Arabidopsis* and humans. (**A**) There are 26 cell-cycle-regulated genes with ultradian-regulated paralogues and 18 ultradian-regulated genes with cell cycle paralogues (supplementary Table S5 online). (**B**) There are 15 cell-cycle-regulated genes with diurnal-regulated paralogues and 15 diurnal-regulated genes with cell cycle paralogues (supplementary Table S6 online). The number of cell-cycle- or diurnal-regulated proteins that do not have diurnal- or cell-cycle-regulated paralogues, respectively, are indicated in white circles. Within the dashed-line white circles are proteins of paralogous groups with cell-cycle- or diurnal-regulated members that do not cycle themselves. The number of genes that are regulated in both cycles is indicated in the grey circles (supplementary Table S3 online). Both in *Arabidopsis* and humans, these genes are not significantly over-represented in our periodic lists. The numbers of diurnal-regulated genes with cell-cycle-regulated paralogues and vice versa are highlighted in black circles.

Many recent studies have shown that the mode of duplication —that is, SSD compared with WGD—has an important role in the functional divergence of paralogues (Maere *et al*, 2005). *S. cerevisiae* is a degenerated tetraploid resulting from WGD after

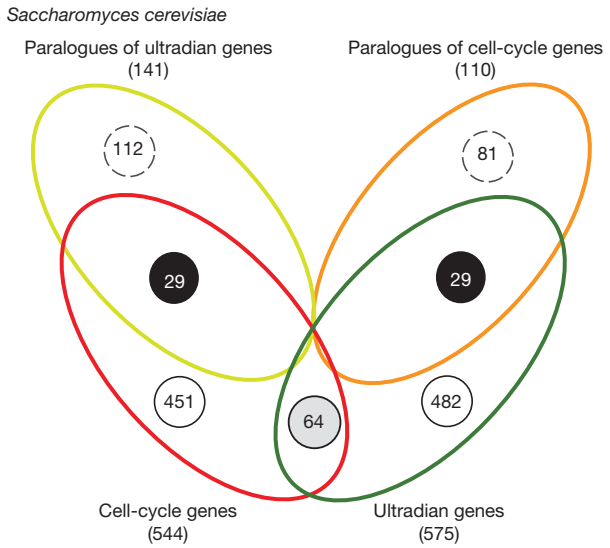


Fig 2 | Four-way Venn diagram of cell-cycle-regulated, ultradian-regulated genes and their paralogues in budding yeast. The number of genes that are regulated by both cycles is indicated in the grey circle. The numbers of ultradian-regulated genes with cell-cycle-regulated paralogues and vice versa are highlighted in the black circles (supplementary Table S7 online). There is an over-representation of cell-cycle-regulated genes with ultradian-regulated paralogues, and vice versa. The number of cell-cycle- or ultradian-regulated proteins that do not have ultradian- or cell-cycle-regulated paralogues, respectively, are indicated in white circles. Within the dashed-line white circles are proteins of paralogous groups with cell-cycle- or ultradian-regulated members that do not cycle themselves.

the divergence of *Saccharomyces* from *Kluyveromyces,* followed by extensive gene loss (Wolfe & Shields, 1997). We decided to explore the origin of cell cycle/ultradian paralogue pairs and ask whether there is a preferred mode of duplication for temporal sub/neo-functionalization. Of a total of 416 paralogous groups (supplementary Table S10 online) in yeast that we identified (see Methods), we subtracted 651 WGD paralogues identified by Byrne & Wolfe (2005), leaving 248 SSD paralogues. For both WGD and SSD paralogues, we observed a significant over-representation of cell-cycle/ultradian-regulated paralogues (supplementary Table S9 online). Although WGD contributes the highest number (31 periodically expressed paralogues), SSD cell-cycle/ultradian-regulated paralogues (a total of 27 paralogues) are more enriched compared with random expectation. Although a more detailed analysis is needed, this implies a stronger selection on SSDs. In any case, both modes of duplication contributed to a lineage-specific functional repertoire of periodic divergent paralogues.

Functional analysis of cell cycle/ultradian paralogue pairs and their mapping to the metabolic network of *S. cerevisiae* (Fig 3) revealed that cell cycle/ultradian sub/neo-functionalization has occurred frequently in paralogues that regulate important metabolic substrates, such as glucose, pyruvate and sulphate. For example, glucose is transported by the major facilitator super-family of transporters (HXT), a few of which compose a cell-cycle/ultradian-regulated paralogous group. *S. cerevisiae* grows in a variety of glucose concentrations because of the presence of several *HXT* genes, which show glucose transport with dual kinetics (high-glucose and low-glucose affinity) and change their expression levels in response to culture conditions (Verwaal *et al*, 2002). Yeast proliferates fast in a glucose-rich environment, wherein low-affinity transporters are expressed (for example, the cell-cycle-regulated *HXT2* gene), but the cell cycle slows down markedly after glucose depletion, upon which high-affinity transporters are induced (for example, the cell-cycle-regulated *HXT7* gene; Ozcan & Johnston, 1999). Trehalose and glycogen—reserve carbohydrates—have been reported to accumulate under low growth rate conditions. Interestingly, they have a dual role: their degradation maintains the ATP flux in *S. cerevisiae* when glucose deteriorates, but they can also fuel the cell to enter the S phase of the cell cycle when culture conditions improve (Silljé *et al*, 1999). The ultradian-rhythm-regulated *HXT5* gene is not affected by glucose concentration in the environment, similarly to *HXT2* or *HXT7*, but it is expressed highly during low growth rate (Verwaal *et al*, 2002). It has been suggested that *HXT5* regulates the uptake of glucose for production of trehalose, which is in accordance with its ultradian role in balancing the redox (ATP) status and helping the cell enter the S phase. In contrast to the cell-cycle-regulated *HXT2* and *HXT7* genes, which sense their glucose-sufficient environment and drive the culture to cell cycle—which is an energy-demanding process—the ultradian-regulated *HXT5* gene senses the glucose-insufficient environment and stores energy, indicating that temporal sub/neo-functionalization accompanies functional divergence.

Conclusion

Here, we report, for the first time, that diverged temporal regulation under the cell cycle and diurnal or ultradian rhythm of newly emerged paralogues seems to be an efficient way in which to orchestrate cellular response to extrinsic and intrinsic signals. This temporal sub/neo-functionalization of paralogues under the cell cycle and diurnal/ultradian rhythm occurs more frequently than expected by chance and spans different lineages (*Arabidopsis, Homo sapiens* and *S. cerevisiae*). As the functional repertoires of duplicated genes in the three organisms studied are different, it seems that the temporal sub/neo-functionalization of duplicated genes has evolved independently in plants, animals and yeasts to distinguish cell-cycle regulation from other periodic processes, perhaps even to coordinate those processes. Further analysis of the functional repertoires of cell-cycle/ultradian-regulated paralogues in yeast indicates that they have arisen through both WGD and SSD and that in the yeast linage are enriched in metabolic functions. Thus, we could show that a large-scale (meta) analysis of duplications in several species reveals details on the evolution of cellular periodicity and provides the first insight into temporal sub/neo-functionalization at the cellular level.

METHODS

Analysis of microarray expression data and benchmarking. To enable a comparison of cell-cycle-regulated genes that have been identified previously (Jensen *et al*, 2006), we reanalysed all microarray expression time courses (supplementary Table S1 online) using the same permutation-based algorithm (de Lichtenberg *et al*, 2005). The resulting lists of periodic transcripts during

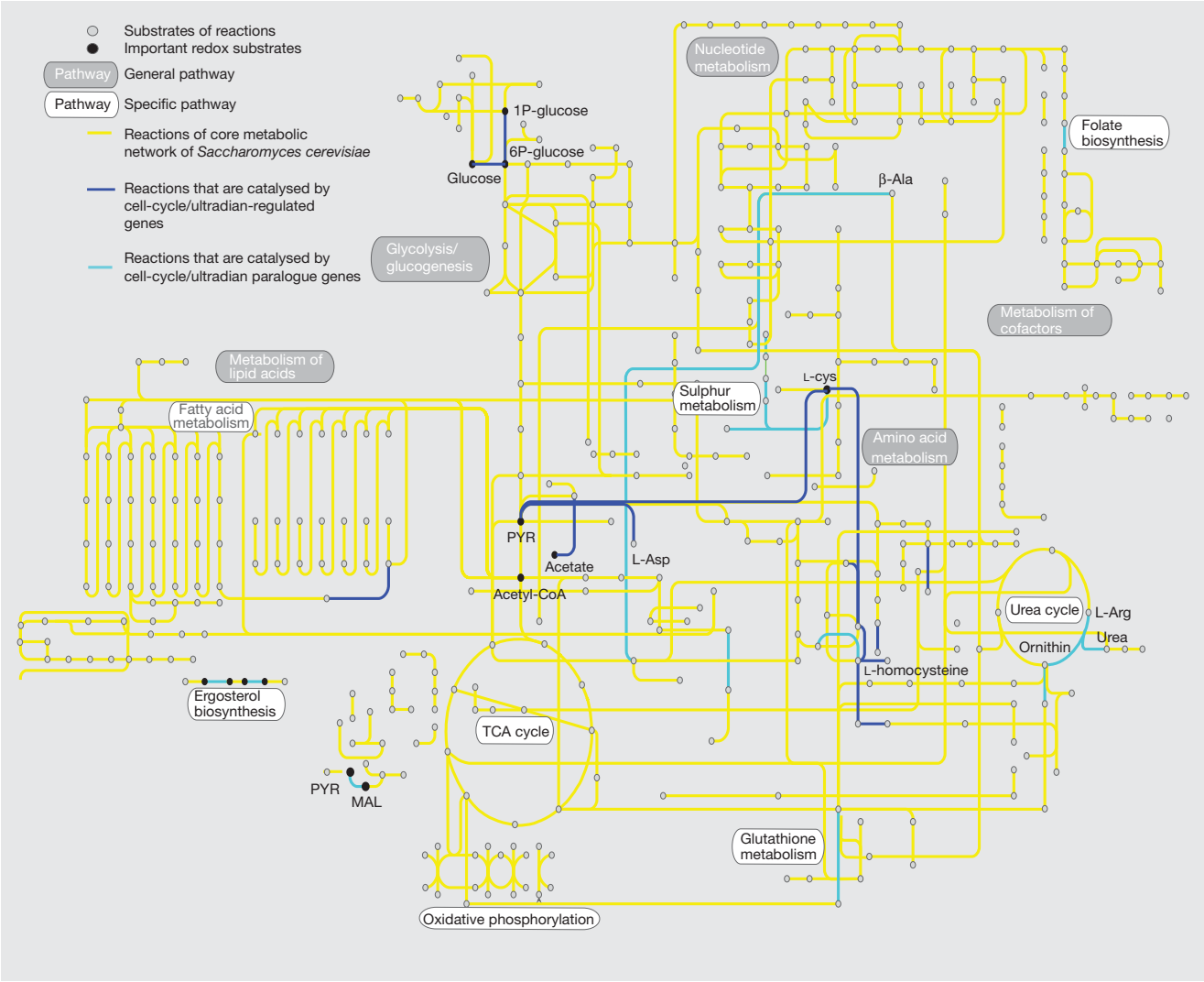


Fig 3 | Core metabolic network of cell-cycle-regulated and ultradian-rhythm-regulated genes in *Saccharomyces cerevisiae*. The core metabolic network of *S. cerevisiae* is shown in yellow. Reactions that are catalysed by cell-cycle/ultradian-regulated genes are highlighted with light blue and cell-cycle/ultradian-regulated paralogues are mapped with dark blue lines. Metabolic substrates that are under cell cycle and ultradian regulation are marked by black circles. It seems that there is a tight regulation of cell cycle and ultradian metabolism judging by the number of common regulated genes/products and the number of paralogues. A few of the common cell-cycle-regulated and ultradian-regulated substrates, such as glucose-6-phosphate and sulphate, are important for glycolysis and redox equilibrium, respectively, in budding yeast. The custom metabolic map shown here was generated by using iPath (Letunic *et al*, 2008).

diurnal rhythm were benchmarked against lists of known diurnal-regulated genes compiled from review articles and The Arabidopsis Information Resource database (supplementary information online). We kept the top 600 diurnal-regulated genes both in *Arabidopsis* and mouse, as these lists capture 75–90% of the known diurnal and cell-cycle-regulated genes (supplementary Fig S2 online). In order to compare the cell cycle and diurnal rhythm genes in humans, we used 13,648 pairs of 1:1 human to mouse orthologues (Hubbard *et al*, 2007) to transfer the mouse list to human diurnal genes.

Identification of eukaryotic orthologous/paralogous groups. Human, *Arabidopsis* and budding yeast proteins were categorized into orthologous groups by an automatic procedure (von Mering *et al*, 2005) similar to the original cluster of orthologous groups procedure (Tatusov *et al*, 2003); all-against-all Smith–Waterman similarities were computed and orthology was then assigned through reciprocal best matches and subsequent triangular linkage clustering (von Mering *et al*, 2005). To perform within-species comparison, we focused on paralogous groups resulting from the first step of orthology assignment (supplementary information online), after which 10,947 proteins were clustered in 3,761 paralogous groups (supplementary Table S2 online).

Supplementary information is available at *EMBO reports* online (http://www.emboreports.org).

ACKNOWLEDGEMENTS
We thank the members of the Bork group for helpful discussions. K.T. is grateful to Takuji Yamada and Ivica Letunic for help with iPath exploration. K.T. is supported by the European Union FP6 Program Contract number LSH-2004-1.1.5-3. The work carried out in this study was supported in part by the Novo Nordisk Foundation Center for Protein Research.

CONFLICT OF INTEREST
The authors declare that they have no conflict of interest.

REFERENCES
Bassham S, Canestro C, Postlethwait JH (2008) Evolution of developmental roles of *Pax2/5/8* paralogs after independent duplication in urochordate and vertebrate lineages. *BMC Biol* **6:** 35–41
Borisjuk L, Rolletschek H, Radchuk R, Weschke W, Wobus U, Weber H (2004) Seed development and differentiation: a role for metabolic regulation. *Plant Biol* **6:** 375–386
Bourdon A *et al* (2007) Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. *Nat Genet* **39:** 776–780
Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15:** 1456–1461
Chabes AL, Pfleger CM, Kirschner MW, Thelander L (2003) Mouse ribonucleotide reductase R2 protein: a new target for anaphase-promoting complex-Cdh1-mediated proteolysis. *Proc Natl Acad Sci USA* **100:** 3925–3939
Dallman PR, Spirito RA, Silmes MA (1974) Diurnal patterns of DNA synthesis in the rat: modification by diet and feeding schedule. *J Nutr* **104:** 1234–1241
de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, Brunak S (2005) Comparison of computational methods for the identication of cell cycle-regulated genes. *Bioinformatics* **21:** 1164–1171
Delaunay F, Laudet V (2002) Circadian clock and microarrays: mammalian genome gets rhythm. *Trends Genet* **18:** 595–597
Fincher GB (1989) Molecular and cellular biology associated with endosperm mobilization in germinating cereal grains. *Annu Rev Plant Physiol Plant Mol Biol* **40:** 305–346
Force A, Lynch M, Pickett FB, Amores A, Yi-lin Y, Postlethwait JH (1999) Preservation of duplicated genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545
Gu Z, Nicolae D, Lu HH, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **12:** 609–613
Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000) Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* **290:** 2110–2113
He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169:** 1157–1164
Hittinger CT, Caroll SB (2007) Gene duplication and the adaptive evolution of a classic switch. *Nature* **440:** 677–681
Hubbard TJP *et al* (2007) Ensembl 2007. *Nucleic Acids Res* **35:** D610–D617
Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443:** 594–597
Kleinjan DA *et al* (2008) Subfunctionalization of duplicated zebrafish *pax6* genes by *cis*-regulatory divergence. *Plos Genet* **4:** e29
Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc Natl Acad Sci USA* **101:** 1200–1205
Lepisto A, Kangasjarvi S, Luomala EM, Brader G, Sipari N, Keranen M, Keinanen M, Rintamaki E (2009) Chloroplast NADPH thioredoxin reductase interacts with photoperiodic development in *Arabidopsis thaliana*. *Plant Physiol* **149:** 1261–1276
Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci* **33:** 101–103
Lloyd D, Murray DB (2007) Redox rhythmicity: clocks at the core of temporal coherence. *BioEssays* **29:** 465–473
Maere S, de Bodt S, Raes J, Casneuf T, van Montagu M, Kuiper M, van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* **102:** 5454–5459
Matés JM, Segura JA, Alonso FJ, Márquez J (2008) Intracellular redox status and oxidative stress: implications for cell proliferation, apoptosis, and carcinogenesis. *Arch Toxicol* **82:** 273–299
Merritt TJS, Quattro JM (2002) Negative charge correlates with neural expression in vertebrate aldolase isozymes. *J Mol Evol* **55:** 674–683
Odom DT *et al* (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39:** 730–732
Ogawa M, Hanada A, Yamauchi Y, Kuwahara A, Kamiya Y, Yamaguchi S (2003) Gibberellin biosynthesis and response during *Arabidopsis* seed germination. *Plant Cell* **15:** 1591–1604
Ohno S (1970) *Evolution by Gene Duplication*. New York, USA: Springer
Ozcan S, Johnston M (1999) Function and regulation of yeast hexose transporters. *Microbiol Mol Biol Rev* **63:** 554–569
Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesc JB (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* **109:** 307–320
Rowicka M, Kudlicki A, Tu BP, Otwinowski Z (2007) High-resolution timing of cell cycle-regulated gene expression. *Proc Natl Acad Sci USA* **104:** 16892–16897
Scanell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440:** 341–345
Schmidt TR, Doan JW, Goodman M, Grossman LI (2003) Retention of a duplicate gene through changes in subcellular targeting: an electron transport protein homologue localizes to the golgi. *J Mol Evol* **57:** 222–228
Silljé HH, Paalman JW, ter Schure EG, Olsthoorn SQ, Verkleij AJ, Boonstra J, Verrips CT (1999) Function of trehalose and glycogen in cell cycle progression and cell viability in *Saccharomyces cerevisiae*. *J Bacteriol* **181:** 396–400
Smith AM, Zeeman SC, Smith SM (2005) Starch degradation. *Annu Rev Plant Biol* **56:** 73–98
Tatusov RL *et al* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41
Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* **310:** 1152–1158
Verwaal R, Paalman JW, Hogenkamp A, Verkleij AJ, Verrips CT, Boonstra J (2002) HXT5 expression is determined by growth rates in *Saccharomyces cerevisiae*. *Yeast* **19:** 1029–1038
von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33:** D433–D437
Wagner A (2002) Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol* **10:** 1760–1768
Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713
Zeeman SC, Smith SM, Smith AM (2007) The diurnal metabolism of leaf starch. *Biochem J* **401:** 13–28

Supplementary Information for:

Evolution and regulation of cellular periodic processes: A role for paralogs

Kalliopi Trachana [1], Lars Juhl Jensen [1,2*] & Peer Bork [1,3 *]

1.EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

2.Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, Blegdamsvej 3b, 2200 Copenhagen N, Denmark

3. Max-Delbrueck-Center for Molecular Medicine, Berlin-Buch, Germany

∗ Corresponding authors:

Lars Juhl Jensen; Tel. + +45 30 506 479; E-mail: lars.juhl.jensen@cpr.ku.dk

Peer Bork; Tel. +49 6221 387 526; E-mail: bork@embl.de

Supplementary methods:

Microarray time-series analysis

In this study, we analyze microarray expression time courses from 3 different organisms. The whole procedure is presented in Figure S1 and the datasets are summarized in Table S1. The identification of diurnal regulated genes is particularly complicated, as there is high biological variance that should be taken into account. In order to eliminate variance that is related to tissue-specific diurnal regulation and microarray reproducibility (Delaunay & Laudet, 2002), we decided to analyze multiple datasets (multiple tissues) (Table S1) and compare the predicted gene lists (benchmarking procedure). More specifically, to identify diurnal regulated genes, for Arabidopsis and mouse, we used five different datasets of microarray time-series (Oster et al, 2006; Stroch et al, 2002; Miller et al, 2007; Edwards et al, 2006; Blasing et al, 2005). To identify ultradian regulated genes in budding yeast, we analyzed the dataset by Klevecz et al. (2004), which describes ultradian rhythm (a ~40 min cycle), while expression data also exists for the related ~4-5h yeast metabolic cycle (YMC), which takes place under glucose-limited conditions (Tu & McKnight, 2006). However, we excluded this dataset because the a YMC-synchronized culture of budding yeast is inherently also synchronized with respect to the cell cycle (Rowicka et al, 2007), making the distinction between cell-cycle- and YMC-regulated genes unclear.

We analyzed the above datasets using the same methodology as for cell cycle study (Jensen et al., 2006), in order to enable the comparison between cell cycle and diurnal/ultradian rhythm and to ensure that any observed differences are due to biological factors. The method that is based on a permutation algorithm has been illustrated elsewhere (de Lichtenberg, 2005), as well as its performance in cell cycle analysis (Jensen et al., 2006 - Supplementary Info). This method was applied to microarray data for the three organisms and we ran a benchmarking procedure for refining the lists of periodic genes.

Benchmarking of diurnal regulated genes

For quality assessment and control of diurnal regulated genes, we benchmark each list of periodic genes to well-known diurnal regulated genes. There are 25 Arabidopsis and just 9 mouse genes that have been reported to cycle during diurnal rhythm in small-scale analyses. Figure S2 shows the fraction of each benchmark identified as periodic by our analysis. For *A.thaliana*, we selected the top-600 cyclic genes based on where the curves break and show no further enrichment over random expectation (Figure 2S).

Because of the very small benchmark set for mouse, we selected top-600 as the cutoff for mouse diurnal-regulated genes based not only on the benchmark curve, but also by the number of genes that are common among the three different datasets (Oster et al, 2006; Storch et al, 2002 and Miller et al, 2007 dataset) (Figure S2). The top-600 genes encompass 75% and 100% of the benchmarking gene sets of Arabidopsis and mouse, respectively. Similar benchmark plots for cell cycle genes have been published elsewhere (Jensen et al., 2006).

Detection of orthologous/paralogous groups

Orthology assignment between genes in the three organisms is important for any inter- and intra-species comparison. We built our orthologous groups using an automatic procedure similar to the original COG procedure (Tatusov et al, 2003). The pipeline has already been published by our group (von Mering et al, 2005). Briefly, we grouped recently duplicated sequences within genomes into 'in-paralogous groups' to be treated as single sequences subsequently. There was no fixed cutoff in similarity, but instead we started with a stringent similarity cutoff and relaxed it step-wise, until all in-paralogs were joined, satisfying the following criteria: all members of a group have to be more similar to each other than to any other protein in any other species and all members of the group have to give hits that overlap by at least 20 residues. This procedure resulted in 1765, 1326, and 471 paralogous groups for Arabidopsis, human and budding yeast, respectively. The distributions of the size of paralogous groups for each organism are presented in Table S2.

Statistical analysis

To test the enrichment of cell cycle/diurnal (or cell cycle/ultradian) regulated paralogs and the statistical significance of it, we used Fisher's exact test. For each contingency table we calculated: 1) the total number of genes that have been tested by cycle-A (e.g. cell cycle) and cycle-B (e.g. diurnal rhythm) microarray experiments, 2) the number of cycle-A regulated genes that belong to first gene category, 3) the number of paralogs of cycle-B regulated genes that belong to first gene category, 4) the number of cycle-A regulated genes that belong to the paralogs of cycle-B regulated genes and vice versa. For each test, the relevant numbers are presented in Table S4.

As is obvious (Table S4), the statistical support of temporal sub-/neo-functionalization becomes much stronger as biological variation is reduced in our dataset; the enrichment of cell cycle/ultradian

regulated paralogs in *S.cerevisiae* is bigger than the enrichment of cell cycle/diurnal regulated paralogs in *A.thaliana* –where we compared different tissues in same organism- which, in turn, is bigger than the enrichment of cell cycle/diurnal regulated paralogs in human –where we compared different tissues from different organisms. The small number of cell cycle/diurnal regulated paralog pairs both in Arabidopsis and human is related to suboptimal datasets. In the case of Arabidopsis, we have a suboptimal cell cycle gene list (Jensen et al, 2006- Supplementary Info); the resulting list captures only 50% of experimentaly verified genes in *A.thaliana*. Given that we only have 50% and 75% sensitivity (benchmarking procedure) for the Arabidopsis cell cycle and diurnal gene list, respectively, we gained overall 37,5% (75*50) of the expected paralog pairs. Thus, 18/(0.75*0.50)=48 is the expected number if the dataset was optimal, pointing to 30 extra pairs of cell cycle/diurnal regulated paralogs that we could not identify. Despite the undentified cell cycle/diurnal regulated paralogs, we obtained high statistical significance based on the 18 predicted paralog pairs (P-value<10^{-5}, 3.4-folds).

Unfortunately, the number of cell cycle/diurnal regulated paralogs, as well as the statistical support, are lower for the human comparison. We compared datasets of cell cycle and diurnal rhythm from human and mouse, respectively, assuming that at least one of two processes, the cell cycle, is comparable between human and mouse -there is no reason to believe that cell cycle regulation is fundamentally different between the two animals. We mapped mouse diurnal-regulated genes to their 1:1 human orthologs and we ended up with 491 human diurnal regulated genes. Of course, we could have done the opposite transfer as well, meaning to map the human cell cycle regulated genes into murine cell cycle regulated genes through 1:1 orthology (Methods). We preferred the first comparison for our convenience, since we had assigned the orthology between human, Arabidopsis and budding yeast in an older study from our laboratory (Methods). Apart from the diurnal regulated genes, we had also to use 1:1 orthology to transfer the mouse genes that were tested commonly in microarray experiments; this is important for identifying the total number of genes that have been tested by cell cycle and diurnal rhythm microarray experiments, which in its turn influence the statistical result. Apart from the 1:1 orthology assignment –estimated at 90% between the two species- that reduce the signal of cell cycle/diurnal sub-/neo-functionalization, there is an additional biological factor: the rapid transcriptional evolution between the two species (Odom et al, 2007). The conservation of regulation of transcription between mouse and human depends on tissue and transcription factors (Odom et al, 2007); however, we can estimate on average that 80% of transcription

regulatory sites are conserved between mouse and human. In the end, we find only 15 cell cycle/diurnal paralogous pairs, but given that we only have 90% sensitivity for the diurnal list, 80% sensitivity for the cell cycle gene list, 90% of genes have 1:1 orthology between mouse and human, and 80% of these are expected to have conserved regulation (Odom et al. 2007), we can state that our best estimate is that there are in reality about twice as many such pairs as we identify. The expected number of genes under optimal conditions is: 15/(0.9*0.8*0.9*0.8)=29.

Family size distribution

Finally, we checked the size of gene families that the cell cycle/ultradian regulated paralog pairs belong and we compared their distribution to this of yeast gene families that contain at least one periodic gene in their members (Table S10, figure S4). The cell cycle, ultradian, and total distribution are very similar, but the distribution of gene families that have both cell cycle and ultradian members is slightly skewed towards larger numbers, which is not surprising, since a larger family is obviously more likely to contain both types of regulated genes. There are a few paralogous groups where one member is regulated by one periodic process (e.g. cell cycle) and multiple members are regulated by the other process (e.g. ultradian rhythm). Those families are highlighted in red in Table S7 (the table that presents the cell cycle/ultradian paralogs). Of 58 cycling genes belonging to paralogous groups containing both cell-cycle-regulated and ultradian-regulated genes, only 8 genes stem from paralogous groups of 5 or more genes. We thus conclude that large gene families are not responsible for the observed signal.

Measurement of the contribution of WGD and SSD in cell cycle/ultradian paralogs regulation

It has been reported that paralogs arose by WGD or SSD have different properties and participate in different functions of the cell (Maere et al, 2005). Due to this, we decided to test the contribution of each category to cell cycle/ultradian regulation and if there is any biases of our results. We used as WGD paralogs the ones have been assigned by Byrne et al (2005), while every other gene that exists in our 'in-paralog' groups was considered as SSD paralog. Both categories of paralogs (WGD and SSD) contribute significantly to periodic sub-/neo-functionalization (Table S9); the p-values (Fisher's test) and the enrichment for each test are presented in Table S9. The large number of WGD paralogs indicates that whole genome duplication is an important mechanism for the emergence of cell cycle/ultradian divergent

regulation. In addition, paralogs that belong in our cell cycle/ultradian regulated list are statistically significantly over-represented in the SSD genes. However, due to the small pool of SSD genes one cannot conclude that one mechanism is more important for the cell cycle/ultradian sub-/neo-functionalization.

Functional analysis of cell cycle/ultradian regulated paralogs

To test if the observed signal is not an artifact of certain functional classes of genes being preferentially duplicated, we analyzed the respective gene sets for overrepresented Gene Ontology terms (Supplementary table S8). Specifically, we made the following comparisons: 1) Calculate enriched terms of cell-cycle-regulated genes relative to all genes; 2) Calculate enriched terms of ultradian-regulated genes relative to all genes; 3) Calculate enriched terms of all duplicated genes relative to all genes; 4) Calculate enriched terms of SSD genes relative to all genes; 5) Calculate enriched terms of WGD genes relative to all genes. No GO terms were found to be overrepresented among duplicated, cell-cycle-regulated, and ultradian-regulated genes. The observed signal is thus not a consequence of functional biases. We used the web-tool FATIGO (Al-Shahrour et al, 2005) that annotates the queries list to GO categories, applies Fisher's exact test for each GO term and adjusts p-value via FDR correction.

References:

Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L. & Dopazo, J. (2005), Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments, Nucleic Acids Research, 33, W460-W464

Blasing, O. E. et al. (2005) Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. *Plant Cel l* **17**, 3257–3281

Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**: 1456-1461

Delaunay F, Laudet V. (2002) Circadian clock and microarrays: mammalian genome gets rhythm. *Trends Genet*. **18**(12):595-7

de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, Brunak S (2005) Comparison of computational methods for the identification of cell cycle regulated genes. *Bioinformatics* **21**: 1164–1171

Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000)

Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* **290**: 2110-2113

Hubbard TJP *et al* (2007) Ensembl 2007. *Nucleic Acids Res* **35**: D610–D617

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and posttranslational cell cycle regulation. *Nature* **443**: 594–597

Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. *PNAS*: 1200-1205

Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci*. **33**:101-10

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes *PNAS* **102**:5454-9

Miller BH, McDearmon EL, Panda S, Hayes KR, Zhang J, Andrews JL, Antoch MP, Walker JR, Esser KA, Hogenesch JB, Takahashi JS (2007) Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation. *PNAS* **104**: 3342-7

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*. **39**(6):730-2

Oster, H. et al. (2006) The circadian rhythm of glucocorticoids is regulated by a gating mechanism residing in the adrenal cortical clock. *Cell Metab*. **4**, 163–173

Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesc JB (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* **109**: 307–320

Rowicka M, Kudlicki A, Tu BP, Otwinowski Z (2007) High-resolution timing of cell cycle-regulated gene expression. *PNAS* **104**:16892-16897

Storch, K.-F. et al. (2002) Extensive and divergent circadian gene expression in liver and heart. *Nature* **417**, 78–83

Tu BP, Mohler RE, Liu JC, Dombek KM, Young ET, Synovec RE, McKnight SL, (2007) Cyclic changes in metabolic state during the life of a yeast cell. *PNAS* **104**: 16886-16891

Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41 (2003)

Supplementary figures:

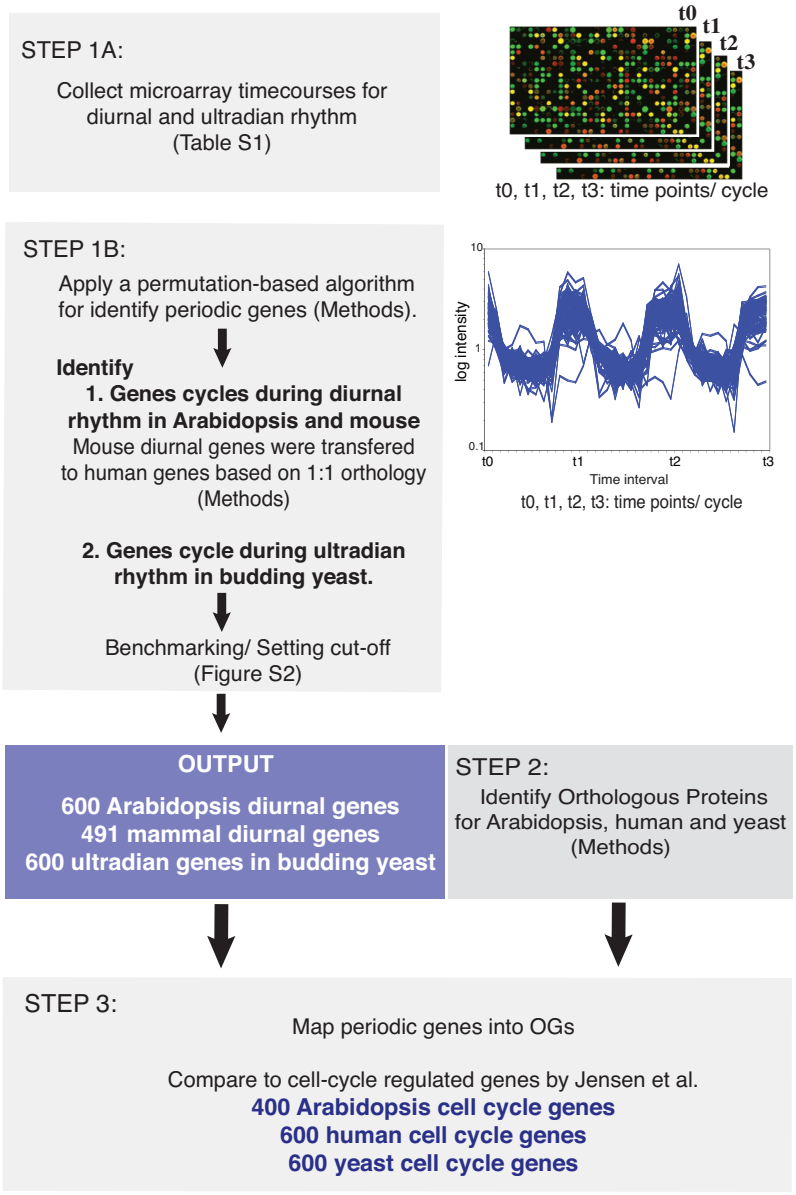


Figure S1: Flowchart of analysis. We compare cellular periodic processes in different organism across the tree of life and try to understand the principles that underlie their evolution. As previous microarray analyses have already revealed cell-cycle-regulated genes in Arabidopsis, human and budding yeast (Jensen et al, 2006), we focused on other periodic processes in the same organisms. Our method is summarized in 3 steps: 1) retrieval and analysis of microarray time courses for diurnal and ultradian rhythm, 2) identification of Arabidopsis, human and budding yeast orthologous groups (OGs) and 3) mapping of diurnal/ultradian regulated genes to OGs and comparison with regulated genes of cell cycle and diurnal/ultradian rhythm.

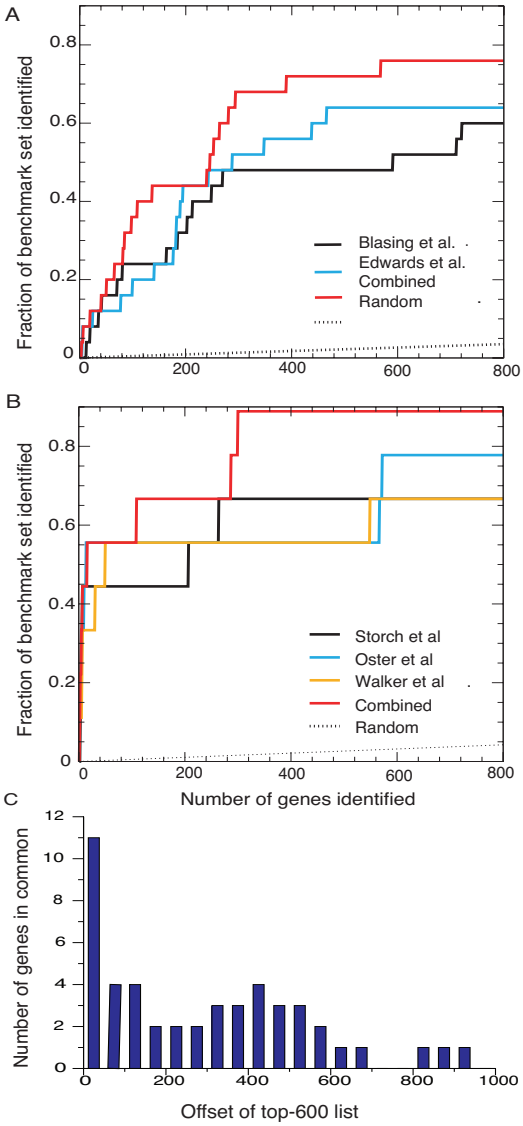


Figure S2: Benchmarking of the *A.thaliana* (A) and *M.musculus* (B, C) diurnal regulated genes. For each organism, we compare our lists of diurnal regulated genes to a species-specific benchmark set. Combined lists that contain both genes from different microarray time courses perform better than lists based only a single study. The fraction of each benchmark set correctly identified is plotted as a function of the number of genes suggested to be periodically expressed. For *A. thaliana*, we selected the top-600 cyclic genes based on where the curve of combined list shows no further enrichment over random expectation. On the other hand, the combined list of mammal diurnal regulated genes is stable after the top-300 genes. However, if we calculate the number of benchmarking genes that are common among the three lists, we expand our list to the top-600 genes.

Figure S3: Comparison of Arabidopsis and human metabolic pathways with diurnal regulated genes. The diurnal regulation is poorly conserved between Arabidopsis and human, as it is related to the biology and the environment of each organism (Harmer et al, 2000; Panda et al, 2002). This figure illustrates the low conservation between plant and animal diurnal regulated metabolism. The light grey line shows the metabolic network of human and Arabidopsis, respectively. The diurnal regulated genes are highlighted in yellow for Arabidopsis and in magenta for human. The majority of diurnally regulated genes are not conserved between the two species, although there might be substrates that are diurnally regulated in both organisms, e.g. the oxaloacetate in the TCA cycle (is indicated by red arrows). There are differences in metabolic networks of diurnal regulation due to either species-specific pathways, such as photosynthesis,

that exists only in plants (continuous-line red box) and not in animals (dashed- line red box), or species-conditional pathways that exist on both species but are diurnally regulated only in one of the two. Continuous- or dashed-line boxes illustrate if the pathway is «on» or «off», respectively. We exemplified steroid biosynthesis (in green) and porphyrin biosynthesis (in light blue) as human and Arabidopsis conditional pathways. As is obvious, periodicity due to diurnal rhythm, similarly to cell cycle periodicity (Jensen et al., 2006), is rarely conserved during evolution. The custom metabolic map shown here was generated using iPath (Letunic et al, 2008).

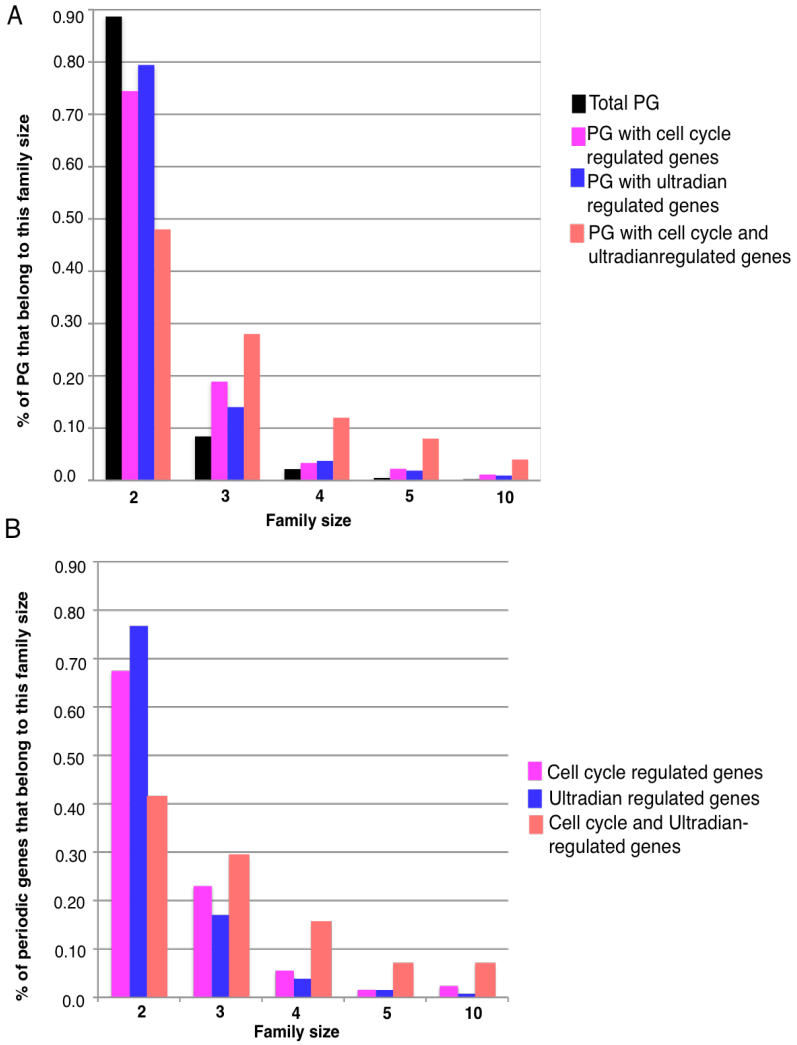


Figure S4: Family size distribution of yeast paralogous groups. (A) The family size distribution of total yeast paralog families is shown in black, while the paralog families that contain at least one cell cycle or ultradian regulated member are shown in magenta and blue, respectively. The distributions of gene families that contain either cell cycle or ultradian regulated members are similar to the distribution of total yeast families. The distribution of gene families that contain both cell cycle and ultradian regulated genes (red) is slightly skewed towards larger family sizes, which is not surprising since a larger family is obviously more likely to contain both types of regulated genes. (B) This panel depicts the periodic members that are present on the paralogous groups of panel A. More specifically, cell cycle regulated genes that belong to cell cycle regulated paralogous groups are depicted in mangenta, while ultradian regulated genes that belong to ultradian regulated paralogous groups are depicted in blue. The red bars correspond to the periodic genes

(both cell cycle and ultradian) that exist on the cell cycle/ultradian regulated paralogous groups. As is obvious, the vast majority of periodic genes belong to 2-member or 3-member families.

Supplementary tables:

Table S1: Summary of diurnal/ultradian rhythm microarray experiments analyzed in this study. The table summarizes the organisms from which the microarray data sets were extracted and the periodic processes that are studied: diurnal and ultradian rhythm. The first lasts 24hours, while the second only 40 min. The time intervals between time points are related to cycle duration and phase (more details on reference studies). We also cited the tissues were the samples were extracted from and the source of the data sets.

Organism	**Periodic Process**	**Time Intervals**	**Tissue**	**Group/Studies**
M.musculus	Diurnal rhythm	4h/cycle	adrenal gland	Oster et al, 2006
M.musculus	Diurnal rhythm	4h/cycle	heart and liver	Storch et al, 2002
M.musculus	Diurnal rhythm	4h/cycle	muscle and liver	Miller et al, 2007
A.thaliana	Diurnal rhythm	4h/cycle	seedlings	Edwards et al, 2006
A.thaliana	Diurnal rhythm	4h/cycle	Leaf	Blasing et al, 2005
S.cerevisiae	Ultradian rhythm	4min/cycles	culture in oxygen	Klevecz et al, 2004

Table S2: Family size distribution of paralogous groups (PG). The table presents the number of gene families and their membership after our paralogy assignment for each of the three organisms (Arabidopsis, human and budding yeast).

Members/PG	*A.thaliana*	*H.sapiens*	*S.cerevisiae*
2	914	758	510
3	420	290	40
4	171	117	14
5	92	73	3
6	57	37	2
7	39	23	1
8	18	14	-
9	22	4	-
10	5	1	-
11	6	1	-
12	7	2	1
13	2	1	-
14	2	1	-
15	3	1	-
16	2	-	-
17	2	1	-
18	1	-	-
19	-	1	-
27	1	-	-
33	1	-	-

Table S3: Genes that are regulated both during the diurnal rhythm and the cell cycle. The table is separated into 2 parts: the first one presents Arabidopsis double regulated genes, while the second one presents the mammalian cell cycle and ultradian regulated genes. The first column of the table contains the gene annotation of each gene. If it is not known, we provide the annotation of the Orthologous Group (OG) to extend the functional analysis. Orthologous groups and genes without annotation are marked with '-'.

Gene/ Orthologous Group Annotation	Gene ID	Gene symbol
Histidine-containing phosphotransfer protein	AT3G16360	AHP4
Hydroxyindole-O-methyltransferase and related SAM-dependent methyltransferases	AT1G21120	T22I11.5
Tyrosine aminotransferase	AT5G53970	K19P17.14
-	AT2G15890	*MEE14*
-	AT2G01660	CRRSP12
-	AT5G28910	F7P1.2
-	AT1G18990	F14D16.14
Histone H1.2 (Histone H1d)	ENSP00000339566	HIST1H1C
Ran GTPase-activating protein 1	ENSP00000216243	RANGAP1
Dynein light chain type 1	ENSP00000242577	DNCL1
Kinesin heavy chain	ENSP00000345045	KIF5B
Splicing factor, arginine/serine-rich 5	ENSP00000216538	SFRS5
Insulin induced protein (growth response protein)	ENSP00000340310	INSIG2
Calcipressin-1	ENSP00000320768	DSCR1

Transcription factor BTEB1 (Basic transcription element binding protein 1)	ENSP00000238031	BTEB1
Mitogen-activated protein kinase 6	ENSP00000243449	MAP2K6
Cyclin-dependent kinase inhibitor	ENSP00000244741	CDKN1A
T-box transcription factor TBX3	ENSP00000257566	TBX3
Mitotic kinesin-like protein 1	ENSP00000260363	KIF23
WD repeat and SOCS box containing protein 1	ENSP00000262394	WSB1
Arginine-Rich protein	ENSP00000273628	ARMET
Ras-related small GTPase, Rho type	ENSP00000296731	RHOBTB3
Tyrosin-protein kinase receptor	ENSP00000301177	AXL
Cyclin G2	ENSP00000315743	CCNG2
TGFbeta receptor signaling protein SMAD and related proteins	ENSP00000332973	SMAD3
Runt-related transcription factor 1	ENSP00000340690	RUNX1
Nuclear factor I/C	ENSP00000342859	NFIC
Heat-Shock protein 105 kDa	ENSP00000318687	HSPH1
3-Hydroxy-3-Methylglutaryl-Coenzyme A reductase	ENSP00000287936	HMGCR

Table S4: Contingency tables for testing cell cycle – diurnal/ultradian paralog pairs. The sub-tables contain the test for each category (e.g. cell cycle regulated genes that belong to ultradian regulated list) for the three organisms The four numbers correspond to 1) the number of cycle A (e.g. cell cycle) regulated genes that have been tested for periodicity both by cell cycle and ultradian microarray experiments, 2) the number of paralogs of the genes that are regulated by cycle B (e.g. ultradian rhythm) and have been tested for periodicity both during cell cycle and ultradian microarray experiments, 3) the number of cycle A (e.g. cell cycle) regulated genes that belong to the latter category and 4) the total number of genes that have been tested both by cell cycle and ultradian microarray experiments.

A)

Organism	**Cell cycle regulated genes**	**Diurnal/Ultradian Paralogs**	**Cell cycle regulated genes that belong to diurnal/ultradian paralogs**	**Total number of genes**	**p-value (Fisher's test)**	**Fold**
A.thaliana	399	355	25	20701	2.9e-08	3.8
H.sapiens	478	197	15	9456	0.042	1.5
S.cerevisiae	544	141	29	6106	1.3e-05	2.3

B)

Organism	**Diurnal/ultradian regulated genes**	**Cell cycle Paralogs**	**Diurnal/ultradian regulated genes that belong to cell cycle paralogs**	**Total number of genes**	**p-value (Fisher's test)**	**Fold**
A.thaliana	547	209	18	20701	1.19e-05	3.4
H.sapiens	413	128	15	9456	0.0004	2.7
S.cerevisiae	575	110	29	6106	1.8e-07	2.8

Table S5: Cell cycle and diurnal- regulated pairs of paralog pairs in Arabidopsis. In this table, we report the 26 cell cycle regulated genes and their diurnal regulated paralogs in Arabidopsis. There are

diurnal regulated genes (e.g. HTB3) with many cell cycle regulated paralogs (e.g. HTB7, HTB11). The orthologous groups are marked with distinctive colors. The first column of the table contains the gene annotation of each gene. If it is not known, we provide the annotation of the Orthologous Group (OG) to extend the functional analysis. Orthologous groups and genes without annotation are marked with '-'.

Gene/ OG Annotation	**Cell cycle- regulated genes**		**Diurnal-regulated genes**	
	TAIR_ID	**Gene symbol**	**TAIR_ID**	**Gene symbol**
Histone H2B	AT2G37470	HTB10	AT2G28720	HTB3
	AT3G09480	HTB7	AT2G28720	HTB3
	AT3G45980	HTB9	AT2G28720	HTB3
	AT3G46030	HTB11	AT2G28720	HTB3
	AT3G53650	HTB6	AT2G28720	HTB3
	AT5G22880	HTB2	AT2G28720	HTB3
	AT5G59910	HTB4	AT2G28720	HTB3
Catalase	AT1G20620	CAT3	AT4G35090	CAT2
Myo-inositol-1-phosphate synthase	AT5G10170	IPS3	AT2G22240	IPS1
Asparagine synthase	AT5G10240	ASN3	AT3G47340	ASN1
Peptidyl-prolyl cis-trans isomerase	AT4G38740	ROC1	AT2G21130	CYP2
Alpha-amylase	AT4G25000	AMY1	AT1G69830	AMY3
Glutathione S-transferase	AT2G47730	GST6	AT1G49860	ATGSTF14
Serine carboxypeptidases (lysosomal cathepsin A)	AT3G45010	SCPL48	AT3G10410	SCPL49
Cis-prenyltransferase	AT5G58784	-	AT5G58770	-
Aquaporin	AT4G00430	PIP1E	AT3G61430	PIP1A
	AT1G01620	PIP1C	AT3G61430	PIP1A
Udp-3-O-[3-Hydroxymyristoyl] N-acetylglucosamine deacetylase	AT4G05080	-	AT1G66490	-
Sexual differentiation process protein ISP4	AT4G16370	ATOPT3	AT4G27730	ATOPT6
Stress responsive protein	AT2G24040	T29E15.24	AT4G30660	T10C21.10
			AT4G30650	F17I23.10
Uncharacterized conserved protein	AT4G22120	-	AT4G15430	DL3760W
Actin depolymerizing factor	AT2G31200	ADF6	AT4G34970	-
Microtubule-associated anchor protein involved in autophagy and membrane trafficking	AT4G16520	ATG8F	AT2G45170	ATG8E
-	AT3G06020	F2O10.2	AT5G22390	MWD9.19
	AT5G19260	F7K24.10	AT5G22390	MWD9.19

Table S6: Cell cycle and diurnal- regulated paralog pairs in mammals. In this table, we report the 16 cell cycle regulated genes and their diurnal regulated paralogs in human. For this comparison, we mapped mouse diurnal genes to human genes based on 1to1 orthology. Similar to Arabidopsis, there are diurnally regulated genes (e.g. ENSP00000259799) with multiple cell cycle regulated paralogs are indicated with grey, and cell-cycle-regulated genes with multiple diurnal-regulated genes with yellow.

.

Gene/ OG Annotation	**Cell cycle- regulated genes**		**Diurnal-regulated genes**	
	Ensembl_ID	**Gene symbol**	**Ensembl_ID**	**Gene symbol**
Histone H2B	ENSP00000289316	H2B.1 B	ENSP00000244601	H2BFR
DNA topoisomerase II (alpha isomerase)	ENSP00000344734	TOP2A	ENSP00000264331	TOP2B
Tubulin beta-2 chain	ENSP00000259818	TUBB2B	ENSP00000259799	-
Tubulin beta-5 chain	ENSP00000259925	-	ENSP00000259799	-
Myosin heavy chain	ENSP00000300036	MYH11	ENSP00000226207	MYH1
	ENSP00000226209	MYH3		
Ribonucleoside-diphosphate reductase M2 chain	ENSP00000302955	RRM2	ENSP00000251810	RRM2B
Ser/Thr-protein kinase SGK	ENSP00000237305	SGK1	ENSP00000340608	SGK2
DnaJ homolog	ENSP00000254322	DNAJB1	ENSP00000294629	DNAJB4
Peptidyl-prolyl cis-trans isomerase	ENSP00000338160	FKBP5	ENSP00000001008	FKBP4
Regulator of G-protein	ENSP00000259406	RGS3	ENSP00000271579	RGS16
			ENSP00000319308	RGS5
Alpha-crystallin C chain	ENSP00000281938	HSPB8	ENSP00000248553	HSPB1
Heat shock 70 kDa	ENSP00000302961	HSPA4	ENSP00000296464	HSPA4L
Heterogeneous nuclear ribonucleoprotein	ENSP00000257767	SYNCRIP	ENSP00000304405	hnRNP R
Regulator of nuclear mRNA	ENSP00000337476	-	ENSP00000217394	-

Table S7: Cell cycle and ultradian- regulated paralog pairs in budding yeast. In this table, we summarize the paralogous groups of *S.cerevisiae* paralogs with cell cycle and ultradian expression pattern. Duplicated genes that have arisen by whole genome duplication (14 pairs) are highlighted with bold letters and those that contribute to glucose-and oxygen-rich adaptation are indicated with blue letters. Paralogous families that have one member regulated by cell cycle and multiple by ultradian rhythm and vice versa are indicated in red boxes.

Gene/ OG Annotation	Members/ Paralogous Group	Ultradian regulated genes	Cell cycle regulated genes
Mitochondrial aldehyde dehydrogenase	2	YER073W	YOR374W
S-adenosylmethionine synthetase	**2**	**YLR180W**	**YDR502C**
Phosphoglucomutase	**2**	**YMR105C**	**YKL127W**
Branched-chain amino acid aminotransferase	**2**	**YHR208W**	**YJR148W**
Cystathionine beta-lyase	2	YGL184C	YAL012W
G2/Mitotic-Specific cyclin	**2**	**YGR108W**	**YPR119W**
Cytochrome protein	**2**	**YEL039C**	**YJR048W**
Glycogen [starch] synthase isoform	**2**	**YFR015C**	**YLR258W**
Transport protein	**2**	**YPR156C**	**YGR138C**
Protein kinase inhibitor	2	YPL004C	YGR086C
Glucokinase	**2**	**YDR516C**	**YCL040W**
Homeobox protein	**2**	**YDR451C**	**YML027W**
Long-Chain-Fatty-Acid-Coa-Ligase	3	YMR246W	YIL009W
Predicted hydrolase/acyltransferase	**3**	**YDR125C** YGR110W	**YLR099C**
emp24/gp25L/p24 family of membrane trafficking proteins	3	YGL002W	YHR110W
Homocysteine S-methyltransferase	**3**	YMR321C **YPL273W**	**YLL062C**
Amino acid transporters	3	YCL025C	YBR069C
Amino acid transporters	3	YBR068C	YKR039W

-	**3**	**YDR033W** YCR021C	**YBR054W**
Polyphosphate (E.C 3.1.3.2)	4	YDL024C	YAR071W YBR092C
Zinc-binding oxidoreductase	4	YCR102C YLR460C	YNL134C
Involved in cell wall protein	**4**	**YJL116C**	**YKR042W** YIL123W
ATP-dependent permease	**5**	**YOR328W**	**YOR153W**
Aspartyl protease	5	YDR144C	YLR121C
Predicted transporter (major facilitator superfamily)	**10**	**YHR096C**	**YDR342C** YFL011W YMR011W

Table S8: Functional analysis of cell cycle regulated genes, ultradian regulated genes, duplicated genes, WGD paralogs and SSD paralogs. We compare the GO Terms that are enriched (adjusted p-value <0.05) in the above categories of genes and if they overlap. No GO terms were found to be overrepresented among the aforementioned gene pools, suggesting that the cell cycle/diurnal sub-/neo-functionalization is not a consequence of functional biases.

A) Ultradian regulated genes vs total genes (575 vs 6106) **Term**	**P-value**	**Adjusted P-value (FDR)**
GO biological process at level 3		
nitrogen compound metabolic process (GO:0006807)	9.13E-14	4.66E-12
biosynthetic process (GO:0009058)	7.73E-05	0.00197174
GO biological process at level 4		
organic acid metabolic process (GO:0006082)	9.14E-19	8.59E-17
amino acid and derivative metabolic process (GO:0006519)	2.58E-15	1.21E-13
amine metabolic process (GO:0009308)	5.62E-14	1.76E-12
generation of precursor metabolites and energy (GO:0006091)	1.41E-08	3.31E-07
alcohol metabolic process (GO:0006066)	5.65E-07	8.86E-06
sulfur metabolic process (GO:0006790)	5.29E-07	8.86E-06
urea cycle intermediate metabolic process (GO:0000051)	8.21E-06	0.000110261
vitamin metabolic process (GO:0006766)	1.61E-05	0.000189481
carbohydrate metabolic process (GO:0005975)	0.00038584	0.00402989
lipid metabolic process (GO:0006629)	0.000848622	0.00797705
pigment metabolic process (GO:0042440)	0.00173554	0.013595
cellular biosynthetic process (GO:0044249)	0.00159451	0.013595
heterocycle metabolic process (GO:0046483)	0.00503475	0.0364051
GO biological process at level 5		
carboxylic acid metabolic process (GO:0019752)	3.36E-19	6.61E-17
nitrogen compound biosynthetic process (GO:0044271)	1.62E-14	1.59E-12
sulfur compound biosynthetic process (GO:0044272)	3.02E-06	0.000198014
water-soluble vitamin metabolic process (GO:0006767)	1.33E-05	0.000523216

electron transport (GO:0006118)	2.29E-05	0.000752188
cellular lipid metabolic process (GO:0044255)	0.000333173	0.00820438
vitamin biosynthetic process (GO:0009110)	0.000509976	0.0111628
carbohydrate catabolic process (GO:0016052)	0.00118943	0.0206672
pigment biosynthetic process (GO:0046148)	0.00125892	0.0206672
alcohol catabolic process (GO:0046164)	0.00115856	0.0206672
carbohydrate transport (GO:0008643)	0.00279291	0.0423233
GO biological process at level 6		
amino acid metabolic process (GO:0006520)	9.00E-16	2.58E-13
amine biosynthetic process (GO:0009309)	2.86E-15	4.09E-13
monocarboxylic acid metabolic process (GO:0032787)	2.32E-06	0.000220996
sulfate assimilation (GO:0000103)	8.47E-06	0.000605549
cellular carbohydrate metabolic process (GO:0044262)	2.58E-05	0.00147437
steroid metabolic process (GO:0008202)	3.50E-05	0.00166723
water-soluble vitamin biosynthetic process (GO:0042364)	0.000342673	0.0140006
GO biological process at level 7		
amino acid biosynthetic process (GO:0008652)	3.10E-16	1.06E-13
aspartate family amino acid metabolic process (GO:0009066)	2.61E-09	4.44E-07
sulfur amino acid metabolic process (GO:0000096)	5.98E-08	6.77E-06
glutamine family amino acid metabolic process (GO:0009064)	1.49E-07	1.27E-05
steroid biosynthetic process (GO:0006694)	2.94E-05	0.00166324
serine family amino acid metabolic process (GO:0009069)	2.94E-05	0.00166324
sterol metabolic process (GO:0016125)	3.55E-05	0.00172196
nonprotein amino acid metabolic process (GO:0019794)	0.000256108	0.0108846
cellular carbohydrate catabolic process (GO:0044275)	0.000714671	0.0269987
GO biological process at level 8		
aspartate family amino acid biosynthetic process (GO:0009067)	1.32E-09	5.72E-07
sulfur amino acid biosynthetic process (GO:0000097)	1.22E-07	2.15E-05
methionine metabolic process (GO:0006555)	1.49E-07	2.15E-05
arginine metabolic process (GO:0006525)	4.58E-06	0.000396087
serine family amino acid biosynthetic process (GO:0009070)	4.36E-06	0.000396087
cysteine metabolic process (GO:0006534)	3.38E-05	0.00243156
sterol biosynthetic process (GO:0016126)	6.23E-05	0.00384241
glutamine family amino acid biosynthetic process (GO:0009084)	8.23E-05	0.00444226
ornithine metabolic process (GO:0006591)	0.000340795	0.0163581
glutamine family amino acid catabolic process (GO:0009065)	0.000860192	0.0371603
lysine metabolic process (GO:0006553)	0.00128056	0.0461
monosaccharide catabolic process (GO:0046365)	0.00119228	0.0461
GO biological process at level 9		
methionine biosynthetic process (GO:0009086)	5.73E-08	1.68E-05
cysteine biosynthetic process (GO:0019344)	1.39E-05	0.00203638
arginine biosynthetic process (GO:0006526)	5.03E-05	0.0049143
glucose metabolic process (GO:0006006)	0.000152661	0.0111824
lysine biosynthetic process (GO:0009085)	0.000975986	0.0476607
hexose catabolic process (GO:0019320)	0.000897688	0.0476607
GO molecular function at level 3		
oxidoreductase activity (GO:0016491)	6.80E-12	5.37E-10
cofactor binding (GO:0048037)	1.35E-06	5.33E-05
lyase activity (GO:0016829)	6.53E-05	0.00171935

transferase activity (GO:0016740)	0.000831098	0.0164142
organic acid transporter activity (GO:0005342)	0.00277006	0.0364724
vitamin binding (GO:0019842)	0.00262545	0.0364724

B) Cell cycle regulated genes vs total genes (544 vs 6106)

Term	P-value	Adjusted P-value (FDR)
GO biological process at level 3		
cell cycle (GO:0007049)	7.66E-15	3.91E-13
cell division (GO:0051301)	8.45E-12	2.15E-10
conjugation (GO:0000746)	1.40E-06	2.39E-05
sexual reproduction (GO:0019953)	3.12E-06	3.98E-05
chromosome segregation (GO:0007059)	6.87E-05	0.000700637
reproductive process (GO:0022414)	0.000156255	0.00132817
response to endogenous stimulus (GO:0009719)	0.00117311	0.00854691
filamentous growth (GO:0030447)	0.0013718	0.00874521
anatomical structure development (GO:0048856)	0.00550621	0.0312018
GO biological process at level 4		
mitotic cell cycle (GO:0000278)	1.48E-12	1.39E-10
cell cycle process (GO:0022402)	2.88E-11	1.35E-09
cytokinesis (GO:0000910)	2.27E-05	0.000711211
external encapsulating structure organization and biogenesis (GO:0045229)	0.000373515	0.00877761
reproduction of a single-celled organism (GO:0032505)	0.000479285	0.00901056
regulation of gene expression, epigenetic (GO:0040029)	0.000870061	0.0136309
reproductive cellular process (GO:0048610)	0.00163677	0.0219795
response to DNA damage stimulus (GO:0006974)	0.00219252	0.0228996
negative regulation of biological process (GO:0048519)	0.00198744	0.0228996
regulation of developmental process (GO:0050793)	0.00397388	0.0373545
anatomical structure morphogenesis (GO:0009653)	0.00534394	0.0456664
GO biological process at level 5		
cell cycle phase (GO:0022403)	1.25E-08	2.47E-06
regulation of cell cycle (GO:0051726)	7.71E-07	7.60E-05
DNA metabolic process (GO:0006259)	2.11E-06	0.00013864
cytokinetic process (GO:0032506)	0.000255196	0.0125684
cell wall organization and biogenesis (GO:0007047)	0.000343692	0.0135415
negative regulation of metabolic process (GO:0009892)	0.00167697	0.0307716
reproductive process in single-celled organism (GO:0022413)	0.00151165	0.0307716
organelle fusion (GO:0048284)	0.00126533	0.0307716
negative regulation of cellular process (GO:0048523)	0.0011537	0.0307716
chromosome organization and biogenesis (GO:0051276)	0.00171821	0.0307716
gene silencing (GO:0016458)	0.00221096	0.0335046
negative regulation of gene expression, epigenetic (GO:0045814)	0.00221096	0.0335046
cytoskeleton organization and biogenesis (GO:0007010)	0.00287512	0.0404571
GO biological process at level 6		
DNA replication (GO:0006260)	3.77E-10	1.08E-07
regulation of progression through cell cycle (GO:0000074)	5.87E-07	8.39E-05
M phase (GO:0000279)	1.22E-06	0.000116072
conjugation with cellular fusion (GO:0000747)	2.22E-06	0.000158887

microtubule-based process (GO:0007017)	4.89E-06	0.000232905
DNA strand elongation (GO:0022616)	4.37E-06	0.000232905
interphase (GO:0051325)	9.89E-05	0.00404211
karyogamy (GO:0000741)	0.000275522	0.00875547
biopolymer glycosylation (GO:0043413)	0.000265263	0.00875547
glycoprotein metabolic process (GO:0009100)	0.000479745	0.0137207
regulation of cell morphogenesis (GO:0022604)	0.000821366	0.0213555
chromosome organization and biogenesis (sensu Eukaryota) (GO:0007001)	0.00104772	0.0249706
negative regulation of cellular metabolic process (GO:0031324)	0.00136804	0.030097
DNA repair (GO:0006281)	0.00156395	0.0319493
GO biological process at level 7		
M phase of mitotic cell cycle (GO:0000087)	8.96E-09	3.05E-06
DNA-dependent DNA replication (GO:0006261)	2.34E-08	3.98E-06
sister chromatid segregation (GO:0000819)	7.38E-07	8.37E-05
sister chromatid cohesion (GO:0007062)	4.35E-05	0.00369679
karyogamy during conjugation with cellular fusion (GO:0000742)	0.000107592	0.00731625
interphase of mitotic cell cycle (GO:0051329)	0.000132288	0.00749634
microtubule cytoskeleton organization and biogenesis (GO:0000226)	0.000214349	0.0104112
glycoprotein biosynthetic process (GO:0009101)	0.00053018	0.020029
negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0045934)	0.00052899	0.020029
postreplication repair (GO:0006301)	0.000669569	0.0227653
nucleotide-excision repair (GO:0006289)	0.000809149	0.0229259
cytoskeleton-dependent intracellular transport (GO:0030705)	0.00080059	0.0229259
regulation of cell shape (GO:0008360)	0.000915253	0.0239374
regulation of DNA metabolic process (GO:0051052)	0.00160296	0.038929
response to pheromone during conjugation with cellular fusion (GO:0000749)	0.00216898	0.0491636
GO biological process at level 8		
mitosis (GO:0007067)	1.56E-08	6.75E-06
DNA strand elongation during DNA replication (GO:0006271)	7.51E-06	0.00162295
chromatin assembly or disassembly (GO:0006333)	0.00010947	0.0157636
DNA replication initiation (GO:0006270)	0.000256867	0.0277416
protein amino acid glycosylation (GO:0006486)	0.000398228	0.0344069
microtubule-based movement (GO:0007018)	0.000532931	0.038371
GO biological process at level 9		
mitotic sister chromatid segregation (GO:0000070)	3.33E-06	0.000741557
lagging strand elongation (GO:0006273)	5.06E-06	0.000741557
chromatin assembly (GO:0031497)	3.73E-05	0.0036456

C) SSD paralogs vs total genes (248 vs 6106)		
Term	**P-value**	**Adjusted P-value (FDR)**
GO biological process at level 3		
regulation of biological quality (GO:0065008)	0.000120333	0.00613701
GO biological process at level 4		

generation of precursor metabolites and energy (GO:0006091)	2.91E-05	0.00273947
carbohydrate metabolic process (GO:0005975)	0.000396357	0.0186288
GO biological process at level 5		
carbohydrate transport (GO:0008643)	3.35E-05	0.00660043
GO biological process at level 6		
monosaccharide transport (GO:0015749)	2.61E-05	0.00746332
cell redox homeostasis (GO:0045454)	0.000167221	0.0239126
GO biological process at level 7		
hexose transport (GO:0008645)	2.34E-05	0.00795988
GO molecular function at level 3		
oxidoreductase activity (GO:0016491)	2.95E-05	0.00232737
carbohydrate transporter activity (GO:0015144)	0.000146817	0.00579928
GO molecular function at level 5		
monosaccharide transporter activity (GO:0015145)	6.96E-05	0.0217114
GO molecular function at level 6		
hexose transporter activity (GO:0015149)	6.32E-05	0.0317955
GO molecular function at level 7		
glucose transporter activity (GO:0005355)	8.94E-05	0.0253946
fructose transporter activity (GO:0005353)	0.000499953	0.0473289
mannose transporter activity (GO:0015578)	0.000499953	0.0473289
D) Total paralogs vs total genes (889 vs 6106)		
There are not significant GO terms		
E) WGD paralogs vs total genes (651 vs 6106)		
There are not significant GO terms		

Table S9: Fisher's test for WGD and SSD paralogs. It seems that both categories contribute significally to cell cycle/ultradian sub-/neo-funactionalization. However, although the number of cell cycle/ultradian paralog pairs (table S7) that have arisen either through WGD or through SSD contributed equally, due to the very small nuber of SSD paralogs the statistical significance of this group is larger.

A)						
Category of paralogs	**Cell cycle regulated genes that belong to ultradian paralogs**	**Ultradian paralogs**	**Cell cycle regulated genes**	**Total number of genes**	**p-value (Fisher's test)**	**Fold**
WGD	15	73	544	6106	0.001	2.3
SSD	14	37	544	6106	1.4e-06	4.5
B)						
Category of paralogs	**Ultradian regulated genes that belong to cell cycle paralogs**	**Cell cycle Paralogs**	**Ultradian regulated genes**	**Total number of genes**	**p-value (Fisher's test)**	**Fold**
WGD	15	87	575	6106	0.015	1.8
SSD	14	54	575	6106	0.0004	2.6

Table S10: Size distribution of paralogous groups in budding yeast. Comparison of the four different categories of paralogous families: 1) total number, 2) those that contain at least one cell cycle regulated member, 3) those that contain at least one ultradian regulated member and 4) those that are presented in table S7. Please notice that the total number of paralogous groups in budding yeast that is presented below is different from that one on table S2 (this is because we took into account as members of each group those proteins that have been tested by both cell cycle and ultradian microarrays). The numbers of cell cycle, ultradian and cell cycle/ultradian members correspond to the number of periodic genes we measured in each category of the aforementioned paralogous groups.

Family size	**Total PG**	**Cell cycle PG**	**Cell cycle members**	**Ultradian PG**	**Ultradian members**	**Cell cycle & Ultradian PG**	**Cell cycle & Ultradian members**
2	369	67	85	85	99	12	24
3	35	17	29	15	22	7	17
4	9	3	7	4	5	3	9
5	2	2	2	2	2	2	4
1	1	1	3	1	1	1	4
Total	416	90	126	107	129	25	58