THE UNIVERSITY OF CHICAGO

FUNCTIONAL EVOLUTION OF YOUNG RETROGENES WITH REGULATORY ROLES IN *DROSOPHILA*

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY

IURI MATTEUZZO VENTURA

CHICAGO, ILLINOIS

JUNE 2019

To Léo

Grant me, O Lord, the courage and the joy
to ascend to the summit of this day

Jorge Luis Borges

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am very grateful to many people who contributed to my personal and professional development during my years as a graduate student. The challenges of grad school and the experiences of living in a different country certainly helped shape the person I am today.

I am grateful to my advisor Manyuan Long, who has been encouraging, patient and supportive throughout the years. His awareness towards his students' well being makes the lab a collaborative and relaxed environment to work in.

I am thankful to past and present members of the Long Lab, for helpful discussions and harmonious company: Maria Vibranovski, Ben Krinsky, Patrick Landback, Claus Kemkemer, Qian Yang, Chengjun Zhang, Wenyu Zhang, Li Zhang, Andrea Gschwend, Alex Advani, UnJin Lee, Emily Mortola, and Jianhai Chen. I am especially in debt to Nick VanKuren, for his diligent and friendly support in numerous stages of my work; Shengqian Xia, for his collaboration with the CRISPR knockouts; and Grace Lee, for her generosity in inviting me to collaborate with her in an exciting project.

I thank the members of my committee for helpful suggestions and discussions: Marty Kreitman, Urs Schmidt-Ott, Vinny Lynch and Joe Thornton. I am also thankful to Ilya Ruvinsky, Marty Kreitman and Marcus Kronforst, whose courses I had the pleasure to contribute to as a teaching assistant.

I thank the helpful staff in the Department of Ecology and Evolution for their patience and attention: Mary Johnson, Bonnie Brown, Connie Homan, Jeff Wisniewski, and especially Audrey Aronowsky and Alison Anastasio, for their friendly support and commitment with the students.

I acknowledge my funding sources, the Science without Borders scholarship (BEX18816/12-6) and the Hinds Funds for Graduate Student Research.

I thank my fellow students in the Darwinian cluster for the friendly environment and insightful conversations. I cannot emphasize enough how crucial the support from my friends was during my years in grad school. I am especially in debt to Debora Sobreira, Darli Massardo, Arvind Pillai, Ayse Tenger-Trolander, Marcos Vieira, Aarti Venkat, Matthew Michalska-Smith, Joe Mihaljevic, Jeff Thorburn, and Yoseop Yoon. I also thank the support of longtime friends for reminding me that academia is just a small part of the world, in special Marilia Lopes Justino, Nara Barbieri, Marcos Moraes, Gustavo Fattori, Jader Armanhi, and Rafael Soares de Souza.

I am grateful to my mom, dad and sister, for their affection, unconditional support, and trust in me. I truly appreciate their encouragement throughout my years in academia, even though this meant being physically (and sometimes emotionally) distant.

Above all, I am grateful to my husband Léo, for reminding me every day about the things that really matter. He has been the most loving and supportive partner I could wish for. Even though we had rich and exciting experiences in grad school, being apart from each other for the past years was the hardest (and dumbest) decision we have made. Reuniting with him will be the best reward after our adventure in the United States.

Abstract

The set of genes in the genome of a species is the result of a dynamic net balance between gain and loss events that happen over evolutionary time. The origination and divergence of new genes are major sources of genomic novelty, and has the potential to generate substantial raw material for the evolution of functional innovations. It has been recently recognized that young genes can take over fundamental functions in basic cellular processes and be essential for the survival of an organism, even when they are restricted to a few species in a phylogeny. It is not clear, however, how newly duplicated genes are integrated into ancestral networks or to what extent they diverge from their parents at the functional level. In the two study cases reported here, I investigate the evolution of three relatively young duplicated genes with regulatory functions, and only found in some fly species. Using computational analyses and experimental assays, I show that a diverse suite of factors was responsible for the functional divergence of the duplicated genes after their origination from their conserved parental genes. After their origination through retrotransposition events in different branches of the *Drosophila* phylogeny, the genes acquired a restrict expression pattern, and rapidly diverged in sequence from their parents, which remained essentially conserved. I further show their phenotypic importance for viability and male fertility, and demonstrate that the duplicated genes diverged in several aspects from their parents, including their protein interactions, genomic binding pattern and impact on global gene regulation. Our results show how young elements can be integrated into conserved processes in different ways, and illustrate the complex nature of evolution driven by new gene origination.

Chapter 1

Introduction

Describing the astonishing diversity at many levels of natural organization, as well as understanding its functioning rules, has been a central goal of biology. Finding out the main principles that underlie complex and diverse living systems has captivated researchers, from traditional naturalists (such as Darwin and Wallace themselves) to contemporary biologists in the genomic era, who are equipped with sophisticated experimental and analytical tools.

In this chapter, I will briefly argue that the origination of new genes (i.e. the formation of new genetic loci) is a topic of fundamental importance in biology, which can shed light on fundamental questions in evolutionary biology. In addition, I will describe the main gene origination mechanisms, as well as the patterns that emerged from large-scale genomic studies, in special regarding gene movement between chromosomes and male-bias expression. I will also summarize case studies that showed how genes restricted to few species can be integrated into fundamental roles, and have an impact in basic cellular functions. Finally, I will conclude by articulating these broad concepts with the studies described in the next chapters.

New genes for old questions

The notion that genomes from distinct species are constituted by a different set of genes, and that new genes can originate and evolve in function over evolutionary time seems commonplace now, but have come a long way to become accepted. The long and rich history behind those ideas intertwines with the very development of genetics in the

twentieth century (reviewed in Taylor and Raes, 2004). For instance, when studying small duplications of chromosome fragments in fruit flies at Thomas Morgan's famous lab, Muller speculated that such duplications could be a way of increasing gene number, which could produce redundant gene copies and diverge over time (Muller, 1935). Initial hypotheses regarding the creation and accumulation of new genes were also in the center of heated debates, such as those concerning the causes of increasing morphological complexity in living organisms (Goldschmidt, 1940), the relative importance of mutations in regulatory versus protein coding regions (Britten and Davidson, 1969), among several others.

Perhaps the early author most associated with concepts regarding the evolution of duplicated genes is Susumo Ohno. In his seminal book (Ohno, 1970), he combined emerging evidences from molecular biology, biochemistry and population genetics to argue that gene duplication is the major contributor to genomic evolution. Ohno's reasoning is that most genes are usually strongly constrained, and new mutations that alter their function are usually removed from populations by purifying selection. Nevertheless, when a gene duplicates, the new copy is redundant, and is free to accumulate beneficial mutations that alter its function without disrupting the ancestral one. Over time, the new copy can evolve a new activity, and both copies are maintained in the genome. Ohno's vigorous promotion of his *neofunctionalization* model became widely popular, and inspired many studies on the evolution of new genes. This model emphasizes the creative role played by gene duplication and, along with the notion that evolution is a tinkering process, which modifies already existing proteins and finds new uses for them (Jacob, 1977), became an influential view of gene function evolution. In an early review on the emerging

principles in molecular evolution, for instance, Kimura and Ohta (1974) go far to state that "gene duplication must always precede the emergence of a gene having a new function".

Because beneficial mutations are rare and should require a long time to accumulate on the duplicated copies, other models were proposed as attempts to explain the large number of duplicated genes retained in the genomes over time. A leading model proposes that new duplicated copies can be retained and diverge through a *subfunctionalization* path (Force et al., 1999). The basic idea is that, after duplication, the paralogs are redundant, purifying selection is relaxed, and degenerative mutations are allowed to accumulate on both copies. With time, the ancestral function of the parental gene may be partitioned between the copies, which complement each other. In addition, a suite of alternative models were proposed to explain the retention of duplicated genes under various scenarios. For instance, a category of models posits the preservation of both copies at all stages of their evolution, due to benefits of increasing the dosage of a gene product (e.g. the adaptive radiation model; Francino, 2005), or by independently refining and optimizing multifunctional genes to fulfill distinct tasks (e.g. escape from adaptive conflict, Bergthorsson *et al.*, 2007; and innovation-amplification-divergence, Piatigorsky, 1991; reviewed in Conant and Wolfe, 2008).

Despite the difficulty of distinguishing between models in real case studies due to analytical and experimental limitations, or simply because of the inherent complexity of biological systems (Hahn, 2009; Innan and Kondrashov, 2010), these models are useful for providing a framework for the investigation of new gene copies. For example, it is pertinent to understand to what extent the function of a duplicated gene and its parent overlap at expression and protein function; as well as to determine in what processes they act on, and

what phenotypes they impact. The systematic dissection of recently originated genes provides more than curious or anecdotal study cases. Although many evolutionary mechanisms can be elucidated by the study of ancient genes, some questions are better tackled by investigating evolutionary young elements. Those genes offer a particularly convenient study model to address certain evolutionary questions because, given their recent origination, they still carry the evolutionary signatures that shaped their formation and evolution (Long et al., 2013). Furthermore, by applying computational and experimental analyses to compare ancestral genes and their duplicates, we can connect their genetic and functional divergence to obtain a detailed view of gene evolution. With this framework in mind, long-standing questions in evolutionary biology can be addressed and illuminated, such as the relative importance of evolutionary processes to genomic evolution, the number and nature of mutations responsible for adaptive traits, the pace and mechanisms underlying genetic adaptation, among others (Kaessmann, 2010; Losos et al., 2013; Stern and Orgogozo, 2008).

New gene origination mechanisms

The ability of generating and analyzing large genomic datasets developed in the last decades enabled studies to transfer their attention from few anecdotal examples of gene duplicates to large lists of young genes. By systematically comparing genomes, such studies were able to identify loci that are present in a group of closely-related species but absent in all others (reviewed in Hardison, 2003; Ventura and Long, 2017). This comparative approach is valuable for elucidating the general picture and common patterns related to

new gene origination, as well as for providing appropriate instances that are worth being investigated experimentally.

Large-scale studies demonstrated that the total gene number in the genome of a species represents a dynamic net balance between gain and loss events that happen over evolutionary time. The origination of a gene (i.e. the formation of a locus that did not exist previously) can occur through various molecular mechanisms (reviewed in detail in Kaessmann *et al.*, 2009; Chen *et al.*, 2013). The most common process is DNA-based duplication, which simply "copy and paste" a DNA sequence from a genomic region to another during replication. Duplications can occur on a wide range of scales, from parts of genes to extreme cases such as whole-genome duplications (Sémon and Wolfe, 2007). Another common process of duplication is retrotransposition (or RNA-based duplication), in which an mRNA molecule is reverse-transcribed and inserted back into the genome. Retrogenes can be easily identified in genome sequences due to, for example, their lack of introns compared to its parental copy and traces of flanking short direct repeats and poly-A tail at the 3' end, although several factors may confound or erase those hallmarks (Casola and Betrán, 2017). Remarkably, retrocopies are inserted into random regions of the genome, and readily recruit or evolve new regulatory elements in a different epigenomic environment (Zhang and Zhou, 2018), leading to fast divergence in their expression pattern compared to the parental copies.

Besides the duplication-based mechanisms described above, new gene structures can also be created by gene fission and fusion (Snel et al., 2000), transposable element domestication (Jangam et al., 2017), and even evolve *de novo*, when mutations in a non-coding region generate a new coding sequence (Tautz, 2014). Perhaps unsurprisingly since

this is biology, the combination of several mechanisms or events are often involved in the origination of a gene. An interesting example is illustrated by *jingwey*, one of the first young genes to be rigorously characterized in *Drosophila* (Long and Langley, 1993). The gene is present in only two species, and it was created after the fusion of exons from two different genes, a DNA-duplicated copy of *yellow-emperor*, and a retrotransposed copy of the *Adh* gene (Long et al., 2003; Wang et al., 2000). This process combined protein domains from the two ancestral genes to create a new chimeric gene.

Patterns of new gene evolution

Comparative genomic studies carried out in the last decades provided a general picture of gene origination events in different clades. Here, I will focus on three specific results that emerged from large-scale comparisons: the rates of gene origination, gene movement (traffic) between chromosomes, and male-bias expression of new genes.

The availability of genomes from closely related species allows us to investigate the rates of new gene origination in particular lineages with high confidence. For example, analyses of the 12 *Drosophila* genomes suggest that around 947 duplicated genes arouse in the subgenus *Sophophora* after its split with the subgenus *Drosophila*, resulting in an overall rate of 15 duplicates per million years (Zhang *et al.*, 2010a), the majority of which (78%) generated by DNA-based duplication. In vertebrates, between 25-30 new duplicates were generated per million years, with a wide variation between lineages (Zhang *et al.*, 2010b; Zhang *et al.*, 2011). In an early study that compared genomes from distant eukaryotes, a conservative estimate of the average rate of new gene origination was around 0.01 per gene per million years, although the estimate is quite variable (Lynch and Conery,

2000). Regardless of the precise rate of origination, it is clear from these reports that gene duplication has the potential to generate substantial raw material for the evolution of genetic novelties.

Along with the general rates of new gene origination, initial comparative studies revealed a striking pattern for retroduplicated genes regarding their non-random chromosomal distribution and biased expression. The reports showed that there is a significant excess of new retrogenes inserted on autosomes that derived from X-linked parental genes (the so-called X→A movement). Furthermore, such derived retrogenes disproportionately exhibit male-biased expression (i.e. they are only expressed in male tissues, or at significantly higher levels in males compared to females). This pattern was extensively reported for *Drosophila* (Betrán *et al.*, 2002; Dai *et al.*, 2006; Bai *et al.*, 2007; Zhang *et al.*, 2010a), *Anopheles* (Toups and Hahn, 2010), and mammals (Emerson *et al.*, 2004; Potrzebowski *et al.*, 2008; Zhang *et al.*, 2010b; Carelli *et al.*, 2016). Several factors may contribute to this pattern (Casola and Betrán, 2017), such as potential compensation for the inactivation of parental genes on the X chromosome during meiosis by their autosomal duplicates (Vibranovski *et al.*, 2009), sexual antagonism (Charlesworth et al., 1987), and even insertion bias due to permissive nuclear organization in the germline (Díaz-Castillo and Ranz, 2012). These hypotheses are reviewed in more detail in the *Gene Traffic* section in Chapter 2.

New genes for old processes

The large-scale comparative studies mentioned above demonstrated beyond doubt that gene origination is a process that have been continuously shaping species genomes. It

remained to be elucidated, however, what is the functional contribution and phenotypic importance of newly originated copies, which became possible with detailed genetic manipulation tools. A traditional view is that basic cellular functions are encoded by a set of old, conserved genes, since some essential processes and pathways are shared by evolutionary distant organisms (Jacob, 1977; Miklos and Rubin, 1996; Lewin et al., 2011). Young genes, on the other hand, would contribute to more dispensable activities, or would be restricted to the adaptation to novel environments (Kondrashov, 2012), but would hardly fulfill essential functions (Ashburner et al., 1999). Moreover, it has been suggested that gene duplications are not only an unnecessary ingredient for developmental processes, but are actually strongly selected against, because of their disruptive effects on gene dosage-sensitive developmental pathways (Carroll, 2008).

Recent experiments challenged this traditional assumption showing that relatively young genes can be integrated into basic cellular processes and have dramatic phenotypic importance (reviewed in Chen *et al.*, 2013). For example, using constitutive RNAi-knockdown to silence the expression of a large number of genes in *D. melanogaster*, Chen *et al.* (2010) showed that a significant proportion of young genes are essential for fly development (~30%). More surprisingly, the fraction of genes that were essential was roughly constant across the phylogeny, regardless of their ages (Xia and Long, unpublished data).

Several functional case studies have shown that new genes can be integrated into basic and conserved cellular processes, and become critical for development and reproduction (extensively reviewed in Chen *et al.*, 2013). In particular, it is worth noting, for instance, several described young duplicates whose silencing was shown to

dramatically reduce male fertility in *Drosophila* (Chen et al., 2012; Ding et al., 2010; Gubala et al., 2017; Nurminsky et al., 1998; Saleem et al., 2012; Yeh et al., 2012; Zheng et al., 2018), and mouse (Bradley et al., 2004; Heinen et al., 2009).

A mechanism by which duplicated genes, even if evolutionary young, can become indispensable for essential roles is by interacting with pre-existing members (Kim et al., 2012). For example, it was shown that the perturbation of young genes can lead to widespread regulatory changes, even in conserved processes (Chen et al., 2012; Ding et al., 2010; Lee et al., 2018; Zhang et al., 2015). The integration of new elements into ancestral pathways is possible because regulatory networks are not stagnant over time, but are recurrently rewired during evolution (Johnson, 2017; Wilinski et al., 2017), which allows new elements to become crucial players in old cellular processes.

New genes for new functions

Besides being integrated into ancestral processes, new genes can increase the functional repertoire of an organism by encoding proteins with different properties when compared to the ancestral gene, and/or by acquiring different expression patterns. For example, careful molecular dissection and comparison of protein activities have shown that duplications have diversified cytochrome P450 enzymes in plants, leading to the creation of a new phenolic pathway only found in one plant family (Matsuno et al., 2009). More recently, Vazquez *et al.* (2018) showed that a retroduplicated gene (*LIF6*) encodes an intracellular protein that induces apoptosis in response to DNA damage in elephants, in contrast to its parental protein, which functions as an extracellular cytokine. Several other

examples of altered protein functions were explored in diverse organisms (extensively reviewed in Ding *et al.*, 2012; Chen *et al.*, 2013; Long *et al.*, 2013).

The extent to which duplicated genes diverge in expression from their parental copies has been thoroughly studied in recent years, stimulated by the availability of whole genome sequences and high-throughput transcriptome data for different tissues, and were carried out in in yeast (Gu et al., 2005; Wagner, 2002), *Arabidopsis* (Casneuf et al., 2006), *Drosophila* (Assis and Bachtrog, 2013; Gu et al., 2004) and mammals (Assis and Bachtrog, 2015; Lan and Pritchard, 2016; Makova and Li, 2003), among others. Particularly in flies, the emerging conclusions from such studies are that the divergence in gene expression after gene duplication tends to be strongly asymmetrical, with the expression of the duplicated copies diverging at higher rates compared to their parents. Moreover, the expression diversity boosted by duplicated genes is especially apparent during development and in reproductive tissues, suggesting that particular processes are more liable to integrate new elements (Gu et al., 2004). In an extreme example, VanKuren and Long (2018) showed that two young duplicates evolved sex-specific reproductive functions in male and females in just 200 thousand years, in order to mitigate antagonistic effects when expressed in the opposite sex. Thus, divergence in expression of duplicated genes may be the first steps leading to the their specialization, as suggested by early models of gene duplication (Force et al., 1999; Ohno, 1970). The most conspicuous examples of such dynamics are observed in retrogenes that acquire male-biased functions, which will be discussed below.

Two case studies of retrogenes with regulatory functions

As mentioned earlier, the excess of autosomal retrogenes duplicated from X-linked copies has been described in *Drosophila* as soon as genomic and transcriptome data became available for such analyses (Betrán *et al.*, 2002; Dai *et al.*, 2006; Zhang *et al.*, 2010a). More recently, Assis and Bachtrog (2013) employed a phylogenetic approach for comparing expression patterns between duplicates, their parents and outgroups and showed that copies originated through RNA-duplication were more likely to exhibit an expression divergence consistent with neofunctionalization or specialization, with strong asymmetry between duplicates and their parents, and significant localization to the testes. Such asymmetry is expected because retrogenes are inserted in random genomic locations, and lack parental cis-regulatory sequences.

Consistent with the emerging pattern from large-scale studies, previous studies have reported the existence of specific versions of several housekeeping genes only expressed in male reproductive tissues. Some of these genes are involved in basic cellular processes such as transcription regulation (Hiller, 2004), translational machinery (Baker and Fuller, 2007), and proteasome subunits (Zhong and Belote, 2007). It has been suggested that these duplicated copies might represent specialized versions of their parental genes, better adapted to accomplish their functions in a specific tissue. Alternatively, it has also been speculated that new genes may be preferentially expressed in the testis during their early stages of evolution, due to promiscuous transcription and fast evolution features. Then, at later stages, such genes may acquire broader expression patterns (the so-called “out of the testis” hypothesis; Vinckenbosch *et al.*, 2006). It is not

clear, however, to what extent duplicated copies diverge in function from their parents (Belote and Zhong, 2009), or their phenotypic relevance.

This thesis explores the evolution and functional divergence of young retrogenes found in *Drosophila* in two different study cases, and illustrates the complexity of evolution driven by new gene origination. First, in Chapter 2, I summarize several examples of duplicated genes involved in basic cellular processes in diverse organisms to argue that even ancient and conserved pathways often integrate new elements during evolution.

In Chapter 3, I describe the evolution of a relatively young gene only found in the subgenus *Sophophora* (*CG2053*), which we named *Poseidon.* In this study, we compare its functional impact with that of its parent *CAF40*, an ancient gene involved in regulatory pathways, and *Zeus*, an even younger duplicate from the same gene, which was shown to have a role in male germline gene regulation. We investigate their divergence in expression pattern, protein interactions, and impact on global gene regulation, and assay their phenotypic effect on viability and male fertility. These analyses were performed in collaboration with Annamaria Sgromo and Elisa Izaurralde, from the Max Planck Institute for Developmental Biology in Tübingen, Germany.

In Chapter 4, I describe my contributions to a study performed in collaboration with Grace Lee, a former post-doc in our lab, now a faculty at the University of California in Irvine. In this study, we investigate the evolution of *Cocoon* (*CG7804*), a young retrogene only found in three *Drosophila* species. Our assays show that the gene is essential for the survival of flies at several developmental stages, and in different tissues from its parental gene, *TBPH*. Along with our knockout and knockdowns assays, we performed functional genomic analyses to show that *Cocoon* essential functions are the result of multiple

interactions with critical genes during development. Other contributing authors in this study are Gavin Rice (UC – Davis) and Don-Yuan Chen (UC – Berkeley).

Finally, in Chapter 5 I summarize the similarities observed in both case studies described here, and suggest that some patterns can be particularly common after gene duplication, in special for genes with regulatory roles.

Chapter 2

Connecting evolutionary genomics to cell biology

The work described in this chapter was published as: Ventura IM and Long M (2017) "Connecting evolutionary genomics to cell biology" in the Encyclopedia of Cell Biology, Vol. 4: 153-159, organized by Bradshaw RA and Stahl PD.

Abstract

Biologists are now able to investigate intricate biological systems by exploring the most fundamental source of biological information – the genome. In the last decades, there has been tremendous advance in sequencing and analyzing genomes from diverse organisms. Systematic comparative analyses between genomes from different species coupled with careful experimental dissection provide a powerful strategy to understand how genomes function and evolve. Here, we argue that studying the origin of recently originated genes can shed light on fundamental evolutionary questions. We illustrate this point by summarizing studies of new genes implicated in basic cellular processes, such as regulatory networks, cell division and protein traffic.

Genomics for old and new questions

Among the long-standing aims of biology, understanding how organisms function, develop, and adapt occupies a central position in the field. In modern biology, researchers pursue some of the same standard questions, but with a remarkable advancement: they are now able to connect intricate biological processes to the most fundamental source of

biological information – the genome. As it has been known for decades, the genome of a species contains the instructions required for its evolutionary perpetuation, including the development, functioning and reproduction of the organism carrying it. The 20th century has witnessed a step-wise advancement in our comprehension of the genome structure and function; from the role of DNA in hereditary, to its biochemical structure, genetic coding logic and regulatory mechanisms.

The launch of the Human Genome Project in the 1990's gave the start to several genome consortiums aiming to sequence model organisms' genomes (e.g. yeast, fruit fly, mouse; see Table 2.1). In a simplified overview, a genome project consists of sequencing the nucleotides of short fragments of DNA, covering the whole organism genome; followed by the computational assembly of these short segments into large sequences representing the chromosomes. Finally, the genome annotation step identifies different functional elements in the assembled sequence (such as protein coding regions and regulatory motifs). In recent years, there has been an explosion in the number of genomic studies published, encouraged by the advance in the sequencing technologies (the so-called 'next-generation sequencing') and bioinformatic analysis tools. The massively parallel sequencing implemented nowadays rapidly generates high amounts of genome-wide data, and allows genomics to pursue biological questions in a unprecedented way, connecting variation at the genome level to its functional consequences (Koboldt et al., 2013)

The simple description given above may create the deceptive impression that the link between nucleotide sequences in the genome and their function is direct and unambiguous. However, the genetic information encoded in the DNA may not be evident when a single genome is read. To put it in another way, when one tries to understand what

portions of the genomes are functionally relevant and how the intricate biological processes arise from them, the 'book of life' (to use a commonplace metaphor) resembles more an encrypted file.

Table 2.1. A selection of landmark genome projects that have contributed to the understanding of genome structure, function and evolution.

Organism	Remark	Reference
Haemophilus influenza (bacterium)	First complete prokaryotic genome	(Fleischmann et al., 1995)
Saccharomyces cerevisiae (yeast)	First complete eukaryotic genome	(Goffeau et al., 1996)
Escherichia coli (bacterium)	Most studied prokaryotic organism	(Blattner et al., 1997)
Caenorhabditis elegans (worm)	Model organism, first multicellular organism genome	(Consortium, 1998)
Drosophila melanogaster (fruit fly)	Important model organism	(Adams et al., 2000)
Arabidopsis thaliana (flowering plant)	First plant genome, important model organism	(Arabidopsis Genome Initiative, 2000)
Homo sapiens	A complex mammal	(International Human Genome Sequencing Consortium et al., 2001; Venter et al., 2001)
Mus musculus (mouse)	Mammal model organism	(Mouse Genome Sequencing Consortium et al., 2002)
12 *Drosophila* genomes	Comparative analysis of fly species	(Drosophila 12 Genomes Consortium et al., 2007)
1000 human genomes	Characterization of human genetic variation across populations	(1000 Genomes Project Consortium et al., 2010)

In order to identify functionally relevant elements, the systematic comparison of genome sequences from different species provides a powerful strategy. The justification behind such comparative framework is that portions of the genome that encode for biologically important functions tend to be more conserved across different species than non-functional regions (Hardison, 2003). The rationale is simple: genomes are continually evolving; therefore, if sequences from different species remain with high similarity, despite their independent evolutionary trajectories since the last common ancestor, it is

reasonable to assume that they were spared by evolution because they play a functional role, which is usually subject to purifying selection, whereas non-functional portions are free to accumulate mutations, or even be deleted from the genome. This approach can be used to identify genes, and it is particularly useful for finding regulatory regions, whose structure is more heterogeneous than that of protein coding sequences. Therefore, when researchers find conserved sequences from evolutionary distant species (such as human, mouse and fly), they conclude that sequence is a good candidate for a functional element, because it retained similar information even in species that have been diverging for millions of years. It must be noted, on the other hand, that this conservative approach may fail to identify genes that underwent intense divergence in a short evolutionary period, such as that driven by rapid positive selection, for example. Thus, complementary annotation approaches, such as measurements of protein-coding potential, probably improve genome analyses (Zhang et al., 2012).

In this context, it is not surprising that the human genome sequencing was followed by similar projects in other model organisms (Table 2.1 shows that several pioneer genome projects were published within a decade). Even if one is interested in exploring genome function of a single species (e.g. identifying genes contributing to human diseases), evolutionary comparisons with genomes from diverse taxa are imperative (Collins et al., 2003). Nevertheless, identifying functional elements in the genome is just the first step for investigating how genes work, interact and evolve. Detailed functional experiments are required to confirm and determine the specific biological functions of genetic elements (Miklos and Rubin, 1996).

Besides the role of identifying functional elements in genomic studies, comparative genomics is also a powerful strategy to investigate evolutionary questions (Hardison, 2003). Comparison of genomes in a phylogenetic framework, ranging from closely related species to taxa in different kingdoms, for example, may help to elucidate exciting questions: What is the basic set of genes common to diverse taxa? How often new genes are created? What are the evolutionary processes involved in diversification and adaptation? An illustrative scheme is provided in Figure 2.1. In the other extreme, comparison of genomes from individuals of the same species can also be applied to pinpoint specific genetic variations underlying relevant phenotypes (such as adaptive traits or disease factors) in genome-wide association studies. A complementary set of questions we are now able to investigate relates to the origin of genetic novelties. In the past decades, many studies have explored the creation of new genetic elements, how they integrated in genetic systems, and their roles in phenotypic evolution (Long et al., 2013).

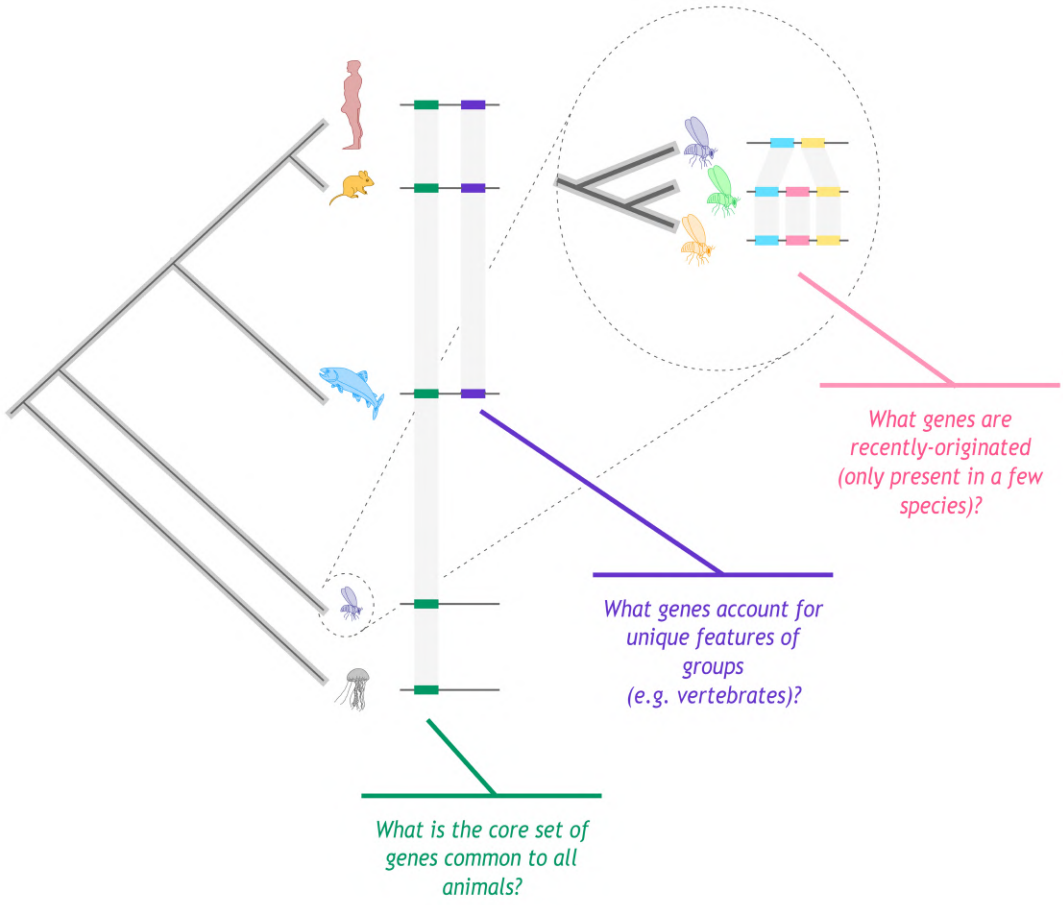


Figure 2.1. Schematic representation of different questions addressed in a comparative genomics approach (based on Hardison, 2003). At left, colored boxes represent genes present in different animal groups in the phylogeny (cnidarian, fly, fish, mouse, and human); at top right, a phylogeny of closely related fly species illustrating a syntenic alignment used to identify a new gene (pink box).

Case studies: new genes for old and new functions

In the rest of this chapter, we will focus on studies that took advantage of genomic data to identify new genes and then investigate fundamental questions about their evolution, especially those related to cell biology functions. We refer to new genes as distinct loci that originated at a certain time in the evolutionary history of a species and that were not present in the ancestors (Chen et al., 2013). New genes offer a particularly convenient study model to address evolutionary questions because, given their relatively

recent origin, they still carry the evolutionary signatures that shaped their formation and molecular evolution (Long et al., 2013). Furthermore, due to the great advance of computational and molecular biology techniques, we can now explore some evolutionary problems in a precise and mechanistic way, identifying patterns, elaborating hypotheses and testing their predictions, in order to achieve a comprehensive view of new genes evolution and connect genetic to functional divergence.

New genes can originate through diverse mechanisms, such as DNA duplication, retrotransposition (RNA-based duplication), gene fission or fusion, and *de novo* creation (reviewed in Chen et al., 2013). New loci can be identified through syntenic alignments of genomes from different species. Genes present in some (or even a single) species but absent in all outgroup species are good candidates for having a recent origination (see Figure 2.1). Furthermore, structural features in the gene and flanking regions can be used to infer the mechanisms of its formation. After identifying a new gene candidate, studies typically seek to confirm its functionality (analyzing its open reading frame length and transcription products, for instance) and explore signatures of molecular evolution (such as synonymous/non-synonymous substitution rates). Ideally, the ultimate goal is to connect the evolution at the nucleotide sequence to its phenotypical impact, evaluating the functional and adaptive impact of the new gene.

Basic cellular functions have been considered to be encoded by a set of old, conserved genes, since some essential processes and pathways are found in evolutionary distant organisms (Miklos and Rubin, 1996), whereas new genes would be involved in minor, dispensable activities. Nevertheless, an increasing number of studies have been reporting the contribution of new genes to basic cellular processes. They show that new

genes can be integrated into conserved pathways, and even become essential for viability and reproduction (Chen et al., 2010). Next, we will explore how the study of new genes can be connected to cellular biology. We will focus on examples of gene movement between chromosomes, as well as the integration of new genes into basic cellular functions (namely, gene regulatory networks, cell division and protein traffic) to argue that the evolution of new genes can shed light on fundamental cellular mechanisms. With that, we hope to stimulate future insightful connections between the genomics and cell biology fields.

Gene traffic

Cytogenetics played a central role in the early decades of genetic studies. Suffice is to mention, for instance, that Thomas Morgan and his students were pioneers in connecting chromosome structure to heredity; and Theodosius Dobzhansky and his students were able to use structural variation (such as chromosomal inversions) as molecular markers for studying genetic variability in natural populations of *Drosophila*, all with elementary microscopy techniques. With the recent advancement of molecular biology techniques, however, evolutionary studies have changed its focus from large variations of structure to the specific changes on nucleotide sequences. In this context, genomic analyses allow the investigation of biological questions at a finer scale. However, basic structural data can be surprisingly informative for testing evolutionary hypotheses in a comparative genomics framework. We will illustrate this point using recent findings of new genes movement patterns in mammals (e.g. humans and rodents) and insects (e.g. flies, mosquitoes and silkworms).

Several studies demonstrated that the distribution of retrogenes (new loci created through RNA-based duplication) and their parental copies in the genomes are not random with respect to their chromosomal position; instead, new genes movement follows a preferential direction (gene traffic). Analyses focusing on the *D. melanogaster* genome (Betrán et al., 2002) and more recent comparative investigations using genomes from 12 *Drosophila* species (Meisel et al., 2009; Vibranovski et al., 2009a), for example, revealed that retrogenes derived from X-linked genes preferentially move to autosomes (X→A), contrasting to random expectations. More interestingly, the large majority of these X→A retrogenes are expressed in testis, suggesting a connection between male-biased function and chromosomal location. To some extent, a similar pattern is observed in human and mouse genomes (Emerson et al., 2004), in which there is an excess of X→A movement of retrogenes, as well as male expression bias. However, in these mammal genomes there is also a significant excess of retrogenes moving in the opposite direction (A→X), which present either female or unbiased expression. Interestingly, a somewhat symmetrical pattern is observed in silkworms (in which females are ZW and males are ZZ), whose retrogenes tend to move out of the Z chromosome and have ovary-biased expression (Wang et al., 2012). Additionally, recent comparative analyses of 16 mosquito genomes also found a significant excess of gene translocations from the X to the autosomes in these species (Neafsey et al., 2015).

How can these findings be explained in an evolutionary perspective? Gene traffic patterns are probably explained by adaptive and mechanistic forces acting on the origination and insertion of new genes. Several models have been proposed to explain the specific roles of each force, and their relative importance is still debated (summarized in

Long *et al.*, 2012). Among the plausible models for explaining why new genes have a preferential movement across the genome, we find the sexual antagonism hypothesis, which predicts that dominant beneficial mutations to females but detrimental to males have higher probability of fixation by natural selection on the X chromosome, while recessive ones have higher probability of fixation on the autosomes. Conversely, dominant and recessive mutations that are beneficial to males but detrimental to females have higher probability of fixation on the autosomes and on the X chromosome, respectively (Charlesworth et al., 1987; Rice, 1984). Other, more mechanistic, explanations are the meiotic sex chromosome inactivation (Vibranovski, 2014; Vibranovski et al., 2009b) and the dosage compensation models (Vicoso and Charlesworth, 2009). Both hypotheses suggest that processes occurring in the X chromosome reduce the expression of male-biased genes (such as meiotic sex-chromosome inactivation and dosage compensation) and, as a consequence, natural selection favors the relocation of these genes to autosomes. Finally, a recently proposed hypothesis (Díaz-Castillo and Ranz, 2012) suggests that the accumulation of testis-expressed retrogenes in *Drosophila* autosomes can also be explained by the nuclear organization of chromosome domains in the male germ line. They argue that the preferential integration of male-expressed retrogenes on autosomes is caused by the increased accessibility of open chromatin domains with testis expression during spermatogenesis, due to distinct chromosome interactions with the nuclear lamin. The lower fraction of accessible domains on the X chromosome in the male germ line would also explain the paucity of retrogenes in this chromosome. This model illustrates how data regarding gene locations can be combined with structural information to provide new explanatory hypotheses.

There is still intense debate regarding the limitations and explanatory power of each model. The general patterns of retrogene movement are probably caused by a combination of factors, each one promoted by a different model. Further understanding will certainly emerge from the collection of more genomic, expression and functional data. In any case, these findings show how basic structural information (in this case, comparison of the chromosomal distribution of genes and their parents) can be connected to genomic data to reveal compelling patterns. We believe that future investigations connecting gene distribution to other chromosome structures (such as chromatin states, chromosomal domains and epigenetic markers) may provide exciting evolutionary insights.

Gene regulatory networks

Gene regulation is a key process for proper cellular function. Spatial and temporal transcriptional activation of genes is determined by complex and extensive regulatory networks, often involving signaling cascades, protein complexes assembly and transcription factors binding to gene regulatory motifs. Recent large-scale genomic analyses have described general patterns regarding the intricate regulatory networks from some model organisms (Mackay, 2014), which allows us to investigate the number of functional elements, connectivity and logic of networks topologies. However, little is known about how regulatory networks actually evolve at a high resolution, how they can be rewired and expanded, and how their constituent elements adapt in concert. A good way to explore that is by investigating how the addition of new elements impact preexisting networks. Next, we will illustrate how the recent integration of gene duplicates provides insights about network evolution.

In one example, the gene *Zeus* (*Drcd-1r*) originated through retrotransposition from an X-linked parental gene, *CAF40* (*Drcd-1*), in the ancestor of *D. melanogaster*, *D. simulans*, *D. mauritiana* and *D. sechellia*, around 4-6 million years ago (Quezada-Díaz et al., 2010). *CAF40* encodes a highly conserved DNA-binding protein and takes part in a transcriptional and post-transcriptional regulatory complex. Interestingly, Zeus protein sequence underwent intense divergence from its parent after duplication, whereas the parental copy remained essentially conserved (Chen et al., 2012). Also, *Zeus* regulation also diverged, acquiring testis-expression, and being subject to histone modifications (Arthur et al., 2014), while the parental *Caf40* remained ubiquitously expressed. Using chromatin immunoprecipitation followed by microarray assays (ChIP-chip), Chen et al. (2012) demonstrated that *Zeus* protein binding sites across the genome have diverged to large extent from the parental gene, promoting the upregulation of male-biased genes and downregulation of female-biased genes. Furthermore, *Zeus* knockout shows a dramatic negative effect on male fertility, evidencing the functional relevance of *Zeus* integration into the regulatory network.

A similar pattern was investigated with another regulator recently duplicated in the *D. melanogaster* species complex. *Nsr* (novel spermatogenesis regulator) originated through DNA duplication from the *kep1* locus in the same phylogenetic branch of *Zeus*, and exhibits evidence of positive selection. The parental gene encodes a conserved RNA-binding regulator, with ubiquitous expression, whereas *nsr* expression is restricted to testis. Alike the previous example, *nsr* knockout also decreases male fertility. Detailed electron microscope revealed that *nsr* knockout results in loss of the outer dynein arms of the sperm axoneme, generating abnormal sperm structures (Ding et al., 2010). More

interesting, the authors suggest that Nsr functional divergence may be related to its subcellular localization, which is much broader than that of Kep1 protein. Therefore, according to their hypothesis, *nsr* acquired new male reproductive function through the regulation of mRNA of pre-existing genes required for spermatogenesis. Curiously, the sperm maturation processes affected here are very conserved during fly evolution.

Both examples mentioned above exemplify how new genes can be integrated into pre-existing regulatory networks and acquire important functions. In this process, the new gene can substantially diverge from the parental locus, changing its interaction partners, expression pattern and subcellular localization. This integration can impact the overall network architecture, and promote its rewiring, which can explain, for example, why some new genes quickly acquired crucial functions in *Drosophila* and even became essential for development and other conserved processes, despite their recent origination (Chen et al., 2010; Kemkemer and Long, 2014).

Cell division

Another cellular process that, despite being highly fundamental and evolutionary conserved, may also be impacted by new gene recruitment is cell division. The intricate cellular mechanisms driving the accurate distribution of a duplicated genome during cell division are critical to ensure proper cellular viability and development. Chromosome alignment and segregation require the interaction between kinetochore proteins associated with the centromere and the spindle microtubules, usually involving conserved protein complexes (Cheeseman and Desai, 2008).

In a detailed investigation, Ross *et al.* (2013) showed that the new duplicated gene *Umbrea*, originated through DNA duplication 12-15 million years ago in *Drosophila*, evolved an essential role for chromosome dynamics during mitotic division. Comparing sequences from a range of species and manipulating their protein structure, they were able to reconstruct the stepwise trajectory of *Umbrea* evolution. After duplication from the parental *HP1B* (which encodes a heterochromatin protein involved in gene regulation), *Umbrea* divergence involved the loss of an ancestral heterochromatin-localizing domain, and other changes that rewired its protein interaction network and led to centromere localization. Surprisingly, whereas the parental *HP1B* gene is dispensable for viability, *Umbrea* developed an essential function, as demonstrated by the high proportion of mitotic errors found in knockdown cells (Chen et al., 2010; Ross et al., 2013).

The authors propose that recurrent changes at centromeric regions may explain the recruitment of new duplicated genes coding for centromeric proteins. *Umbrea* constitutes a notable example of how new genes can be integrated into preexisting cellular networks and impact its overall architecture, even in fundamental cellular processes, such as chromosome segregation.

Protein traffic

As a final example, we will discuss a study case that challenges our concepts of conservation and homology of subcellular components. Many complex cellular structures, such as organelles, are found in phylogenetic distant groups of organisms. The parsimonious interpretation is that those shared features are homologous, i.e. they are ancient structures inherited from a common ancestor. There are at least some cases,

however, in which the finding of complex structures in distant organisms is better explained by independent convergent evolution, rather than by shared inheritance.

The hypothesis of independent origins was explored with specialized secretory vesicles called dense core granules (DCGs). These vesicles accumulate in the cytoplasm and function as storage reservoirs, until extracellular signals cause their fusion with the plasma membrane and the extracellular release of their content. Because DCGs are found in a wide range of eukaryotes (e.g. mammals, fungi, protozoans), the prevailing hypothesis was that they were already present in the early eukaryotic ancestor. However, analyses of the molecular composition of DCGs in mammals and ciliates (groups that diverged more than 1 billion years ago) failed to find homologous proteins, and some molecular mechanisms involved in DCGs processes are fundamentally different in these groups. Such observations support the idea that these secretory granules had independent origins in distinct phylogenetic lineages (Elde et al., 2007).

A supplementary example comes from clathrin proteins involved in endocytosis. This process is highly conserved in eukaryotes, and there is strong evidence for the shared inheritance of the underlying genes. Nevertheless, recent duplications of clathrin genes during chordate evolution generated distinct isoforms with tissue-specific function (Elde et al., 2007). This instance illustrates how the inheritance of ancient genes, combined with lineage-specific duplications, can lead to functional divergence and adaptation. The possibility that other common cellular features are the result of convergent processes, rather than representing shared inheritance, provides a stimulating alternative view of the evolutionary history of complex structures.

Perspectives

The recruitment of newly originated genes to play a role in fundamental cellular processes may be a general and recurrent theme in evolution, rather than represent exceptional events. Table 2.2 summarizes several new genes documented in *Drosophila* with basic cellular functions. Note that those genes originated in diverse phylogenetic branches in the *Drosophila* history, ranging from 0 to 25 million years ago (for the sake of comparison, this upper limit corresponds to the time that separates humans from our primate cousins *Rhesus*, a relatively recent event in evolutionary terms). In the near future, the collection of additional complete genomes will improve our ability to discover and date the origin of genes more precisely. Allied to experimental molecular biology techniques, this knowledge will allow a mechanistic understanding of the functional evolution of new genes, and of biological functions in general. We believe that future cases of novel genes implicated in basic cellular processes will emphasize their relative importance for adaptation.

Table 2.2. Examples of new genes involved in basic cellular processes reported in *Drosophila* species.

Gene name	Gene age (MY)	Cellular process
jingwei	0-3	Alcohol metabolism
p24-related-2	0-3	Golgi vesicle transport
Cyp9f2	3-6	Intracellular oxidoreductase activity
Xcbp1	3-6	Chaperone
TfIIA-S-2	6-11	Transcription factor
Ribosomal protein S28a	6-11	Ribosomal structural constituent
methuselah-like 6	6-11	Membrane G-protein coupled receptor
ACXA	11-25	Intracellular signal transduction
Chemosensory protein B 93b	11-25	Detection of pheromones
Scavenger receptor class C, type III	11-25	Cellular internalization of foreign material
Lysozyme X	11-25	Lysozyme activity, defense against bacteria
Ribosomal protein L37b	11-25	Ribosomal structural constituent

The evolutionary relevance of new genetic elements also points to the appealing perspective of their significance in shaping human-specific traits, as well as contributing to diseases. Despite variable estimates, we can conservatively conclude that at least 300 genes are specific of the human lineage, and other 1000 are only found in primates (Zhang and Long, 2014). Strikingly, transcriptome analyses revealed that a large proportion of new genes are expressed in human brains during development when compared to mouse, and also suggested that those genes originated in the same evolutionary period during which the human neocortex was expanding (Zhang et al., 2011). Such findings demonstrate how genome-wide expression comparisons can be used to shed light on long-standing questions regarding human traits. Additionally, detailed experimental investigations have shown the implication of young genes in diseases (reviewed in Zhang et al., 2012). Some remarkable examples include genes related to tumors, such as the hominoid-specific *TBC1D3* (Wainszelbaum et al., 2008), which is involved in cellular signaling and trafficking; and the primate-specific *CT45A1*, which upregulates oncogenic and metastic genes (Shang et al., 2014). Once more, these examples illustrate how evolutionary and cellular approaches can be combined to explore relevant questions.

The finding that distantly related taxa share many common cellular processes, and even ancient genes, may give the false impression that basic functions, and the genes underlying them, are not affected by evolution, and remain unaltered at the molecular level. This view is not in accordance with our current view of evolution. The existence of deeply conserved processes does not imply that their underlying components have been static over evolutionary time. Cellular processes are highly dynamic, and so is their evolution. Even when a biological process is conserved across long evolutionary periods

(such as the cell division mechanisms mentioned above), the molecular components responsible for them may undergo a dynamic turnover. With these examples, we want to emphasize the idea that, far from providing mere anecdotal reports, the evolution of new genes can help to illuminate main evolutionary mechanisms and understand fundamental biological processes and structures.

Acknowledgements

I.M.V was supported by the Science without Borders scholarship (BEX18816/12-6), and M.L. was supported by U. S. National Institutes of Health (R01GM100768-01A1), National Science Foundation (NSF1051826) and Endowment of Edna K. Papazian Distinguished Professorship at the University of Chicago. We thank all lab members for discussions, and Marilia Lopes Justino for helping with the figure.

Chapter 3

The young retrogene *Poseidon* impacts gene expression through interactions with an ancient regulatory complex in *Drosophila*

The work described in this chapter is summarized in a manuscript under preparation as: Ventura IM, Xia S, Sgromo A, Izaurralde E, Long M. "The young retrogene Poseidon *impacts gene expression through interactions with an ancient regulatory complex in* Drosophila*".*

Abstract

The origination and divergence of new genes are a major source of genomic and functional novelty. It has been recently recognized that such young genes can take over fundamental functions in basic cellular processes. It is not clear, however, how newly duplicated genes are integrated into ancestral networks or to what extent they diverge from their parents at the functional level. Here, we describe the evolution of a relatively young gene only found in some *Drosophila* species (*CG2053*), which we named *Poseidon*. We compare its functional impact with that of its parent *CAF40*, an ancient gene involved in regulatory pathways, and *Zeus*, an even younger duplicate, which was shown to have a role in male germline gene regulation. We show that, after their origination through retrotransposition, both duplicates rapidly diverged at the sequence level, and exhibit distinct timing of expression in different testis cell types. Using knockdown and knockout assays, we show that the three genes have an importance for fly viability and male fertility. Moreover, our transcriptome analyses demonstrate that the three genes have a broad and distinct effect in the expression of hundreds of genes, with almost half of the differentially

expressed genes being perturbed exclusively by one paralog, but not the others. Finally, we show that Poseidon protein, but not Zeus, conserved CAF40 ability to interact with an ancient protein regulatory complex. Our results show how young elements can be integrated into conserved processes in different ways, illustrating the complex nature of evolution driven by new gene origination.

Introduction

The complex regulation of gene expression is essential for proper cell function. Accurate spatial and temporal transcriptional activation and repression, as well as post-transcriptional regulation, is determined by intricate regulatory networks, often involving extensive signaling processes and the recruitment of multi-protein regulatory complexes. Thus, describing the components of such regulatory circuits, as well as understanding their role in the evolution of gene regulation can extend our comprehension of how organisms adapt and diversify over time (Erwin and Davidson, 2009; Halfon, 2017; Wilson et al., 1977). Fundamental cellular functions, including basic regulatory processes common to distantly related organisms, are often assumed to be carried out by old conserved elements, whereas evolutionary young genes would be involved in more restrict, even dispensable, activities (Kondrashov, 2012; Miklos and Rubin, 1996). However, recent case studies have challenged this view, showing that young genes can be incorporated into ancestral regulatory networks, with major impact in the expression of numerous genes (Chen et al., 2012; Ding et al., 2010). The importance of such integration of new elements into fundamental cellular processes is illustrated by dramatic examples of young genes which were experimentally shown to have acquired indispensable roles in development or

reproduction in a short evolutionary time, even when present in a single species (Chen et al., 2010; Lee et al., 2018; Ross et al., 2013; Saleem et al., 2012; VanKuren and Long, 2018).

The incorporation of new elements, in particular evolutionary young genes, into ancestral regulatory networks remains elusive and underexplored (Abrusán, 2013; Zhang et al., 2015). In particular, little is known about the molecular mechanisms involved in the evolution of regulatory networks driven by recently evolved gene duplicates. For instance, it is not clear to what extent the regulatory role of a young gene diverges from that of the parental copy, as well as what specific cellular processes and phenotypes they impact. In this context, the detailed comparison of old genes and their young duplicated paralogs can shed light on the mechanisms leading to the integration of new elements into preexisting cellular processes.

Despite the fact that newly originated genes are integrated into diverse functions (Long et al., 2013), extensive comparative genomic analyses have reported an intriguing pattern: there is a strong excess of parental genes on the X chromosome that produced autosomal duplicated genes with specific expression in the male germline (Kaessmann et al., 2009). This pattern was confirmed for various organisms such as flies (Bai et al., 2007; Betrán et al., 2002; Dai et al., 2006; Zhang et al., 2010a), mosquitos (Toups and Hahn, 2010), and mammals (Carelli et al., 2016; Emerson et al., 2004). The preferential fixation of male-biased duplicated copies into autosomes likely reflects the fact that, during the meiotic stage, there is a suppression of expression of genes in the X chromosome. As a result, natural selection favors the fixation of autosome-inserted duplicated copies that escape the X chromosome and compensate for the expression of its parental gene (Vibranovski *et al.*, 2009, 2012; reviewed in Casola and Betrán, 2017). In addition, other

factors are thought to contribute to this scenario, such as the fact that the testis is a rapidly evolving organ, prone to the accumulation of new elements, consistent with intense sexual selection (Harrison et al., 2015); and that, during late spermatogenesis stages, the transcription of new gene copies is facilitated by the permissive chromatin state, which may facilitate the promiscuous transcription, insertion, and subsequent evolution of newly arisen genes (Kaessmann, 2010).

Consistent with the pattern described above, previous studies have reported that male reproductive tissues express specific versions of several housekeeping genes involved in basic cellular processes, such as the proteasome (Zhong and Belote, 2007), transcription (Hiller, 2004), and translational machineries (Baker and Fuller, 2007). It was suggested that these duplicated copies may represent specialized versions of their parental genes, required to accomplish the intense and coordinated changes in gene expression observed during spermatogenesis (White-Cooper, 2010). It is not clear, however, why the duplicated, specific copies are so abundant, or to what extent they diverge from their parental ones (Belote and Zhong, 2009).

In this study, we investigate the evolution and functional impact of *Poseidon* and *Zeus*, two relatively young genes expressed in the testes, which independently duplicated from *CAF40*, and are only found in some *Drosophila* species (Zhang et al., 2010a). *CAF40* is an ancient gene, broadly expressed in all fly tissues. The locus encodes a highly conserved protein in eukaryotes, with orthologs identified and experimentally studied from mammals (e.g. mouse and human) to insects (e.g. *Drosophila*) to fungi (e.g. yeast) (Collart and Panasenko, 2017). *CAF40* is a core member of the also conserved CCR4-NOT regulatory complex, a multi-protein assembly involved in transcriptional, post-transcriptional and

translational regulation (Miller and Reese, 2012). The complex promotes the deadenylation and degradation of mRNA targets, a critical regulatory process for proper spermatogenesis (Legrand and Hobbs, 2018). By integrating several regulatory processes, the complex is considered a key regulator of eukaryotic gene expression (Collart, 2016). CAF40 function in the complex involves the recruitment and stabilization of the protein assembly, and it also acts independently of the complex, interacting with transcription factors and altering their activation potential (Garapaty et al., 2008).

Zeus was already shown to play a role in male fertility in *Drosophila*, by binding and regulating the expression of a large set of target genes, many not shared with the parental *CAF40* (Chen et al., 2012). Here, we first describe the divergence between *CAF40* and its two retroduplicates, *Poseidon* and *Zeus*, in gene sequence and expression patterns. Second, using RNAi-knockdown and CRISPR-Cas9-deletions, we further explored their phenotypic importance for viability and male fertility. Third, we demonstrate that Poseidon protein, but not Zeus, retained the ability to interact with the CCR4-NOT complex. Finally, our RNA-seq data demonstrate that the independent silencing of each paralog impacts the regulation of a distinct set of genes, likely due to a misbalance between regulatory processes in which the paralogs are integrated to. Together, our data show that both young duplicates play a role in *Drosophila* male germline regulatory pathways, interact with highly conserved regulatory mechanisms, and impact the gene expression network in different ways.

Results

Below, we describe the diverse computational and experimental assays we employed to investigate *Poseidon* and *Zeus* evolution and function, and to what extent they

diverge from those of the parental gene *CAF40*. Briefly, we explore their: i) origination process and sequence divergence; ii) expression patterns; iii) phenotypic impact for viability and male fertility; iv) interaction with the CCR4-NOT complex; and v) impact on global gene expression.

Poseidon origination and rapid evolution

Comparative genomic analyses from public databases had previously identified *Zeus* as a diverged copy of *CAF40* (Zhang et al., 2010a), which prompted its functional description as a duplicated gene (Chen et al., 2012; Quezada-Díaz et al., 2010). Curiously, such analyses had also revealed the presence of another annotated related gene in flies, although it had not been studied until now (*CG2053* in *D. melanogaster*, which we named *Poseidon*, as a reference to Zeus' brother in the Greek mythology).

Poseidon intact ORFs are present in the third chromosome of 18 *Drosophila* species (Fig. 3.1A), all from the subgenus *Sophophora*. Reciprocal BLAST searches did not find any other significant match in eukaryotes besides *CAF40* and *Zeus* orthologs (e.g. PSI-BLAST E-value = $7x10^{-56}$, coverage = 81% between CAF40 and Poseidon from *D. melanogaster*), which suggests this is a relatively young gene only present in some fly species.

The presence of introns is useful for determining the mechanisms of origination of new genes. *CAF40* orthologs have between 4 and 6 introns in *Drosophila* species. *Poseidon*, on the other hand, only exhibits one small intron in its 3' end in some species, which is unrelated in sequence or position to any intron found in *CAF40*. The lack of the ancestral introns in the duplicated gene suggests that *Poseidon* likely originated through an X to

autosome retrotransposition event, with the insertion of the duplicate in the third chromosome, and subsequent gain of a new intron (Fig. 3.1B).

Our phylogenetic analyses support the hypothesis that *Poseidon* and *Zeus* originated through two independent duplication events from *CAF40* in different branches of the *Drosophila* phylogeny (Fig. 3.1C). *Zeus* phylogenetic position is consistent with its retroduplication after the split of the most recent common ancestor of *D. melanogaster* and *D. yakuba* (6-11 MYA), as previously reported (Quezada-Díaz et al., 2010).

Our phylogenetic analyses strongly support the hypothesis that *Poseidon* duplication event happened after the split of the common ancestor of *Drosophila* and other Diptera (e.g. genera *Culex* and *Musca* in Fig. 3.1), around 132 MYA, which is in agreement with the fact that this gene is not found in any species outside the genus *Drosophila* (Fig. 3.1A). However, *Poseidon* exact origination period within *Drosophila* is harder to precise, because its phylogenetic distribution among species slightly conflicts with its estimated phylogenetic position. On one side, our phylogeny suggests that the duplication happened before the split of the *Sophophora* and *Drosophila* subgenera (50 MYA), although the support for this node is not very strong (Fig. 3.1C). The duplication event in that branch would require the loss of *Poseidon* in flies from the *Drosophila* subgenus clade, since it is not present in any of the four species with available genomic data (*D. virilis, D. mojavensis, D. albomicans, D. grimshawi*; Fig. 3.1A). Alternatively, the presence of *Poseidon* in 18 species from the *Sophophora* clade but absence in species from the *Drosophila* subgenus suggest that the duplication might have occurred after the split of the two subgenera, in the common ancestor of *D. melanogaster* and *D. willistoni* (50 MYA). In this case, the early split of *Poseidon* and *CAF40* in our estimated phylogeny might be the result of an artifact, such as

long-branch attraction (Bergsten, 2005), since *Poseidon* sequences are rapidly diverging compared to the parental *CAF40*.

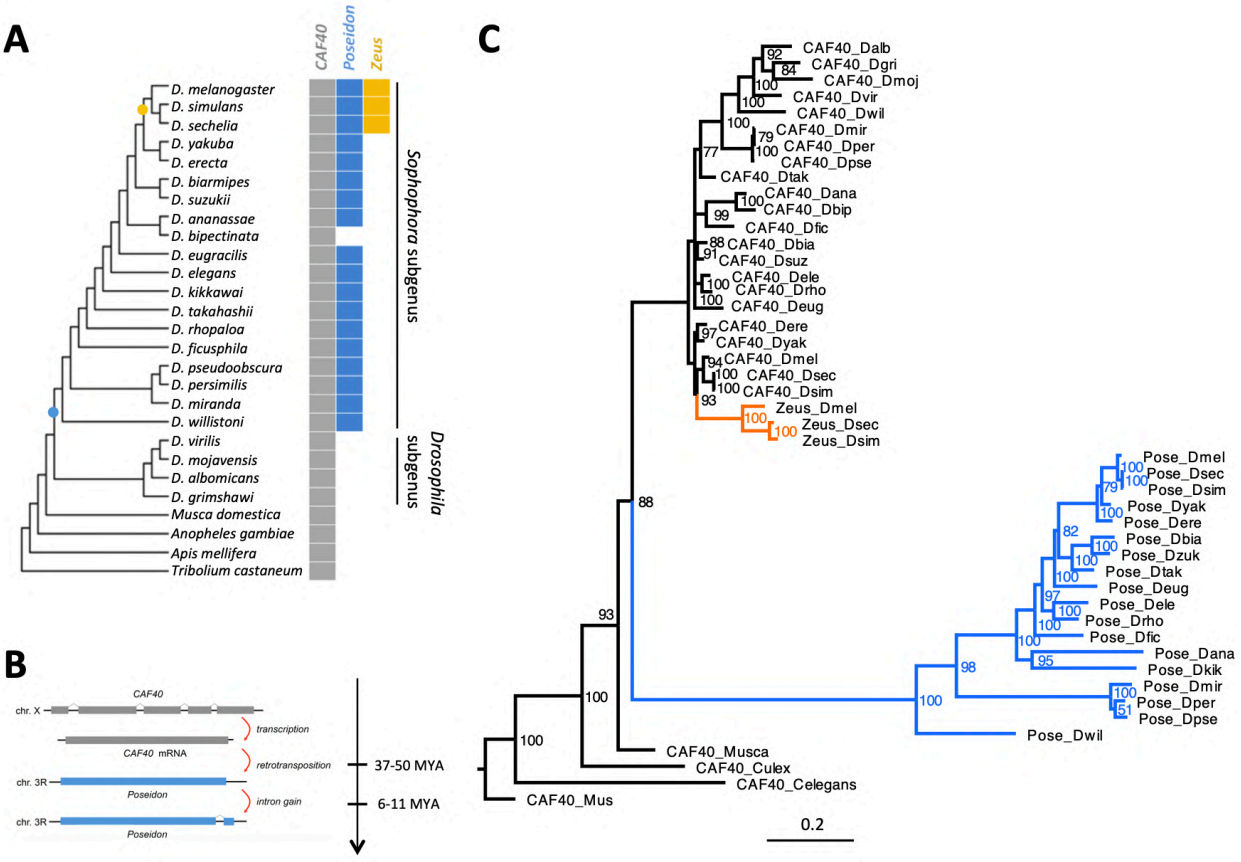


Figure 3.1. Origination of *Poseidon* and *Zeus* from *CAF40* in *Drosophila.* A) Distribution of the three paralogs among *Drosophila* species, with other insects as outgroups. Filled boxes represent the presence of each gene. B) Scheme depicting *Poseidon* origination process through retroduplication, and insertion into an autosome. C) Phylogenetic relationship among the three paralogs reconstructed through Bayesian method (*CAF40* homologs in black, *Poseidon* in blue, and *Zeus* in orange; numbers represent the Bayesian posterior probability of each node).

Regardless of the exact time of the duplication event, our analyses support the conclusion that *Poseidon* and *Zeus* are two independent, relatively young duplicates from the same parental gene *CAF40*, and are only present in some *Drosophila* species. Also, they

show that both duplicates have substantially diverged at the sequence level from *CAF40*, which remained highly conserved. As an illustration, CAF40 protein sequences from *D. melanogaster* and *D. willistoni* (which split around 50 MYA) diverged in only 2.3% of the sites, whereas Poseidon and Zeus diverged 56.8% and 30.6%, respectively, from CAF40 in *D. melanogaster*.

The duplicates diverged at highly conserved sites

In order to understand whether the duplicated proteins accumulated replacements at conserved residues in the ancestral protein, or merely at the highly variable termini of the protein, we estimated the Shannon entropy (H) for each residue in an alignment of CAF40 homologs from 56 eukaryotes, and contrasted it with the replaced residues in the duplicates (Fig. S3.1). We found that amino acid replacements in Poseidon and Zeus occurred even at extremely conserved sites of CAF40 (Fig. S3.2 depicts replacements mapped onto the parental protein structure). In both duplicates, replacements are distributed throughout the protein structure, including the charged groove formed by the conserved armadillo-repeat domain (Garces et al., 2007), which was shown to be important for CAF40 interactions (Chen et al. 2014). For the sake of comparison, out 49 residues that are completely conserved in CAF40 among eukaryotes (H=0; Fig. S3.1), Poseidon diverged in 25, and Zeus, in 11 of them. Furthermore, several residues that were experimentally shown to be functionally relevant for CAF40 interaction with its protein partners in previous studies were replaced in the duplicates (Table S3.1; Chen et al. 2014; Mathys et al. 2014; Sgromo et al. 2017). The extensive replacement of amino acids that are highly

constrained in the parental protein suggests that Poseidon and Zeus biochemical properties may have diverged substantially from CAF40.

The duplicates acquired a restrict expression pattern

Part of the phenotypic divergence between duplicated genes and their parents may result from their differential expression pattern. We used extensive transcriptome data from publicly available databases to investigate to what extent *Poseidon* and *Zeus* diverged from *CAF40* at the expression level. First, a comparison of expression of these genes in several tissues from *D. melanogaster* evidences that both *Poseidon* and *Zeus* have acquired a narrower expression pattern when compared to *CAF40* (summarized in Fig. 3.2A). The duplicates are only expressed at larval imaginal discs, adult male reproductive tissues, and pupae at low or intermediate levels, in sharp contrast to *CAF40*, which is expressed at all assayed tissues and development times, from intermediate to high levels. We confirmed the distinct expression of the three paralogs in *D. melanogaster* larvae, head, and testis through RNA extraction followed by RT-PCR (Fig. S3.3).

We also compared the expression of each paralog at different testis cell types in *D. melanogaster*. Taking advantage of a previous study that isolated cells from different spermatogenesis phases (mitosis, meiosis, post-meiosis) (Vibranovski et al., 2009b), we show that the three paralogs have distinct expression dynamics in that organ (Fig. 3.2B). While *CAF40* expression levels are consistently high across the three phases, with the lowest expression in meiotic cells, *Poseidon* exhibits a peak of expression in the meiotic stage, a common pattern for autosomal retrogenes expressed in the testis (Vibranovski et al., 2009b). *Zeus* is even more variable, with a strong peak of expression in the early mitotic

phase, and subsequent drop of expression in the meiotic and post-meiotic stages. Fast divergence in the expression pattern is often observed for retrogenes, which are inserted in genomic contexts diverse from the parental copy, and may promptly acquire and/or evolve new cis-regulatory elements (Bai et al., 2008). The distinct expression dynamics of each paralog suggests that they may not be functionally redundant, even when expressed in the same male reproductive organ. Their phenotypic impact for male fertility is addressed below.

Finally, we analyzed transcriptome data from other six *Drosophila* species in order to understand whether the duplicates male-biased expression pattern is conserved across the phylogeny. For all the additional six species, *CAF40* is expressed at intermediate or high levels in adults, consistent with its role as a housekeeping gene. In contrast, all the other species exhibit significant male-biased expression of the duplicated genes (four additional species for *Poseidon*, and one for *Zeus*), similar to the pattern observed in *D. melanogaster* (Fig. 3.2C).

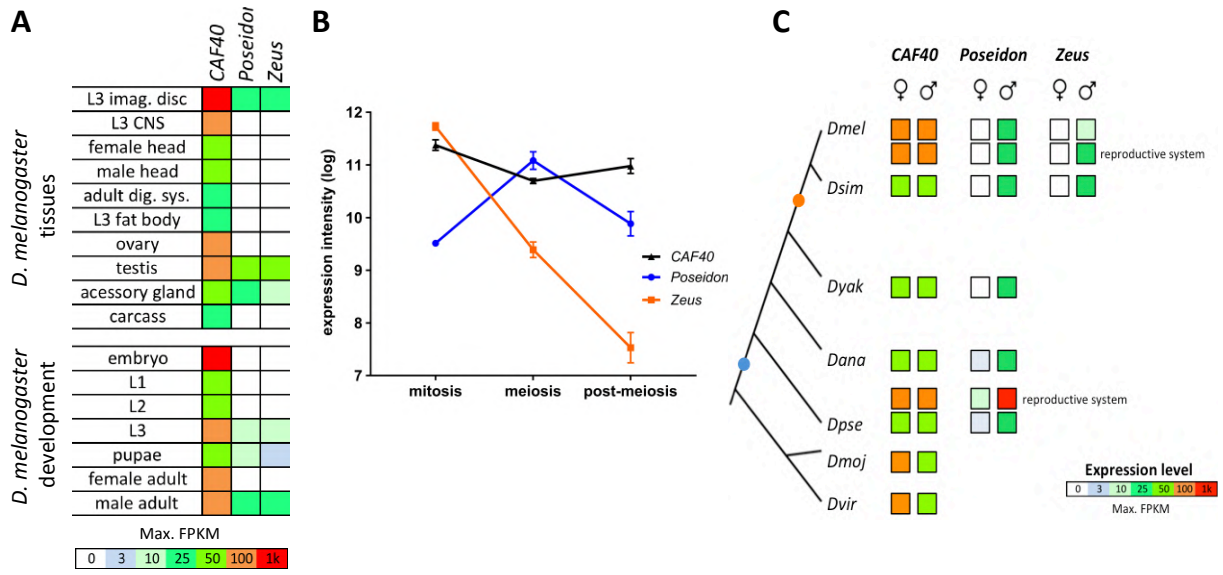


Figure 3.2. *CAF40*, *Poseidon* and *Zeus* have distinct expression patterns. A) Summary of the expression intensity of each paralog in *D. melanogaster* tissues and development times. Expression summarized as average FPKM. *CAF40* broad expression contrasts with the duplicates restricted expression pattern. B) Expression level at three different cell types in *D. melanogaster* testes (mitosis, meiosis, post-meiosis). Error bars indicate SD. The expression dynamics of each paralog diverges according to the cell type. C) Summary of expression intensity in female/male adults for *Drosophila* species with available data for each paralog. Circles in the phylogeny represent the duplication event for *Poseidon* (blue) and *Zeus* (orange). Male-biased expression for *Poseidon* and *Zeus* are conserved across the phylogeny. Data extracted from (Brown et al., 2014; VanKuren and Vibranovski, 2014; Vibranovski et al., 2009b).

Poseidon and *Zeus* impact viability and fertility

The duplicates restrict pattern of expression, along with their conserved sex-specific expression across fly species, suggest that the duplicates may have been integrated into developmental and/or reproductive processes. First, we tested this hypothesis by using RNAi-mediated knockdown and CRISPR-based knockout to assay the importance of the three paralogs on *D. melanogaster* egg to adult viability.

RNAi knockdown using both a ubiquitous (*Tub84B*>GAL4) and an imaginal disc-specific driver (*T80*>GAL4) confirmed that *CAF40* expression is essential for fly survival (less than 3% of the flies developed into adults when the gene was silenced with the

ubiquitous driver, t-test, $p < 0.0001$ for the comparison with the control; Fig. 3.3A). *CAF40* essential role in cellular processes is also observed in distant eukaryotes, as evidenced by knockout experiments in *C. elegans* (Kamath et al., 2003) and humans (Wang et al., 2015).

Both knockdown and knockout assays evidenced that *Poseidon* and *Zeus* also have a significant phenotypic impact on fly viability, although less dramatic than the parental gene. When RNAi-silenced, these genes reduced the relative fly viability in around 20%, with a slightly stronger effect of the ubiquitous driver over the imaginal disc one (t-test, $p < 0.05$ for all the comparisons with the controls; Fig. 3.3A). Such a difference could be due to a higher efficiency of the ubiquitous driver, or to the fact that the genes may be expressed at tissues other than the imaginal discs, such as the pupae stage. A similar negative impact on fly viability (reduction of ~25% compared to the control) was achieved when the genes were knocked out, as evidenced by the reduced viability of lines homozygous to two independent frameshift deletions for each of the two genes (t-test, $p < 0.05$ for both genes compared to the control; Fig. 3.3B). The similar effect achieved with knockdown and knockout assays are consistent with the high efficiency of the RNAi strategy in lowering the expression level of the three genes, as measured through RT-PCR (Fig. S3.4).

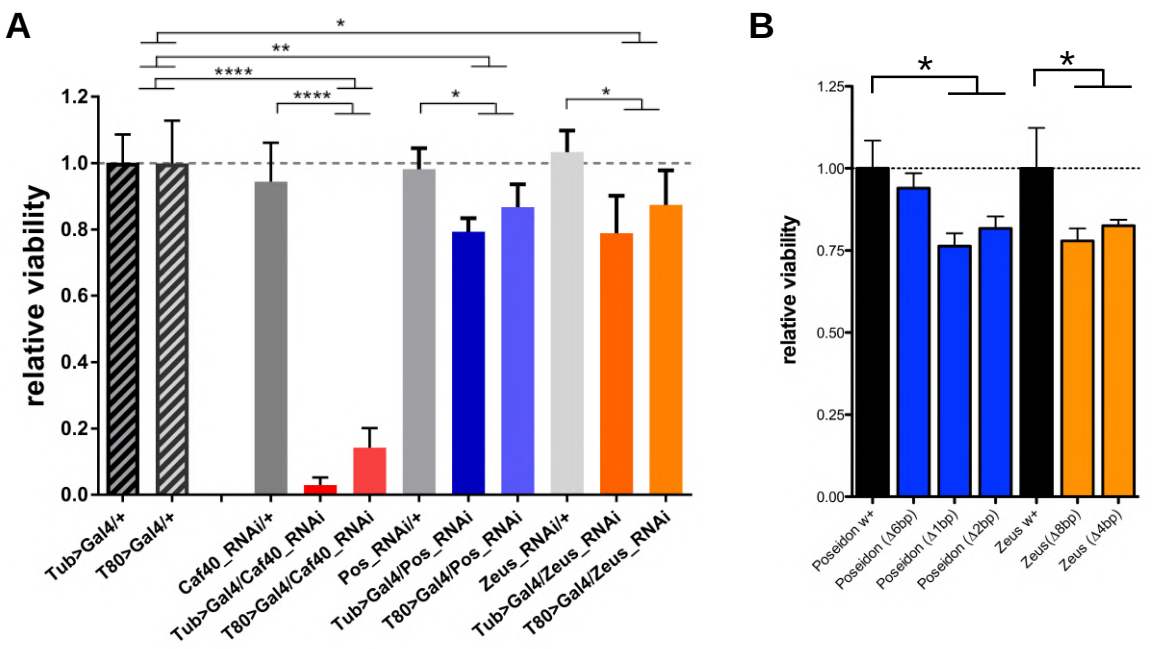


Figure 3.3. *CAF40*, *Poseidon* and *Zeus* impact on viability. A) Viability of RNAi-expressing flies, relative to the control genotype from each individual cross. The two hatched bars in the left show the relative viability of the two control GAL4-expressing drivers, *Tub* (ubiquitous) and *T80* (imaginal-discs). *CAF40*-RNAi-expressing flies in red; *Poseidon* in blue and *Zeus* in orange. B) Viability of homozygous flies for different CRISPR-Cas9 deletions for *Poseidon* (blue) and *Zeus* (orange) relative to the control from the same background strain (black). Notice the non-significant effect of the only non-frameshift deletion for *Poseidon* (first blue bar on the left); t-test: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$. Error bars indicate SD.

Given the conserved expression of the duplicates on the testes of several fly species, we also assayed the effect of silencing each gene on male fertility. We used RNAi-knockdown with two different testis-specific drivers: *nanos*>GAL4 (typically expressed at spermatogonia and male germline stem cells), and *Bam*>GAL4 (expressed at late spermatogonia and early spermatocytes stages), which allowed us to assay the genes importance at different spermatogenesis phases (White-Cooper, 2012). Independent silencing of the three genes had a negative impact on male fertility (Fig. 3.4). *CAF40* knockdown had the strongest effect (drop of ~45% on relative fertility using *nanos*; t-test p

< 0.0001), in accordance with its function as a fundamental housekeeping gene. *Poseidon* strongest effect was observed with the *Bam* driver (reduction of ~39% on relative fertility, t-test p < 0.001), although the effect is also significant with *nanos* (drop of ~28% on relative fertility; t-test p < 0.001). The effect of *Zeus* knockdown, on the other hand, was only significant when the *nanos* driver was employed (drop of ~36% on relative fertility; t-test p < 0.0001), but not the Bam-driver (t-test p = 0.59). These results are consistent with the timing of expression of each paralog at different spermatogenesis stages mentioned before (Fig. 3.2B). In particular, *Zeus* effect is only observed when the gene is knocked down at early stages, consistent with its peak of expression in mitotic cells. Finally, *Poseidon* and *Zeus* impact on male fertility was also confirmed using CRISPR-generated frameshift deletions, reducing relative fertility in around 17% and 32%, respectively, compared to the control from the same background strain (t-test p < 0.05 for both comparisons; Fig. 3.4C).

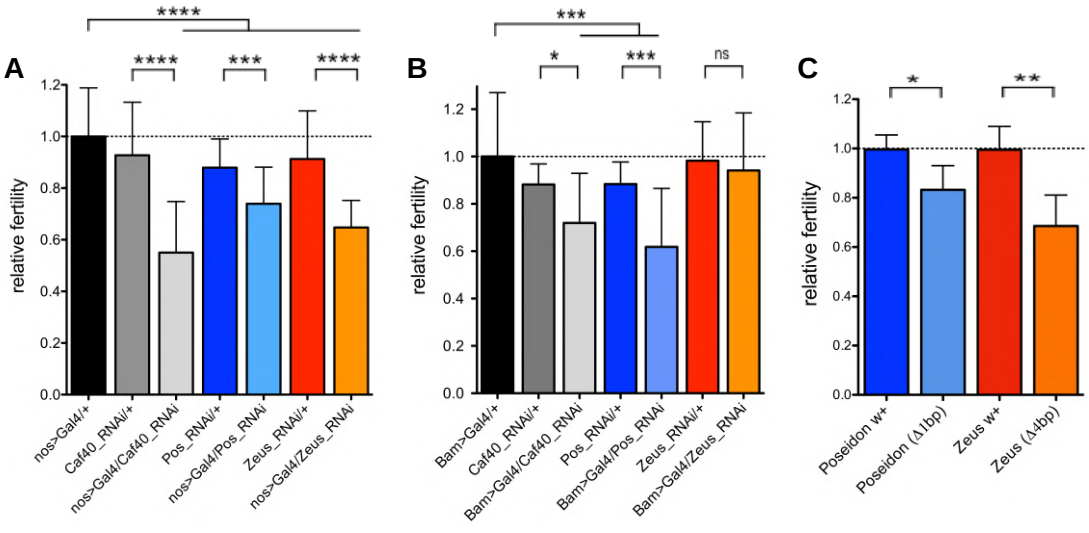


Figure 3.4. *CAF40*, *Poseidon* and *Zeus* impact on male fertility. Relative fertility of males expressing RNAi using testis-specific drivers: A) *nanos*>GAL4 – early

(Figure 3.4, continued) spermatogenesis; and B) *Bam*>GAL4 – late spermatogenesis. Black bars represent the controls, *CAF40*-RNAi-expressing flies in grey, *Poseidon* in blue, *Zeus* in orange. C) Fertility of males homozygous for frameshift CRISPR-Cas9 deletions for *Poseidon* (blue) and *Zeus* (orange) compared to the control of the same background strain (darker bars). t-test: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$. Error bars indicate SD.

Poseidon and Zeus interaction with the CCR4-NOT complex

Given CAF40 conserved role as a core member of the CCR4-NOT regulatory complex in eukaryotes, we were tempted to test whether the two duplicates maintained this ancestral function, despite their intense protein sequence divergence. First, we independently expressed a GFP-tagged version of each paralog in Dm S2 cells, and assayed their interaction with HA-tagged NOT1 through co-immunoprecipitation followed by Western blotting analysis. NOT1 was selected because it is the only known interaction partner of CAF40 in the CCR4-NOT complex (Collart and Panasenko, 2017). We found that Poseidon conserved the ability to interact with NOT1, whereas Zeus either lost it, or binds NOT1 only weakly (Fig. 3.5A).

Second, we also tested the paralogs ability to interact with two additional post-transcriptional regulators, Bam and Roquin, which were recently shown to recruit the CCR4-NOT complex in *D. melanogaster* through direct interactions with CAF40 (Sgromo et al., 2017, 2018). Again, Poseidon retained the ancestral binding ability, in contrast to Zeus, which failed to bind to either ancestral partner (Fig. 3.5B-C).

The conservation of ancestral interaction with NOT1 observed for Poseidon suggests that it should be able to recruit the CCR4-NOT regulatory complex, in contrast to Zeus. If this is true, Poseidon should have conserved the repressive effect on targeted mRNAs extensively observed for CAF40 (Sgromo et al., 2017). We tested this hypothesis by

measuring each paralog ability to repress a luciferase reporter mRNA in a λN-tethering assay. In agreement with previous reports, our assays show that CAF40 promotes a strong repression of the tethered reporter activity, compared to the control, consisting of a similar reporter transcript lacking the binding sites for the λN-tagged protein (Fig. 3.5D). Accordingly, Poseidon is also able to reduce the reporter expression to similar levels (around 10% of the control level, Fig. 3.5D). In contrast, Zeus exhibits a weaker repressive ability compared to the parental protein, although it is clearly significant (around 35% of the control level).

Taken together, these results suggest that Poseidon conserved CAF40 ancestral ability to interact with the CCR4-NOT complex through interactions with NOT1, as well as recruit two known regulatory partners of the complex, leading to the repression of targeted transcripts. Zeus, on the other hand, had its CCR4-NOT recruitment ability lost or weakened. Nevertheless, the fact that Zeus is still able to promote the reporter repression to a lesser extent suggests that it either evolved new protein interactions involved in mRNA repression, or conserved other ancestral, not yet described, CAF40 protein interactions related to this activity.

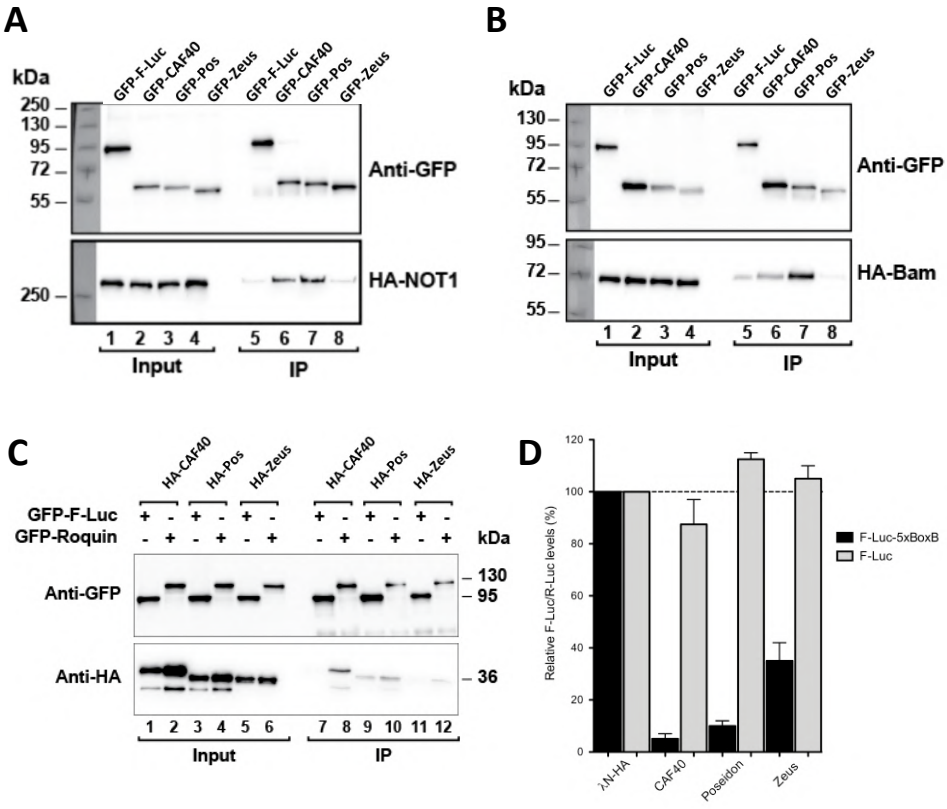


Figure 3.5. CAF40, Poseidon and Zeus protein interaction with the CCR4-NOT complex. Western blotting assaying the interaction between each paralog and three different protein partners, according to co-immunoprecipitation and pull-down assays in Dm S2 cells. In all panels, GFP-F-Luc served as a negative control. Input samples consist of 3% for the GFP-tagged proteins and 1% for the HA-tagged proteins, and immunoprecipitate samples of 10% for the GFP-tagged proteins and 30% for the HA-tagged proteins. Protein size markers are shown on the right in each panel. A) Cells expressing GFP-tagged versions of CAF40, Poseidon and Zeus, and HA-tagged NOT1. Notice the positive bands for the three paralogs in the input sample (left), and positive bands for HA-NOT1 in the co-immunoprecipitated sample for CAF40 and Poseidon, but not Zeus. B) Cells expressing GFP-CAF40, Poseidon and Zeus and HA-Bam. C) Cells expressing HA-CAF40, Poseidon and Zeus, and GFP-Roquin. D) Luciferase mRNA reporter repressing ability of each paralog assayed through a λN-tethering assay. A plasmid expressing R-Luc served as a transfection control, and an F-Luc reporter lacking the binding sites for λN-HA (BoxB) was used as control (grey bars). F-Luc activity levels were normalized to those of the R-Luc control and set to 100% in cells expressing the λN-HA peptide. Error bars indicate SD of five replicates. CAF40 and Poseidon exhibit similar abilities of repressing the luciferase reporter (black bar) compared to the control (grey bar). On the other hand, Zeus exhibits lower, though still significant, repression ability.

CAF40, *Poseidon* and *Zeus* impact on gene regulation

Given *CAF40* central role in several cellular regulatory processes, we asked what is the impact of the perturbation of the three paralogs for global gene expression. Moreover, since the two duplicates are highly diverged at their protein sequence and expression dynamics from the parent, they offer a good opportunity to assay to what extent the impact on gene regulation differs among the paralogs.

We conducted genome-wide transcriptome analysis of adult testes to assay the impact of *CAF40*, *Poseidon* and *Zeus* germline-specific knockdown on global gene expression, compared to controls from the same background strain. Our transcriptome data showed that RNAi-silencing was effective and specific against each paralog, reducing mRNA levels of each gene in at least 60% compared to the control, while not impacting the other paralogs (Table S3.2). Knocking down the three genes had a significant impact on the regulation of hundreds of genes. Figure 3.6A shows the number of genes whose expression level was up- or down-regulated more than 1.5-fold for each knockdown, compared to the control. At this threshold, the knockdowns affected the expression of 2,622 genes, which corresponds to more than a fifth of the genes mapped in our transcriptome (11,491) (Fig. S3.4-5). Such a widespread effect on gene expression suggests that *Poseidon* also conserved a role on gene regulation in male germline, as previously shown for *CAF40* and *Zeus* (Chen et al., 2012).

We also investigated to what extent the impact on gene regulation differs upon each paralog knockdown. A large set of genes had their expression affected by all the three knockdowns compared to the control. For instance, a common set of 670 genes was perturbed when the three paralogs were individually silenced, which corresponds to

25.5% of all differentially expressed genes at >1.5-fold (Fig. 3.6B). On the other hand, a substantial set of genes (46.3%) was perturbed by only one of the knockdowns, but not shared with the other two, which reveals the distinct impact that each paralog has in the global regulatory network.

Remarkably, our gene ontology analyses revealed that, among the differentially expressed genes, the most significant enrichment was found among the 670 perturbed genes shared by the three knockdown samples. Particularly, there was a high proportion of genes involved in catabolic functions (Table S3.3), such as serine-type endopeptidase activity (GO:0004252, adj. $p < 10^{-28}$), serine hydrolase activity (GO:0017171, adj. $p < 10^{-27}$), and catalytic activity (GO:0003824, adj. $p < 10^{-6}$). Accordingly, the cellular processes with the most significant enrichment were proteolysis (GO:0006508, adj. $p < 10^{-7}$) and reproduction (GO:0032504, adj. $p < 10^{-7}$) (Tables S3.3).

The number of perturbed genes shared with the parental *CAF40*-KD differed between *Poseidon*-KD and *Zeus*-KD. Among all the genes differentially expressed in the *Poseidon*-KD sample, 63.7% are shared with the *CAF40*-KD, whereas the same is true for only 53.3% of the genes affected by *Zeus*-KD (fig. 3.6C), which suggests that *Zeus* impact on gene regulation is more diverged than that of *Poseidon* ($\chi^2 = 50.7$, $p < 10^{-10}$ for the comparison of shared/not shared with *CAF40*-KD). Similar proportions are observed when only differentially expressed genes higher than 2- or 5-fold are considered (Fig. S3.7). For instance, 37.5% of >5-fold differentially expressed genes in *Zeus*-KD are exclusive (not shared with the other knockdowns), whereas the same is true for only 17.7% of the genes impacted by *Poseidon*-KD ($\chi^2 = 7.69$, $p < 0.01$; Fig. S3.7).

Despite the differences in the set of genes that are perturbed by each knockdown, for the vast majority of genes that were commonly affected by *CAF40* and one of the duplicates knockdown, the perturbation occurred in the same direction (i.e. up- or down-regulation) on both samples (Fig. 3.6D). Yet, once again *Zeus*-KD had a larger proportion of genes that were perturbed in the opposite direction of that of *CAF40*-KD when compared to *Poseidon*-KD (9.1% of perturbed genes in *Zeus*-KD versus 3.8% in *Poseidon*-KD were in the opposite direction of that of *CAF40*-KD, $\chi^2 = 21.3$, $p < 10^{-5}$ for the opposite/same direction as *CAF40*-KD comparison), consistent with previous observations by (Chen et al., 2012).

Finally, we classified the genes with perturbed expression compared to the controls in each knockdown sample as male or female-biased, based on two independent *Drosophila* databases (Assis et al., 2012; Gnad and Parsch, 2006). For all the three knockdowns, there was an enrichment of genes with male-biased expression among the differentially expressed sets, following the classification of both databases (χ^2 test, $p < 0.01$ for all the male/female-biased genes comparisons; Table S3.4). The impact of the three knockdowns on the regulation of male-biased expressed genes is in agreement with the importance of the three paralogs for male fertility, as described above.

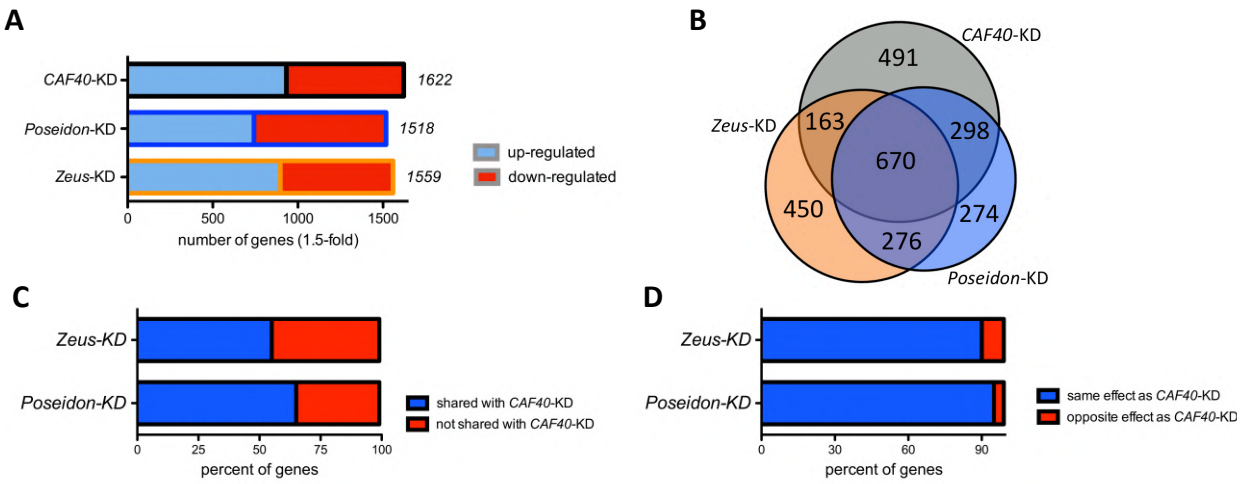


Figure 3.6. Impact of *CAF40*, *Poseidon* and *Zeus* knockdowns on global gene expression. A) Number of genes with >1.5-fold change in expression compared to the control in the knockdown of each paralog measured through RNA-seq. Up-regulated genes shown in blue, down-regulated shown in red. B) Venn diagram representing the number of genes differentially expressed upon the knockdown of each paralog (*CAF40* in grey, *Poseidon* in blue, *Zeus* in orange) and their intersections. Areas in the figure are proportional to the number of genes in each category. Notice the large number of genes (670) commonly affected by the three knockdowns, as well as genes exclusively perturbed by one paralog (1,215). C) Proportion of genes differentially expressed upon *Poseidon* and *Zeus* knockdown that are also affected by *CAF40* knockdown. Notice that *Poseidon* affects a higher proportion of genes shared with *CAF40* compared to *Zeus* (63.7% versus 53.3%; χ^2 = 50.7, $p < 10^{-10}$). D) Proportion of genes differentially expressed upon the knockdown of *Poseidon* and *Zeus* that are affected in the same direction (i.e. up- or down-regulated) in the *CAF40*-KD. Notice that, despite the fact that the large majority of genes are affected in the same direction by the duplicates and the parental knockdown, *Zeus* knockdown affects a larger proportion of genes in the opposite direction of that of *CAF40* when compared to *Poseidon* (9.1% of perturbed genes in *Zeus*-KD versus 3.8% in *Poseidon*-KD were in the opposite direction of *CAF40*-KD, χ^2 = 21.3, $p < 10^{-5}$).

Discussion

The analyses presented here allow us to investigate the evolution of two young genes after their independent retroduplication from the same parental gene in *Drosophila*. We showed that both *Poseidon* and *Zeus* are functional genes that have substantially diverged from *CAF40* at the protein sequence shortly after duplication, whereas the parental gene remained highly conserved (Fig. 3.1C). Furthermore, even residues that have

been conserved in CAF40 for a long evolutionary time (namely, amino acids common to all eukaryotic homologs in our alignment) were replaced in the duplicates, which may impact conserved functions of the protein.

Poseidon and *Zeus* have also diverged at the expression pattern in a similar fashion. Both genes have independently acquired restricted regulation, being expressed at low or intermediate levels during larval development and at male reproductive tissues, whereas *CAF40* remained broadly expressed. The duplicates male-biased expression is conserved throughout the *Drosophila* phylogeny (Fig. 3.2C), suggesting that both genes either quickly acquired expression in male reproductive tissues after duplication, which then remained conserved across species, or repeatedly evolved a similar expression pattern, which seems less likely (Bai et al., 2008; Sorourian et al., 2014).

Despite their overlapping expression in the testes, the three paralogs have distinct timing of expression in the cell types of that organ. We showed that *Poseidon* expression peaks at meiotic cells, the same stages in which *CAF40* has its lowest expression (Fig. 3.2B). Such dynamics was extensively reported for X-linked parental genes and their autosome duplicated copies, and suggests an expression compensation between the paralogs due to the inactivation of genes on the X chromosome during meiosis (Casola and Betrán, 2017; Vibranovski et al., 2009b). On the other hand, *Zeus* peak of expression happens at early spermatogenesis stages (mitotic cells), and drops during and after meiosis. Such difference in expression is consistent with our fertility experiments, which showed that *Zeus* negative effect on male fertility is only significant when knocked down with an early germline driver, but not with a mid-spermatogenesis one (Fig. 3.4). In contrast, *CAF40* and *Poseidon* knockdown effects are observed at both early and mid-stages. Therefore, despite the

overlap of expression of the three paralogs in the testes, they exhibit distinct dynamics in different cell types, and may be required for different moments of spermatogenesis.

Likewise, our functional assays also demonstrated that *Poseidon* and *Zeus* have an importance for egg to adult viability, consistent with their expression during larval development. Although less strong than silencing *CAF40*, which is an essential gene, knockdowns and knockouts of both duplicates yielded lower viability (Fig. 3.3). Because the independent perturbation of each copy had a detrimental impact on viability and fertility, it suggests that the duplicates are not completely redundant, i.e. the divergence observed at expression and/or protein function make them functionally distinct, to the extent that the presence of two wild-type paralogs did not automatically compensate for the silencing or knockout of a third paralog.

Our co-immunoprecipitation assays showed that Poseidon protein conserved CAF40 ability to interact with NOT1 in the CCR4-NOT complex, as well as to recruit Bam and Roquin. Moreover, such interactions are consistent with CAF40 and Poseidon strong repressive effect of a tethered reporter transcript (Fig. 3.5). On the other hand, Zeus failed to bind, or binds only weakly, to NOT1 and the other two assayed proteins. Again, this difference is in agreement with its weaker repressive ability in our tethering assay (Fig. 3.5). Although significantly lower, Zeus retained a significant repressive ability, which probably reflects either Zeus weak binding to the CRR4-NOT complex, the conservation of ancestral, not yet described interactions or the evolution of new interactions involved in repressive activities. Taken together, these results suggest that, whereas Poseidon probably inherited CAF40 role in the CCR4-NOT complex (although with a distinct impact on gene regulation, as discussed below), Zeus likely acts independently of this complex.

One possibility, for instance, is that Zeus main mechanism of action consists solely of binding the genome and regulating transcription, which is in agreement with previous analyses that showed that Zeus presents a distinct binding pattern in the genome, and regulates a different set of genes compared to CAF40 (Chen et al., 2012). If that is the case, Zeus protein function consists of a subfunction of the ancestral CAF40, consistent with the partitioning of functions postulated by subfunctionalization models (Force et al., 1999).

Lastly, our genome-wide transcriptome analyses demonstrated that the independent perturbation of the three paralogs impacts the regulation of hundreds of genes in the testes (Fig. 3.6), in agreement with *CAF40* well established role in transcriptional and post-trancriptional regulation (reviewed in Collart 2016), as well as previously assayed for *Zeus* (Chen et al., 2012). On one hand, the comparison of the genes that are differentially expressed in the knockdown samples revealed that, among the large set of genes commonly perturbed by the paralogs knockdown compared to the control, there is a strong enrichment for genes related to cellular catabolic processes (Fig. 3.6, Table S3.3). Such enrichment suggests that the knockdown of the paralogs affects the regulatory balance between transcription, translation and degradation of numerous downstream genes, given the importance of the parental gene in coordinating and integrating different regulatory pathways (Miller and Reese, 2012). In addition, for the vast majority of genes co-affected by more than one paralog knockdown, the direction of the impact (i.e. up- or down-regulation) was the same on both samples (Fig. 3.6D), suggesting that the paralogs share a similar effect on a large portion of the target genes.

On the other hand, almost half the differentially expressed genes are exclusively perturbed when one of the paralogs is silenced, but not the others, which reflects the

distinct impact of each paralog in the regulatory network (Fig. 3.6B). Such differences in the sets of genes impacted can be explained by the fact that each paralog has a distinct timing of expression in the testis (Fig. 3.2B), and/or has different binding affinities to their transcript and genomic targets, which was not explored in this study.

Curiously, despite the fact that *Zeus* is younger and less diverged from *CAF40* when compared to *Poseidon*, it seems to be more functionally distinct from the parental gene, with respect to expression pattern (Fig. 3.2B), protein interaction (Fig. 3.5A-C), and repression activity (Fig. 3.5D). Moreover, such distinctions are consistent with the observation that *Zeus* knockdown impacts the expression of a larger proportion of genes not impacted by *CAF40* knockdown (Fig. 3.6B-D). These observations reinforce the importance of experimental assays for determining the impact of divergence on gene function, since sequence similarity alone is not a reliable predictor.

Taken together, the analyses presented here show that *Poseidon* and *Zeus*, despite their relatively recent origination in *Drosophila*, were integrated into fundamental cellular processes with profound impact in the regulatory network and phenotype. Such integration was accompanied by divergence at protein sequence, timing of expression and protein interactions, emphasizing the complex nature of evolution driven by new gene origination (Chen et al., 2013). The fixation of two retroduplicated copies from the same conserved parental gene with extensive regulatory functions in a relatively short evolutionary period follows the intriguing pattern of the recurrent origination of specialized copies of housekeeping genes with male-biased expression (Belote and Zhong, 2009; Kaessmann, 2010). It is possible that, for the case studied here, evolution takes advantage of *CAF40* constrained role in a highly pleiotropic regulatory complex to generate

specialized versions of the same structure (Pavlicev and Wagner, 2012), which may help to promote the fast and coordinate changes in gene regulation through different transcriptional and post-transcriptional mechanisms required during spermatogenesis (Legrand and Hobbs, 2018; White-Cooper, 2010).

Methods

Molecular evolutionary analyses

Poseidon (*CG2053*) had been previously computationally identified as a putative young gene (Zhang et al., 2010a). Gene and protein sequences were retrieved from Flybase and NCBI, aligned with MUSCLE (Edgar, 2004) and manually curated. Reciprocal PSI-BLAST (NCBI) searches were employed to survey for *CAF40*, *Poseidon* and *Zeus* orthologs in eukaryotes. The proper substitution model for the alignment (GTR+G) was selected through a likelihood ratio test using jModeltTest (Posada, 2008), and the phylogenetic relationship among the paralogs was inferred through Bayesian analysis in MrBayes (Ronquist et al., 2012). MCMC analysis was run with 4 chains for 2 million generations, with trees begin sampled every 500 generations, and the first 25% of samples were discarded as burn-in. Shannon's entropy was calculated for an alignment of CAF40 orthologous protein sequences from 56 eukaryotes (Fig. S3.1), and the entropy value (H) for each residue was plotted onto CAF40 protein structure from *D. melanogaster* (Sgromo et al., 2017) using PyMOL (Fig. S3.2).

For the divergence of expression analysis, we retrieved the summary of expression data (average FPKM [fragments per kilobase per million mapped reads]) from modENCODE and public RNA sequencing data from various fly species (Brown et al. 2014;

Chen et al. 2014; VanKuren and Vibranovski 2014). Expression values of each paralog at different spermatogenesis stages in *D. melanogaster* was compared using data from the SpermPress database (Vibranovski et al., 2009b).

Knockdown and knockout phenotypes

In order to assay the knockdown effect of each paralog on egg to adult viability, homozygous UAS-TRiP RNAi lines (Perkins et al., 2015) were crossed to a balanced constitutive driver line (*Tub84B*>GAL4/TM3) and an imaginal disc-specific driver line (*T80*>GAL4/CyO) (Table S5 shows the list of lines used). At least 10 independent replicates of 3 couples were allowed to cross and lay eggs for 7 days at 23°C. All F1 adults in the progeny were scored, and the proportion of wild/balancer phenotypic markers for all replicates was compared to control crosses (TRiP background line BDSC 36303 crossed to the driver lines). Male fertility effects were assayed for the three paralogs by driving GAL4 expression using male-germline-specific *nanos*-GAL4 and *Bam*-GAL4 drivers, which are expressed in early and late spermatogenesis, respectively (Table S3.5). At least 15 replicates with 3-5 days old knockdown males were individually crossed to two virgin females from background line BDSC 36303 for one day. Females were allowed to lay eggs for 7 days, and all the F1 adults were counted. Knockdown efficiency for each paralog was confirmed through RT-PCR (Table S3.2). RNA samples were extracted in triplicate using Quagen RNeasy kit, digested with DNAse (Invitrogen) to remove genomic DNA contamination, and reverse-transcribed with SuperScript III Reverse Transcriptase (Invitrogen) using oligo(dT) primers. RT-PCR was performed using iTaq Universal SYBR Green Supermix (Biorad), with three technical replicates for each biological replicate.

Quantitative PCR values were normalized using the $\Delta\Delta C_T$ method to the *Rp49* control product.

CRISPR-Cas9 frameshift deletions were induced for *Poseidon* and *Zeus*. Guide RNAs were designed using FlyCRISPR Optimal Finder to target early portions of the exon, and injected (500 ng/ul) along with Cas9 protein (PNA Bio Lab) into embryos from the BDSC 25710 line, following (Bassett and Liu, 2014) (Tables S3.5-6). F1 mutant individuals were screened and crossed to balancer lines (w^+;Sb/TM3; and w^+;Sco/CyO, respectively). Small frameshift deletions were confirmed through Sanger sequencing, and created early stop-codons in the transcribed genes (Fig. S3.8). Viability and male fertility assays were performed with knockout flies as described above, using the injected line BDSC 25710 as control.

Co-immunoprecipitation assay

DNA constructs with the coding region of *D. melanogaster* genes *CAF40*, *NOT1*, *Roquin* and *Bam* used were described before (Sgromo et al., 2017, 2018). Plasmids encoding *Poseidon* and *Zeus* were generated by inserting the corresponding cDNA (Thermo Scientific) into the pAc5.1-λN-HA and pAc5.1-GFP vectors (Rehwinkel et al., 2005; Tritschler et al., 2007) using HindIII and XhoI restriction sites. All constructs were confirmed by Sanger sequencing. For co-immunoprecipitation assays in *D. melanogaster* S2 cells (ATCC), 2.5×10^6 cells were seeded per well in 6-well plates and transfected using Effectene transfection reagent (Qiagen), using 1 – 1.8 μg of plasmids expressing the HA- or GFP-tagged versions of the assayed proteins.

Cells were harvested 3 days after transfection, and co-immunoprecipitation assays were performed using RIPA buffer [20 mM HEPES (pH 7.6), 150 mM NaCl, 2.5 mM MgCl2, 1% NP-40, 1% sodium deoxycholate supplemented with protease inhibitors as previously described (Sgromo et al., 2018; Tritschler et al., 2008)]. All co-immunoprecipitation assays in S2 cell lysates were performed in the presence of RNaseA as previously described (Sgromo et al., 2017). All Western blots were developed using an ECL western blotting detection system (GE Healthcare). The antibodies used in this study are listed in Table S3.7.

Luciferase assay

For the λN-tethering assays in *D. melanogaster* S2 cells, 2.5×10^6 cells per well were seeded in six-well plates and transfected using Effectene (Qiagen). The transfection mixtures contained the following plasmids: 0.1 μg of Firefly luciferase reporters (F-Luc-5BoxB or F-Luc-V5), 0.4 μg of the Renilla Luciferase (R-luc) transfection control and 0.1 μg of plasmids expressing the λN-HA-tagged paralogs. The plasmids for tethering assays in S2 cells (F-Luc-5BoxB, F-Luc-V5 and R-Luc) were previously described (Behm-Ansmant et al., 2006; Zekri et al., 2013). Cells were harvested 3 days after transfection and Firefly and Renilla luciferase activities were measured by using a Dual-Luciferase Reporter Assay System (Promega) with three replicates.

RNA-seq analysis

Total RNA was extracted with Arcturus PicoPure Isolation kit (Qiagen) from testes of 3-5 days old knockdown males and controls, with three biological replicates. A total amount of 1 μg RNA per sample was used to construct the cDNA library, using NEBNext

Ultra RNA Library Prep Kit for Illumina (NEB) following manufacturer's recommendations. Briefly, poly(A) mRNA was purified from total RNA using oligo(dT)-attached magnetic beads, reverse-transcribed to double-stranded cDNA with random primers, end-repaired and ligated with NEB adaptors for Illumina, before sequencing (HiSeq 4000, University of Chicago Genomics Core Facility).

Raw reads were processed and mapped to *D. melanogaster* reference genome (dm6) using STAR with default parameters (Dobin et al., 2013), and evaluation of transcriptional expression was carried out using featuresCounts (Liao et al., 2014). For the differential expression analysis, methods DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010), and limma (Ritchie et al., 2015) were independently employed. Genes were considered as differentially expressed if they were consensually called by the three methods, with an expression fold change of at least 1.5 compared to the control at false discovery rate less than 0.05. For differentially expressed genes, enriched biological processes and molecular functions were identified using GOrilla (Eden et al., 2009), with p-values < 10^{-4}, and a false discovery rate of 0.05. The analyses of differentially expressed genes with male/female-biased expression followed the classification of two independent *Drosophila* databases (Assis et al., 2012; Gnad and Parsch, 2006).

Author contributions

IMV and ML designed the study. IMV performed the genetic analyses and functional experiments. SX and IMV generated CRISPR-Cas9 knockouts. AS and EI designed and performed Co-IP experiments. All authors discussed the results. IMV and ML wrote the manuscript.

Chapter 4

The young retrogene *Cocoon* is essential for *D. melanogaster* survival

The work described in this chapter summarizes my contributions to the project developed in collaboration with Grace Lee (now at University of California – Irvine). The manuscript reporting the whole project is under review as: Lee YCG, Ventura IM, Rice GR, Chen DY, Long M. "Rapid evolution of essential developmental functions of a young gene via interactions with other essential genes".

Abstract

New genes originated relatively recently and are only present in a subset of species in a phylogeny. Accumulated evidence suggests that new genes, like old genes that are conserved across species, can also take on important functions and be essential for the survival and reproductive success of organisms. Here, we focused on a young retro-duplicated gene (*CG7804,* which we named *Cocoon*), which originated less than four million years ago and is found in three *Drosophila* species. We found that, unlike its evolutionarily conserved parental gene*, Cocoon* has diverged rapidly in sequence and expression since its birth. Despite its young age, *Cocoon* is essential for the survival of *D. melanogaster* at multiple developmental stages, and in different tissues from its parental gene. Functional genomic analyses found that Cocoon protein binds to a subset of genomic sites bound by the parental gene. Importantly, Cocoon binding predominantly locates at genes that have

other essential functions and/or have multiple gene-gene interactions, suggesting that *Cocoon* is essential to survival through interactions with genes that have large impact on the regulatory network. Our study is an important step towards deciphering the evolutionary trajectory by which new genes functionally diverge from their parents and play essential roles.

Introduction

The genetic basis of the diversity of life remains a central question in evolutionary biology. Nucleotide substitutions or indels that change protein-coding or regulatory sequences are often observed to contribute to functional, phenotypic, and behavioral polymorphism and divergence within and between species (e.g. (Ding et al., 2016; Linnen et al., 2013; Rost et al., 2004; Wittkopp et al., 2009), reviewed in (Barrett and Hoekstra, 2011; Wray, 2007). However, in addition to gradual change at pre-existing genes, gene composition turns over rapidly even between closely related species (e.g. (Demuth et al., 2006; Zhang et al., 2010a, 2011) and reviewed in (Kaessmann, 2010). Indeed, while humans and chimpanzees have only diverged 1.5% in their orthologous coding sequences (Consortium, 2005), they differ by at least 6% of their gene contents (Demuth et al., 2006).

The origination of new genes is an important evolutionary process contributing to the dynamic turnover of genes in genomes over the phylogeny. This dynamic gene turn over has been widely documented in *Drosophila* (Zhang et al., 2010a), primates (Demuth et al., 2006; Zhang et al., 2011), and plants (Moore and Purugganan, 2005). Because of their recent origin, new genes are only present in a subset of species in a phylogeny and the prevailing view was that they have dispensable functions and are not essential to an

organism's fitness (e.g. (Ashburner et al., 1999; Jacob, 1977). However, recent evidence in a variety of eukaryotic species shows that new genes can quickly become essential for an organism's viability and fertility (Charrier et al., 2012; Chen et al., 2010; Cooper and Kehrer-Sawatzki, 2011; Hazelett et al., 2012; Ranz and Parsch, 2012; Reinhardt et al., 2013; Ross et al., 2013; VanKuren and Long, 2018), suggesting that new genes unique to few species can also have essential functions similar to those of highly conserved genes across the phylogeny.

The main mechanism by which new genes arise is through duplication, in which a copy of a gene is created through either DNA or RNA intermediates in the genome. Many evolutionary fates have been predicted for the duplicated (new) and original (parental) genes, grossly pseudo-functionalization, neo-functionalization, or sub-functionalization (Innan and Kondrashov, 2010; Lynch and Conery, 2000; Ohno, 1970). Despite the convenient conceptual distinction, it is often challenging to distinguish between these alternative models due to the fact that the past evolutionary trajectories are usually unknown or hard to decipher. Several in-depth analyses of the evolutionary steps leading to gained novel fertility functions of duplicated genes have shed light on the initial evolutionary processes of gained essential function supported by new genes (Chen et al., 2012; Ding et al., 2010; Heinen et al., 2009; Loppin et al., 2005; Yeh et al., 2012). In contrast, few studies have focused on viability (e.g. (Ross et al., 2013)). Many genes responsible for essential viability functions (e.g. development of body plan in *Drosophila* embryos (Stauber et al., 1999)) are identified as ancient gene duplicates (reviewed in (Chen et al., 2013)), suggesting new genes indeed can gain a critical role in the most essential and core functions of organisms. Yet, the past evolutionary trajectories of gained essential viability

function by new genes and whether that is similar to those of essential fertility function still need further investigation.

A potential mechanism by which duplicated genes become essential is by integrating into the cellular genetic network, through either multiple protein-protein or protein-nucleic acids interactions with pre-existing genes. Indeed, new genes with essential fertility functions were discovered to have locally or globally reshaped the regulatory network (Chen et al., 2012; Ding et al., 2010). Similarly, a new gene could quickly became essential for survival by gaining interaction partners in a gene network, a hypothesis consistent with the observations that genes with many interaction partners (hub genes) are more likely to have essential functions (Batada et al., 2007; Blomen et al., 2015; Jeong et al., 2001; Yu et al., 2004) and reviewed in (Barabási and Oltvai, 2004; Barabási et al., 2011). However, comparisons of ancient orthologous genes report that the accumulation of gene-gene interactions is a slow evolutionary process (Kim et al., 2012). Whether and how, in a short evolutionary time, new genes can acquire a far-reaching impact on gene interaction network is still an open question.

In this study, we characterized the evolutionary history and function of a young duplicated gene that quickly became essential for the *survival* of *Drosophila melanogaster.* This young gene (*CG7804,* which we named *Cocoon*) duplicated from another essential gene (*TBPH,* also known as *TDP-43 human homolog* or *CG10327*) through retrotransposition less than four million years ago (Zhang et al., 2010a), and is present in few *Drosophila* species. The especially young age of *Cocoon* offers a rare opportunity to investigate the initial evolutionary steps of gaining essentiality by new genes. The parental gene, *TBPH,* is highly conserved among animals (Ayala et al., 2005; Li et al., 2010), its null mutant was found to

be lethal in *Drosophila* (Feiguin et al., 2009; Hazelett et al., 2012; Lin et al., 2011), and a mutant allele is associated with neuronal diseases in human (Sreedharan et al., 2008). *TBPH* was also shown to bind to nucleic acids (Kuo et al., 2009), influencing the splicing (Ayala et al., 2006; Bose et al., 2008; Buratti and Baralle, 2001) and transcriptional regulation (Ayala et al., 2008) of many genes. On the other hand, little is known about the duplicated gene, *Cocoon.* We found that unlike its evolutionarily conserved parental gene, *Cocoon* has rapidly evolved since its birth. Despite its young age, our functional assays show that it has a role in male fertility, and it is essential for the survival of *D. melanogaster* at multiple developmental stages, and in different tissues from *TBPH.* Furthermore, our functional genomic analyses (not described in detail in this chapter) showed that that *Cocoon* underwent multiple gains and losses of genomic binding sites compared to *TBPH*. Remarkably, *Cocoon* DNA binding targets are enriched for genes that have essential functions (i.e. mutant lethal) and/or engage in a large number of protein-protein/gene-gene interactions. Our study is an important step towards deciphering the evolutionary trajectory by which duplicated genes functionally diverge from the parental gene and become essential for development.

Results

Choice of studying *CG7804*

Our study began with a question: whether and how a young gene became essential for survival in a short evolutionary time through influencing gene interaction network? We proposed to address one of the many plausible scenarios: a young gene acquired multiple nucleic acid bindings and thus has a global influence on gene regulatory network in non-

reproductive tissues. Accordingly, we used several criteria to narrow down our candidate young duplicated genes whose evolutionary trajectories would help addressing the question. We decided to focus on duplicated genes that have a good annotated gene model (according to Flybase), are less than four million years old (according to (Zhang et al., 2010a)), and have a parental gene with nucleic acid binding activity. We further narrowed down our list to *CG7804* by choosing young genes that were suggested to exhibit lethal and/or semi-lethal phenotypes in previous *in vivo* RNAi screens (Mummery-Widmer et al., 2009; Neely et al., 2010; Schnorrer et al., 2010), and performed genomic and functional analysis in order to test our hypotheses.

Cocoon is a young retrogene

The *Cocoon* locus (on chr3L) is only present in *D. melanogaster, D. simulans,* and *D. sechellia* (Zhang et al., 2010a), which suggests it originated between 1.5 and 3.5 million years ago. A PSI-BLAST search revealed that its closest match in *D. melanogaster* is *TBPH* (E-value = 10^{-167}). Because the gene sequence lacks the introns present in *TBPH* (on chr2R), it likely originated through an RNA-based duplication event (retrotransposition) (Fig. 4.1A). Despite the fact that the entire coding sequence of *TBPH* was inserted to create the duplicate, *Cocoon* sequences from the three species exhibit a premature stop codon, resulting in a considerably shorter protein (318 versus 531 amino acids in *D. melanogaster*). *Cocoon* is predicted to possess the same two nucleic acid binding domains (residues 109-174 and 194-239) found in *TBPH* (Fig. 4.1B). Such RRM (RNA recognition motif) domains have been demonstrated to bind to both RNA and DNA (Kuo et al., 2009), and are crucial to *TBPH* function. On the other hand, Cocoon lost the C-terminal glycine-rich

region found in the parental protein (Fig. 4.1B), which was shown to be critical to TBPH function as a splicing factor in *Drosophila* and humans (Ayala et al., 2005).

Despite its short evolutionary history, we detected a burst of 77 amino acid substitutions in the *D. melanogaster* lineage after the origination of *Cocoon* using a maximum-likelihood method (Yang, 2007), in sharp contrast with TBPH, which only accumulated 4 amino acid replacements in the same period (Fig. 4.1C). The fast evolution of *Cocoon* since its origination resulted in the high divergence in amino acid sequences between it and *TBPH* (28.6%). As a comparison, *TBPH* paralogs from *D. melanogaster* and *D. yakuba,* a species in which *Cocoon* is absent, diverge only 5.0% in amino acid sequences. We hypothesized that Cocoon intense divergence at sequence and protein size might have impacted its function when compared to the parental protein, which we explore below.

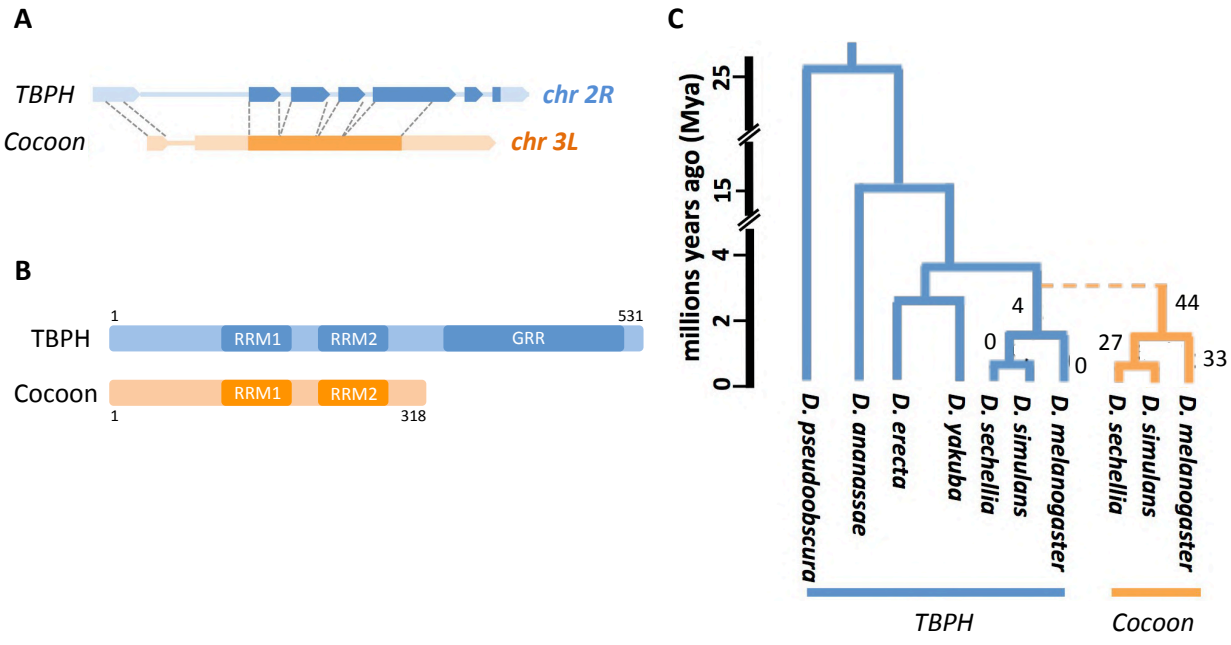


Figure 4.1. Structural and evolutionary history of *Cocoon* and *TBPH*. (A) Exon-intron structure of *TBPH* (blue) and *Cocoon* (orange). Filled boxes represent exons (darker color - coding sequence; lighter color - UTRs) while lines represent introns. Because *Cocoon* originated through a retrotransposition event, it lacks the introns of *TBPH.* (B) The structures of the protein sequences of *TBPH* and *Cocoon* are shown. Darker colors

(Figure 4.1, continued) represent the predicted structures of the RRM (RNA-recognition motifs) and GRR (glycine-rich region) predicted using Phyre (Kelley et al., 2015). Notice that Cocoon peptide is considerably shorter, and lacks the GRR domain of the parental protein. (C) The duplication event of *Cocoon* (orange) from *TBPH* (blue) is denoted as a dashed line in the phylogeny. The number of amino acid substitutions inferred by PAML (see text) is denoted at right to branches.

TBPH and *Cocoon* are essential for the survival of *D. melanogaster*

In order to assay *Cocoon* phenotypic role, we first used extensive transcriptome data from public databases to investigate its expression. According to modENCODE expression reports in *D. melanogaster* (Brown et al., 2014; Graveley et al., 2011), *Cocoon* expression is restricted to larval imaginal discs, pupae, and male reproductive tissues, whereas *TBPH* exhibits a broader expression pattern at multiple tissues and developmental stages, both in males and females (Fig. 4.2A). Such male-biased expression pattern for *Cocoon* is also found in *D. simulans*, whereas *TBPH* is expressed in adults of both sexes in other *Drosophila* species in which *Cocoon* is absent (Fig. 4.2B; VanKuren and Vibranovski, 2014; Brown *et al.*, 2014).

We employed GAL4/UAS system to knockdown the expression of *Cocoon* and *TBPH* individually, first using ubiquitous GAL4 drivers (*Act5C*-GAL4 and *Tub*-GAL4). Consistent with previous studies (Feiguin et al., 2009; Hazelett et al., 2012; Lin et al., 2011), *TBPH* knockdown showed that the gene is essential for the survival of *D. melanogaster* (lethality rate: 97.5% with *Act5C*-GAL4 and 86.6% with *Tub*-GAL4). Surprisingly, despite originating recently on an evolutionary timescale, expression knockdown of *Cocoon* also led to very low survival rate (lethality rate: 94.7% with *Act5C*-GAL4 and 95.5% with *Tub*-GAL4, Fig. 4.2C). For *Cocoon* knockdown, the lethality happens at larval and pupal stages (Figure 4.2C). Indeed, we found that flies could not develop pass the pharate adult stage and

identified many eclosion lethal incidences (i.e. flies could not emerge and were found stuck and dead half way in pupal cases, which inspired the gene naming, Fig. 4.2D).

We then used tissue-specific *GAL4-drivers* to knockdown the expression of *CG7804* and *TBPH* in different tissues. Expression knockdown of *TBPH* using larval neuronal-specific *elav* GAL4-driver leads to much lower survival rate than expression knockdown of *CG7804* using the same driver (Fig. 4.2E), which is consistent with the previously identified role of *TBPH* in neuronal functions (Feiguin et al., 2009; Hazelett et al., 2012). On the other hand, expression knockdown of *CG7804* using *Dll* (leg imaginal disc) and *salm* (imaginal discs) led to lower survival rates than those of *TBPH* knockdown with the same GAL4 *drivers* (Fig. 4.2E). These disparities in tissue-specific knockdown effects suggest that the expression of *CG7804* is essential for viability at different tissues from those of its parental gene, *TBPH.*

We also used CRISPR/CAS9 system (Gratz et al., 2013) to generate null mutants of *CG7804* (see Methods). Consistent with the results using GAL4/RNAi expression knockdown, *CG7804* knockout leads to extremely high lethality rate (99.4%). In addition, another mutant of *CG7804* that was generated by a different approach (insertion of a MIMIC construct in the coding region (Venken et al., 2011) also shows extremely high lethality (99.21%), and the two mutants cannot complement each other. We didn't find any effect of *Cocoon* knockout on the developmental time (egg to adult) compared to control of the same background strain (Fig. 4.2F). Overall, both our expression knockdown and null mutant analyses support the conclusion that *CG7804* is highly essential for the survival of *D. melanogaster,* despite being young and only present in few species.

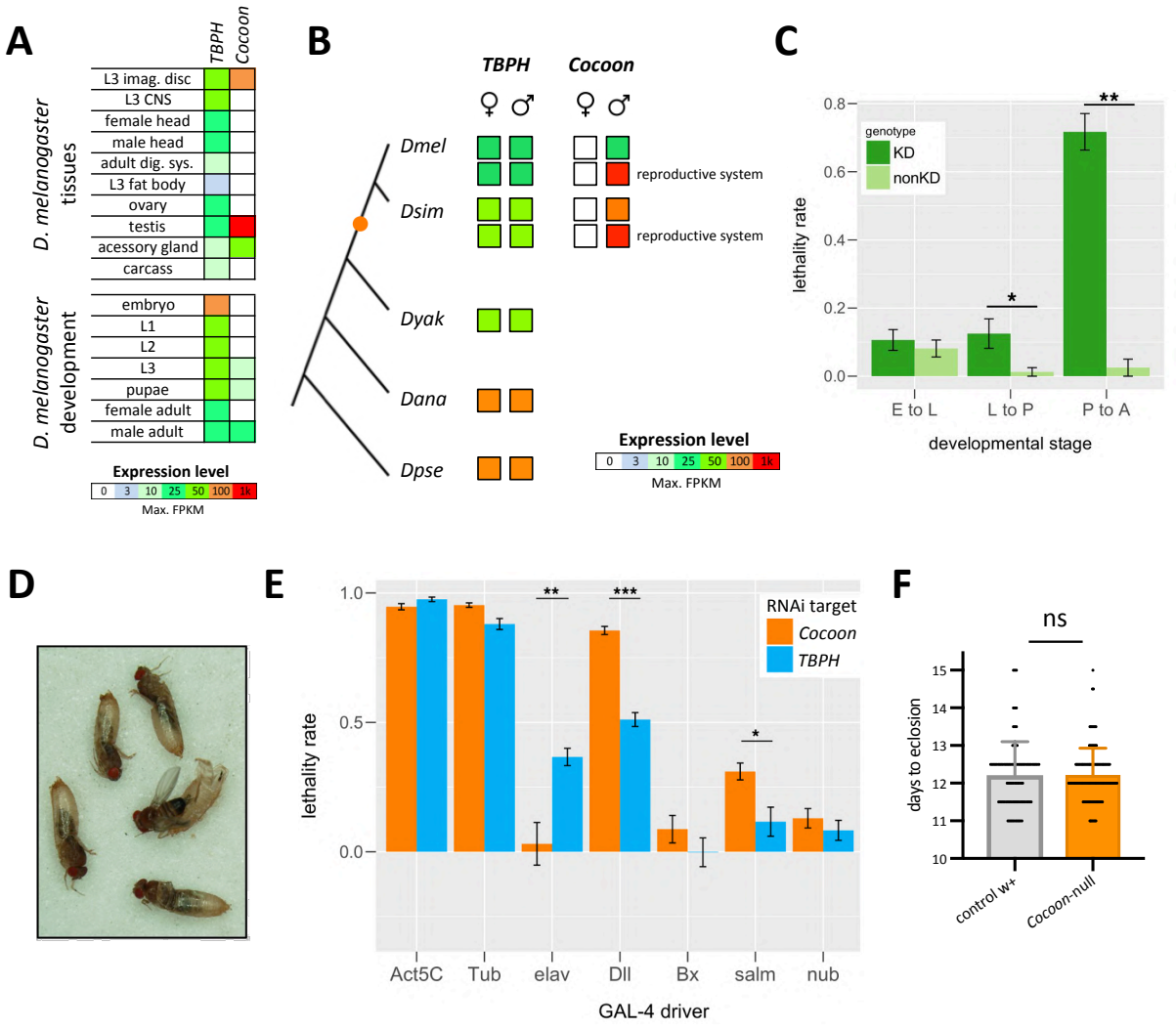


Figure 4.2. *Cocoon* and *TBPH* expression and phenotypic effect. A) Summary of the expression intensity of *Cocoon* and *TBPH* in *D. melanogaster* tissues and development times. Expression summarized as maximum expression (FPKM). *Cocoon* has a more restricted expression pattern compared to *TBPH*. B) Summary of expression level in female/male adults for different *Drosophila* species. Adult male-biased expression for *Cocoon* is observed in *D. melanogaster* and *D. simulans*. Data extracted from (Brown et al., 2014; Chen et al., 2014b; VanKuren and Vibranovski, 2014). (C) Expression knockdown of *Cocoon* using Tub-GAL4 driver results in different lethality rates at different developmental stages. (D) Expression knockdown of *Cocoon* leads to eclosion lethal at the pupal stage. (E) *Cocoon* expression is essential in different tissues from those of its parental gene, *TBPH.* Lethality rate is significantly different between the genes knockdown when using *elav, Dll,* and *salm* GAL4 drivers. (F) The developmental time from egg to eclosed adults is not different between *Cocoon* knockout and wildtype individuals (CRISPR background strain; $p = 0.83$). E: embryo; L: third instar larvae (L3); P: pupae; A: adult. KD: individuals with *Cocoon* knockdown genotype; nonKD: wildtype individuals; Mann-Whitney U test: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TBPH and *Cocoon* impact male fertility

Given the high expression of *Cocoon* in male reproductive tissues, it is natural to wonder whether the gene also impact male fertility. We used germline specific GAL4 driver (Bam-GAL4) to knockdown the expression of *Cocoon* and *TBPH* in the testis and tested whether that influence male fertility. A similar negative effect on the mean offspring number was observed for *Cocoon* and *TBPH* knockdowns (Fig. 4.3). Thus, despite its evolutionary recent origination, *Cocoon* knockdown also affects male fertility, although the effect is less dramatic than that observed in viability.

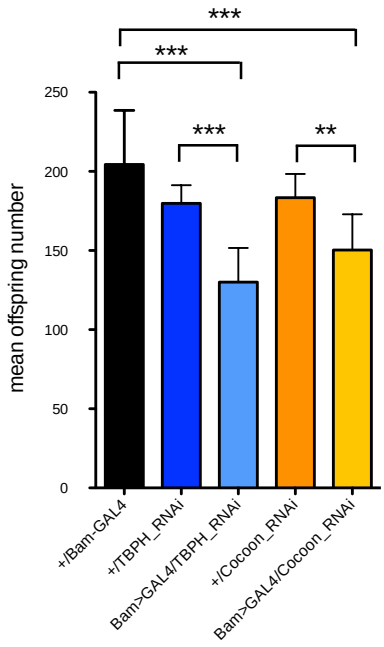


Figure 4.3. **Role of *Cocoon* and *TBPH* in male fertility**. Mean offspring number of males with *Cocoon* and *TBPH* knockdown is lower than background genotypes (Cocoon-RNAi strain, TBPH-RNAi strain, and Bam-GAL4). Mann-Whitney U test: *p < 0.05, **p < 0.01, ***p < 0.001.

Cocoon genomic binding pattern and impact on global gene expression

In order to further investigate the mechanisms by which *Cocoon* is essential for the survival of *D. melanogaster* at multiple developmental stages, we compared the impact of

Cocoon knockdown on global gene expression, through RNA-seq. In addition, we generated GFP-tagged versions of *Cocoon* and *TBPH*, and compared their genomic binding sites through ChIP-seq. Because these functional genomics analyses were designed and performed by Grace Lee, they will not be described here in detail, but the main findings are described below. A preliminary version of the complete manuscript is available at *www.biorxiv.org/content/10.1101/226936v1*.

Briefly, our transcriptome analyses revealed that *Cocoon* knockdown significantly perturbed the expression of hundreds of genes, consistent with the well-known role of the parental *TBPH* in gene regulation. It is worth noting that the rate of development from eggs to adults is not significantly different between *Cocoon* knockout and wildtype individuals (as described in Fig. 4.2F), suggesting that the observed global transcriptome differences are not merely driven by shifts in the developmental rates.

Remarkably, in contrast with a previous study that found that *TBPH* knockout leads to alternative splicing of hundreds of transcripts (Hazelett et al., 2012), our analyses did not detect transcripts whose alternative spliced forms significantly differs between wildtype background and *Cocoon* knockout, suggesting that the duplicate may not function as a splicing factor. This result is consistent with the loss of the C-terminal glycine-rich domain in *Cocoon* (Fig. 4.1B), since this is the region that was shown to interact with other splicing factors (Ayala et al., 2005).

Finally, our ChIP-seq analyses revealed that TBPH and Cocoon proteins exhibit a distinct binding pattern across the genome. Out of all TBPH genomic binding sites (1864 genes), Cocoon binds to only a subset of them (529 genes, 28.3%), as well as to other 120 sites not bound by the parental protein (18.5% of Cocoon sites). Moreover, the genes

bound by Cocoon have more experimentally validated protein-protein interactions and reported genetic interactions than an average gene, as well as are more likely to exhibit lethal phenotypes according to previous reports. These results suggest that, after duplication, Cocoon protein lost most of the binding sites from the parental protein, but remained essential for development through regulating the expression of other essential and highly connected genes.

Discussion

The origination of new genes is a major contributor to the dynamic gene turnover observed across evolutionary time. Despite being young and restricted to few species on a phylogeny, new genes have diverse essential functions and may play important roles for adaptation (reviewed in (Chen et al., 2013; Kaessmann, 2010; Kondrashov, 2012; Ventura and Long, 2017). Yet, how, in a short evolutionary time, new genes become essential is still an open question, and detailed functional dissection of new genes and their parents is a important step to address this phenomenon.

Despite its young age, and the fact that the gene is only present in three fly species (Fig. 4.1), our focused duplicated gene, *Cocoon* (*CG7804*), was found to be essential for *D. melanogaster* viability through expression knockdown and gene knockout experiments (Fig. 4.2). Our analyses demonstrate that Cocoon and its parent TBPH diverged considerably in several aspects. *Cocoon* has acquired a more restrict expression pattern, in contrast with *TBPH* broad expression. Such divergence is commonly observed for retrogenes, which recruit or evolve new regulatory elements upon their origination (Zhang and Zhou, 2018). Furthermore, divergence in expression has critical functional

consequences for these genes, since our tissue-specific expression knockdown assays showed that *Cocoon* and *TBPH* are required at different tissues (Fig. 4.2E).

Our functional genomic analyses (not described here in details, see Results) revealed that *Cocoon* has a widespread effect on transcriptome regulation, consistent with the observed essentiality of the gene for development. Interestingly, Cocoon protein probably lost the ancestral function as a splicing factor observed for TBPH, which is consistent with the loss of the C-terminal glycine-rich region due to a premature stop codon in the retrogene (Fig. 4.1B). In addition, our ChIP-seq analysis suggests that *Cocoon* is essential by engaging in interactions with and regulating the expression of multiple pre-existing genes, in particular, those with reported lethal phenotypes, or that have many protein-protein/genetic interaction partners (hub genes). Moreover, the genomic sites bound by Cocoon represent only a subset of all sites bound by the parental protein (28.3%), suggesting that the duplicated protein may have evolved distinct genomic binding preferences.

The perturbation of highly connected genes/proteins is expected to have more widespread influence on the overall network than perturbations of peripheral nodes with few links (for biological networks, reviewed in (Barabási and Oltvai, 2004)). Indeed, in gene-gene interaction network, hub genes are often found to be essential (Blomen et al., 2015; Jeong et al., 2001; Yu et al., 2004). In particular, the binding of Cocoon to other essential and/or hub genes may explain why it plays an essential role for fly development.

Overall, our observations show that Cocoon evolution involved fast divergence at several different aspects of gene function, in contrast to the parental gene high conservation. For instance, *TBPH* human ortholog (*TDP-43*) is able to complement a loss of

function *TBPH* mutant in *D. melanogaster* (Li et al., 2010). As a contrast, the duplicated gene evolution involved divergence in expression, protein structure (i.e. loss of a protein domain), and genomic binding pattern. Moreover, despite its young age, the gene has an essential role on global gene expression, in particular on the regulation of highly connected genes expressed during development.

Methods

Evolutionary analysis of *Cocoon* and *TBPH*

We used coding sequence of *Cocoon* (*CG7804*) and *TBPH* of 12 *Drosophila* species from (Clark et al., 2007) and aligned using MUSCLE (Edgar, 2004), followed by manual curation. We used CODEML program in PAML (Yang, 2007) to estimate the number of amino acid replacements on each branch. Domains of Cocoon protein were predicted by Pfam (Finn et al., 2016) and Phyre (Kelley et al., 2015). For the divergence of expression analysis, we retrieved the summary of expression data (FPKM [fragments per kilobase per million mapped reads]) from modENCODE and public RNA sequencing data of diverse fly species (Brown et al., 2014; Chen et al., 2014b; VanKuren and Vibranovski, 2014).

Generation of transgenic strains and mutants

Design and injection of guide RNA, and screen for CRISPR mutants were performed by Genetic Service Inc. (Sudbury, MA). *Cocoon* mutant has a 2 bp deletion in the coding sequence, which is confirmed by Sanger sequencing (Fig. S1). To further confirm that our CRISPR mutant is a true null mutant, we used another mutant of *Cocoon* to perform complementation tests. We used a strain (BDSC 36014) that has a MIMIC construct

(Venken et al., 2011) inserted in the coding sequence of the gene. The presence of the MIMIC insertion was confirmed by PCR. We found CRISPR and the MIMIC strain do not complement each other, suggesting that our CRISPR *Cocoon* is a null mutant. We balanced *Cocoon* mutants over balancer chromosomes with ubiquitously expressed GFP for developmental stage-specific lethality analysis (see Table S1 for strains used). It is worth mentioning that, even for those few knockout individuals that survived to adult, we were able to detect expression of *Cocoon* either through RT-PCR or RNA-seq. However, these detected transcripts all have the frameshift deletion.

Essentiality analysis

Virgin females of RNAi strain (homozygous) were crossed to males of GAL4 strain. The GAL4 strain is heterozygous for GAL4 construct, which is balanced with visible markers and/or construct of ubiquitously expressed GFP. Expression knockdown and wildtype offspring were recognized by visible markers (adult) or presence/absence of GFP (embryo, larva, and pupa). All comparisons are within crosses (RNAi/+; GAL4/+ vs RNAi/+; +/balancer). For each cross, the expected number of knockdown individuals was estimated using the number of individuals with other genotypes and with the assumption that alleles were inherited following Mendelian proportions. The survival rate of knockdown individuals was estimated as observed number of knockdown individuals divided by the expectation, and the lethality rate is one minus the survival rate. At least 10 independent crosses that have at least 20 adults in each cross were counted. For tracking stage-specific lethality, 20 embryos/larvae/pupa of each genotype were collected and placed on fresh medium and the numbers of next-stage individuals were counted after 5 and 10 days. We

collected embryos of mixed stages through standard apple juice plate, larvae at L3 stage, and white pre-pupa. This experiment was repeated at least four times for a specific genotype or specific developmental stage. GAL4 and RNAi strains used in the study can be found in Table S1. Estimation of survival rate and tracking of stage-specific lethality rate for knockout individuals used the same methods as experiments with knockdown individuals. All flies were reared with standard *Drosophila* medium at 25 °C with 12/12 light and dark cycle.

Male fertility assay

Expression of *Cocoon* or *TBPH* was knocked down using a germline-specific GAL4 driver (Bam>GAL4, from G. Findlay lab, see Table S1), and their progeny was counted. In details, sets of 10 virgin females from the GAL4 driver were crossed to 10 males of RNAi strain (homozygous) to obtain males with *Cocoon* or *TBPH* knockdown. Those 3-5 days old virgin males were allowed to cross to two females from strain BDSC36304 (the background strain from which RNAi strains were generated) for two days, and then crossed again for two days with two additional virgin females. Females were allowed to lay eggs for 7 days, and total progeny was counted after 20 days. Ten to fifteen crosses were used for each genotype tested. In addition to males with *Cocoon* or *TBPH* knockdown (CG7804-RNAi/Bam>GAL4 or TBPH-RNAi/Bam>GAL4), males with genotypes Bam>GAL4/+ and RNAi/+ were tested as controls.

Developmental rate analysis

To investigate whether the *Cocoon* knockout affects the fly development time (which could potentially confound the developmental stage-specific RNA-seq analysis), the egg-adult development time was compared between the knockout and the wildtype background individuals. 80-100 inseminated females were allowed to lay eggs in an agar plate for 1h, and then 20 eggs were transferred to food vials, where they developed at 25°C. Adult eclosion was scored twice a day until all adults had eclosed.

RNA-seq experiment and analysis

Despite the fact that the functional genomics analyses performed in this study are not described in detail in the Results, their methodology will be briefly summarized here, in order to offer a context for some of the discussed conclusions. Homozygous knockout individuals were collected from the progeny of the crossing between heterozygous individuals for the null allele (*Cocoon*-null/GFP-balancer). For the wildtype counterparts, we used the Cas9 strain from which the knockout mutant was generated. We collected 0-24hr embryos using standard apple juice plate, wandering L3, and white pre-pupa. For all three stages, we used mixed-sex individuals and have two biological replicates for each genotype at each developmental stage. Total RNAs were extracted from collected materials using RNeasy Plus kit (Qiagen). RNA-Seq sequencing library was prepared using Illumina TruSeq and sequenced on Illumina Hi-Seq with 100bp, paired-end reads (IGSB Sequencing core, the University of Chicago). Processed reads were mapped to *D. melanogaster* release 6 genome, and DESeq2 (Love et al., 2014) was used to normalize and estimate expressional

fold enrichment between two *D. melanogaster* genotypes. Only genes with at least 10 mean read counts were included in the DESeq2 analysis.

ChIP-seq experiment and analysis

ChIP was performed using modEncode protocol (http://www.modencode.org/) using anti-GFP antibody (from Kevin White's laboratory) with two biological IP replicates for each genotype (Cocoon-GFP and TBPH-GFP). ChIP-Seq sequencing library was prepared using NuGen Ovation Ultralow Library Systems V2 (San Carlos, CA) and sequenced on Illumina Hi-Seq with 100bp, paired-end reads (IGSB Sequencing core, the University of Chicago). Processed reads were mapped to *D. melanogaster* reference genome release 6, and we used IDR analysis to identify enrichment peaks with lower than 1% irreproducibility rate between replicates. Genes overlapping with enrichment peaks were identified using Bedtools (Quinlan and Hall, 2010).

Analysis of gene properties

Degree of each gene in protein-protein network was estimated as the number of experimentally validated, non-redundant protein-protein interaction using data from BioGrid 3.4 (Stark et al., 2006). The degree of each gene in genetic interaction network was estimated as the number of reported genetic interaction on Flybase (release 2017, Feb). Phenotypic data for all genes annotated were downloaded from Flybase, which were based on either knockout mutants or expressional knockdown analysis.

Acknowledgment

We thank Jennifer Moran of Genome Engineering core of the University of Chicago, and Alec Victor, Matt Kirkey, Jeffrey Gersch from Kevin White's lab for hosting YCGL for ChIP experiments. We thank Dr. Findlay for generously Bam>GAL4 strain and Dr. White for sharing GFP antibody. Mia Levine, Claus Kemkemer, Benjamin Krinsky, and Nicholas VanKuren provided helpful discussions of the project. We are also grateful to Josie Reinhardt, Maria Vibranovski, and Li Zhao for critically reading the manuscript. YCGL was supported by NIH NRSA F32 GM109676 and Chicago Biomedical Consortium Postdoctoral Research Award PDR-043. IMV was supported by the Science without Borders scholarship (BEX18816/12-6). GRR was supported by NSF Graduate Research Fellowship. ML was supported by NSF1051826 and NIH R01GM116113.

Author contributions

YCGL, IMV and ML designed the study. YCGL and IMV performed the genetic analyses and knockdown viability experiments. IMV performed the knockout viability, fertility and developmental rate experiments. YCGL performed the functional genomic analyses. GRR and DYC performed imaging analyses. YCGL, IMV and ML wrote the manuscript.

Chapter 5

Conclusions

The two study cases described in Chapter 3 (*CAF40*, *Poseidon* and *Zeus*) and Chapter 4 (*TBPH* and *Cocoon*) allow us to investigate the mechanisms underlying the evolution of two relatively young duplicated genes in *Drosophila*, as well as their functional roles and phenotypic impact. Broadly speaking, the two studies demonstrate that a diverse suite of factors was responsible for the functional divergence of the duplicated genes after their origination from their conserved parental genes. Despite not being directly associated, the duplicated genes evolution in the two systems exhibits remarkable similarities, and suggests that some patterns can be particularly common after gene duplication, in special for genes with regulatory roles. Below, I will summarize the general picture drawn from these examples regarding their origination mechanisms, expression pattern, protein divergence, and regulatory impact.

First, it is intriguing to observe that the three new genes described here originated through independent retrotransposition (RNA-based duplication) events, even though this is not the most common origination mechanism of new genes in flies. In the *D. melanogaster* subgroup, for example, retrotransposition accounts for around 10% of the retained new genes, in contrast to 78% of genes that originated through DNA-based duplication (Zhou et al., 2008). When choosing a duplicated gene to investigate, we often focus on those that present promising features to dissect, such as high divergence at expression and/or sequence. Notably, retrogenes are able to readily diverge in expression from their parents upon origination, since they are inserted in different genomic regions

from their parents. Such origination mechanism creates gene copies that lack the ancestral cis-regulatory elements, epigenetic environment, and even potential regulatory introns (Zhang and Zhou, 2018). Indeed, in both of the systems studied here, the duplicated genes acquired a narrower expression pattern, being restricted to larval imaginal discs and male reproductive tissues, whereas the parental genes remained broadly expressed.

The divergence in expression between the duplicates and their parents has critical implications for their function. As we demonstrated in our expression knockdowns assays, the parental genes and their duplicates do not have a completely redundant expression pattern across different tissues and/or cell types. As a result, the phenotypic impact of each paralog depends on the tissues in which they are knocked down, which differs between the parental and duplicated genes. The prompt divergence in expression, likely facilitated by the random genomic insertion that created the new copies, may help to explain why the duplicates were retained in the genome, and not lost or pseudogenized over time (Zhang and Zhou, 2018).

Another key aspect common to both systems studied here is the rapid and asymmetrical divergence in protein sequence after duplication. While both *CAF40* and *TBPH* sequences remained highly conserved in the *Drosophila* phylogeny, their duplicates extensively diverged, even at conserved residues. Such divergence seems to be significant for their functions as well. For instance, we showed that Zeus protein fails to recruit partners from the CCR4-NOT complex; and Cocoon has a distinct genomic binding pattern compared to its parent, besides lacking the ancestral protein domain involved in mRNA splicing.

Such asymmetrical divergence in protein sequence was already reported for duplicates of other conserved genes, such as homeobox genes that regulate the development of metazoans (Holland et al., 2017; Leite et al., 2018). Remarkably, both parental genes studied here, *CAF40* and *TBPH*, are ancient genes, deeply conserved across eukaryotes and metazoans, respectively. Given the role of both parents in crucial cellular functions, it is possible that the mutations accumulated in the duplicates were not allowed in the constrained parental proteins because of potential deleterious pleiotropic effects on their structure. Once they were duplicated, though, the new copies were released from such strict negative selective pressure and allowed to diverge and increase the expression diversity of the processes they act on (Gu et al., 2004).

We also demonstrated that the duplicated genes have a widespread impact on global gene regulation. In both systems, the knockdown of the duplicates perturbed the expression of hundreds of genes, directly and indirectly. Again, this result is consistent with the well-described importance of their parental genes in important regulatory pathways. Our results show that the duplicates inherited the ability to engage in important regulatory processes, although their effects on gene expression are not identical to the parental genes, likely due to differences in their timing of expression, binding pattern in the genome, and/or protein function. These results are consistent with the idea that gene regulatory networks are constantly expanding and reshaping to accommodate novel genes (Voordeckers et al., 2015). In addition, the timing of expression of *Poseidon* is consistent with a potential compensation for its X-linked parental gene inactivation caused by meiotic sex chromosomal inactivation, as reported for other retrogenes (Vibranovski et al., 2009b).

On the other hand, the expression and functional divergence suggest that meiotic compensation is unlikely for *Zeus* and *Cocoon*.

In this context, it is notable that the three duplicated genes studied here independently acquired expression in the very same tissues, namely larval imaginal discs and testes. Notably, the critical developmental processes that take place in those tissues require fast and coordinated changes in gene regulation (Beira and Paro, 2016; White-Cooper, 2010). These observations suggest that processes involving the coordination of gene expression networks, despite being crucial and highly pleiotropic, are liable to the accumulation of new elements, such as newly arisen duplicated genes. Again, these results foster the idea that the duplication of key conserved elements may be common despite their predicted negative effects on developmental processes (Carroll, 2008), and that evolutionary young genes can carry out important, and even essential, functions in fundamental cellular processes.

Bibliography

1000 Genomes Project Consortium, T., Durbin, R.M., Altshuler (Co-Chair), D., Durbin (Co-Chair), R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061.

Abrusán, G. (2013). Integration of new genes into cellular networks, and their structural maturation. Genetics *195*, 1407–1417.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The Genome Sequence of Drosophila melanogaster. Science (80-.). *287*, 2185 LP-2195.

Arabidopsis Genome Initiative, T. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796.

Arthur, R.K., Ma, L., Slattery, M., Spokony, R.F., Ostapenko, A., Nègre, N., and White, K.P. (2014). Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. Genome Res. *24*, 1115–1124.

Ashburner, M., Misra, S., Roote, J., Lewis, S.E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., et al. (1999). An Exploration of the Sequence of a 2.9-Mb Region of the Genome of Drosophila melanogaster: The Adh Region. Genetics *153*, 179–219.

Assis, R., and Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila; Proc. Natl. Acad. Sci. *110*, 17409 LP-17414.

Assis, R., and Bachtrog, D. (2015). Rapid divergence and diversification of mammalian duplicate gene functions. BMC Evol. Biol. *15*, 1–7.

Assis, R., Zhou, Q., and Bachtrog, D. (2012). Sex-biased transcriptome evolution in Drosophila. Genome Biol. Evol. *4*, 1189–1200.

Ayala, Y.M., Pantano, S., D'Ambrogio, A., Buratti, E., Brindisi, A., Marchetti, C., Romano, M., and Baralle, F.E. (2005). Human, Drosophila, and C. elegans TDP43: Nucleic Acid Binding Properties and Splicing Regulatory Function. J. Mol. Biol. *348*, 575–588.

Ayala, Y.M., Pagani, F., and Baralle, F.E. (2006). TDP43 depletion rescues aberrant CFTR exon 9 skipping. FEBS Lett. *580*, 1339–1344.

Ayala, Y.M., Misteli, T., and Baralle, F.E. (2008). TDP-43 regulates retinoblastoma protein phosphorylation through the repression of cyclin-dependent kinase 6 expression. Proc. Natl. Acad. Sci. U. S. A. *105*, 3785–3789.

Bai, Y., Casola, C., Feschotte, C., and Betrán, E. (2007). Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol. *8*, 1–9.

Bai, Y., Casola, C., and Betrán, E. (2008). Evolutionary origin of regulatory regions of retrogenes in Drosophila. BMC Genomics *9*, 1–9.

Baker, C.C., and Fuller, M.T. (2007). Translational control of meiotic cell cycle progression

and spermatid differentiation in male germ cells by a novel eIF4G homolog. Development *134*, 2863–2869.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68.

Barrett, R.D.H., and Hoekstra, H.E. (2011). Molecular spandrels: tests of adaptation at the genetic level. Nat. Rev. Genet. *12*, 767–780.

Bassett, A., and Liu, J.-L. (2014). CRISPR/Cas9 mediated genome engineering in Drosophila. Methods *69*, 128–136.

Batada, N.N., Urrutia, A.O., and Hurst, L.D. (2007). Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. TIG *23*, 480–484.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. *20*, 1885–1898.

Beira, J. V, and Paro, R. (2016). The legacy of Drosophila imaginal discs. Chromosoma *125*, 573–592.

Belote, J.M., and Zhong, L. (2009). Duplicated proteasome subunit genes in Drosophila and their roles in spermatogenesis. Heredity (Edinb). *103*, 23–31.

Bergsten, J. (2005). A review of long-branch attraction. Cladistics *21*, 163–193.

Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007). Ohno ' s dilemma : Evolution of new genes under continuous selection. *104*.

Betrán, E., Thornton, K., and Long, M. (2002). Retroposed new genes out of the X in Drosophila. Genome Res. *12*, 1854–1859.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The Complete Genome Sequence of Escherichia coli. Science (80-.). *277*, 1453 LP-1462.

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096.

Bose, J.K., Wang, I.-F., Hung, L., Tarn, W.-Y., and Shen, C.-K.J. (2008). TDP-43 overexpression enhances exon 7 inclusion during the survival of motor neuron pre-mRNA splicing. J. Biol. Chem. *283*, 28852–28859.

Bradley, J., Baltus, A., Skaletsky, H., Royce-Tolland, M., Dewar, K., and Page, D.C. (2004). An X-to-autosome retrogene is required for spermatogenesis in mice. Nat. Genet. *36*, 872.

Britten, R.J., and Davidson, E.H. (1969). Gene Regulation for Higher Cells: A Theory. Science (80-.). *165*, 349 LP-357.

Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J.,

Park, S., Suzuki, A.M., et al. (2014). Diversity and dynamics of the Drosophila transcriptome. Nature *512*, 393.

Buratti, E., and Baralle, F.E. (2001). Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. J. Biol. Chem. *276*, 36337–36343.

Carelli, F.N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., and Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. *26*, 301–314.

Carroll, S.B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell *134*, 25–36.

Casneuf, T., De Bodt, S., Raes, J., Maere, S., and Van de Peer, Y. (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol. *7*, R13.

Casola, C., and Betrán, E. (2017). The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? Genome Biol. Evol. *9*, 1351–1373.

Charlesworth, B., Coyne, J.A., and Barton, N.H. (1987). The Relative Rates of Evolution of Sex Chromosomes and Autosomes. Am. Nat. *130*, 113–146.

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.-E., Lambert, N., de Marchena, J., Jin, W.-L., Vanderhaeghen, P., Ghosh, A., Sassa, T., et al. (2012). Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation. Cell *149*, 923–935.

Cheeseman, I.M., and Desai, A. (2008). Molecular architecture of the kinetochore–microtubule interface. Nat. Rev. Mol. Cell Biol. *9*, 33.

Chen, S., Zhang, Y.E., and Long, M. (2010). New genes in Drosophila quickly become essential. Science (80-.). *330*, 1682–1685.

Chen, S., Ni, X., Krinsky, B.H., Zhang, Y.E., Vibranovski, M.D., White, K.P., and Long, M. (2012). Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. EMBO J. *31*, 2798–2809.

Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. Nat. Rev. Genet. *14*, 645–660.

Chen, Y., Boland, A., Kuzuoğlu-Öztürk, D., Bawankar, P., Loh, B., Chang, C. Te, Weichenrieder, O., and Izaurralde, E. (2014a). A DDX6-CNOT1 Complex and W-Binding Pockets in CNOT9 Reveal Direct Links between miRNA Target Recognition and Silencing. Mol. Cell *54*, 737–750.

Chen, Z.X., Sturgill, D., Qu, J., Jiang, H., Park, S., Boley, N., Suzuki, A.M., Fletcher, A.R., Plachetzki, D.C., FitzGerald, P.C., et al. (2014b). Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res. *24*, 1209–1223.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203–218.

Collart, M.A. (2016). The Ccr4-Not complex is a key regulator of eukaryotic gene expression. Wiley Interdiscip. Rev. RNA *7*, 438–454.

Collart, M.A., and Panasenko, O.O. (2017). The Ccr4-Not Complex: Architecture and Structural Insights BT - Macromolecular Protein Complexes: Structure and Function. J.R. Harris, and J. Marles-Wright, eds. (Cham: Springer International Publishing), pp. 349–379.

Collins, F.S., Green, E.D., Guttmacher, A.E., and Mark, S. (2003). A vision for the future of genomics research. Nature *422*, 15–17.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job : How duplicated genes find new functions. *9*.

Consortium, C. elegans S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Consortium, T.C.S. and A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

Cooper, D.N., and Kehrer-Sawatzki, H. (2011). Exploring the potential relevance of human-specific genes to complex disease. Hum. Genomics *5*, 99–107.

Dai, H., Yoshimatsu, T.F., and Long, M. (2006). Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes. Gene *385*, 96–102.

Demuth, J.P., Bie, T. De, Stajich, J.E., Cristianini, N., and Hahn, M.W. (2006). The Evolution of Mammalian Gene Families. PLoS One *1*, e85.

Díaz-Castillo, C., and Ranz, J.M. (2012). Nuclear Chromosome Dynamics in the Drosophila Male Germ Line Contribute to the Nonrandom Genomic Distribution of Retrogenes. Mol. Biol. Evol. *29*, 2105–2108.

Ding, Y., Zhao, L., Yang, S., Jiang, Y., Chen, Y., Zhao, R., Zhang, Y., Zhang, G., Dong, Y., Yu, H., et al. (2010). A young drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. PLoS Genet. *6*, 1–12.

Ding, Y., Zhou, Q., and Wang, W. (2012). Origins of New Genes and Evolution of Their Novel Functions. Annu. Rev. Ecol. Evol. Syst. *43*, 345–363.

Ding, Y., Berrocal, A., Morita, T., Longden, K.D., and Stern, D.L. (2016). Natural courtship song variation caused by an intronic retroelement in an ion channel gene. Nature *536*, nature19093.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Drosophila 12 Genomes Consortium, T., Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 1–7.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Elde, N.C., Long, M., and Turkewitz, A.P. (2007). A role for convergent evolution in the secretory life of cells. Trends Cell Biol. *17*, 157–164.

Emerson, J.J., Kaessmann, H., Betrán, E., and Long, M. (2004). Extensive Gene Traffic on the Mammalian X Chromosome. Science (80-.). *303*, 537 LP-540.

Erwin, D.H., and Davidson, E.H. (2009). The evolution of hierarchical gene regulatory networks. Nat. Rev. Genet. *10*, 141.

Feiguin, F., Godena, V.K., Romano, G., D'Ambrogio, A., Klima, R., and Baralle, F.E. (2009). Depletion of TDP-43 affects Drosophila motoneurons terminal synapsis and locomotive behavior. FEBS Lett. *583*, 1586–1592.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (80-.). *269*, 496 LP-512.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary , Degenerative Mutations. Genetics 1531–1545.

Francino, M.P. (2005). An adaptive radiation model for the origin of new gene functions. Nat. Genet. *37*, 573–577.

Garapaty, S., Mahajan, M.A., and Samuels, H.H. (2008). Components of the CCR4-NOT complex function as nuclear hormone receptor coactivators via association with the NRC-interacting factor NIF-1. J. Biol. Chem. *283*, 6806–6816.

Garces, R.G., Gillon, W., and Pai, E.F. (2007). Atomic model of human Rcd-1 reveals an armadillo -like-repeat protein with in vitro nucleic acid binding properties. Protein Sci. *16*, 176–188.

Gnad, F., and Parsch, J. (2006). Sebida: A database for the functional and evolutionary analysis of genes with sex-biased expression. Bioinformatics *22*, 2577–2579.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 Genes. Science (80-.). *274*, 546 LP-567.

Goldschmidt, R. (1940). The Material Basis of Evolution. (John Wiley & Sons, Ltd).

Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J., and O'Connor-Giles, K.M. (2013). Genome Engineering of *Drosophila* with the CRISPR RNA-Guided Cas9 Nuclease. Genetics *194*, 1029–1035.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of

Drosophila melanogaster. Nature *471*, 473–479.

Gu, X., Zhang, Z., and Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc. Natl. Acad. Sci. U. S. A. *102*, 707–712.

Gu, Z., Rifkin, S.A., White, K.P., and Li, W.-H. (2004). Duplicate genes increase gene expression diversity within and between species. Nat. Genet. *36*, 577.

Gubala, A.M., Schmitz, J.F., Kearns, M.J., Vinh, T.T., Bornberg-Bauer, E., Wolfner, M.F., and Findlay, G.D. (2017). The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen de Novo. Mol. Biol. Evol. *34*, 1066–1082.

Hahn, M.W. (2009). Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. J. Hered. *100*, 605–617.

Halfon, M.S. (2017). Perspectives on Gene Regulatory Network Evolution. Trends Genet. *33*, 436–447.

Hardison, R.C. (2003). Comparative Genomics. PLOS Biol. *1*, 156–160.

Harrison, P.W., Wright, A.E., Zimmer, F., Dean, R., Montgomery, S.H., Pointer, M.A., and Mank, J.E. (2015). Sexual selection drives evolution and rapid turnover of male gene expression. Proc. Natl. Acad. Sci. *112*, 4393–4398.

Hazelett, D.J., Chang, J.-C., Lakeland, D.L., and Morton, D.B. (2012). Comparison of Parallel High-Throughput RNA Sequencing Between Knockout of TDP-43 and Its Overexpression Reveals Primarily Nonreciprocal and Nonoverlapping Gene Expression Changes in the Central Nervous System of Drosophila. G3 Genes|Genomes|Genetics *2*, 789–802.

Heinen, T.J.A.J., Staubach, F., Häming, D., and Tautz, D. (2009). Emergence of a New Gene from an Intergenic Region. Curr. Biol. *19*, 1527–1531.

Hiller, M. (2004). Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. Development *131*, 5297–5308.

Holland, P.W.H., Marlétaz, F., Maeso, I., Dunwell, T.L., and Paps, J. (2017). New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *372*, 20150480.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

International Human Genome Sequencing Consortium, T., Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860.

Jacob, F. (1977). Evolution and tinkering. Science (80-.). *196*, 1161–1166.

Jangam, D., Feschotte, C., and Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet. *33*, 817–831.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

Johnson, A.D. (2017). The rewiring of transcription circuits in evolution. Curr. Opin. Genet.

Dev. *47*, 121–127.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: Mechanistic and evolutionary insights. Nat. Rev. Genet. *10*, 19–31.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.

Kemkemer, C., and Long, M. (2014). New genes important for development. EMBO Rep. *15*, 460–461.

Kim, J., Kim, I., Han, S.K., Bowie, J.U., and Kim, S. (2012). Network rewiring is an important mechanism of gene essentiality change. Sci. Rep. *2*.

Kimura, M., and Ohta, T. (1974). On some principles governing molecular evolution. Proc. Natl. Acad. Sci. U. S. A. *71*, 2848–2852.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell *155*, 27–38.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B Biol. Sci. *279*, 5048–5057.

Kuo, P.-H., Doudeva, L.G., Wang, Y.-T., Shen, C.-K.J., and Yuan, H.S. (2009). Structural insights into TDP-43 in nucleic-acid binding and domain interactions. Nucleic Acids Res. *37*, 1799–1808.

Lan, X., and Pritchard, J.K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science (80-.). *352*, 1009 LP-1013.

Lee, Y.C.G., Ventura, I.M., Rice, G.R., Chen, D.-Y., and Long, M. (2018). Rapid evolution of gained essential developmental functions of a young gene via interactions with other essential genes. BioRxiv 226936.

Legrand, J.M.D., and Hobbs, R.M. (2018). RNA processing in the male germline: Mechanisms and implications for fertility. Semin. Cell Dev. Biol. *79*, 80–91.

Leite, D.J., Baudouin-Gonzalez, L., McGregor, A.P., Iwasaki-Yokozawa, S., Akiyama-Oda, Y., Oda, H., Pisani, D., Lozano-Fernandez, J., Turetzek, N., Prpic, N.-M., et al. (2018). Homeobox Gene Duplication and Divergence in Arachnids. Mol. Biol. Evol. *35*, 2240–2253.

Lewin, B., Krebs, J. E., Kilpatrick, S. T., Goldstein, E. (2011). Lewin's Genes X.

Li, Y., Ray, P., Rao, E.J., Shi, C., Guo, W., Chen, X., Woodruff, E.A., Fushimi, K., and Wu, J.Y. (2010). A Drosophila model for TDP-43 proteinopathy. Proc. Natl. Acad. Sci. *107*, 3169–3174.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose

program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lin, M.-J., Cheng, C.-W., and Shen, C.-K.J. (2011). Neuronal Function and Dysfunction of Drosophila dTDP. PLoS One *6*, e20371.

Linnen, C.R., Poh, Y.-P., Peterson, B.K., Barrett, R.D.H., Larson, J.G., Jensen, J.D., and Hoekstra, H.E. (2013). Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. Science (80-.). *339*, 1312–1316.

Long, M., and Langley, C.H. (1993). Natural selection and the origin of jingwey, a chimeric processed functional gene in Drosphila. Science (80-.). *260*, 91–95.

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: Glimpses from the young and old. Nat. Rev. Genet. *4*, 865–875.

Long, M., Vibranovski, M.D., and Zhang, Y.E. (2012). Evolutionary interactions between sex chromosomes and autosomes. In Rapidly Evolving Genes and Genetic Systems, (Oxford: Oxford University Press), p.

Long, M., VanKuren, N.W., Chen, S., and Vibranovski, M.D. (2013). New Gene Evolution: Little Did We Know. Annu. Rev. Genet. *47*, 307–333.

Loppin, B., Lepetit, D., Dorus, S., Couble, P., and Karr, T.L. (2005). Origin and Neofunctionalization of a Drosophila Paternal Effect Gene Essential for Zygote Viability. Curr. Biol. *15*, 87–93.

Losos, J.B., Arnold, S.J., Bejerano, G., Iii, E.D.B., Hibbett, D., Moritz, C., Orr, H.A., Hoekstra, H.E., and Mindell, D.P. (2013). Evolutionary Biology for the 21st Century. PLoS Biol. *11*, e1001466.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lynch, M., and Conery, J.S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science *290*, 1151–1155.

Mackay, T.F.C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat. Rev. Genet. *15*, 22–33.

Makova, K.D., and Li, W.-H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. *13*, 1638–1645.

Mathys, H., Basquin, J., Ozgur, S., Czarnocki-Cieciura, M., Bonneau, F., Aartse, A., Dziembowski, A., Nowotny, M., Conti, E., and Filipowicz, W. (2014). Structural and Biochemical Insights to the Role of the CCR4-NOT Complex and DDX6 ATPase in MicroRNA Repression. Mol. Cell *54*, 751–765.

Matsuno, M., Compagnon, V., Schoch, G.A., Schmitt, M., Debayle, D., Bassard, J.-E., Pollet, B., Hehn, A., Heintz, D., Ullmann, P., et al. (2009). Evolution of a novel phenolic pathway for pollen development. Science *325*, 1688–1692.

Meisel, R.P., Han, M. V, and Hahn, M.W. (2009). A complex suite of forces drives gene traffic from Drosophila X chromosomes. Genome Biol. Evol. *1*, 176–188.

Miklos, G.L.G., and Rubin, G.M. (1996). The role of the genome project in determining gene

function: Insights from model organisms. Cell *86*, 521–529.

Miller, J.E., and Reese, J.C. (2012). Ccr4-Not complex: The control freak of eukaryotic cells. Crit. Rev. Biochem. Mol. Biol. *47*, 315–333.

Moore, R.C., and Purugganan, M.D. (2005). The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol. *8*, 122–128.

Mouse Genome Sequencing Consortium, T., Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.W., Mardis, E.R., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520.

Muller, H.J. (1935). The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetica *17*, 237–252.

Mummery-Widmer, J.L., Yamazaki, M., Stoeger, T., Novatchkova, M., Bhalerao, S., Chen, D., Dietzl, G., Dickson, B.J., and Knoblich, J.A. (2009). Genome-wide analysis of Notch signalling in Drosophila by transgenic RNAi. Nature *458*, 987–992.

Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science (80-.). *347*, 1258522.

Neely, G.G., Hess, A., Costigan, M., Keene, A.C., Goulas, S., Langeslag, M., Griffin, R.S., Belfer, I., Dai, F., Smith, S.B., et al. (2010). A Genome-wide Drosophila Screen for Heat Nociception Identifies α2δ3 as an Evolutionarily Conserved Pain Gene. Cell *143*, 628–638.

Nurminsky, D.I., Nurminskaya, M. V, De Aguiar, D., and Hartl, D.L. (1998). Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature *396*, 572–575.

Ohno, S. (1970). Evolution by gene duplication. (John Wiley & Sons, Ltd).

Pavlicev, M., and Wagner, G.P. (2012). A model of developmental evolution: Selection, pleiotropy and compensation. Trends Ecol. Evol. *27*, 316–322.

Perkins, L.A., Holderbaum, L., Tao, R., Hu, Y., Sopko, R., McCall, K., Yang-Zhou, D., Flockhart, I., Binari, R., Shim, H.-S., et al. (2015). The Transgenic RNAi Project at Harvard Medical School: Resources and Validation. Genetics *201*, 843–852.

Piatigorsky, J. (1991). The recruitment of crystallins: new functions precede gene duplication. Science (80-.). *252*, 1078 LP-1079.

Posada, D. (2008). jModelTest: Phylogenetic Model Averaging. Mol. Biol. Evol. *25*, 1253–1256.

Potrzebowski, L., Vinckenbosch, N., Marques, A.C., and Kaessmann, H. (2008). Chromosomal Gene Movements Reflect the Recent Origin and Biology of Therian Sex Chromosomes. PLoS Biol. *6*, 6:e80.

Quezada-Díaz, J.E., Muliyil, T., Río, J., and Betrán, E. (2010). Drcd-1 related: a positively selected spermatogenesis retrogene in Drosophila. Genetica *138*, 925–937.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ranz, J.M., and Parsch, J. (2012). Newly evolved genes: Moving from comparative genomics to functional studies in model systems. BioEssays *34*, 477–483.

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D., and Izaurralde, E. (2005). A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. Rna *11*, 1640–1647.

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J., and Jones, C.D. (2013). De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLOS Genet. *9*, e1003860.

Rice, W.R. (1984). SEX CHROMOSOMES AND THE EVOLUTION OF SEXUAL DIMORPHISM. Evolution (N. Y). *38*, 735–742.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47–e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol. *61*, 539–542.

Ross, B.D., Rosin, L., Thomae, A.W., Hiatt, M.A., Vermaak, D., Cruz, A.F.A. de la, Imhof, A., Mellone, B.G., and Malik, H.S. (2013). Stepwise Evolution of Essential Centromere Function in a Drosophila Neogene. Science *340*, 1211–1214.

Rost, S., Fregin, A., Ivaskevicius, V., Conzelmann, E., Hörtnagel, K., Pelz, H.-J., Lappegard, K., Seifried, E., Scharrer, I., Tuddenham, E.G.D., et al. (2004). Mutations in *VKORC1* cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature *427*, nature02214.

Saleem, S., Schwedes, C.C., Ellis, L.L., Grady, S.T., Adams, R.L., Johnson, N., Whittington, J.R., and Carney, G.E. (2012). Drosophila melanogaster p24 trafficking proteins have vital roles in development and reproduction. Mech. Dev. *129*, 177–191.

Schnorrer, F., Schönbauer, C., Langer, C.C.H., Dietzl, G., Novatchkova, M., Schernhuber, K., Fellner, M., Azaryan, A., Radolf, M., Stark, A., et al. (2010). Systematic genetic analysis of muscle morphogenesis and function in Drosophila. Nature *464*, 287–291.

Sémon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. Curr. Opin. Genet. Dev. *17*, 505–512.

Sgromo, A., Raisch, T., Bawankar, P., Bhandari, D., Chen, Y., Kuzuoglu-Öztürk, D., Weichenrieder, O., and Izaurralde, E. (2017). A CAF40-binding motif facilitates recruitment of the CCR4-NOT complex to mRNAs targeted by Drosophila Roquin. Nat. Commun. *8*.

Sgromo, A., Raisch, T., Backhaus, C., Keskeny, C., Alva, V., Weichenrieder, O., and Izaurralde, E. (2018). Drosophila Bag-of-marbles directly interacts with the CAF40 subunit of the CCR4–NOT complex to elicit repression of mRNA targets. Rna *24*, 381–395.

Shang, B., Gao, A., Pan, Y., Zhang, G., Tu, J., Zhou, Y., Yang, P., Cao, Z., Wei, Q., Ding, Y., et al. (2014). CT45A1 acts as a new proto-oncogene to trigger tumorigenesis and cancer metastasis. Cell Death &Amp; Dis. *5*, e1285.

Snel, B., Bork, P., and Huynen, M. (2000). Genome evolution: gene fusion versus gene fission. Trends Genet. *16*, 9–11.

Sorourian, M., Kunte, M.M., Domingues, S., Gallach, M., Özdil, F., Río, J., and Betrán, E. (2014). Relocation facilitates the acquisition of short cis-regulatory regions that drive the expression of retrogenes during spermatogenesis in Drosophila. Mol. Biol. Evol. *31*, 2170–2180.

Sreedharan, J., Blair, I.P., Tripathi, V.B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J.C., Williams, K.L., Buratti, E., et al. (2008). TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. Science *319*, 1668–1672.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535-539.

Stauber, M., Jäckle, H., and Schmidt-Ott, U. (1999). The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene. Proc. Natl. Acad. Sci. U. S. A. *96*, 3786–3789.

Stern, D.L., and Orgogozo, V. (2008). The Loci of Evolution: how predictable is genetic evolution? Evolution (N. Y). *62*, 2155–2177.

Tautz, D. (2014). The Discovery of De Novo Gene Evolution. Perspect. Biol. Med. *57*, 149–161.

Taylor, J.S., and Raes, J. (2004). Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annu. Rev. Genet. *38*, 615–643.

Toups, M.A., and Hahn, M.W. (2010). Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. Genetics *186*, 763–766.

Tritschler, F., Eulalio, A., Truffault, V., Hartmann, M.D., Helms, S., Schmidt, S., Coles, M., Izaurralde, E., and Weichenrieder, O. (2007). A Divergent Sm Fold in EDC3 Proteins Mediates DCP1 Binding and P-Body Targeting . Mol. Cell. Biol. *27*, 8600–8611.

Tritschler, F., Eulalio, A., Helms, S., Schmidt, S., Coles, M., Weichenrieder, O., Izaurralde, E., and Truffault, V. (2008). Similar Modes of Interaction Enable Trailer Hitch and EDC3 To Associate with DCP1 and Me31B in Distinct Protein Complexes . Mol. Cell. Biol. *28*, 6695–6708.

Vankuren, N.W., and Long, M. (2018). Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. Nat. Ecol. Evol. *2*, 705–712.

VanKuren, N.W., and Vibranovski, M.D. (2014). A novel dataset for identifying sex-biased genes in Drosophila. J. Genomics *2*, 64–67.

Vazquez, J.M., Sulak, M., Chigurupati, S., and Lynch, V.J. (2018). A Zombie LIF Gene in Elephants Is Upregulated by TP53 to Induce Apoptosis in Response to DNA Damage. Cell Rep. *24*, 1765–1776.

Venken, K.J.T., Schulze, K.L., Haelterman, N.A., Pan, H., He, Y., Evans-Holm, M., Carlson, J.W.,

Levis, R.W., Spradling, A.C., Hoskins, R.A., et al. (2011). MiMIC: a highly versatile transposon insertion resource for engineering Drosophila melanogaster genes. Nat. Methods *8*, 737–743.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science (80-.). *291*, 1304 LP-1351.

Ventura, I., and Long, M. (2017). Connecting evolutionary genomics to cell biology. In Encyclopedia of Cell Biology, (Elsevier), pp. 153–159.

Vibranovski, M.D. (2014). Meiotic sex chromosome inactivation in Drosophila. J. Genomics *2*, 104–117.

Vibranovski, M.D., Zhang, Y., and Long, M. (2009a). General gene movement off the X chromosome in the Drosophila genus. Genome Res. *19*, 897–903.

Vibranovski, M.D., Lopes, H.F., Karr, T.L., and Long, M. (2009b). Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet. *5*.

Vibranovski, M.D., Zhang, Y.E., Kemkemer, C., Lopes, H.F., Karr, T.L., and Long, M. (2012). Re-analysis of the larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-specific gene expression related to the drosophila X chromosome. BMC Biol. *10*, 49.

Vicoso, B., and Charlesworth, B. (2009). The deficit of male-biased genes on the D. melanogaster X chromosome is expression-dependent: a consequence of dosage compensation? J. Mol. Evol. *68*, 576–583.

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proc. Natl. Acad. Sci. U. S. A. *103*, 3220 LP-3225.

Voordeckers, K., Pougach, K., and Verstrepen, K.J. (2015). How do regulatory networks evolve and expand throughout evolution? Curr. Opin. Biotechnol. *34*, 180–188.

Wagner, A. (2002). Asymmetric functional divergence of duplicate genes in yeast. Mol. Biol. Evol. *19*, 1760–1768.

Wainszelbaum, M.J., Charron, A.J., Kong, C., Kirkpatrick, D.S., Srikanth, P., Barbieri, M.A., Gygi, S.P., and Stahl, P.D. (2008). The hominoid-specific oncogene TBC1D3 activates Ras and modulates epidermal growth factor receptor signaling and trafficking. J. Biol. Chem. *283*, 13233–13242.

Wang, J., Long, M., and Vibranovski, M.D. (2012). Retrogenes moved out of the Z chromosome in the silkworm. J. Mol. Evol. *74*, 113–126.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096 LP-1101.

Wang, W., Zhang, J., Alvarez, C., Llopart, A., and Long, M. (2000). The Origin of the Jingwei Gene and the Complex Modular Structure of Its Parental Gene, Yellow Emperor, in Drosophila melanogaster. Mol. Biol. Evol. *17*, 1294–1301.

White-Cooper, H. (2010). Molecular mechanisms of gene regulation during Drosophila spermatogenesis. Reproduction *139*, 11–21.

White-Cooper, H. (2012). Tissue, cell type and stage-specific ectopic gene expression and RNAi induction in the Drosophila testis. Spermatogenesis *2*, 11–22.

Wilinski, D., Buter, N., Klocko, A.D., Lapointe, C.P., Selker, E.U., Gasch, A.P., and Wickens, M. (2017). Recurrent rewiring and emergence of RNA regulatory networks. Proc. Natl. Acad. Sci. *114*, E2816–E2825.

Wilson, A.C., Carlson, S.S., and White, T.J. (1977). Biochemical Evolution. Annu. Rev. Biochem. *46*, 573–639.

Wittkopp, P.J., Stewart, E.E., Arnold, L.L., Neidert, A.H., Haerum, B.K., Thompson, E.M., Akhras, S., Smith-Winberry, G., and Shefner, L. (2009). Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. Science *326*, 540–544.

Wray, G.A. (2007). The evolutionary significance of *cis*-regulatory mutations. Nat. Rev. Genet. *8*, nrg2063.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yeh, S.-D., Do, T., Chan, C., Cordova, A., Carranza, F., Yamamoto, E.A., Abbassi, M., Gandasetiawan, K.A., Librado, P., Damia, E., et al. (2012). Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition. Proc. Natl. Acad. Sci.

Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., and Gerstein, M. (2004). Genomic analysis of essentiality within protein networks. Trends Genet. *20*, 227–231.

Zekri, L., Kuzuoğlu-Öztürk, D., and Izaurralde, E. (2013). GW182 proteins cause PABP dissociation from silenced miRNA targets in the absence of deadenylation. EMBO J. *32*, 1052–1065.

Zhang, J., and Zhou, Q. (2018). On the regulatory evolution of new genes throughout their life history. Mol. Biol. Evol. msy206-msy206.

Zhang, Y.E., and Long, M. (2014). New genes contribute to genetic and phenotypic novelties in human evolution. Curr. Opin. Genet. Dev. *29*, 90–96.

Zhang, W., Landback, P., Gschwend, A.R., Shen, B., and Long, M. (2015). New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol. *16*, 202.

Zhang, Y.E., Vibranovski, M.D., Krinsky, B.H., and Long, M. (2010a). Age-dependent chromosomal distribution of male-biased genes in Drosophila. Genome Res. *20*, 1526–1533.

Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A.B., and Long, M. (2010b). Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. *8*.

Zhang, Y.E., Landback, P., Vibranovski, M.D., and Long, M. (2011). Accelerated Recruitment of New Brain Development Genes into the Human Genome. PLoS Biol *9*, e1001179.

Zhang, Y.E., Landback, P., Vibranovski, M., and Long, M. (2012). New genes expressed in human brains: Implications for annotating evolving genomes. BioEssays 982–991.

Zheng, Y., Bi, J., Hou, M.-Y., Shen, W., Zhang, W., Ai, H., Yu, X.-Q., and Wang, Y.-F. (2018). Ocnus is essential for male germ cell development in Drosophila melanogaster. Insect Mol. Biol. *27*, 545–555.

Zhong, L., and Belote, J.M. (2007). The testis-specific proteasome subunit Pros 6T of D. melanogaster is required for individualization and nuclear maturation during spermatogenesis. Development *134*, 3517–3525.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in Drosophila. Genome Res. *18*, 1446–1455.

Appendix

Supplementary Figures for Chapter 3

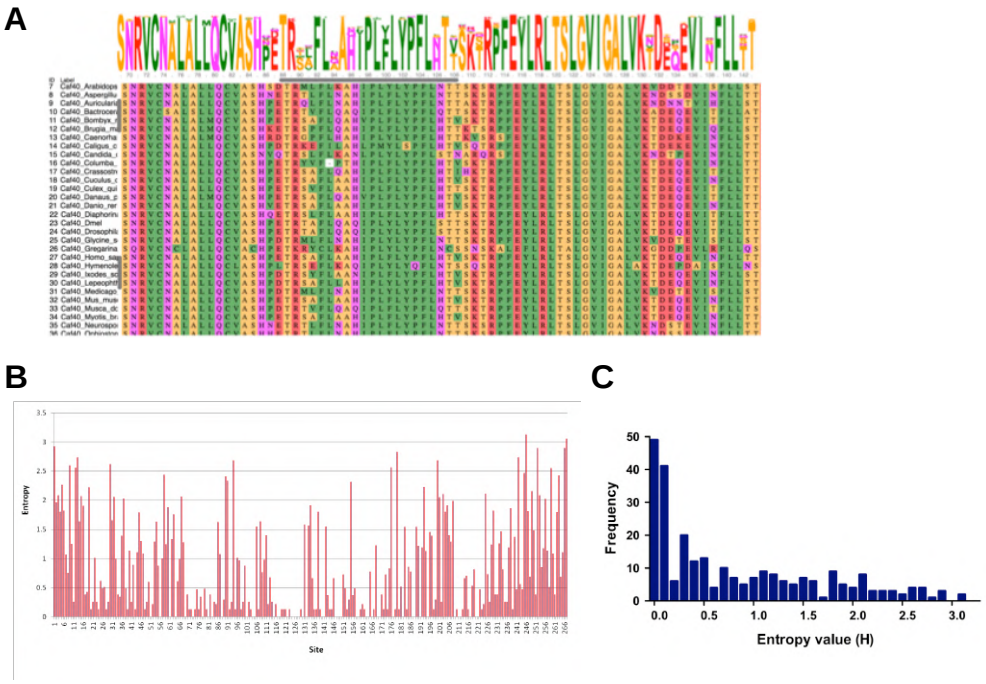


Figure S3.1. Shannon's entropy (H) for each residue in the CAF40 alignment from eukaryotic orthologs. A) Part of the alignment among CAF40 orthologs from 56 eukaryotes, with conserved and variable residues. B) Entropy values (H) for each residue in CAF40. C) Histogram plot of entropy values calculated for CAF40. The two bars on the left represent very low values, i.e. highly conserved residues in the alignment.

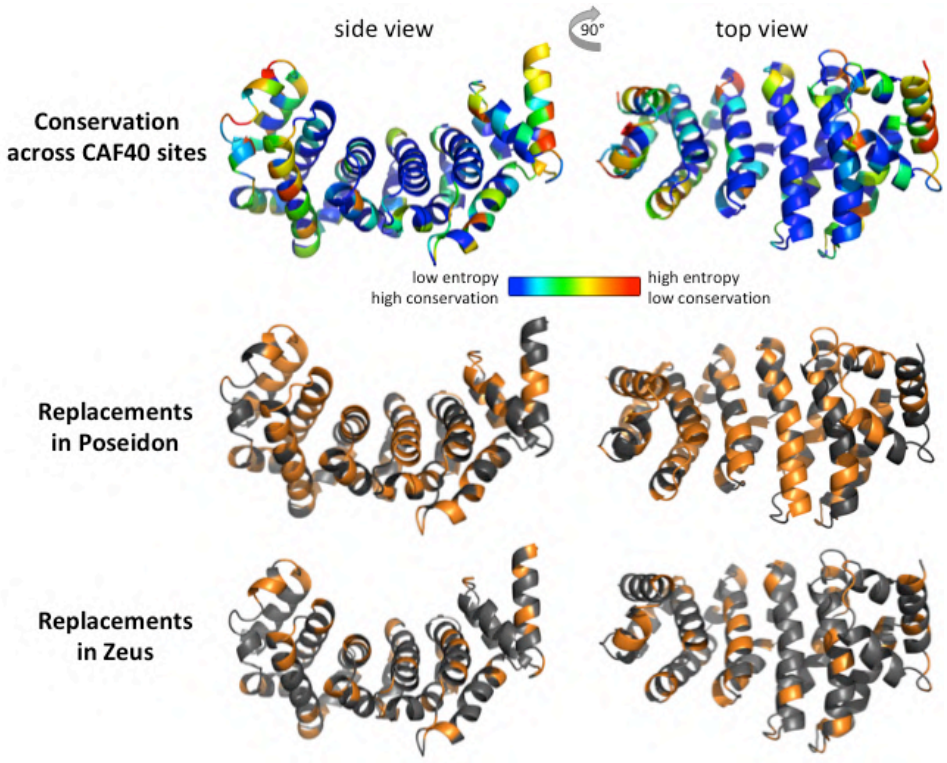


Figure S3.2. **The duplicates diverged even at highly conserved sites in CAF40**. At the top row, entropy values for each site in an alignment of CAF40 homologs from 56 eukaryotes were mapped into the known protein structure of CAF40 in *D. melanogaster* (left and right are two different views; blue represents the lowest entropy, i.e. high conservation, whereas red depicts the highest entropy). At the middle and bottom rows, sites that diverged between CAF40 and Poseidon and Zeus, respectively, are highlighted in orange. Notice that replacements are distributed throughout the protein structure, including the groove, which interacts with other proteins and nucleic acids.

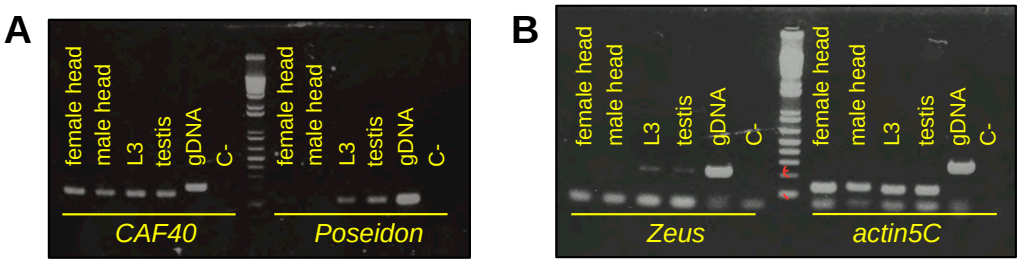


Figure S3.3. Agarose gels confirming the expression of each paralog at different tissues from *D. melanogaster*. RNA extraction was followed by DNAse treatment and reverse transcribed for each sample. Primers for *actin5C* were used as positive control. Small differences between the size of reverse-transcribed and gDNA bands are due to the presence of introns in *CAF40* and *actin5C*. A) *CAF40* and *Poseidon*; B) *Zeus* and *actin5C* control.

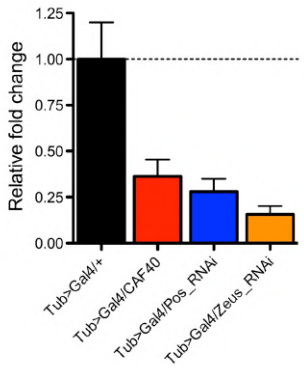


Figure S3.4. Constitutive RNAi-knockdown efficiency measured through quatitative PCR. Expression levels of each target paralog relative to the normalized control (+/Tub>GAL4 driver), in larvae with the indicated genotypes. Quantitative PCR results were normalized using the $\Delta\Delta C_T$ method to Rp49 product. Bars show means and SD for three replicates.

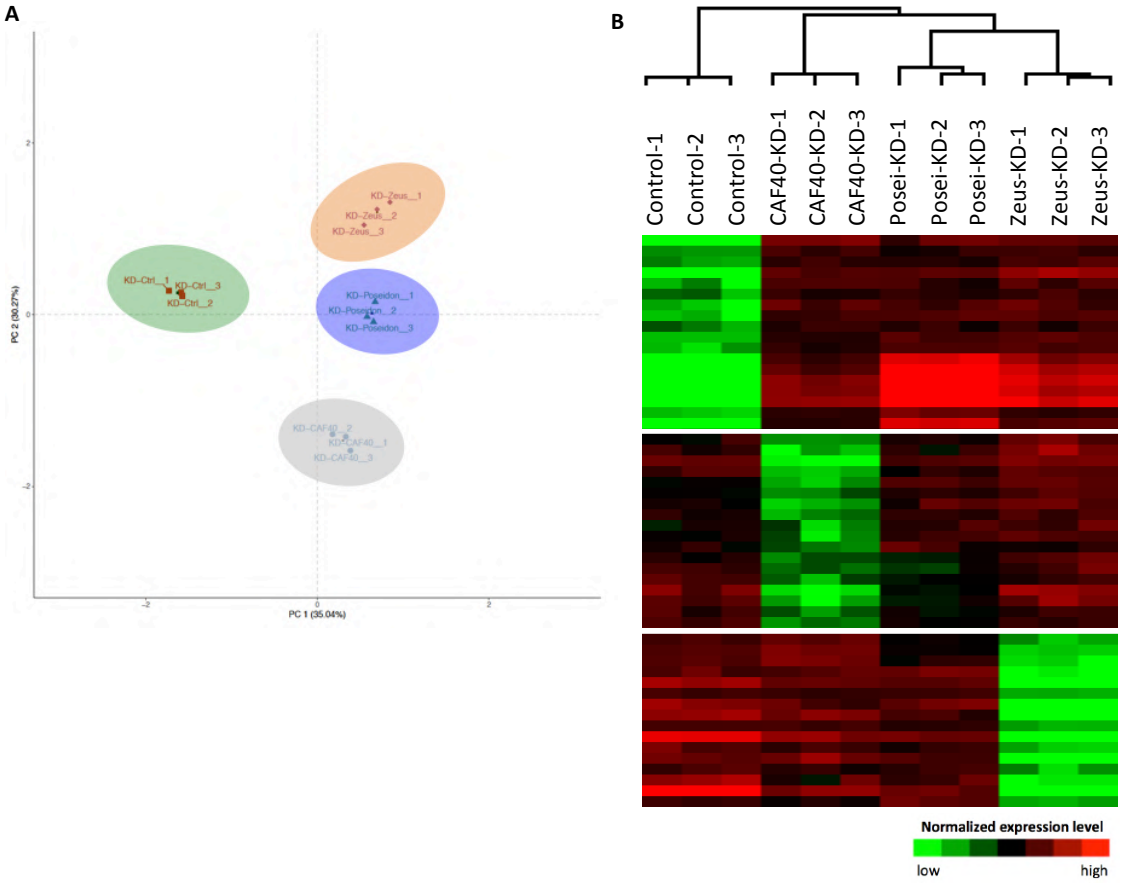


Figure S3.5. Clustering of samples upon knockdown in the RNA-seq assay. A) PCA plot of all the replicates in the control (green), *CAF40*-KD (grey), *Poseidon*-KD (blue) and *Zeus*-KD (orange). The x-axis represent PC1, and y-axis, PC2. B) Heatmap showing

(Figure S3.5, continued) normalized expression values of each sample for selected genes upon *CAF40*, *Poseidon* and *Zeus* independent knockdown, compared to the control. Each row represents a different gene, selected among all the differentially expressed ones, and each column represents a different replicate. Top panel: genes commonly affected by the knockdown of the three paralogs relative to the control; middle panel: genes only affected by *CAF40* knockdown; bottom panel: genes only affected by *Zeus* knockdown.

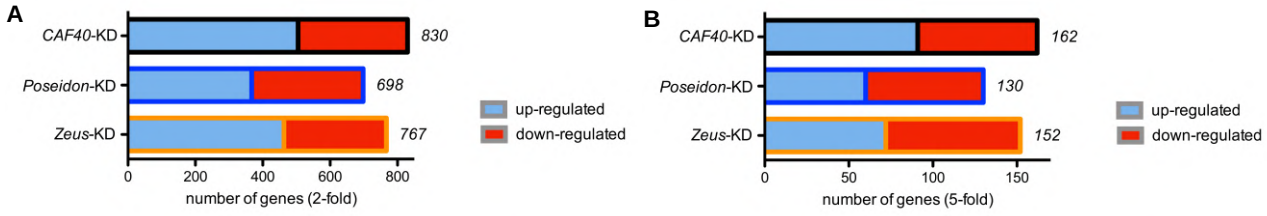


Figure S3.6. Impact of *CAF40*, *Poseidon* and *Zeus* knockdowns on global gene expression. Number of genes with >2-fold change (panel A) and >5-fold change (panel B) in expression compared to the control in the knockdown of each paralog measured through RNA-seq. Up-regulated genes shown in blue, down-regulated shown in red.

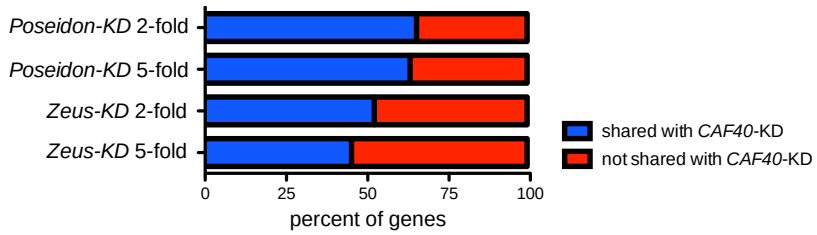


Figure S3.7. Proportion of genes differentially expressed upon *Poseidon* and *Zeus* knockdown that are also impacted by *CAF40*. Notice that *Poseidon* affects a higher proportion of genes shared with *CAF40* at both 2-fold and 5-fold, compared to *Zeus.*

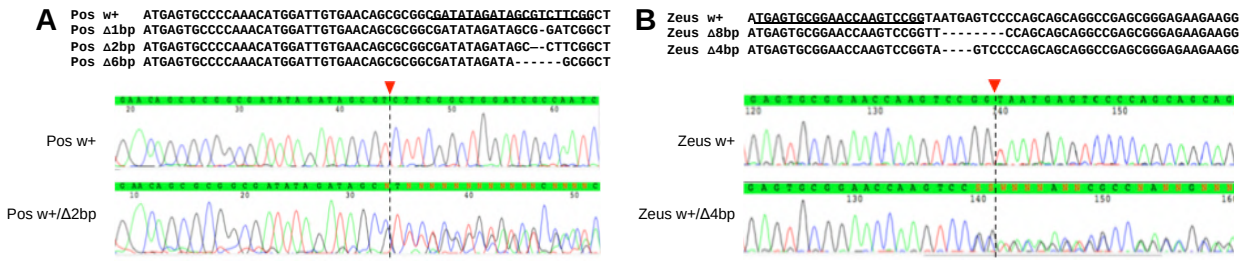


Figure S3.8. Knockout mutants for *Poseidon* (A) and *Zeus* (B) generated through CRISPR-Cas9. Top panel: wild-type and mutant gene sequences for both duplicates confirmed through Sanger-sequencing. Bottom panel: sequencing chromatogram depicting a wild-type (top) and a heterozygous mutant (bottom). Red triangle represents the expected cut site. Notice the multiple peaks downstream of the cut sites, reflecting heterozygous alleles.

Supplementary Tables for Chapter 3

Table S3.1. Poseidon and Zeus accumulated mutations at functionally important residues in CAF40. Below, replaced residues in Poseidon and/or Zeus proteins that were experimentally assayed in previous studies, and shown to be important for CAF40 function (Chen et al., 2014a; Mathys et al., 2014; Sgromo et al., 2017) **.**

Residue in CAF40 (position in *Dmel*)	Residue in Poseidon	Residue in Zeus	Structural role
N(58)	H	K	interaction interface with NOT1
F(60)	F	V	interaction interface with NOT1
C(64)	S	C	interaction interface with NOT1
T(78)	Y	T	interaction interface with NOT1
P(79)	Q	T	interaction interface with NOT1
L(117)	C	M	interaction interface with NOT1
C(200)	S	C	interaction interface with NOT1
Y(203)	L	H	interaction interface with NOT1
H(208)	H	Q	interaction interface with NOT1
A(248)	L	A	interaction interface with NOT1
A(84)	G	A	interaction interface with Roquin
N(88)	H	N	interaction interface with Roquin
R(130)	R	E	interaction interface with Roquin
P(131)	A	P	interaction interface with Roquin
Y(134)	H	Q	interaction interface with Roquin
G(144)	N	C	interaction interface with Roquin
L(177)	M	L	interaction interface with Roquin
T(180)	I	V	interaction interface with Roquin
V(181)	L	A	interaction interface with Roquin
F(184)	F	S	interaction interface with Roquin

Table S3.2. RNAi-knockdown efficiency for reducing the expression of each paralog in our RNA-seq assays. Notice that each knockdown significantly reduced the expression of its targeted gene, while not affecting the expression of the other two paralogs.

Knockdown target (RNAi)	Effect on paralog	log2fold change	fold change	p value	q value	significant?
***CAF40*-KD**	*CAF40*	-1.4304	0.3710	2.337 E-49	4.331 E-47	Significant
	Poseidon	-0.1075	0.9285	0.2484	0.4077	NotSignificant
	Zeus	0.0091	1.0060	0.9536	0.9733	NotSignificant
***Pos*-KD**	*CAF40*	0.0833	1.0592	0.3739	0.5230	NotSignificant
	Poseidon	-1.2700	0.4147	7.541 E-37	3.094 E-34	Significant
	Zeus	0.1030	1.0739	0.5034	0.7671	NotSignificant
***Zeus*-KD**	*CAF40*	0.0563	1.0396	0.5483	0.6847	NotSignificant
	Poseidon	-0.1517	0.9002	0.0987	0.2012	NotSignificant
	Zeus	-3.3262	0.0997	6.184 E-63	1.974 E-60	Significant

Table S3.3. List of Gene Onthology terms with the top 10 most significant enrichment among the genes differentially expressed upon *CAF40*, *Poseidon* and *Zeus* knockdown according to our RNA-seq analysis. Only genes with p-values lower than 10^{-4} and false discovery rate q-value lower than 0.05 are shown (N=total number of genes; B=total number of genes associated with a GO term; n=number of genes with differential expression; b=number of genes in the intersection).

DEG sample	GO Term	Description	p-value	q-value	Enrich.	N	B	n	b
CAF40-KD: Process	GO:0006508	proteolysis	5.21E-11	3.76E-07	1.67	9210	635	1270	146
	GO:0006631	fatty acid metabolic process	1.09E-07	3.93E-04	2.49	9210	102	1270	35
	GO:0017085	response to insecticide	7.39E-07	1.33E-03	4.99	9210	16	1270	11
	GO:0006030	chitin metabolic process	1.33E-06	1.91E-03	2.83	9210	59	1270	23
	GO:0009593	detection of chemical stimulus	3.81E-06	4.59E-03	3.11	9210	42	1270	18
	GO:0006633	fatty acid biosynthetic process	1.08E-05	9.77E-03	2.67	9210	57	1270	21
	GO:0006040	amino sugar metabolic process	2.43E-05	1.95E-02	2.38	9210	73	1270	24
	GO:1901071	glucosamine-containing compound metabolic process	3.61E-05	2.61E-02	2.38	9210	70	1270	23
	GO:0006022	aminoglycan metabolic process	4.67E-05	3.06E-02	2.22	9210	85	1270	26

Table S3.3: Continued from previous page.

DEG sample	GO Term	Description	p-value	q-value	Enrich.	N	B	n	b
CAF40-KD: Process	GO:0006508	proteolysis	5.21E-11	3.76E-07	1.67	9210	635	1270	146
	GO:0016042	lipid catabolic process	6.90E-05	4.15E-02	2.21	9210	82	1270	25
CAF40-KD: Function	GO:0008236	serine-type peptidase	6.35E-37	1.62E-33	3.87	9210	180	1270	96
	GO:0017171	serine hydrolase activity	1.18E-36	1.51E-33	3.85	9210	181	1270	96
	GO:0004252	serine-type endopeptidase	1.41E-35	1.21E-32	4.01	9210	159	1270	88
	GO:0008233	peptidase activity	2.15E-25	1.10E-22	2.44	9210	405	1270	136
	GO:0004175	endopeptidase activity	1.29E-22	5.48E-20	2.67	9210	277	1270	102
	GO:0016787	hydrolase activity	5.04E-11	1.84E-08	1.41	9210	144	1270	280
	GO:0004180	carboxypeptidase activity	3.79E-08	1.21E-05	4.83	9210	21	1270	14
	GO:0032450	maltose alpha-glucosidase	1.54E-07	4.38E-05	6.53	9210	10	1270	9
	GO:0090599	alpha-glucosidase activity	7.44E-07	1.58E-04	5.93	9210	11	1270	9
	GO:0003824	catalytic activity	1.02E-06	2.01E-04	1.17	9210	324	1270	524
CAF40-KD: Component	GO:0044421	extracellular region part	3.98E-09	4.97E-06	1.62	9210	592	1270	132
	GO:0031226	intrinsic component of plasma membrane	3.17E-08	1.98E-05	1.84	9210	303	1270	77
	GO:0005887	integral component of plasma membrane	4.26E-07	1.77E-04	1.78	9210	294	1270	72
	GO:0005615	extracellular space	6.89E-07	2.15E-04	1.58	9210	472	1270	103
	GO:0044459	plasma membrane part	1.37E-06	3.42E-04	1.51	9210	561	1270	117
	GO:0005576	extracellular region	1.74E-06	3.63E-04	1.74	9210	288	1270	69
	GO:0016021	integral component of membrane	9.62E-06	1.50E-03	1.35	9210	918	1270	171
	GO:0005777	peroxisome	3.45E-05	4.79E-03	2.29	9210	79	1270	25
	GO:0042579	microbody	3.72E-05	4.65E-03	2.24	9210	84	1270	26
	GO:0031012	extracellular matrix	4.14E-05	4.70E-03	2.16	9210	94	1270	28
Pos-KD: Process	GO:0006508	proteolysis	7.63E-14	5.50E-10	1.81	9210	635	1170	146
	GO:0007606	sensory perception of	5.61E-06	2.02E-02	2.4	9210	92	1170	28
	GO:0050909	sensory perception of taste	1.04E-05	2.50E-02	3.94	9210	24	1170	12
	GO:0006030	chitin metabolic process	2.11E-05	3.81E-02	2.67	9210	59	1170	20
	GO:0006022	aminoglycan metabolic process	9.87E-05	1.42E-01	2.22	9210	85	1170	24
Pos-KD: Function	GO:0008236	serine-type peptidase	8.90E-34	1.14E-30	3.89	9210	180	1170	89
	GO:0017171	serine hydrolase activity	1.54E-33	1.32E-30	3.87	9210	181	1170	89
	GO:0004175	endopeptidase activity	4.22E-26	2.70E-23	2.93	9210	277	1170	103
	GO:0008233	peptidase activity	1.44E-24	6.13E-22	2.49	9210	405	1170	128
	GO:0016787	hydrolase activity	9.75E-13	3.56E-10	1.47	9210	144	1170	269
Pos-KD: Component	GO:0044421	extracellular region part	3.03E-11	3.78E-08	1.74	9210	592	1170	131
	GO:0005615	extracellular space	2.50E-10	1.56E-07	1.8	9210	472	1170	108
	GO:0005576	extracellular region	1.73E-07	7.20E-05	1.86	9210	288	1170	68
	GO:0031226	intrinsic component of plasma membrane	4.02E-05	1.25E-02	1.64	9210	303	1170	63

Table S3.3: Continued from previous page.

DEG sample	GO Term	Description	p-value	q-value	Enrich.	N	B	n	b
Pos-KD: Compon.	GO:0005887	integral component of plasma	5.66E-05	1.41E-02	1.63	9210	294	1170	61
	GO:0031012	extracellular matrix	7.52E-05	1.57E-02	2.18	9210	94	1170	26
Zeus-KD: Process	GO:0006508	proteolysis	4.20E-09	3.03E-05	1.6	9210	635	1230	136
	GO:0050830	defense response to Gram-positive bacterium	7.74E-08	2.79E-04	3.57	9210	42	1230	20
	GO:0006022	aminoglycan metabolic	8.64E-06	2.08E-02	2.38	9210	85	1230	27
	GO:0042335	cuticle development	1.76E-05	3.18E-02	2.3	9210	88	1230	27
	GO:0055085	transmembrane transport	1.97E-05	2.84E-02	1.52	9210	437	1230	89
	GO:0040003	chitin-based cuticle	1.99E-05	2.39E-02	2.37	9210	79	1230	25
	GO:0006629	lipid metabolic process	2.74E-05	2.83E-02	1.57	9210	368	1230	77
	GO:0010025	wax biosynthetic process	2.98E-05	2.68E-02	4.99	9210	12	1230	8
	GO:0010166	wax metabolic process	2.98E-05	2.39E-02	4.99	9210	12	1230	8
	GO:0006030	chitin metabolic process	4.41E-05	2.89E-02	2.54	9210	59	1230	20
Zeus-KD: Function	GO:0008236	serine-type peptidase activity	5.54E-25	7.08E-22	3.33	9210	180	1230	80
	GO:0017171	serine hydrolase activity	8.68E-25	7.40E-22	3.31	9210	181	1230	80
	GO:0004175	endopeptidase activity	5.48E-18	2.80E-15	2.49	9210	277	1230	92
	GO:0008233	peptidase activity	1.11E-17	4.71E-15	2.18	9210	405	1230	118
	GO:0005506	iron ion binding	1.46E-10	5.34E-08	2.7	9210	122	1230	44
	GO:0016705	oxidoreductase activity	1.13E-09	3.63E-07	2.5	9210	138	1230	46
	GO:0016787	hydrolase activity	1.44E-09	4.09E-07	1.38	9210	144	1230	266
	GO:0008010	structural constituent of chitin-based larval cuticle	1.31E-08	3.35E-06	4.39	9210	29	1230	17
	GO:0005214	structural constituent of chitin-based cuticle	1.92E-08	4.47E-06	3.95	9210	36	1230	19
	GO:0020037	heme binding	3.19E-08	6.81E-06	2.57	9210	105	1230	36
Zeus-KD: Component	GO:0044421	extracellular region part	2.10E-15	2.62E-12	1.87	9210	592	1230	148
	GO:0005615	extracellular space	8.43E-13	5.26E-10	1.89	9210	472	1230	119
	GO:0005576	extracellular region	2.52E-10	1.05E-07	2.03	9210	288	1230	78
	GO:0031012	extracellular matrix	7.73E-06	2.41E-03	2.31	9210	94	1230	29
	GO:0009986	cell surface	1.12E-05	2.80E-03	2.93	9210	46	1230	18
	GO:0031226	intrinsic component of	1.46E-05	3.03E-03	1.66	9210	303	1230	67
	GO:0005887	integral component of plasma	7.07E-05	1.26E-02	1.6	9210	294	1230	63

Table S3.3: Continued from previous page.

DEG sample	GO Term	Description	p-value	q-value	Enrich.	N	B	n	b
overlap CAF, Pos, Zeus-KD: Process	GO:0006508	proteolysis	6.17E-14	4.57E-10	2.33	9842	675	532	85
	GO:0043252	sodium-independent organic anion transport	8.69E-06	3.22E-02	13.21	9842	7	532	5
overlap CAF, Pos, Zeus-KD: Function	GO:0004252	serine-type endopeptidase activity	3.64E-29	9.78E-26	6.2	9842	161	532	54
	GO:0017171	serine hydrolase activity	6.92E-28	6.20E-25	5.66	9842	183	532	56
	GO:0008233	peptidase activity	9.19E-23	4.94E-20	3.36	9842	441	532	80
	GO:0004175	endopeptidase activity	2.46E-22	1.10E-19	3.97	9842	298	532	64
	GO:0016787	hydrolase activity	2.95E-17	1.13E-14	1.89	9842	1555	532	159
	GO:0003824	catalytic activity	1.97E-07	6.63E-05	1.29	9842	353	532	247
	GO:0016298	lipase activity	1.37E-06	4.08E-04	4.54	9842	57	532	14
	GO:0004553	hydrolase activity,	1.54E-06	4.13E-04	4	9842	74	532	16
	GO:0008235	metalloexopeptidase activity	2.08E-06	5.07E-04	4.72	9842	51	532	13
	GO:0008237	metallopeptidase activity	3.03E-06	6.79E-04	3.11	9842	125	532	21
overlap CAF, Pos, Zeus-KD: Process (ranked)	GO:0032504	multicellular organism	6.71E-08	1.84E-04	27.26	727	16	10	6
	GO:0000003	reproduction	6.71E-08	9.20E-05	27.26	727	16	10	6
	GO:0006508	proteolysis	9.28E-08	8.48E-05	1.72	727	96	282	64
overlap CAF, Pos, Zeus-KD: Function (ranked)	GO:0004252	serine-type endopeptidase activity	2.23E-08	1.81E-05	1.7	727	54	387	49
	GO:0004175	endopeptidase activity	5.67E-08	2.30E-05	1.65	727	69	371	58
	GO:0008233	peptidase activity	2.05E-07	5.54E-05	1.55	727	90	371	71
	GO:0070011	peptidase activity, acting on L-amino acid peptides	2.05E-07	4.15E-05	1.55	727	90	371	71
	GO:0008236	serine-type peptidase activity	8.03E-07	1.30E-04	1.66	727	58	371	49
	GO:0017171	serine hydrolase activity	8.03E-07	1.09E-04	1.66	727	58	371	49
	GO:0016787	hydrolase activity	6.00E-06	6.97E-04	1.37	727	190	319	114
	GO:0140096	catalytic activity, acting on a protein	2.94E-05	2.99E-03	1.42	727	115	356	80

Table S3.4. Proportion of differentially expressed genes (DEG) upon knockdown of *CAF40*, *Poseidon* and *Zeus* that exhibit significant either female- or male-biased expression according to two different databases (Assis et al., 2012; Gnad and Parsch, 2006). Notice that the knockdown of the three paralogs affects a significantly higher proportion of male-biased genes when compared to the total number of genes in our dataset.

			comparison to Gnad et al 2006		comparison to Assis et al 2012	
sample	**female-biased genes**	**male-biased genes**	**χ^2**	**p-value**	**χ^2**	**p-value**
all genes according to Gnad et al 2006	786	603				
all genes according to Assis et al 2012	1396	1750				
DEG in *CAF40*-KD	44	96	32.43	<0.000001	6.72	0.0095
DEG in *Poseidon*-KD	34	84	33.82	<0.000001	15.8	0.00007
DEG in *Zeus*-KD	38	92	35.84	<0.000001	8.72	0.0031

Table S3.5. List of fly strains used in this study.

source	genotype	note
BDSC 4533	w[*]; In(2LR)noc[4L]Sco[rv9R], b[1]/CyO, P{w[+mC]=ActGFP}JMR1	CyO-GFP balancer
BDSC 4534	w[*]; Sb[1]/TM3, P{w[+mC]=ActGFP}JMR2, Ser[1]	TM3, Ser, GFP balancer
BDSC 5138	y[1] w[*]; P{w[+mC]=tub84B-GAL4}LL7/TM3, Sb[1]	Tub84B-GAL4
BDSC 1878	P{w[+mW.hs]=GawB}T80/CyO	T80-GAL4
BDSC 36303	y[1] v[1]; P{y[+t7.7]=CaryP}attP40	TRiP RNAi background strain
BDSC 25710	X-Cas9 background	background strain for CRISPR injection
BDSC 67988	y[1] sc[*] v[1]; P{y[+t7.7] v[+t1.8]=TRiP.HMS05851}attP40	Poseidon TRiP RNAi
BDSC 67987	y[1] sc[*] v[1]; P{y[+t7.7] v[+t1.8]=TRiP.HMS05850}attP40	CAF40 TRiP RNAi
G. Findlay lab	y[1] w[*]; P{w[+mC]=Bam-GAL4}LL7/TM3, Sb[1]	germline specific Bam-GAL4 driver
BDSC 32180	y[1] w[*]; P{w[+mC]=nanos-GAL4}LL7/TM3, Sb[1]	germline specific nanos-GAL4 driver
BDSC 36683	y[1] sc[*] v[1]; P{y[+t7.7] v[+t1.8]=TRiP.HMS01571}attP2	Zeus TRiP RNAi

Table S3.6. List of primers used in this study.

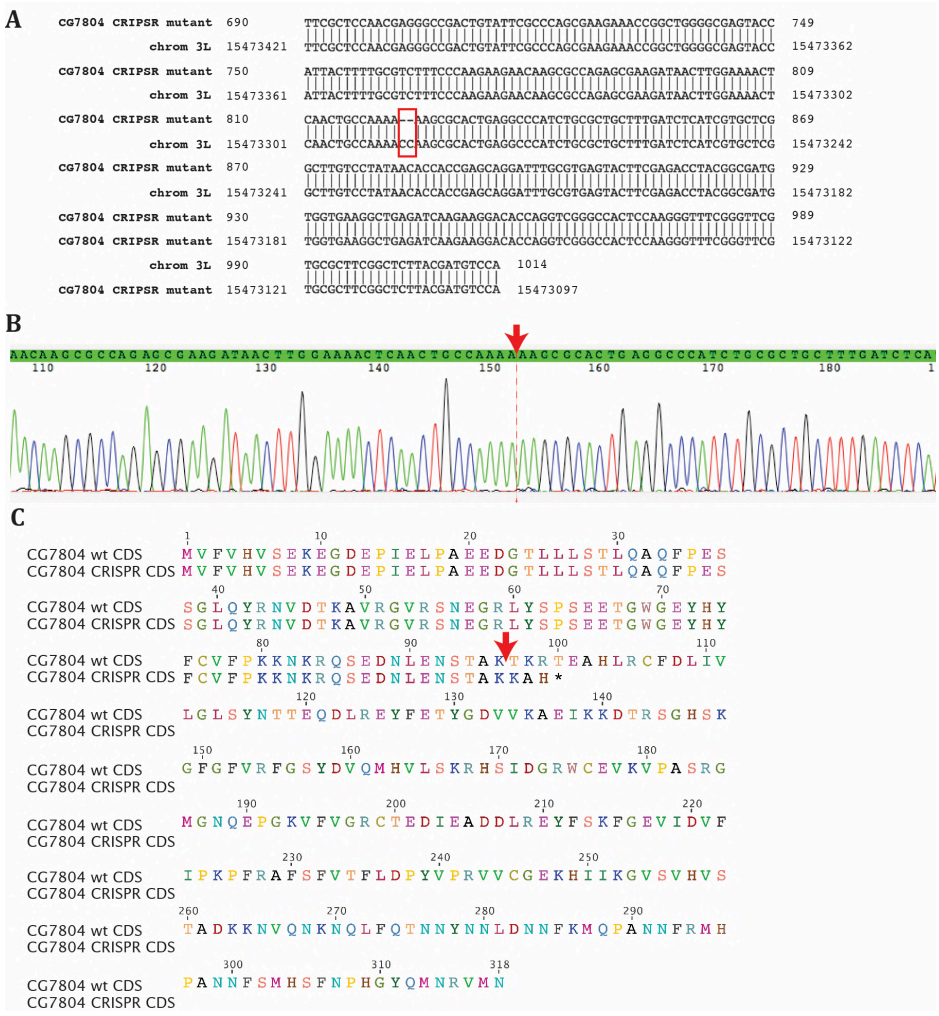
Purpose	primer name	primer sequence
Screen and confirmation of CRISPR mutant	Posei_CRISPR_F	TGAGTGCCCCAAACATGGAT
	Posei_CRISPR_R	GCCCAGAGACGCGTTTCTTTG
	Zeus_CRISPR_F	AATGAGTGCCCCAAACATGG
	Zeus_CRISPR_R	GAGACGCGTTTCTTTGTTGC
qPCR (Zeus)	Zeus_PP11452_F	GGAACCAAGTCCGGTAATGAG
	Zeus_PP11452_R	GGGCGTCGTTATTATGGGGTA
qPCR (CAF40)	CAF40_PP17951_F	AGCAAGAAGCGTGAGACGG
	CAF40_PP17951_R	AAAGGGTACAAGTACAGCGGT
qPCR (Poseidon)	Pos_PP14812_F	ACTCTTGCTATGGCATTCCTTTG
	Pos_PP14812_R	TGTTGGCCGGTTAATTCAATTTC
qPCR (control gene)	qRp49F	CCGCTTCAAGGGACAGTATC
	qRp49R	GACAATCTCCTTGCGCTTCT

Table S3.7. List of antibodies used in the co-immunoprecipitation assays.

Antibody	Source	Catalog Number	Dilution	Monoclonal/ Polyclonal
Anti-HA-HRP	Roche	12 013 819 001	1:5,000	Monoclonal
Anti-GFP (for western blotting)	Roche	11 814 460 001	1:2,000	Monoclonal (mouse)
Anti-GFP (for immunoprecipitation)	In house (Izaurralde's lab)			Polyclonal (rabbit)
Anti-V5	AbD Serotec	MCA1360GA	1:5,000	Monoclonal
Anti-mouse-HRP	GE Healthcare	NA931V	1:10,000	Monoclonal

Supplementary Figures for Chapter 4

Figure S4.1: Confirmation of CRISPR knockout of *Cocoon* homozygous individuals. (A) Alignment of the CG7804 CRISPR line with the reference genome sequence. Notice the 2 bp deletion. (B) Chromatogram of the CG7804 sequence with the induced deletion. Red arrow indicates the predicted cut site. (C) Alignment of the predicted amino acid sequence of wildtype and CRISPR mutant CG7804 protein.



Supplementary Tables for Chapter 4

Table S4.1. List of fly strains used in this study.

source	genotype	note
BDSC 38355	y[1] v[1]; P{y[+t7.7] v[+t1.8]=TRiP.HMS01823}attP40/CyO	CG7804 TRiP, effective RNAi line. Confirmed by qPCR
BDSC 39014	y[1] v[1]; P{y[+t7.7] v[+t1.8]=TRiP.HMS01932}attP40	TBPH TRiP RNAi
BDSC 4533	w[*]; In(2LR)noc[4L]Sco[rv9R], b[1]/CyO, P{w[+mC]=ActGFP}JMR1	Cyo-GFP
BDSC 4534	w[*]; Sb[1]/TM3, P{w[+mC]=ActGFP}JMR2, Ser[1]	TM3, Ser, GFP
BDSC 36014	y[1] w[*]; Mi{y[+mDint2]=MIC}CG7804[MI02615]	An unbalanced MIMIC insertion in CG7804
BDSC 25374	y[1] w[*]; P{Act5C-GAL4-w}E1/CyO	Act5C-GAL4
BDSC 5138	y[1] w[*]; P{w[+mC]=tubP-GAL4}LL7/TM3, Sb[1]	Tub-GAL4
BDSC 458	P{w[+mW.hs]=GawB}elav[C155]	elav-GAL4
BDSC 3038	P{w[+mW.hs]=GawB}Dll[md23]/CyO	Dll-GAL4
BDSC 25706	w[1118] P{w[+mW.hs]=GawB}Bx[MS1096]; P{w[+mC]=UAS-Dcr-2.D}2	Bx-GAL4
BDSC 25755	P{w[+mC]=UAS-Dcr-2.D}1, w[1118]; P{w[+mW.hs]=GawB}salm[LP39]	salm-GAL4
BDSC 25754	P{w[+mC]=UAS-Dcr-2.D}1, w[1118]; P{w[+mW.hs]=GawB}nubbin-AC-62	nub-GAL4
BDSC 25752	P{w[+mC]=UAS-Dcr-2.D}1, w[1118]; P{w[+mW.hs]=en2.4-GAL4}e16E, P{w[+mC]=UAS-2xEGFP}AH2	en-GAL4
BDSC 9748	y1 w1118; PBac{y+-attP-3B}VK00031	VK31, attP strain for TBPH-GFP BAC insertion
BDSC 9752	y1 w1118; PBac{y+-attP-3B}VK00037	VK37, attP strain for CG7804-GFP BAC insertion
BDSC 3605	w[1118]	w1118
BDSC 36304	y[1] v[1]; P{y[+t7.7]=CaryP}attP40	TRiP RNAi background strain
Genetic Service Inc	X-cas9[RFP, GFP]	background strain for CRISPR injection
generated in house	Act5C-GAL4/Cyo-GFP	generated by combining Act5C-GAL4/Cyo and #4533
generated in house	Tub-GAL4/TM3, Ser, GFP	generated by combining Tub-GAL4/TM3, Sb and $4534
G. Findlay lab	Bam-GAL4, UAS-Dicer2/TM3 ; GFP	germline specific GAL4 driver

Table S4.2. List of primers used in this study.

Purpose	primer name	primer sequence
Screen and confirm CRISPR mutant	CG7804_CRISPR_F	GTTTCCGGAATCTAGCGGTCTG
	CG7804_CRISPR_R	GAGAGTACGTGCATCTGGACATC
	TBPH_CRISPR_F	GATCCTGCGGTCTGAAGTACC
	TBPH_CRISPR_R	GAAACCGAAGCCCTTGGACTG
Confirm presence of MIMIC insertion at CG7804	CG7804_MIMIC_3B_F	AGGGTAGGATCCGTTGACCT
	CG7804_MIMIC_3B_R	GTCGATCACTTCGCCAAACT
Confirm absence of MIMIC insertion at CG7804	CG7804_MIMIC_L	GCTTCGGCTCTTACGATGTC
	CG7804_MIMIC_R	GTCGATCACTTCGCCAAACT
RT-PCR (CG7804 CDs)	CG7804_CDs_RA298_F	GAACAAATCCCTAGGCAATCC
	CG7804_CDs_RA1345_R1	GGTTGGGCATATCCAACAGT
qPCR (CG7804)	CG7804_e1to2_L	TGAATTTGCCTGGGTGTG
	CG7804_e1to2_R2	TCGGCGGACTATTGATAAGG
qPCR (TBPH)	TBPH_RE_F1	TCTACTTGCCATACGCAGTCA
	TBPH_RE_R1	GTCTCGATGCGTTTGGTCTT
qPCR (reference)	qRp49F	CCGCTTCAAGGGACAGTATC
	qRp49R	GACAATCTCCTTGCGCTTCT
qPCR (reference 2)	actin_qPCR_F	GCGTCGGTCAATTCAATCTT
	actin_qPCR_R	AAGCTGCAACCTCTTCGTCA
Confirm insertion of BACs	attP-F-new	AGGTCAGAAGCGGTTTTCGGGAGTAGTG
	attP-R-new	GGTCGTAAGCACCCGCGTACGTGTCCAC
	P[acman]-1.0-F	ACGCCTGGTTGCTACGCCTGAATAAGTG
	P[acman]-1.0-R	CCCACGGACATGCTAAGGGTTAATCAAC
Confirm presence of GFP tag (CG7804)	CG7804_GFP_F	GCATGCATTCATTTAATCCACA
	CG7804_GFP_R1	GAACTTCAGGGTCAGCTTGC
Confirm presence of GFP tag (TBPH)	TBPH_GFP_F	CAGAGCAGCGGATCTCAAA
	TBPH_GFP_R1	GAACTTCAGGGTCAGCTTGC