Comparative analysis of gene duplications and their impact on expression levels in nematode genomes

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tbingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von:
Praveen Baskaran
aus Ullikkottai, Tamil Nadu, Indien

Tübingen
2017

Tag der mündlichen Qualifikation: 09-May-2018
Dekan: Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter: Prof. Dr. Daniel Huson
2. Berichterstatter: Prof. Dr. Ralf J. Sommer

ERKLÄRUNG

Hiermit erklre ich, dass ich die Arbeit selbstndig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, October 2017

Praveen Baskaran

Abstract

Gene duplication is a major mechanism that plays a vital role in different evolutionary innovations, ranging from generating novel traits to phenotypic plasticity. Evolutionary impact of gene duplication and the fate of duplicated genes has been studied in detail. However, little is known about the impact of gene duplication on gene expression with respect to different evolutionary time scales. Here, we study genome-wide patterns of gene duplications in nematodes and assess their effect on expression levels. This study encompasses various macroevolutionary comparisons at different time scales and microevolutionary comparisons within the species *Pristionchus pacificus*.

At the macroevolutionary level, by comparing species separated more than 280 million years ago, we found various lineage-specific expansions in multiple gene families along the *Pristionchus* lineage. Moreover, we found that duplicated genes are highly enriched among developmentally regulated genes. Interestingly, the results also show evidence for selection on duplication to increases the gene expression levels in a developmental stage-specific manner.

To gain insights into the microevolution of gene expression levels after gene duplication, we compared different strains of *P.pacificus* and found that an additional gene copy does usually not increase gene expression levels in the different strains. Furthermore, we found a strong depletion of duplicated genes in large parts of the *P. pacificus* genome indicating towards negative selection against gene duplication. This shows that the impact on gene expression levels following gene duplication differs dramatically, where a selection for increased gene dosage dominates macroevolution and negative selection on gene duplication dominates within species level.

This led us to wonder what happens at the intermediate time scale. We compared recent duplicates of *P. pacificus* with their single-copy orthologs in two closely related species and found a pattern similar to the microevolutionary trend. Additionally, comparison of closely related species of the *Strongyloides* genus and its developmental transcriptome also shows overall strong depletion of duplicated genes, similar to the observation at the microevolutionary level. At the same time, a strong enrichment of duplicated genes was found at a developmental stage associated with the parasitic activity of the nematodes. Similar to the macroevolutionary picture of *P. pacificus*, we also found selection for higher gene dosage in parasitism-associated gene families of *S. papillosus*, indicating the adaptive potential of duplicated genes. Even though these studies show widespread selection against both duplication and changes in gene expression, duplications are favoured in some conditions leading to adaptive changes in the organism. Overall this indicates that the regulation of expression levels of duplicated genes was subjected to different selection processes at different time scales, which represent a complex interplay between different evolutionary processes like natural selection, population dynamics, and genetic drift.

Zusammenfasssung

Genduplikation ist der weitverbreitetste, für die Entstehung neuer Gene verantwortliche Mechanismus und trägt damit entscheidend zur morphologischen und biologischen Vielfalt in allen Formen des Lebens bei. Obwohl Genduplikation in verschiedenen Systemen bereits ergiebig erforscht wurde und zur Entwicklung zahlreicher theoretischer Modelle geführt hat, gibt es nur wenige Studien, die Genduplikation in den Genomen von Fadenwürmern (Nematoden) untersucht haben.

Ziel dieser Arbeit ist es, allgemeine Muster der Genduplikation aus Vergleichen von Nematodengenomen abzuleiten, die verschiedene evolutionr̈e Zeitspannen wiederspiegeln. Dabei zeigen phylogenetische Analysen von Multigenfamilien zwischen dem Modellorganismus *Pristionchus pacificus* und anderen entfernt verwandten Nematodenarten, wie z.B. *Caenorhabditis elegans*, dass selbst in Genfamilien, die die gleiche Anzahl von Genen in beiden Arten haben, die meisten Gene aus abstammungslinien-spezifischen Genduplikationen hervorgegangen sind, d.h. die unmittelbaren Vorfahren der meisten Gene sind jünger als die der beiden Arten. Genexpressionsdaten verschiedener Entwicklungsstadien deuten darauf hin, dass in einigen Fällen Duplikationen genutzt wurden, um die absolute Intensität spezieller Expressionsmuster zu erhöhen. Dagegen zeigt die Untersuchung von kürzlich innerhalb einer Art aufgetretenen Duplikationen widersprüchliche Tendenzen auf. Während weite Teile des Genoms frei von Duplikationen zu sein scheinen, führen die existierenden Duplikationen überraschenderweise nicht zu einer Erhöhung der Genexpression. Dies kann als Ergebnis von Selektion gegen die durch Duplikationen erhöhte Transkriptmenge gedeutet werden. Diese Tendenz wird durch den Vergleich weiterer nahe verwandter Arten der Gattung Pristionchus bestätigt. Dahingegen zeigt die Analyse von Genfamilien, deren Größe mit dem Auftreten des Parasitismus in der Familie Strongyloididae massiv zugenommen hat, ähnliche Muster wie sie im Vergleich zwischen *P. pacificus* und *C. elegans* auftreten.

Zusammenfassend stimmen die Ergebnisse dieser Arbeit mit der Ansicht überein, dass die Effekte der meisten Mutationen, einschließlich Duplikationen, entweder neutral oder schädlich sind. Allerdings können bestimmte äußere Umstände, wie z. B. die Koevolution mit dem Immunsystem eines Wirtes oder die schnelle Anpassung an schwankende Umwelteinflüsse, dazu führen, dass Duplikationsereignisse einen selektiven Vorteil erbringen können. Trotz der Seltenheit solcher adaptiven Ereignisse, scheinen sie entscheidend zur Entstehung von Artenvielfalt beigetragen haben, da einige der aus diesen Ereignissen hervorgegangen Genkopien über hunderte von Millionen Jahren stabil geblieben sind, und evolutionäre Vergleichen von entfernt verwandten Arten maßgeblich prägen.

Preface

The work and data that is presented has been published in the following research articles.

Markov, G. V., Baskaran, P., and Sommer, R. J. (2015). **The Same or Not the Same: Lineage-Specic Gene Expansions and Homology Relationships in Multigene Families in Nematodes.** Journal of Molecular Evolution, 80(1):18-36.

Baskaran, P. and Rödelsperger, C. (2015). **Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus.** PloS one, 10(6) e0131136.

Baskaran, P., Rödelsperger, C., Prabh, N., Serobyan, V., Markov, G.V., Hirsekorn, A., and Dieterich, C. (2015). **Ancient gene duplications have shaped developmental stage-specic expression in Pristionchus pacificus.** BMC Evol Biol, 15(1).

Baskaran,P., Jaleta, T., Streit, A. and Rödelsperger C. (2017). **Duplications and Positive Selection Drive the Evolution of Parasitism-Associated Gene Families in the Nematode Stongyloides papillosus**. Genome Biol Evo,l 9(3): 790-801.

Acknowledgements

First, I would like to express my gratitude to my advisors Dr. Christian Rdelsperger and Prof. Ralf Sommer for the continuous support, patience, encouragement, opportunities and immense knowledge. The guidance and insightful comments of Prof. Ralf Sommer helped me over these past years and while writing this thesis. Thank you for giving me the opportunity and freedom to explore different concepts and collaborate with many brilliant minds of sommer lab.

I am grateful to Dr.Christian Rdelsperger for being a super mentor and advisor, without him much of my work is not possible. I have immensely benefited from his expertise in bioinformatics, evolutionary biology, and programming. His suggestions and encouragement have incented me to widen my research from various perspectives.

I would like to Prof. Daniel Huson for accepting to be my co-supervisor and an evaluator of my thesis. Big thanks to Neel Prabh for his useful suggestions and proofreading my thesis.

Thanks to Metta Riebesell for translating my thesis summary into German. I also like to thank Dr. Adrain streit and Dr. Tegegn Jaleta for providing with the opportunities to work on parasitic nematodes.

It was a great pleasure to be working with many sharp minds of sommer lab : Suryesh Namdeo, Bogdan, Dhananjay, Vladisloav, Micheal and Gabriel Markov. I also like to thank Kostadinka and Karin for their administrative help.

Finally, I would like to thanks my wife and my parents for their support throughout my doctorate.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Evolution by gene duplication

De novo formation of new genes, horizontal gene transfer, and duplication of existing genes are the three different mechanisms by which an organism can acquire new genes. Among these, duplication of existing genes has emerged as the major mechanism that plays a role in the evolution of phenotypic complexity and functional diversification of genes [Adler et al., 2014]. Understanding the evolutionary pattern of gene duplication might provide useful insights about the processes generating phenotypic diversity across species or lineages. Gene duplication has been associated with the evolutionary innovation in gene functions. Few classic example for evolutionary significance of gene duplication are i) Evolution of color sensitive retina pigments in the ancestors of apes by duplication of opsin genes, generating additional opsin that can detect a spectrum of color [Dulai et al., 1999] and ii) partition of spatial expression pattern of human beta-globin duplicates, leading to optimization of oxygen binding affinity [Makova and Li, 2003]. The evolution of adaptive immune system in vertebrates is facilitated by several rounds of duplication of immunoglobulin genes. This indicates the absence of gene duplication would limit the genomes' ability to adapt to the changing environment [Zhang, 2003]. Duplication of HOX genes cluster also presents an interesting example of morphological innovations. The effects of HOX genes were first discovered in *Drosophila* for their involvement in the organization of structures along the body axis [Hughes and Kaufman, 2002]. It has been found the HOX genes as organized as genomic clusters and their order in the genome control the body layout of many animals like flies, fish and humans. The number of HOX clusters varies between vertebrates and invertebrates, and the duplication of these clusters was shown to coincide in the adaptive radiation observed in the vertebrates [Wagner et al., 2003]. Soshnikova et al [Soshnikova et al., 2013] found the duplication of HOX clusters has been associated with the diversification of complex organs in vertebrates like heart and kidney. Apart from these examples, widespread existence of gene families acts as primary evidence for the pivotal role of gene duplication in the evolution of gene function and adaptation.

1.1.1 Mechanisms of gene duplication

Duplication processes include frequent small-scale duplication of chromosomal segments and large-scale or even whole genome duplication events. Whole genome duplications (WGD) are rare events, which results in an increase in the number of chromosomes by the factor of 2 and are

also known as polyploidy. Such polyploidization has been reported in Yeast, plants and fungi [Conant and Wolfe, 2008]. Segmental duplication involves duplication of large genomic region roughly (1-100 kb) and the transposition of the duplicated segment to a new genomic location [Eichler, 2001]. Based on the transposition of duplicated copy, segmental duplications are classified as chromosome-specific duplication or trans-chromosomal duplication [Eichler, 2001]. Such segmental duplications have a strong association with genomic instability and chromosomal rearrangement [Bailey and Eichler, 2006]. Segmental duplication also includes, small-scale gene duplication which involves duplication of a genomic region containing partial or complete gene [Kaessmann, 2010]. Products of such small duplication events are inserted near the parent copy and are referred as tandem duplicates.

Homologous recombination between paralogous genomic regions (regions sharing a certain degree of sequence similarity), which results in unequal crossing over, is a major mechanism by which the segmental duplications are generated (Figure 1.1A). Similarly, the replication slippage also results in closely spaced gene duplicates with direct orientation. While unequal crossing over of misaligned homologous genomic regions occurs during meiosis, errors during DNA replication can lead to slippage. Even the repetitive elements with just a few bases are sufficient to cause slippage. These two mechanisms i.e unequal crossing over and replication slippage act as primary sources of gene duplicates in *C. elegans* [Katju and Lynch, 2003]. Retrotransposition is an alternative gene duplication mechanism, by which retrocopies are produced (Figure 1.1B). New copies generated by reverse transcription of mature RNA (mRNA) are inserted at random sites in the genome [Kaessmann, 2010, Hurles, 2004]. Retro-transposons play a vital role in this mechanism by encoding necessary enzyme like reverse transcriptase for transcription. Because of lack of introns and regulatory elements, retrogenes were long been thought as a processed pseudogenes [Kaessmann, 2010].

1.1.2 Microevolution of duplication

On a population level, fixation of the new gene is a rare event. In evolutionary term, for a gene duplication to be considered successful, the new gene must get fixed and maintained over time in the population [Hurles, 2004]. Lynch and Conery, [Lynch and Conery, 2000] estimated that on average 1% of duplicated genes are successfully fixed in a genome every million year. Gene duplications are shown to be favored in many circumstances that include a low rate of evolution, less protein-protein interactions, fewer pleiotropic constraints and smaller fitness defects.

Evolutionary fate of duplicated genes has been studied in greater detail to understand the necessity of duplications. Gene duplication provides both short (increase in gene dosage) and long-term (novel gene function) benefits [Rogozin, 2014]. There are also numerous cases where the products of gene duplications have been subjected to strong evolutionary constraints [Baskaran and Rödelsperger, 2015, Rogozin et al., 2014]. Another critical aspect of duplication is its impact on gene dosage, which plays a vital role in the retention of duplicates. Altered gene dosage by duplication has been associated with several genomic disorders [Emanuel and Shaikh, 2001]. This shows the potentially deleterious effect of duplication events. In some cases, duplication can be neutral by not altering the gene dosage. But in rare cases, duplications result in beneficial effects; in such cases, duplicated genes are rapidly fixed in a population. In general, because of the initial functional redundancy after gene duplication, and high probability of deleterious mutations over functionally beneficial mutations, the frequency of loss of duplicated genes is far greater than retention. The fixation of duplicated genes that escape from the accumulation of degenerative mutations, depends on the

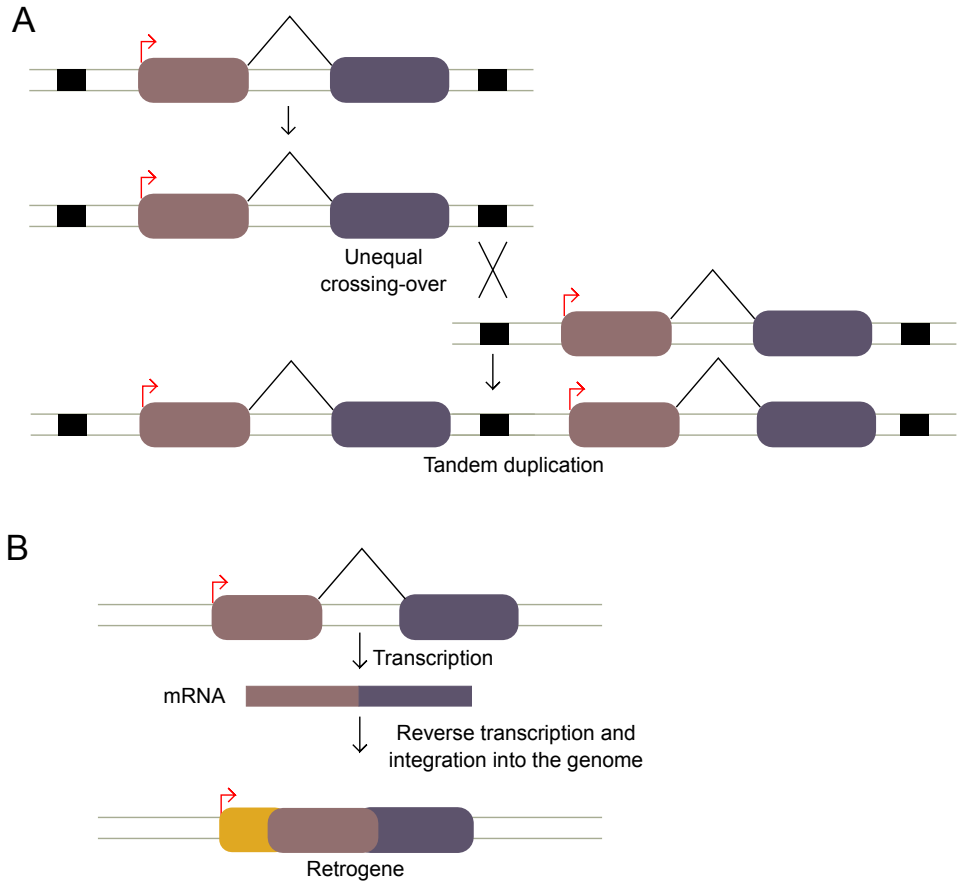


Figure 1.1: **Mechanisms of gene duplication.** Schematic representation of two different mechanism of gene duplications (adapted from [Kaessmann, 2010]). A) Recombination of homologous regulatory region leading to tandem duplication of a gene, where black box represent regulators element and other two rectangle represents the exon with transcription start site (red arrow). B) Retrotransposition of transcribed mRNA to intron less retrogene. The yellow box represents the flanking sequences in the newly inserted region.

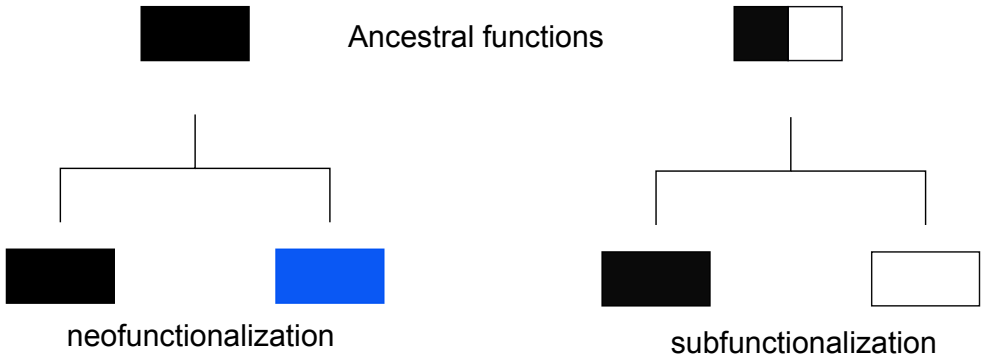


Figure 1.2: **Evolutionary models of duplicated genes.** Schematic representation of different models for retention of duplicated genes (Adapted from [Ohno, 1970]). A) Neofunctionalization of duplicated genes; where one of the duplicated copy acquired new function (blue), the other copy maintains the ancestral function. B) Partition of ancestral function by subfunctionalization. In this case ancestor is multifunctional (black and white) and its functions are partitioned between the duplicated copies. This allows each copy to specialize in their individual function.

effective population size [Lynch and Conery, 2000]. To better understand the evolutionary forces acting on the paralogous genes after duplication, a number of theoretical models have been postulated such as neofunctionalization and subfunctionalization (Figure 1.2). The evolution of gene sequence and its expression can, in theory, be independent. For example, the expression of a duplicated copy in a new tissue or developmental stage can be treated as neofunctionalization, even though; both copies perform the same molecular function (Figure 1.3A). However, the most probable fate of duplicated genes is Pseudogenization, accumulation of deleterious mutation leading to inactivation and eventually loss of one copy [Innan and Kondrashov, 2010, Zhang, 2003].

1.2 Retention of Duplicated genes

1.2.1 Neofunctionalization

According to Susumu Ohno [Ohno, 1970], who proposed the neofunctionalization model of paralog evolution, one of the duplicated genes acquires a novel function. In general, single copy genes are under strong purifying selection against the accumulation of nonsynonymous mutations. This allows the non-duplicated genes to have a certain degree of functional or expression conservation across different species. Neofunctionalization assumes one of the duplicated copies is free from the selection pressure and accumulates mutations. If the mutations are advantageous, this leads to a fixation of mutations and eventually new function (Figure 1.2A). The other copy is maintained by strong purifying selection and retains the ancestral function [Innan and Kondrashov, 2010]. A number of studies have found a period of relaxed selection after duplication, which allows one of the copies to acquire mutations in accelerated rate at the amino-acid levels. This leads to asymmetry in divergence rate or mutation accumulation between gene pairs [Hurles, 2004]. If the gene pairs evolve independently, selective advantage and rate of advantageous mutations determine the probability of neofunctionalization [Lynch and Conery, 2000].

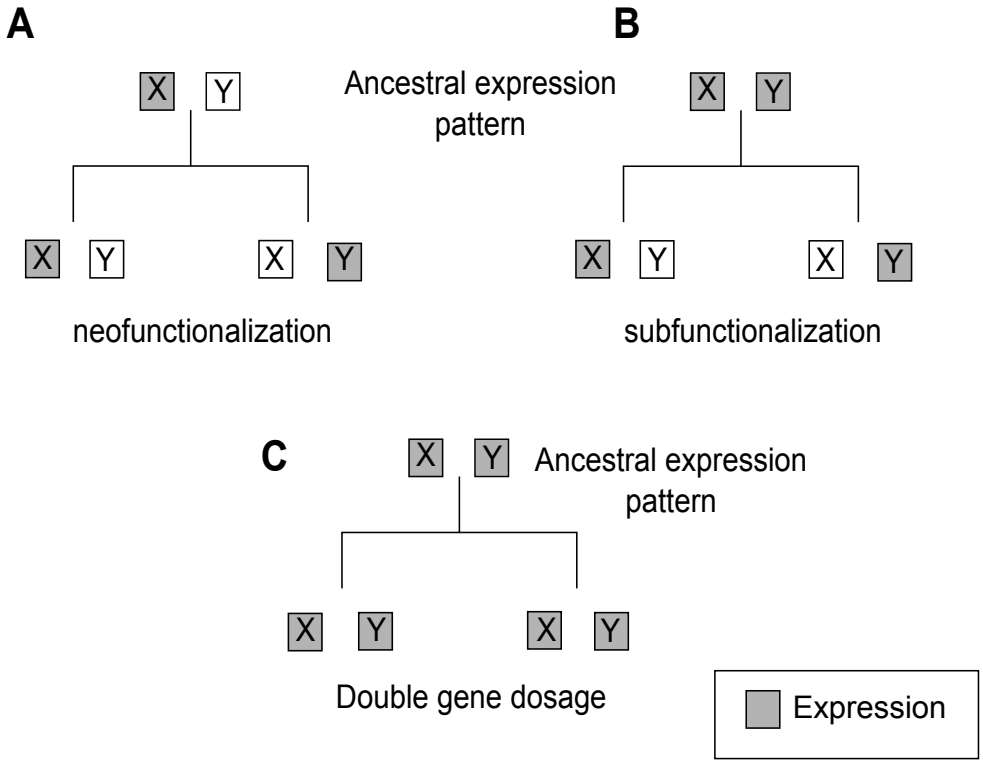


Figure 1.3: **Expression models of duplicated genes.** Schematic representation of expression models of duplicated gene retention. Here, X and Y represents two samples, for example, two different tissues or developmental stages and where grey box indicate evidence for expression in a particular sample. A) Shows the neofunctionalization model of duplicated gene retention. The ancestral gene has expression evidence in the sample X. After duplication one of the copies acquire expression in the sample Y, which was not observed in ancestor. B) Partition of expression pattern between duplicated copies. Here, the ancestral gene has expression in both samples. C) Shows double gene dosage, where the duplication is favoured to increase the expression in X.

Positive Darwinian selection on the sequence level indicates the selective advantage and rate of advantageous mutations. Convincing instances of such positive Darwinian selection was observed in the evolution of snake venom Phospholipase A2 and Disintegrins. A number of studies have reported that the evolution of snake venom is shaped by neofunctionalization of duplicated toxin gene families [Casewell et al., 2011]. Interestingly, Huminiecki et al, [Huminiecki et al., 2009] speculated duplication and neofunctionalization of dauer pathway genes in the common ancestor of Rhabditoid nematodes. Few classic examples of neofunctionalization mediated functional innovation include the evolution of Retinoic acid receptor in vertebrates and the evolution of antifreeze proteins in antarctic zoarchid fish [Escriva et al., 2006, Deng et al., 2010].

At the expression level, expression of a duplicated copy in a new tissue/developmental stage, which was not observed in the ancestor, can be treated as neofunctionalization (Figure 1.3A). Here, both copies can have the same molecular function. Using this approach, it was estimated that large fraction of young duplicates in Drosophila is retained by neofunctionalization [Assis and Bachtrog, 2013].

Even though neofunctionalization is prominent in nature, homogenization by gene conversion plays a vital role in preventing the acquisition of new function. Active gene conversion process can remove advantageous mutations and prevent fixation of neo-functionalized alleles [Teshima and Innan, 2008]. For a stable preservation of the duplicated genes in a population, strong selection against gene conversion is essential [Innan and Kondrashov, 2010].

1.2.2 Subfunctionalization

The acquisition of new function leads to long-term preservation of duplicated genes in the genomes. However, not all the duplication lead to functional innovation. A classic example of duplication without the emergence of a new function but preserved for a longer time is the evolution of human beta-globin genes. Three paralogs of beta-globin show a distinct expression pattern in three developmental stages, with one gene expressed in embryos, another gene in fetuses and finally the third gene in neonates [Hurles, 2004]. All three beta-globin paralogs show differences in O2 binding affinities that are optimized for each developmental stage. This spatiotemporal expression can be explained by subfunctionalization model, an extension of neofunctionalization model, that describes the partition of ancestral functions following duplication [Force et al., 1999] (Figure 1.2B and Figure 1.3B).

Under subfunctionalization, both copies of the duplicated gene accumulate degenerative mutations that impair different functional domains or properties of the ancestral gene. If this is the case, then both copies experience similar selection pressure, which in turn leads to a nearly symmetrical rate of amino acid evolution between the copies [Innan and Kondrashov, 2010]. However, this does not apply to all duplicates evolving under subfunctionalization. Paralogs genes with the signature of subfunctionalization and asymmetric evolutionary rate have been observed in yeast and *C. elegans* [Katju and Bergthorsson, 2013, van Hoof, 2005]. The most parsimonious explanation for asymmetric evolutionary rates is an unequal partition of ancestral sub-functions in case of genes with multiple independent mutable functions.

A number of studies have shown that subfunctionalization is the preferred mechanism for the preservation of duplicated genes [Force et al., 1999, Lynch et al., 2001]. This can be attributed to the observations that the probability of deleterious mutations is higher than the advantageous mutations. Partition of ancestral functions is achieved with the help of deleterious mutations by the loss of functional domains or change in gene structure. Thus, subfunctionalization is a neutral process, that does not always lead to adaptive changes in the organism [Prince and Pickett, 2002].

Another important aspect of splitting the ancestral functions is the removal of pleiotropic constraints acting on a single gene and allowing the natural selection to fine tune duplicated members to perform specific sub-functions [Force et al., 1999]. A gene is said to be pleiotropic if it controls more than one trait or function. The reduction of pleiotropy facilitates genetic analysis of developmentally crucial genes by high-throughput mutagenesis in different model systems. Finally, subfunctionalization, plays a significant role in the development of reproductive incompatibility through unsolved subfunctions at the time of reproductive isolation [Force et al., 1999].

Lynch and force [Lynch et al., 2001] estimated effective population size and mutation rates have a strong impact on the probability of duplicate gene retention. Accordingly, subfunctionalization is more likely to be a preferred mechanism in organisms with small population size and low coding mutation rates. Other factors that influence the retention are the chromosomal location of duplicated genes and the mechanism of gene duplication [Lynch et al., 2001]. However, this study does not account for the intrinsic properties of the genes like, plasticity, network connection and functional category, which can largely affect the retention probability. For a better understanding of evolution by gene duplication, population-based models should incorporate both intrinsic and extrinsic properties (population size and mutation rate) of the genes while estimating the probability of preservation of new gene duplicates.

In case of organisms with a large population size, it is interesting to speculate that subfunctionalization can act as an intermediate state during the early stages of duplication, extending the time of exposure to natural selection before being taken over by other processes [Force et al., 1999]. Taken together, subfunctionalization gives a nice alternative to non-functionalization and neofunctionalization and may account for a large fraction of genes preserved in the genomes.

1.3 Models of transcriptome evolution

Naively, duplication of a gene should lead to double gene dosage and this cannot be explained by the neo and sub-functionalization models. Moreover, the regulation of gene expression is a key factor for controlling many biological processes and the change in gene expression plays an important role in the phenotypic diversity observed in closely related species. Apart from phenotypic changes caused by coding sequences alteration, changes in gene expression provide key steps in the molecular basis of adaptation. A classic instance of adaptive phenotypic evolution due to changes in gene expression levels is the beak morphology of Darwin's Finches [Uebbing et al., 2016, Abzhanov et al., 2004].

Evolution of gene expression between species or between different populations of the same species has long been debated. However, the contribution of neutral selection or by stochastic processes to the evolutionary changes in gene expression is not well known. To better understand how gene expression evolves, numbers of models have been postulated. The neutral model of gene expression evolution assumes linear changes in the rate of expression levels with time [Khaitovich et al., 2004]. Khaitovich et al [Khaitovich et al., 2004] found the rate of change in gene expression levels between human, chimpanzee and orangutan is proportional to their divergence times, which is consistent with the neutral selection. This is in accordance with Kimura's theory of neutral evolution; genes with higher divergence in expression within species tend to vary much between species [Kimura, 1983].

However, the impact of positive selection for advantageous traits should not be ruled out, as it can also change the expression levels over evolutionary time. This positive or directional selection on gene expression will lead to smaller and larger expression variance within species and between

species, respectively. Alternatively stabilizing or negative selection reduces the expression variance between and within species [Uebbing et al., 2016]. At the same time a low rate of neutral evolution can have the same effect [Bedford and Hartl, 2009].

As different processes can lead to similar changes in gene expression, it is not easy to disentangle neutral evolution from the selection. And for this reason, methods or models used to access selection pressure in coding sequences cannot be applied to gene expression data. Bedford and Hartl [Bedford and Hartl, 2009], taking the phylogenetic structure into account, proposed two mathematical models of gene expression divergence to differentiate between neutral and stochastic selection processes. The First model is based on Brownian motions (BM) caused by random mutations that are fixed in an evolving population. BM processes can effectively model selective neutrality or neutral evolution by predicting the degree of variance in gene expression in proportion to time [Bedford and Hartl, 2009].

Since, BM processes are less suitable to model the evolution of traits are subjected to both negative selection and drift, a simple extension of BM model was developed to incorporate selection processes. Ornstein-Uhlenbeck (OU) processes describe a random walk model with some pull towards a particular state. Using the OU framework, it was found that the negative selection restricts the changes in gene expression observed between seven species of *Drosophila* [Bedford and Hartl, 2009].

Even though these models are useful to understand the expression divergence of one-to-one orthologs, it does not account for other sources of variation such as, variations caused by environmental factors and other genomic mechanisms. Recently, Rohlfs et al, [Rohlfs et al., 2014], extended the OU model to incorporate within species variation. As gene duplication is a major mechanism of molecular innovation, models incorporating the effect of gene duplication or loss on expression divergence across species will be beneficial.

1.3.1 Selection for higher gene Dosage

At times, duplication happens in order to produce more of the same or to increase the gene dosage [Ohno, 1970]. The gene dosage hypothesis posits duplicated copies are fixed by the positive selection for augmented gene dosage, which in turn increases the organism fitness [Kondrashov et al., 2004, Innan and Kondrashov, 2010] (Figure 1.3C). For example, increased amount of salivary amylase gene product by duplication is shown to be beneficial to humans by improving digestion of starchy diets [Perry et al., 2007]. This model assumes that expression levels of sub-optimally expressed genes are enhanced by the duplication through expression-enhancing mutations. This model also assumes reduced functional divergence after duplication, as functional divergence reduces the effective dosage of the gene product [Qian and Zhang, 2008]. However, this contradicts the general observation of rapid divergence in either expression or function after gene duplication. This can be explained by the effect of gene conversion on the genes selected for a higher dosage. Gene conversion maintains the sequence identity between duplicated copies, which restricts divergence and leads to long-term retention of duplicates [Hurles, 2004, Sugino et al., 2006]. Genes duplicated in the *Pristionchus* lineage, shows selection for higher gene dosage in developmental stage-specific manner and minor sequence divergence [Baskaran et al., 2015]. Duplication of genes involved in protein-protein interactions is more likely to experience the positive selection for increased dosage owing to the dosage sensitivity [Veitia, 2005].

1.4 Introduction to phylum Nematoda

Insects and nematodes are the most diverse and species-rich groups amongst all animal phyla and its members are successfully adapted to diverse ecological niches [Ugot et al., 2001]. The exact number of species is difficult to estimate, however, more than 25,000 nematodes and a million insect species have been described. Taken together, they account for more than half the number of all living organism described so far. Even though the number of classified insects is larger than nematodes, each insect can act as a host for multiple nematode species. Moreover, the uniform basic anatomy of nematodes makes it difficult to distinguish different species without proper molecular genetic analysis. Taken together the exact number of nematode species could be ranging from 1 to 100 million [Blaxter, 2003]. Nematodes have successfully adapted to various ecological niches on earth ranging from hot springs to polar ice and from bacterivores to obligate parasites and this ubiquitous nature of nematodes indicates their ecological importance [Sanghvi et al., 2016, Coghlan, 2005, De Ley, 2006]. The phylum Nematoda is classified into 5 major clades Rhabditida, Tylenchina, Spirurina, Enoplida, and Dorylaima [Coghlan, 2005] (Figure 1.4). The members of each clade are substantially diverse in almost every aspect ranging from genetics to developmental levels [De Ley, 2006]. The well-studied free-living model organism *Caenorhabditis elegans* belongs the clade Rhabditida, which also consist of parasitic species and the satellite model species *Pristionchus pacificus* [Coghlan, 2005]. One interesting pattern observed from detailed studies of different nematode species is the parallel or independent evolution of important traits such as, parasitism and hermaphroditism. Using a molecular phylogenetic approach, Blaxter et al. [Blaxter et al., 1998] found evidence for the multiple independent origins of parasitism within the phylum Nematoda. It was estimated that the parasitism has independently arisen at least 15 times in plant and animal parasitic nematodes [Blaxter and Koutsovoulos, 2015]. The evolutionary history of plant and animal parasitic nematodes are quite distinct. While horizontal transfer of genes (HGT) from unrelated species are quite common in plant parasitic species, some animal parasitic species like *Brugia malayi* benefits from their endosymbiontic relationship with bacteria like Wolbachia [Blaxter and Koutsovoulos, 2015, Dieterich and Sommer, 2009]. Like parasitism, variation in reproductive modes has been observed in the phylum Nematoda. The nematode species have two modes of reproduction, dioecious (male and female outcrossing) and androdioecious (hermaphroditism). Hermaphroditic nematodes are capable of self-fertilization and are found in different genus including *Caenorhabditis* and *Pristionchus* [Denver et al., 2011, Weadick and Sommer, 2016]. Dioecy was predicted to be the ancestral mode of reproduction in nematodes and switching to hermaphroditism happened multiple times independently [Denver et al., 2011, Weadick and Sommer, 2016]. This transition of reproduction modes has been associated with different consequences such as reduction of male sperm size in *C. elegans* [Weadick and Sommer, 2016, LaMunyon and Ward, 1999] and reduced life span among hermaphroditic nematodes of *Pristionchus* genus [Weadick and Sommer, 2016].

1.4.1 A well-studied model organism *C. elegans*

C. elegans is a small free-living nematode with size ranging from 0.25 mm (hatched larvae) to 1 mm long (adults). Wild-type *C. elegans* feeds on bacteria and can be easily isolated from microorganism-rich rotten vegetable matters [Barrière and Félix, 2006]. *C. elegans* was first used as a genetic model by Sydney Brenner, to understand the fundamental of animal behavior, genetics and development [DH and Fitch, 2005]. Over time, substantial molecular research on *C. elegans*

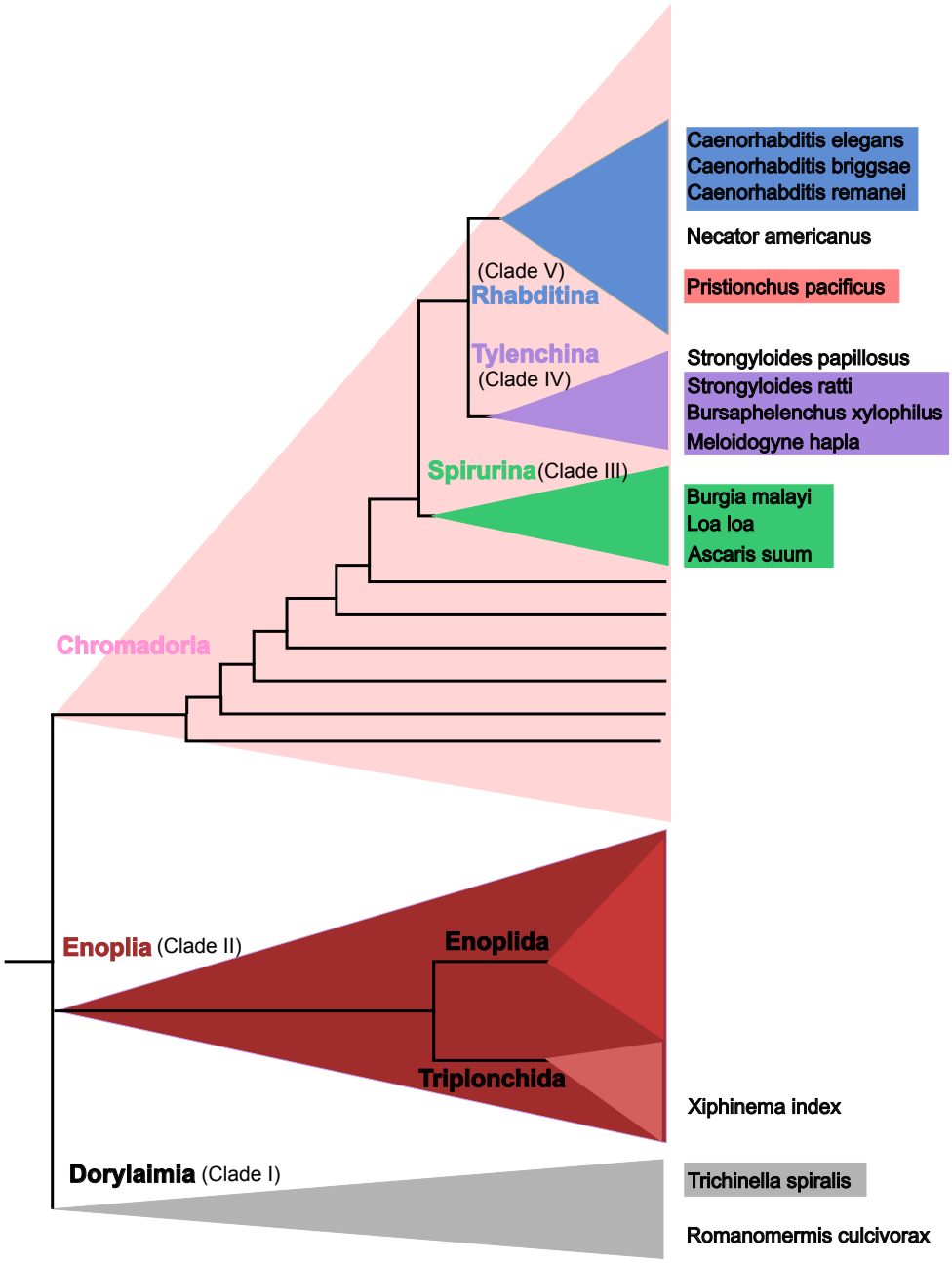


Figure 1.4: **General Phylogeny of phylum Nematoda** Schematic phylogenetic tree of phylum Nematoda, representing five clades of nematodes(Adapted from [Blaxter and Koutsovoulos, 2015]). The phylogeny was reconstructed using small subunit ribosomal RNA [Blaxter and Koutsovoulos, 2015]. *Pristionchus* and *Caenorhabditis* genus belongs to clade V group of nematodes. All the clades contain parasitic and free living nematodes, also it include nematodes that are obligate outcrossers or hermaphordites. Nematode species used for the macroevolutionary analysis of gene duplication (Chapter 3.1) was highlightef with box around species names. Same color code (box color) was used to display genes from different nematode species in the detailed phylogenetic trees of different gene families (see figures 3.1 - 3.3 and Appendix figures A1- A.4).

provides a detailed picture of various mechanisms involved in the host-parasite interaction, dauer formation, innate immunity, evolution of traits and so on [DH and Fitch, 2005, Corsi et al., 2015, Kaletta and Hengartner, 2006]. What makes *C. elegans* the most successful model organism? First, short generation time and small size greatly facilitate the maintenance of worms for experimental use. Second, the self-fertilizing ability of *C. elegans*, a single worm can produce a population. Third, worms can be frozen and reused whenever required. Fourth, fixed number of cells allows tracking the fate of each cell to generate complete cell lineage and finally its ability to grow in both liquid medium and agar plates. Overall, the animal populations are easy to maintain and relatively inexpensive.

Development of *C. elegans* is rapid and it takes three days at 25°C to complete a life cycle. *C. elegans* larvae pass through four larval stages, L1, L2, L3 and L4 before entering the egg-laying adult stage, and this happens when the conditions are favorable. Under unfavorable conditions such as depletion of the food source or overcrowding of nematodes, L2 larvae take an alternative path to a long-lived, non-feeding, growth arrested dauer stage. The nematodes can stay in the dauer stage for months by developing thick cuticles and when conditions are normal, dauer larvae resume development by entering the L4 stage to continue their life cycle.[Corsi et al., 2015]

C. elegans is predominantly a selfing species but exhibit two sexual forms; males, and hermaphrodites. Most of the individuals in a population are self-fertilizing hermaphrodites, which are basically females whose gonads also produce sperm. Hermaphrodites can mate with males and thus cause recombination in different lineages [Corsi et al., 2015]. *C. elegans* is transparent in nature, which allows easy visualization of individual cells and subcellular components in detail. One of the interesting aspects of these nematodes is their anatomically simple body plan, for example, fixed and determined number of cells, adult hermaphrodites consists of 959 somatic cells [Hillier et al., 2005].

1.4.2 *C. elegans* as a model for human diseases

C. elegans has also been increasingly used as a reference to understand the conservation of molecular mechanisms in other closely and distantly related organisms. This can be achieved by identifying orthologs or functionally similar genes between different organisms using comparative genomics approaches. Modeling a molecular mechanism of human disease in non-mammalian models such as *C. elegans* largely depends on the likelihood of functional/targeted gene sequence conservation between two species [Kaletta and Hengartner, 2006]. Shaye and Greenwald,[Shaye and Greenwald, 2011] found that approximately 38% of *C. elegans* protein-coding genes have orthologs in the human genome. 60-80% of human genes have orthologs in *C. elegans*. Out of 2466 genes known to be associated with human diseases, 21% have orthologs in the *C. elegans* genome. A classic instance of major innovation by using *C. elegans* to model novel functional targets of human disease is the identification of presenilin gene. The human ortholog of presenilin is associated with the Alzheimer's disease [Kaletta and Hengartner, 2006]. *C. elegans* was used to identify the interacting partners of genes associated with human Huntington's disease, even though no recognizable orthologues are found in *C. elegans* [Kaletta and Hengartner, 2006, Parker et al., 2004]. Regardless of many benefits, using non-mammalian models such as nematodes has its own limitation. Being a simple eukaryotic model, *C. elegans* does not have all molecular pathways exists in humans and hence, it is not possible to study such pathways in *C. elegans*. Despite its intrinsic limitation, *C. elegans* is considered as a valuable model to study and understand the molecular basis of biological process in other organisms.

1.4.3 *P. pacificus* : A satellite model organism

Although a wide range of studies in *C. elegans* provides useful insight into different evolutionary and development process, it is difficult to generalize the phenomenon for the entire phylum of millions based on one species. In order to better understand these processes in a comparative sense, different nematode models were surveyed. *Pristionchus pacificus* is one such nematode surveyed and established as a satellite model organism for comparative and evolutionary developmental studies [Hong and Sommer, 2006]. It has been estimated that both species have shared last common ancestors between 280 - 430 million years [Dieterich et al., 2008].

C. elegans and *P. pacificus* share similar features that make them successful model organisms. Similar to *C. elegans*, *P. pacificus* is also a selfing species, has a short life span (3.5 days at 25°C), its small size makes it easy to maintain in large collections, transparency and finally, worms can be cryo-preserved. Both species develop through four larval stages and can enter growth arrested dauer stage upon unfavorable conditions.

Evolutionary studies place both *C. elegans* and *P. pacificus* in the rhabditid group but in different families. *P. pacificus* belongs to the family Diplogastridae and spent the first larval stage in the egg. *C. elegans* first larvae hatches in L1, whereas in *P. pacificus* the first larval stage is embryonic molt and juveniles emerges from the egg as J2 stage larvae. The molting stage J1 is not free living and non-feeding. Unlike *C. elegans*, a free-living worm, *P. pacificus* exhibits a necromenic association with scarab beetles. *P. pacificus* worms sit on the beetles as growth arrested dauers, waiting for the beetle to die. Once the beetle died and start to decompose, the nematodes resume development by sensing and feeding the thriving microorganism in the decaying carcass. This association is known as the necromeny [Hong and Sommer, 2006].

Despite morphological similarities, *P. pacificus* lacks the pharyngeal grinder, found in *C. elegans* [Hong and Sommer, 2006]. *C. elegans* is a bacterivore, while *P. pacificus* is an omnivore. Studies on feeding behavior of *P. pacificus* have shown that the nematodes can feed on bacteria, fungi and juveniles of other nematodes [Sanghvi et al., 2016] [Wilecki et al., 2015]. However, *P. pacificus* can also be raised in the laboratory using monoxenic *E. coli*. Because of its omnivorous feeding behavior, *P. pacificus* exhibits two distinct mouth form, narrow mouth (Stenostomatous) for bacterial feeding and wide mouth (Eurystomatous) for predation of other nematodes. The mouth of *P. pacificus* is equipped with teeth like denticles, which is restricted to diplogastridae family and represents an evolutionary innovation that facilitates the predatory feeding behavior [Bento et al., 2010, Sommer and McGaughran, 2013].

Both *C. elegans* and *P. pacificus* have 5 autosomes and one sex chromosome. Syntenic analysis of chromosomal regions revealed extensive intrachromosomal rearrangement between the two species. However, large-scale macrosynteny was observed in 5 out of 6 *P. pacificus* chromosomes to its *C. elegans* counterpart [Sommer, 2006].

P. pacificus is a cosmopolitan species, that has been extensively sampled throughout the world and facilitate the integration of developmental and evo-devo studies with ecology and population genetics. This extensive sampling yielded more than 600 strains of *P. pacificus* from Asia, America, Africa and the Mascareigne islands of the Indian ocean especially the La Reunion island [Sommer and McGaughran, 2013]. Population-based evolutionary studies grouped the *P. pacificus* strains into four major clades A, B, C and D. The La Reunion Island is considered as a biodiversity hotspot and the complete genetic diversity of *P. pacificus* found worldwide was also found on this island alone. McGaughran et al, [McGaughran et al., 2013], found evidence for independent colonization of the the La Reunion Island by members of different clades. Resequencing of 104 isolates from

this island provided the first insight into the evolutionary forces underlying the genetic diversity and population structure [Rödelsperger et al., 2014]. High genetic and phenotypic diversity observed among different populations of *P. pacificus* on La Reunion Island, enabled different ecological studies to understand natural variations in chemosensation [McGaughran et al., 2013], dauer formation [Bose et al., 2014], social behavior [Moreno et al., 2016], cold tolerance [McGaughran and Sommer, 2014] and PH tolerance [McGaughran et al., 2016]. Colonization in diverse environment and exposure to heterogeneous environmental pressures makes *P. pacificus* powerful model to study ecology and evolution.

1.4.4 Genome sequencing of *P. pacificus*

Sequencing the genome of an organism provides a key framework for understanding the molecular changes associated with ecology and evolution. The genome of *P. pacificus* was first sequenced in 2008, with the estimated genome size of 169 MB and more than 26,000 predicted protein-coding genes [Dieterich et al., 2008]. Subsequently, gaps in the initial assembly were filled using 454 read sequencing in 2010 [Borchert et al., 2010]. One of the interesting findings is the identification of genes belonging to Cellulases and Diapausins families in *P. pacificus*. No Cellulases genes have been reported in nematodes except in the plant parasitic nematodes [Sommer and Streit, 2011]. Initial analysis of seven *P. pacificus* cellulases genes showed sequence similarity with slime molds. This indicates that *P. pacificus* has acquired these genes from unrelated donors by horizontal gene transfer (HGT). Comparative genomic and transcriptomic analysis of cellulase genes show that multiple gene duplications and positive selection might have played a vital role in the integration of HGT in nematodes [Mayer et al., 2011]. Another HGT-acquired gene family, Diapausins were reported to be derived from beetles [Dieterich et al., 2008]. Detailed analysis of other *P. pacificus* HGT-acquired genes, based on codon usage and homology search, suggested insects as a potential donor [Rödelsperger and Sommer, 2011]. Given that the *P. pacificus* has an entomophilic association with insects, the genome of the scarab beetle *Oryctes borbonicus* was sequenced to investigate the presence of *P. pacificus* HGT-acquired genes. Detailed phylogenetic analysis of different *O.borbonicus* gene family shows no additional evidence of HGT between beetles and *P. pacificus* [Meyer et al., 2016].

The genome size of *P. pacificus* is larger than that of *C. elegans* but smaller than its sister species *P.exspectatus* and *P.arcanus*. Even though the repetitive elements cover approximately 17% of *P. pacificus* genome, the difference in genome size cannot be fully attributed to repetitive elements [Dieterich et al., 2008]. Comparative genomic analysis of *P. pacificus* gene family with other nematodes provided the first insight into the expansion and loss of gene. More than 20% of the predicted protein-coding genes were present as single copy orthologs in *P. pacificus*, *C. elegans*, *C.briggsae* and *B.malayi*, indicating that these core orthologs existed in the common ancestor. Using 575 one-to-one orthologs, common to all four nematodes, the time since the separation of *P. pacificus* and *C. elegans* was estimated to be between 280-430 million years [Dieterich et al., 2008].

Interestingly, more that 30% of the predicted genes have no sequence similarities to any coding sequences in other nematodes. The large amount of non-conserved genes observed in *P. pacificus* is more common to other nematodes and are called as orphan genes [Rödelsperger et al., 2013]. Detailed analysis of orphan genes based on transcriptomic and proteomic data shows 39-81% of orphan genes are predicted to be protein-coding [Prabh and Rödelsperger, 2016].

Finally, more than 40% of genes have homologous coding sequences in other nematodes, indicating the possible role of lineage-specific gene duplication. These duplication events facili-

tate the expansion of different gene families in *P. pacificus* including Cytochrome P450, UDP-glycosyltransferase, and sulfotransferase. Cytochrome P540 gene family shows a drastic increase in family members with 198 genes in *P. pacificus* and only 67 in *C. elegans*. Additionally, lineage-specific gene loss was also observed in some gene families especially in receptor L domain and seven-transmembrane receptor [Dieterich et al., 2008]. These lineage-specific gene gain and loss events show how the genome changes in response to adaptation in different environments.

1.5 Comparative genomics of nematodes

With ever reducing cost and increased throughput of sequencing technology, the genome of organisms from different clades of the animal kingdom is being sequenced. Among the nematode phyla, *C. elegans* genome was fully sequenced in 1998 and it is also the first multicellular organism that was sequenced completely [The C. elegans Sequencing Consortium et al., 1998]. With ≈100MB in size, the *C. elegans* genome was just 3 % the size of the human genome ≈3GB. However, this dramatic difference in the genome size is not reflected when considering the number of genes in *C. elegans* and humans. The presence of ≈21,000 genes in *C. elegans* indicates a high gene density along the genome [Sommer and Streit, 2011]. Following the success of *C. elegans* genome, its close hermaphroditic relatives, *C.briggsae* and *C.tropicalis* were sequenced in 2003 and 2011 with the estimated genome size of 108MB and 79MB respectively [Sommer and Streit, 2011, Fierst et al., 2015, Stein et al., 2003]. The genome size of hermaphroditic nematodes are smaller than gonochoristic species (obligate out crossers) of *Caenorhabditis* genus such as *C.remanei*, *C.japonica*, and *C,brenneri* each with an estimated genome size larger than 130MB. The reduction of genome size in hermaphroditic nematodes in comparison with outcrossing species was found to be associated with the transition to self-fertilization [Fierst et al., 2015].

In order to understand the genomic basis of host-parasite interaction and evolution of parasitism, a number of parasitic nematode genome projects were initiated. The draft genome of human parasitic nematode *Brugia malayi* was assembled in 2007 with 88MB of the estimated genome size and ≈11,000 predicted protein-coding genes [Ghedin et al., 2007]. This extreme difference in gene count in *B.malayi* is attributed to gene loss events owing to the high predictability of host environment and its dependency towards the endosymbiotic Wolbachia for the development and reproduction [Sommer and Streit, 2011, Ghedin et al., 2007]. Along the same line, Plant parasitic nematodes, *Meloidogyne incognita* and *Meloidogyne hapla* were sequenced in 2008 with the estimated genome size of 82 and 53 MB, respectively [Abad et al., 2008, Opperman et al., 2008]. *Trichinella spiralis*, a vertebrate parasite, also have a small genome (64MB), suggesting the transition to parasitism can also lead to genome size reduction [Mitreva et al., 2011]. Recently, parasitic nematodes of strongyloides genus *S.ratti, S.papillosus, S.venezuelensis* and *S.stercoralis* were sequenced. The estimated genome size of strongyloides nematodes ranges from 43 to 60 MB, further supporting the evolution of parasitism at the cost of genome size reduction [Hunt et al., 2016]. The availability of genome sequences of different nematode species facilitates studying the genomic basis of a wide variety of natural processes by comparative genomics approach.

Research on *P. pacificus* aims to study the evolution of gene function and pathways at a mechanistic level. Comparative genomics can help us to learn more about the evolutionary forces acting on various traits by studying the underlying gene loci. For instance, comparative analysis of *P. pacificus* and *C. elegans* shows a negative selection on, orphan gene dau-1, specific to *Pristionchus* genus, regulates dauer formation by copy number dependent manner [Mayer et al., 2015]. A *P.*

pacificus sulfatase gene eud-1, which controls the mouth form phenotype, shown to be under strong negative selection by comparative analysis of *P. pacificus* and *P.exspectatus* [Ragsdale et al., 2013]. Detailed analysis of eud-1 also revealed an interesting pattern of serial duplications followed by specialization of a single gene that controls the mouth form phenotype [Ragsdale and Ivers, 2016]. Another *P. pacificus* gene, nhr-40 was identified to act downstream of eud-1 in the pathway that controls the mouth dimorphism. Even though nhr-40 belongs to the extremely duplicated nuclear hormone receptor family, phylogenetic analysis revealed a one-to-one ortholog relationship between *P. pacificus* and *C. elegans* nhr-40 gene [Kieninger et al., 2016]. This indicates strong evolutionary constraints prevent the duplication of nhr-40, to preserve functional conservation. The evolutionary implication of these analyses suggests that genes controlling ecologically relevant traits are surprisingly not under positive selection as could be suggested by the fact that the environment changes rapidly.

Pathway genes involved in different biological processes have been studied in detail using functional comparative genomics approach. Genome-wide phylogenetic analysis of *P. pacificus* and *C. elegans* revealed gene duplication and domain switching among *P. pacificus* genes and gene families involved in the Beta-oxidation pathway. Peroxisomal Beta-oxidation pathway is involved in small molecules biosynthesis such as ascaroside and paratosides in nematodes. Detailed analysis of Beta-oxidation pathway genes using CRISPR-Cas9 shows functional conservation and divergence of *P. pacificus* daf-22 paralogs [Markov et al., 2016]. Taken together, comparative genomic approaches provide a unique perspective to understand the evolutionary forces acting on the genetic loci associated with important traits.

1.6 Aim of the thesis

The aim of this thesis is to investigate the forces shaping the evolution of duplicated genes in nematodes with a particular emphasis on changes in gene expression after duplication. In order to achieve this goal, we identified and compared dominant forces at different evolutionary time scales, using developmental transcriptomes and genes affected by structural variants. This involves comparing samples at various different evolutionary time scales spanning micro and macroevolution

At the macroevolutionary level, we identified widespread duplication events in gene families associated with two different metabolic pathways, by comparing *P. pacificus* with 10 other nematode species (including *C. elegans*). We performed a detailed analysis of developmental transcriptomes of *P.pacificus* to investigate the developmental regulation and to understand the impact of lineage-specific duplication on gene expression levels.

To investigate the evolutionary forces at the microevolutionary level, we compared three strains of *P. pacificus* and characterized genes that are affected by structural variants using genomic and transcriptomic data.

Finally, we also performed a detailed investigation of the developmental transcriptome of parasitic nematodes of the *Strongyloides* genus. Additionally, we also investigated the transcriptome of mixed-stage worms of three closely related *Pristionchus* species. These two studies allow us to compare closely related species to gain insight into the selection forces acting on gene expression and protein sequences at an intermediate time scale. In this context, we investigated the duplication pattern and expression profiles of parasitism-associated gene families in the nematode *S.papillosus* and recent duplicates in *P.pacificus*.

Chapter 2

Materials and Methods

This chapter contains content from the following publications. The copyright holder has granted the re-use permission.

Markov, G. V., Baskaran, P., and Sommer, R. J. (2015). The Same or Not the Same: Lineage-Specic Gene Expansions and Homology Relationships in Multigene Families in Nematodes. Journal of Molecular Evolution, 80(1):18-36.

Baskaran, P. and Rödelsperger, C. (2015). Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus. PloS one, 10(6) e0131136.

Baskaran, P., Rödelsperger, C., Prabh, N., Serobyan, V., Markov, G.V., Hirsekorn, A., and Dieterich, C. (2015). Ancient gene duplications have shaped developmental stage-specic expression in Pristionchus pacificus. BMC Evol Biol, 15(1).

Baskaran,P., Jaleta, T., Streit, A. and Rödelsperger C. (2017). Duplications and Positive Selection Drive the Evolution of Parasitism-Associated Gene Families in the Nematode Stongyloides papillosus. Genome Biol Evo,l 9(3): 790-801.

2.1 Phylogenetics Anaylsis

2.1.1 Orthologs, Paralogs and Homologs

In comparative biology, homology is the fundamental concept which describes the relationship between a pair of genes or biological structures that originated from a common ancestor. A classical example of structural homology is the bat's wings and human's hand [Fitch, 1970, Koonin and Galperin, 2003]. Because of the shared ancestry, homologous structure or sequences were assumed to perform same or similar functions. In contrast, analogy refers to structures or sequences with similar function without shared ancestry [Fitch, 1970].

Recently, sequence (nucleotide or amino acid) based methods are widely used to infer homology between different taxa. In sequence-based approach, the similarity between two sequences represents shared ancestry. The two sequences are referred as homologous sequences because they are derived from a common ancestor. The shared ancestry between sequences is the result of either speciation or duplication events. Orthologs are homologous genes found in two different species as a result of speciation event. In other words, the same gene found in different species

[Ward and Moreno-Hagelsieb, 2014]. In contrary to orthologs, paralogs are generated as a result of gene duplication events. Paralogs are further classified into two groups, in-paralogs, and out-paralogs. In-paralogs are the genes found in the same species, as duplication occurred after speciation. Out-paralogs are a pair of genes found in two different species due to duplication before speciation. Schematic representation of orthologs, paralogs and gene duplication is illustrated Figure 2.1. In this work, orthologs were referred as one-to-one orthologs, because we only consider genes that are maintained as a single copy in different species. Whenever a gene has multiple homologs in another species, these are considered as paralogs (out-paralogs).

Identification of orthologs and paralogs is utmost important for comparative functional genomics, as it is based on the assumption that orthologs share similar functions [Studer and Robinson-Rechavi, 2009] The ortholog conjecture assumes that orthologs have the same function because changing the basic function involves loss of original function, which is harmful [Chen and Zhang, 2012]. The most reliable way to distinguish and identify orthologs and paralogs is to reconstruct the phylogenetic tree using all homologs sequences. Trees can also be used to detect lineages specific expansion events, as the duplicated genes often cluster together to forms a sub-tree. In contrary, one-to-one orthologs reflect the species tree. Figure 2.1 shows the schematic representation of lineage-specific expansion and one-to-one orthologs.

2.1.2 Reconstruction of phylogenetic trees

Phylogenetic trees represent the evolutionary relationship between species and genes. Such phylogenetic trees can be reconstructed using distance, parsimony, likelihood and Bayesian inference methods. Distance based methods assume a molecular clock and calculate evolutionary distances based on the number of substitutions per site per time period. One such distance-based method is Neighbor-joining, which group sequences that have the smallest number of changes between them [Durbin et al., 1998]. Parsimony based methods work by assigning a cost and search different trees to find a tree with minimum cost or a minimum number of substitutions that explains the observed sequence data [Durbin et al., 1998]. Maximum likelihood (ML) approach is an alternative to distance and parsimony-based methods. ML methods use an evolutionary model, which assign a probability for a particular substitution, compute likelihood score for each topology and check the probability that the selected model can generate the observed sequences [Durbin et al., 1998]. Finally, phylogeny is inferred from the tree with the highest likelihood. In ML analysis, the maximum likelihood estimate is the combination of parameters that maximize the likelihood function [Nielsen, 2005]. Bayesian inference of phylogeny also uses different models of evolutions but calculates posterior probabilities based on prior probabilities for generating the most likely tree for given sequences. In contrast to ML, Bayesian inferences treat parameters as random variables. Both ML and Bayesian methods take full advantages of all information from a multiple sequence alignment [Nielsen, 2005].

To reconstruct the phylogeny of gene families and homologous gene clusters of *P. pacificus* and *S. papillosus*, protein sequences were aligned using Clustal omega [Sievers et al., 2014] with default options. The multiple sequence alignment results were corrected using trimAl program [Capella-Gutiérrez et al., 2009] and manual inspection in Seaview visualization program [Gouy et al., 2010]. Prottest 2 software was used to identify the best models that can explain the changes in the observed alignment [Darriba et al., 2011]. Substitution models with highest Bayesian inference criteria (BIC) score were treated as the best model. Maximum likelihood phylogenetic trees were reconstructed using the multiple sequence alignment and the best model in phangorn R package [Schliep, 2010] or

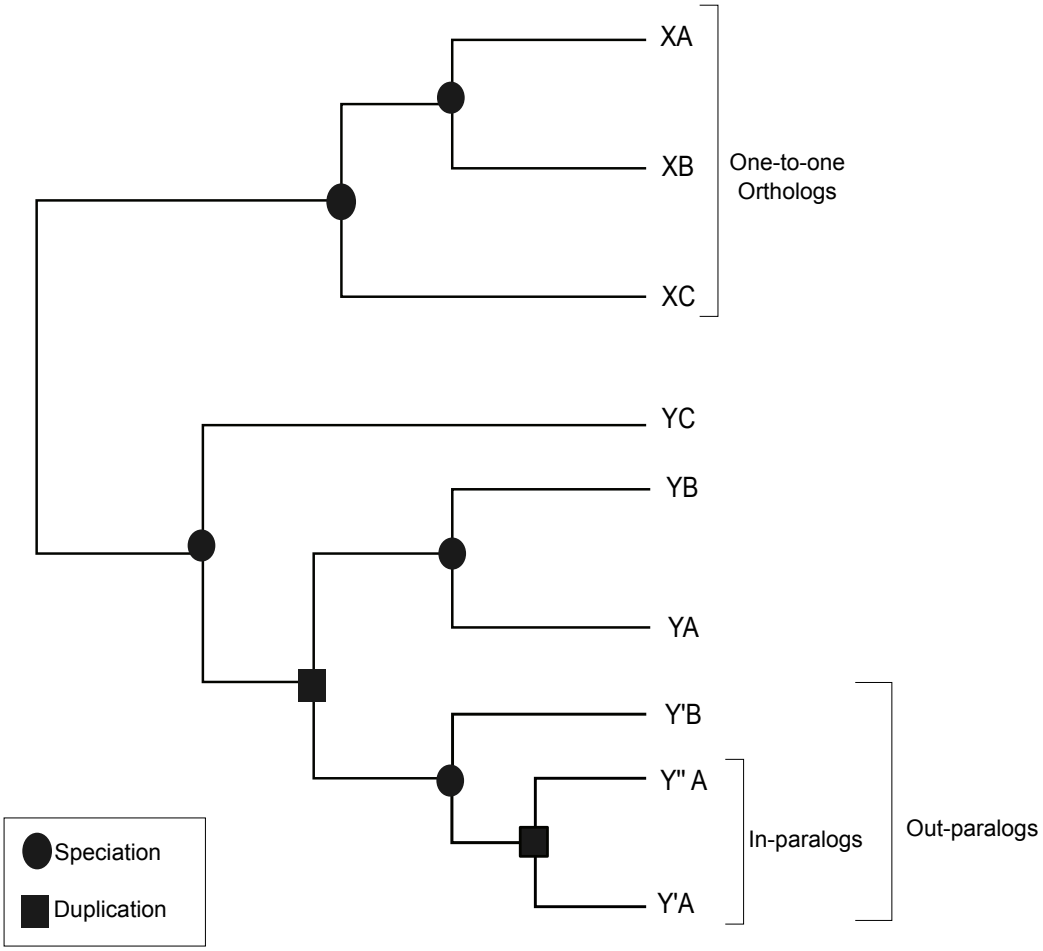


Figure 2.1: **Schematic representation of homology relationship.**
Cartoon representation of phylogenetic relationship between genes of three species A, B and C. Homologous gene X has single copy in all three species and the subtree represents the species tree. The gene X said to have one-to-one orthologous relationship between all three species. Duplication of Y" produced Y"' in species A and these two genes are defined as In-paraogs. Y and Y' are generated as result of duplication before speciation event that separate A and B, so Y'B and Y'A or Y"A genes have Out-paralog relationship.

Raxml [Stamatakis, 2014]. The trees were resampled using bootstrap methods for 100 repetitions.

2.2 Automated prediction of homology relationship

Determining homologues relationship between a pair or family of genes is straightforward using the phylogenetic tree reconstruction methods. At the same time the phylogenetic approach is computationally expensive. So automated orthology detection methods are required to predict homologues relationship on a genome-wide level. To facilitate automated detection of homologous relationship, a number of methods has been developed. These automated approaches make using of techniques ranging from the simple pairwise blast to complex Markov clustering algorithm to groups the seqeunces.

The protein sequences of *C. elegans*, *C. briggsae*, *C. angaria*, *Haemonchus contortus*, *Meloidogyne hapla*, *Brugia malayi*, *Bursaphelenchelus xylophilus*, *Ascaris suum*, *Trichinella spiralis*, *Heterorhabditis bacteriophora*, *Loa loa*, *Wuchereria bancrofti*, *Meloidogyne incognita*, *Panagrellus redivivus*, *Dirofilaria immitis* were compared to identify the homologs of *P. pacificus*. Similarly, *Stongyloides papillosus* homologs were identified using the protein sequences of *S. ratti*, *S. stercoralis* and *S. venezuelensis*. Different methods and tools were used in this work to determine orthologous relationship depending on the number of species investigated.

2.2.1 Best Reciprocal Hits (BRH)

One such method is best reciprocal hits approach, which offers a simple blast framework for orthologous detection between two species. BRH scores, produced by blast algorithm can be used to define orthologs pairs in two genomes, if the pairs are the best hit for each other [Ward and Moreno-Hagelsieb, 2014].

2.2.2 InParanoid

BRH method is quite useful to detect one-to-one orthologs but not suitable to effectively separate inparalogs from out-paralogs. To predict inparalogs, an automatic paralogs calling method, InParanoid was used [Remm et al., 2001]. InParanoid works by clustering orthologs using two-way best pairwise match, applies an algorithm to detect in-paralogs and assigns confidence score for both orthologs and in-paralogs. In this work, the one-to-one pairs between *P. pacificus* and *C. elegans* were predicted using a variant of the best-reciprocal hits approach that takes inparalogs into account [Stein et al., 2003, Sinha et al., 2012b, Mitreva et al., 2011]. More precisely, inparalogs were first defined and then best-reciprocal hits were assigned as one-to-one orthologs, only if neither the *C. elegans* nor the *P. pacificus* protein had any inparalog. This procedure predicted 5985 one-to-one orthologous pairs. The quality of one-to-one orthology predictions was evaluated using a data set of 107 *C. elegans* genes for which orthology relationships were manually investigated using alternative versions of *P. pacificus* gene predictions, TBLASTN [Altschul et al., 1990] searches to complement incomplete gene models, and subsequent phylogenetic analysis including all potential paralogous sequences.

2.2.3 MCL Clustering of multiple species

The Major limitation with BRH and InParanoid is that both methods cannot be applied for multi-species comparisons. Scalable graph based tools such as MultiParanoid and OrthoMCL

are being used extensively for multi-species comparisons. MultiParanoid, an extension of InParanoid, uses multiple Inparanoid comparison results and applies a clustering algorithm to generate multi-species homologous groups [Alexeyenko et al., 2006]. Since MultParanoid depends on the shared ancestry for clustering species, only a few closely related eukaryotic species can be used [Alexeyenko et al., 2006]. In the work on the developmental transcriptome of *S. papillosus*, we used OrthoMCL [Li, 2003]. OrthoMCL works similar to Inparanoid when applied to two species. It works on all pairwise blast results from multiple species to define groups of homologous proteins and uses Markov clustering algorithm (MCL) to resolve many-to-many homologous relationship. The resulting groups have multiple genes from same or different species, from which different homologous relationships between species can be inferred. The clustering of homologous groups depends heavily on blast output, which intern depends on two parameters, Blast e-value cutoff and percentage match cutoff. In this analysis, E-value and percentage match cutoff were set to 10^{-5} and 50%, respectively. Additionally, we also used amino acid sequence length cutoff of 50. After all pairwise BLASTP searches, we loaded the blast hits results into a relational database to find gene pairs using orthomclPairs command. Once the gene pairs were found, MCL algorithm was used to cluster gene pairs with default options. Classification of homologous groups into one-to-one orthologous, one-to-many or many-to-many was performed using custom in-house scripts.

2.2.4 Classification of homologous clusters

We classified the homologous clusters into different classes in a study-specific manner. For the investigation of developmental transcriptomes of *P. pacificus*, homologous gene clusters obtained using the above methods (BRH and InParanoid) were classified into 3 classes, one-to-one orthologs, homologs, and orphans. Clusters with a single gene from the two or more species were treated as one-to-one orthologs. Clusters with more than one gene from any given species were classified as homologs. Clusters with one or more genes, but from a single species were grouped as orphans. Homologs and orphan clusters were further classified into "with paralogs" or "singletons". Homologs with paralogs (Many-to-Many and Many-to-one) genes have homologs both within species and in other species. A homolog singleton (one-to-many) refers to a gene with more than one homolog in at least one other species, but not within species. Similarly, "orphans with paralogs" (Many-to-zero) have homologs only within species and orphan singleton genes have no homologs in any species taken into consideration. In the study on developmental transcriptomes of *S. papillosus*, we used orthoMCL to group sequences from different *Strongyloides* species. The resulting clusters were then classified into one-to-one orthologs, 1:many (single gene in *S. papillosus* with multiple homologs in other species), Many :X (multiple *S. papillosus* genes with 0,1 or multiple homologs in other species) and Singletons (Single gene *S. papillosus* with no homologs in any species taken into consideration).

2.3 Analysis of re-sequencing data

In order to investigate gene duplication and losses between different natural isolates of the nematode *P. pacificus*, we analyzed two isolates RS5410 (sampled in La Reunion) and RS5200 (sampled in India). The data for the two strains were already sequenced as a part of 104 *P. pacificus* resequencing project [Rödelsperger et al., 2014]. *P. pacificus* strain PS312, sampled in California was used as a reference strain to detect structural variants (SVs) in the natural isolates. Even though,

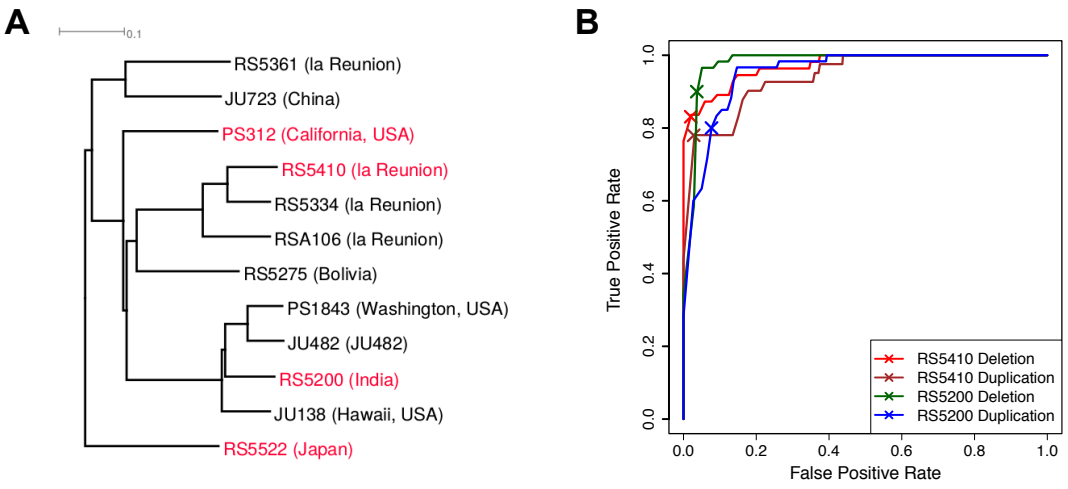


Figure 2.2: **Identification of SVs in *P. pacificus* Strains.** A) General phylogeny of *P. pacificus* strains and the sister species *P. exspectatus* (RS5522). Strains and species used to study the impact of SVs are colored in red. The phlogeny was generated using the neighbor joining methods based on $\approx 450,000$ parsimony informative sites. B) ROC analysis of 100 randomly chosen SVs that are identified by the program cnv-seq and manually evaulated. This analysis allows to choose various p-value and fold-change cutoff by considering true and false positive prediction.

the strains were members of different clades, have 1% genetic diversity with each other. Figure 2.2A shows the phylogenetic relationship between the strains analyzed. Genomic libraries were prepared as described previously [Rödelsperger et al., 2014, Ragsdale et al., 2013] and sequenced in the Illumina Hiseq 2000 platform as 100bp paired end reads. The resequenced reads were quality trimmed and alignment using the program Stampy [Lunter and Goodson, 2011]. *P. pacificus* Hybrid 1 assembly was used a reference genome for the alignment of raw reads. Duplicate reads were removed using samtools [Li et al., 2009] and local realignment was performed using GATK [McKenna et al., 2010].

2.3.1 Copy number variation (CNV)

Structural variations (SVs) are large variations in the genomic regions of an organism, with size ranging from 1KB to several MBs and includes inversions, translocation and copy number variations (such as deletion and duplication). Copy number variations (CNVs) are intermediate scale SVs and are typically greater than 1KB but less than 5 MB. CNVs show a difference in copy number when two or more genomes are compared [Zhao et al., 2013]. CNVs can be detected using a number of methods, which includes paired-end mapping, split read based methods, read depth approach, assembly based approach and also by methods that combine information of any aforementioned methods [Zhao et al., 2013]. To detect copy number variations (CNVs) between the natural isolates, reads aligned to the *P. pacificus* reference assembly was used as input. Differential read depth approach implemented in software cnv-seq was used to detect the differences in copy number [Xie and Tammi, 2009]. cnv-seq uses a sliding window based approach across the genome

to determine the read depth and compares between samples to test the significant difference in read coverage based on a Poisson model. Default options were used to run cnv-seq and sliding window size was set to 1000 bases. The results of cnv-seq were parsed extract fold difference in read coverage, start and end of CNVs, and its statistical significance.

2.3.2 Quality assessment and parameter choice

In order to minimize false SV calls in regions of low assembly quality or assembly gaps, we used an approach that explicitly test the resequencing data of the reference strain (PS312). The P-value and coverage fold-change cutoff were chosen based on the evaluation of 100 duplications and deletions from each strain by manually verifying the alignments using the Integrative Genomics viewer [Thorvaldsdóttir et al., 2013] . Each predicted structural variations were then manually classified as true positive or false positive. Thereby, the predicted SVs were evaluated based on additional features: coverage of the surrounding areas should be different even within the strain of interest, apparent heterozygous variant calls support the presence of a duplication, unusually high read coverage or fragmented islands of coverage in the reference sample indicate towards assembly problems. A receiver operating characteristic (ROC) analysis was performed to examine the quality of predictions depending on varying thresholds for p-value (P) and log2 fold change (Fc) in read coverage (Figure 2.2B) and finally a combination of thresholds that appeared to us as a good tradeoff between true and false positives was chosen. The chosen values were $Fc = 1.01$ and $P < 10^{-15}$ for RS5410 duplications, $Fc = -1.66$ and $P < 10^{-5}$ for RS5410 deletions, $Fc = 0.74$ and $P < 10^{-15}$ for RS5200 duplications, $Fc = -1.63$ and $P < 10^{-5}$ for RS5200 deletions. These values correspond to a true and false positive rate of 83% and 2% for deletion and 78% and 2.8% for duplications in RS5410 and 90% and 3.6% for deletion and 80% and 7.7% for duplications in RS5200. In order to generate a set of duplications with precise information of breakpoints, a split read approach as implemented in the software pindel [Ye et al., 2009] was additionally applied and identified 183 tandem duplications (¿ 2kb) in RS5200 and 117 in RS5410. For polarization of duplications and deletions, the analysis was restricted to the conserved syntenic regions identified by the program CYNTENATOR [Rödelsperger and Dieterich, 2010] between the reference strain and the sister species *P. exspectatus*.

2.4 Expression analysis

2.4.1 Alignment of raw reads

Alignment or mapping of raw reads back to the reference genome is an important standard step in the analysis of next-generation sequencing data, irrespective of the techniques (RNA-seq, Chip-seq or whole genome sequencing). In general, an alignment tool applies a series of insertions, deletions, and mismatches to successfully align the reads and assigns a score to identify the optimized location in the reference genome. Different tools are optimized to align data from different NGS methods, for examples Bowtie [Langmead and Salzberg, 2012] align reads from genomic DNA sequencing data, while STAR software [Dobin et al., 2013] is more specialized to align data from RNA-seq. The difference in the alignment of read is that RNA-seq aligners have to deal with long introns. RNA-seq tools such as Tophat2 and STAR align reads to the reference genome and identifies splice junction between exons by joining the aligned reads [Kim et al., 2013, Dobin et al., 2013].

In this work, TopHat2 [Kim et al., 2013] was used extensively to align RNA-seq reads to the reference genome. TopHat2 internally uses Bowtie, which in-turn uses an indexing method for fast alignment [Kim et al., 2013]. So the *P. pacificus* hybrid assembly was indexed using Bowite2 with default parameters and used as an option for TopHat2. A reference annotation in GFF format file was provided as an additional option for TopHat2 to identify exon-exon junctions. All TopHat2 outputs were saved as coordinate sorted bam file for quantification.

2.4.2 Quantification of gene expression levels

Cufflinks [Trapnell et al., 2012], an RNA-seq read quantification tool from TopHat suite was used estimate the gene expression levels. Multi-mapping reads or reads mapping to multiple genomic locations were corrected using -multi-read-correct option and the read fragment bias was also corrected using -b options in cufflinks. The gene expression levels were normalized and estimated using classical FPKM approach. All other options were set to default.

2.4.3 Differential gene expression analysis

Cuffdiff2 [Trapnell et al., 2012], another tool from TopHat suite, was used estimate transcript-level expression and to identify differential expressed genes. As a comparative tool, Cuffdiff2 takes at least two alignment files and controls for variability between biological replicates for each condition.

Ambiguously mapped reads were corrected using -multi-read-correct option and false discovery rate (FDR) cutoff was set to a maximum 0.05. Samples were normalized using the median of geometric means of fragment counts across all samples using geometric normalization method. Genes with expression difference in log-fold change greater than 1 or less than -1 and p.value (FDR corrected) less than study-specific cutoff were treated as significantly differentially expressed. Different p.value cutoff was used to determine the statistical significance. For the *P. pacificus* and *S. papillosus* developmental transcriptomes studies, 0.05 was used as cutoff. A cutoff of 0.01 was used for when comparing transcriptomes of different strains of *P. pacificus*.

2.4.4 PCA and clustering methods

To gain first insight into the transcriptome of different samples or developmental stages, different clustering methods were used. To estimate the global variability of the transcriptomes, principle component analysis was performed using prcomp function implemented in R [R Core Team, 2014].

In the analysis of the developmental transcriptomes of *P. pacificus*, samples were compared based on the number of differentially expressed genes in each pairwise comparison, Euclidean distances were calculated using the number of genes with significant differences in expression and clustered using hierarchical clustering function "hclust" implemented in R [R Core Team, 2014]. To identify clusters of co-regulated genes, an unsupervised biclustering algorithm SAMBA [Tanay et al., 2002], implemented in the EXPANDER package (V 6.3.1) was used. A matrix of size nm with n genes and m comparisons was generated, with the individual entries indicating whether a gene was significantly differentially expressed ($-1 :=$ *downregulation* and $1 :=$ *upregulation*) or not. SAMBA algorithm was executed using the matrix as relative expression data and with the default setting in the Expander package.

2.4.5 Visualization of expression pattern in phylogenetic trees

To manually inspect and compare the expression levels and sequence divergence of homologous genes, gene expression levels were displayed along with phylogenetic trees. Homologous gene sequences from the orthoMCL clusters, and sequences with protein domains of interest were extracted and aligned using multiple sequence alignment (MSA) software clustal omega [Sievers et al., 2014]. Both Input and output format for the MSA were set to fasta format. Phylogenetic trees were reconstructed using maximum likelihood methods implemented in phangorn R package [Schliep, 2010]. The best substitution model that describes the evolution of homologous genes was identified with prottest software [Darriba et al., 2011]. Expression or differential expression matrix for each gene in the homologous clusters was also generated using base functions in R. The phylogenetic tree and the expression matrix files were loaded onto the interactive tree of life server (itol) [Letunic and Bork, 2011] to display the expression levels and differential expression status in different samples.

For differential expression data, the matrix was reduced to 1,0 and -1, which represents up-regulation, no difference, and down-regulation respectively. This simplified matrix was saved as a text file and was only used to visualize the expression pattern as heatmaps. The text file was then appended with itol heatmap specific options such as field labels, dataset labels, and separators. Detailed list of options can be found in the itol help page. To sort the heatmap columns, a tree in newick format was provided using the "field_tree" option. A hierarchical cluster was generated using the manhattan distances based on the reduced matrix. The hierarchical cluster was converted to phylo object using as.phylo function and saved as newick format using the write.tree function from ape R function [Paradis et al., 2004]. To display colors on each leaf, a text file with three columns was generated. Each row contains information about a gene in the phylogenetic tree, with gene identifier, a keyword "label" and color code in hexadecimal format.

2.5 Evolutionary Analysis

2.5.1 PAL2NAL

PAL2NAL is a program that generates a codon alignment from a multiple sequence alignment of proteins and DNA [Suyama et al., 2006]. It is essential to convert alignment of proteins and corresponding DNA to codon to estimate the number of changes in the DNA sequences that can result in synonymous and non-synonymous changes in the protein sequences. Options -nogap (remove columns with gaps and stop codons within the frame) was used during the execution of PAL2NAL. The Output files were saved in paml format.

2.5.2 Estimation of sequence evolutionary rate

PAML (Phylogenetic Analysis by Maximum Likelihood) is a software package for the analysis of evolutionary constraints acting on the protein or DNA sequences [Yang, 2007]. The PAML suite includes tools such as baseml (for nucleotide sequences) and codeml (for codons) for the estimation of substitution rates between sequences.

Codeml estimates omega value (ω), the ratio of non-synonymous and synonymous substitution rate, which in-turn indicates the strength of natural selection acting on the protein sequences. Omega value = 0 indicates neutral selection, while $\omega < 1$ and $\omega > 1$ indicates negative and positive

selection, respectively. One ratio model was also used to derive an omega estimate each orthoMCL clusters. One ratio model estimates omega value by taking the average of ratios over all sites and all lineages. One omega for whole alignment, summarizes the evolutionary rates act on the genes. Since the ratio is calculated based on the based average of all branches, it cannot exceed a value of one. Protein sequences from each orthoMCL cluster were aligned to generate multiple sequences alignment, converted to codon alignment using PAL2NAL and used as input for codeml. Codeml control file was modified by setting model =0, NSites = 0, and runmode =1. To estimate the general pattern of natural selection action on different gene sets, omega values were estimated by comparing sequences in each gene clusters in a pairwise manner. This analysis was performed by specifying the model=0, NSsites =0 and runmode = -2 in the codeml control files.

2.5.3 Detection of positive selection

Codeml consists of different substitutions models to better understand the evolutionary process. The branch model allows omega value to differ among branches in the phylogenetic tree, which can be used detect positive selection acting on a particular lineage. The site models allow omega values to differ among sites in the protein. There are 6 different site models, which fits different evolutionary scenarios. Model M0 is the basic model and assumes one unique omega value for all sites. M1a model splits alignment into two categories, i.e sites under negative selection ($\omega < 1$) and site under neutral evolution ($\omega = 1$). M2a is similar to M1a but it assumes additional category, sites under positive selection ($\omega > 1$). Alignment sites are divided into ten (M7) or eleven (M8, M8a) classes. Model M7 assumes that all classes of sites are either neutrally evolving or negatively selected ($\omega \geq 1$). Models m8 and m8a add one explicit class of sites with $\omega > 1$ (m8) and $\omega = 1$ (m8a) [Yang, 2007]. Signatures of positive selection can be detected by comparing different models. Finally, the branch-site models, which combines site and branch models, allows omega to vary across all branches and among all sites on the phylogenetic trees. The branch-site models are best suited to detect sites under the strong influence of positive selection across a particular lineage [Yang, 2007]. In this thesis, different site models were compared to identify sites under positive selection. Specially, model m8 vs. m7 and m8a, respectively were compared and defined as evidence for positive selection if both the m7 and m8a models were rejected over m8 in a likelihood ratio test with $p - value < 0.01$. To detect positively selected sites, a ML tree regenerated using the cluster was provided as a treefile in codeml control file. Nsites option in the control file was set to 0, 7 and 8 to estimate likelihood for positively selected sites using models M0, M7 and M8 respectively. Omega values were estimated during the modeling by setting to fix_omega option to 0. To estimate using M8a model, codeml was executed again with Nsites = 8, fix_omega=1 and omega=1.

2.6 Functional characterisation of gene sets

Genes that shows differential expression in a comparison or belongs to set of co-regulated genes or within CNV regions, were further analyzed to gain insights into their putative functions.

2.6.1 Gene ontology

To identify the ontology terms associated with *P. pacificus* genes, one-to-one orthologs between *P. pacificus* and *C. elegans* was identified using methods described in chapter 2.2. Gene ontology

terms associated with *C. elegans* one-to-one orthologs were determined using the David functional annotation webtool [Huang et al., 2009] and the resulting GO terms were transferred to *P. pacificus* one-to-one orthologous genes. Go terms were taken into considerations only if the enrichment score greater than 1 and FDR corrected $p - value < 0.05$.

2.6.2 Protein domain enrichment

Protein domain enrichment analysis was performed using hmmsearch and in-house R scripts. The hmmearch program was used to search for known protein domains in the protein sequences. The search program hmmsearch and the PFAM domain database were both obtained from the HMMER package (version 3.0) [Johnson et al., 2010]. Domain enrichment analysis of differentially expressed genes was performed in R and Fisher's exact test was used to test for statistical significance (FDR corrected $p - value < 0.05$).

2.6.3 Comparison with previous expression profiling studies

The transcriptome of *P. pacificus* has been investigated earlier to understand the mechanism of immunity [Sinha et al., 2012a], lifespan [Rae et al., 2012] and dauer exit signals [Sinha et al., 2012b]. The *P. pacificus* genes differentially expressed in different developmental stages were compared with these three previous gene expression-profiling studies. Enrichment of developmentally regulated genes among differentially expressed genes from these previous studies was calculated using custom R script

Chapter 3

Results

This chapter contains content from the following publications. The copyright holder has granted the re-use permission.

Markov, G. V., Baskaran, P., and Sommer, R. J. (2015). The Same or Not the Same: Lineage-Specic Gene Expansions and Homology Relationships in Multigene Families in Nematodes. Journal of Molecular Evolution, 80(1):18-36.

Baskaran, P. and Rödelsperger, C. (2015). Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus. PloS one, 10(6) e0131136.

Baskaran, P., Rödelsperger, C., Prabh, N., Serobyan, V., Markov, G.V., Hirsekorn, A., and Dieterich, C. (2015). Ancient gene duplications have shaped developmental stage-specic expression in Pristionchus pacificus. BMC Evol Biol, 15(1).

Baskaran,P., Jaleta, T., Streit, A. and Rödelsperger C. (2017). Duplications and Positive Selection Drive the Evolution of Parasitism-Associated Gene Families in the Nematode Stongyloides papillosus. Genome Biol Evo,l 9(3): 790-801.

3.1 Macroevolutionary patterns of gene duplications in nematodes

3.1.1 Summary

One of the intensely debated evolutionary concepts is the functional conservation among orthologs called Orthologous Conjecture (OC) [Nehrt et al., 2011, Chen and Zhang, 2012, Altenhoff et al., 2012, Rogozin et al., 2014]. OC assumes that the single copy orthologs in two different species perform similar function than paralogs. This functional conservation is achieved with the help of a selection, which acts against changes in the orthologous genes. However, upon duplication, the selection constraints are relaxed allowing the paralogs to accumulate changes. Lineage-specific duplications limit one-to-one orthologs and create an opportunity for functional divergence. As the first step in understanding the evolutionary forces acting on the genome, identification of homologous relationship between genes is of utmost importance. In this work, we investigate more than 2000 manually curated protein sequences from 11 nematodes of seven different genera to determine their orthology relationships. The dataset contains seven different gene families, which vary in the number of family members and functions. Detailed phylogenetic analysis of all gene families has highlighted

the underrepresentation of one-to-one orthologs between *C. elegans* and *P. pacificus* . Of all *C. elegans* and *P. pacificus* genes taken into consideration, less than 12 % are one-to-one orthologs. Interestingly, most of the family members were clustered together in different subtrees, which indicates lineage specific duplication events. Estimation of gene birth and death rates show strong increases in gene number in the branch leading *P. pacificus* compared to nematodes of different clades. Taken together, this study shows that the evolutionary history of these gene families in *P. pacificus* is shaped by massive expansion events.

3.1.2 Multigene families from 11 nematodes

In this work, we manually curated and analyzed datasets from several multigene families that are potentially involved in two different substrate metabolism pathways. Given the variable quality and incompleteness of different nematode genomes, manual curation allows creating a comparable quality dataset to precisely evaluate the history of individual gene family. Specifically, we compared gene families such as short-chain dehydrogenases reductases (SDR), cytochrome P450 (CYP), glutathione-S-transferases (GST), UDP-glucuronosyltransferases (UGT), ABC transporters, fatty acid elongases (ELO) and fatty acid desaturases (FAT). The gene families, GST, SDR, CYP, GST and ABC are involved in the Xenobiotic metabolism and ELO and FAT belongs to the PUFA (Polyunsaturated and branched-chain fatty acid) synthesis pathway. The two pathways were chosen because the enzymes involved metabolize two different kinds of substrate [Lindblom and Dodd, 2006, Watts, 2009]. Previous knowledge about basic topology and structural conservation of these families along with reasonable size in terms of family members allows us to root the phylogenetic trees and preciously evaluate different evolutionary events.

For the analysis, we chose 11 nematode species such as *P. pacificus* , *C. elegans* , *C. briggsae*, *C. remanei*, *B. malayi*, *L. loa*, *A. suum*, *T. spiralis*, *B. xylophilus*, *M. hapla*, and *S. ratti*. The species were chosen in such a way that it cover 4 out of 5 distinct nematode clades. The sequences were collected using blast searches against GenBank and wormbase (WS232). The *P. pacificus* data were collected from http://www.pristionchus.org

3.1.3 The GST family Shows eighteen lineage-specific duplication

In order to test the mode of expansions and losses, we first performed a detailed analysis of GST gene family by reconstructing the phylogenetic tree (Figure 3.1 and Appendix figure 6.1). Conserved protein structures [van Rossum et al., 2004, Perbandt et al., 2005, Asojo et al., 2007], reasonable family size (59 in *C. elegans*) and well-studied phylogeny [Sheehan et al., 2001] provided excellent basis for detailed investigation of GST proteins. We found that the number of GST family members varies widely among chosen nematode species, ranging from 59 in *P. pacificus* and *C. elegans* to only two in T.spiralis. GSTs are relatively short proteins (about 200 amino acids) but can be divided into 5 subclasses, GST Kappa, GST omega, GST zeta, GST Pi and GST sigma/nu. Visual inspection of the five subclasses shows that at least one group of paralogous clusters were found in Clade III and Clade IV. The GSTs of clade IV and clade V nematodes show a strong pattern towards lineage-specific expansions that are manifested in the form of paralogous clustering, including 18 expansion events in *P. pacificus* . However, the expansion pattern varies between different subclasses of GST and between species. We found three paralogs of *P. pacificus* GSTs in the omega and zeta, but no expansion in the pi class. By narrowing our analysis to the two focal species, we found that even though *P. pacificus* and *C. elegans* have exactly same number of GST genes (59), only 11

genes were found to have one-to-one orthologous relationship (Figure 3.1 and Appendix figure 6.1).

3.1.4 Lineage-specific expansion of CYP, SDR and UGT families

The members of cytochrome P450 (CYP) family show strong conservation in protein sequences and are involved in the synthesis of endogenous hormones and xenobiotic metabolism [Brown et al., 2008]. Similar to GST, CYP also shows huge variation in gene content across species. By reconstructing the CYP phylogenetic tree, we found 9 different *P. pacificus* specific paralogous groups, indicating that in *P. pacificus* CYP have undergone expansion at least 9 times in the past. Similarly, in the case of SDR, a large family of NAD(P)(H)-dependent oxidoreductases, we also found large-scale expansions specific to clade IV nematodes (17 expansion events). Specifically, we found the SDR family history was shaped by 11 expansion in *P. pacificus* and 4 expansions in *C. elegans* . In contrast to large-scale expansions, only 9 one-to-one ortholog genes between *P. pacificus* and *C. elegans* were found (Appendix figure 6.2).

The UGT family consists of enzymes that are highly diverse and are involved in the addition of glucuronyl residues in xenobiotic metabolism. We found that the UGT family has undergone 19 amplification events in different nematodes including five in *P. pacificus* . Out of five amplification events, two were large-scale lineage-specific expansions in *Pristionchus* lineage encompassing 24 and 54 genes. Similarly, 1 out of 3 amplification events observed in the *Caenorhabditis* genus, was a massive expansion encompassing 154 genes. Even though we observed evidence for large-scale expansions in *Pristionchus* and *Caenorhabditis* genus, only one gene was identified as one-to-one ortholog between *C. elegans* and *P. pacificus* . Taken together, the pattern of lineage-specific expansion of these three gene families was more similar to the observation in GST family. The observed pattern of lineage specific expansions in all four-gene families show that the expansion events observed are not an exception but more likely the rule (Appendix figure 6.2).

3.1.5 Lineage-specific expansion of Desaturases and Elongases in *Pristionchus* genus

In *C. elegans* , desaturases (FAT) and elongases (ELO) are involved in the production of polyunsaturated C:20 fatty acids [Watts, 2009]. By reconstruction of desaturases phylogenetic tree, we found the family tree can be divided into four statistically well-supported subtrees (Figure 3.2). One of the subtrees that contain *C. elegans* FAT-1 and FAT-2, shows that single lineage-specific duplication events occurred independently in *Caenorhabditis* genus, *P. pacificus* , and *B. xylophilus*. The subtrees with *C. elegans* FAT-5, FAT-6, and FAT-7, shows six paralogs of *P. pacificus* clustered together indicating lineage-specific amplification. In contrary, the subtree that contains other *C.elelgans* FAT genes, including functionally uncharacterized *C. elegans* protein F33D4.4 and Y54E5A.1, show strong conservation, indicated by the presence of one-to-one orthologs between *P. pacificus* and *Caenorhabditis* genus.

In the case of the elongase family, we found a high degree of conservation in the *Caenorhabditis* genus, with nine *C. elegans* genes having one-to-one orthologs in other *Caenorhabditis* species (Appendix figure 6.3). Two out of the nine *C. elegans* genes are preserved as one-to-one orthologs with *P. pacificus* . We could not find any potential duplication events in the *C. elegans* elongase family. In contrary, we found at least two different expansions of *P. pacificus* resulting in more than three in-paralogs. While a single *P. pacificus* specific expansion of ELO-9 group resulted in 9 paralogs, expansion of ELO1-/ELO-7 groups resulted in 14 pacificus in-paralogs, from one or

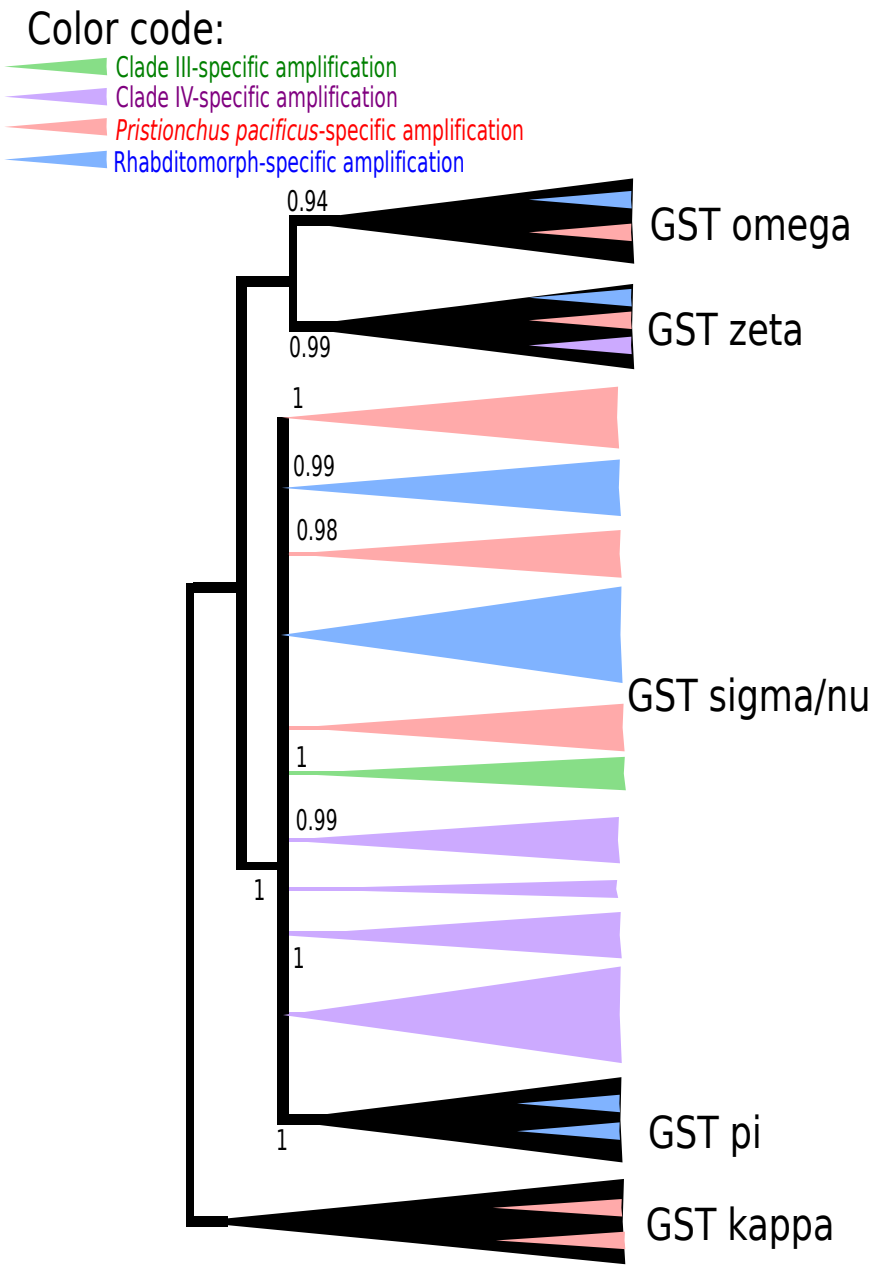


Figure 3.1: **Gene duplication in GST family.** Schematic tree representing phylogenetic relationship among nematode GST genes. Different colors on the tree represents lineage-specific expansion events in different nematode lineages. This simplified view shows three different expansion of GST Sigma/nu in *P.pacificus* lineage. The value on each node represents statistical support based on likelihood-ratio test. Values more than 0.97 are considered fully reliable. Gene from *Ceanorhabditis* species are colored in blue, *P. pacificus* in red, Clade IV nematodes (*B. xylophilus*, *M. hapla* and *S.ratti*), in purple, Clade III nematodes (*A. suum*, *B. malayi* and *L. loa*)in green and clade I nematode (*T. spiralis*) in grey

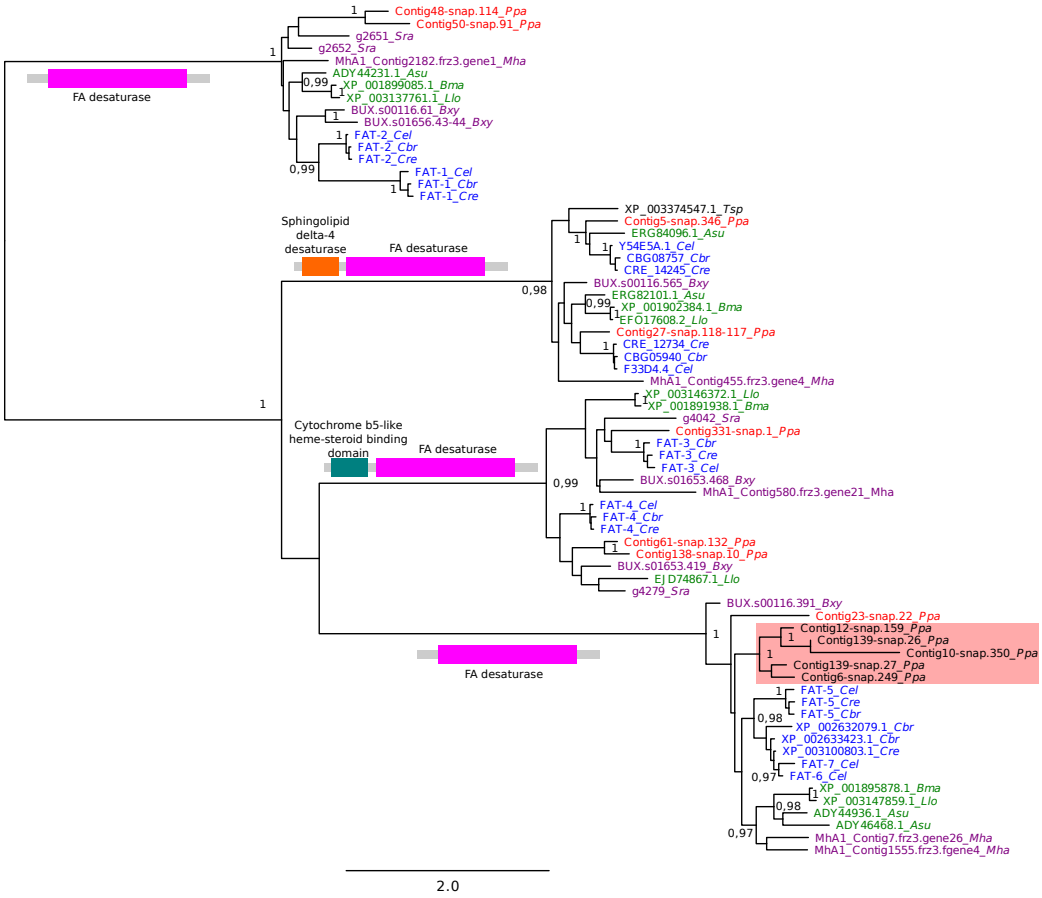


Figure 3.2: **Pattern of gene duplication in Desaturases family.** Maximum likelihood tree reconstructed using desaturases protein sequences of 11 nematode species. The tree shows strong conservation between *Pristionchus* and *Ceanorhabditis* genus with single *P. pacificus*-specific expansion (highlighed in red)

two distinct amplification events. Taken together, we found that gene families involved in PUFA synthesis pathways (FAT and ELO) have a smaller frequency of gene duplication compared to families in the xenobiotic pathway.

3.1.6 High degree of conservation in ABC transporter family

The ABC transporter family plays a central role in the transportation of metabolites between different compartments of the cell or from the cytoplasm to outside of the cell [Sheps et al., 2004]. The subfamilies of ABC transporter were shown to have variations both in the domain number and order [Sheps et al., 2004].

We performed a detailed analysis of the four type of domain combinations found in the nematodes. The ABC transporter subfamilies, ABCA, ABCB, and ABCC belong to the first type, subfamily ABCD belongs to second, ABCF belongs to third and ABCG subfamily belongs to fourth type domain combination. We found a large number of *P. pacificus* genes in all domain types have

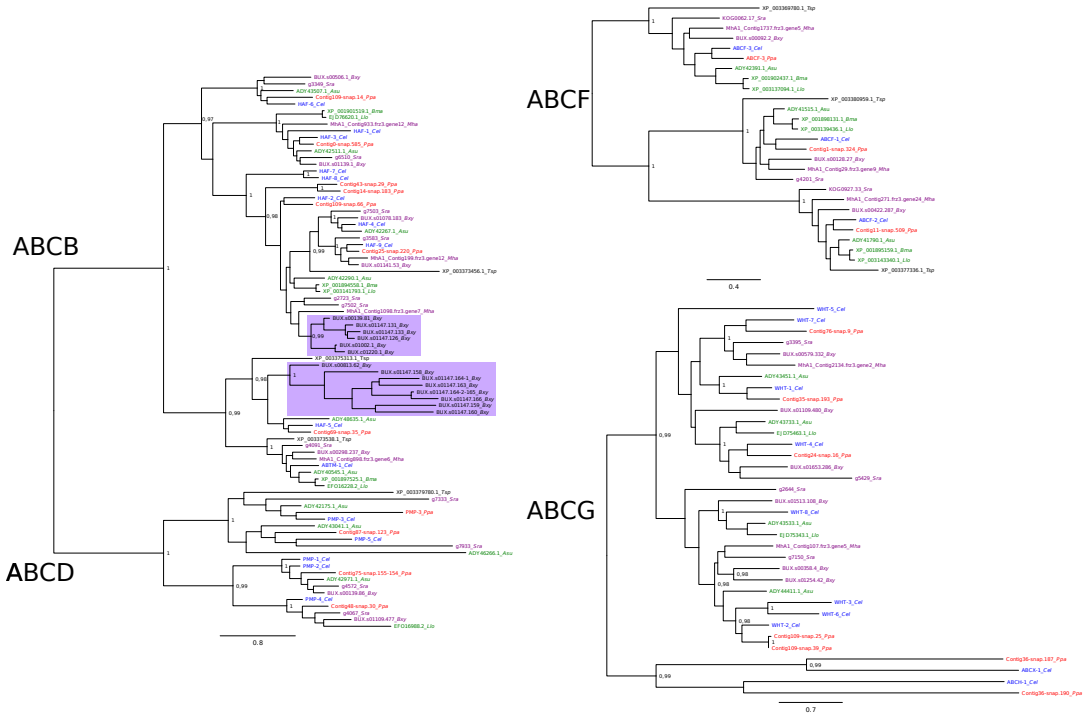


Figure 3.3: **Conservation in ABC transporters subfamiliy between *C. elegans* and *P. pacificus*.** Maximum likelihood tree of four ABC subfamilies shows majority of genes were maintained as one-to-one orthologs between *C. elegans* and *P. pacificus*. Interestingly, ABCB subfamiliy shows two expansion events in *B. xylophilus* (highlighted in purple).

one-to-one orthologs with *C. elegans* . Specifically, we found seven out of 15 *C. elegans* genes have one-to-one orthologs in *P. pacificus* in ABCD and ABCB types. A similar pattern was also observed in type ABCF and ABCG, where all 3 ABCF *P. pacificus* genes and seven out of nine ABCG *P. pacificus* genes have one-to-one orthologs in *C. elegans* . Even though we found *P. pacificus* specific expansion events in first type domain comparison, a larger proportion of genes show one-to-one ortholog relationship between *C. elegans* and *P. pacificus* (Figure 3.3 and Appendix figure 6.4).

Despite varying degree of conservation among the subfamilies of ABC transporter, overall conservation pattern observed in ABC transporter is much higher than in any other multigene families analyzed in this work. This suggests stronger evolutionary constraints acting against changes in the ABC transporter family.

3.1.7 Increase in gene content along the branches leading to *Pristionchus pacificus*

By reconstructing the ancestral gene set profiles, we investigated the general dynamics of gene birth and death rates among all eleven fully sequenced nematodes. We found variations in both at the predicted total gene count and its relative proportions. While there is a general trend of decrease

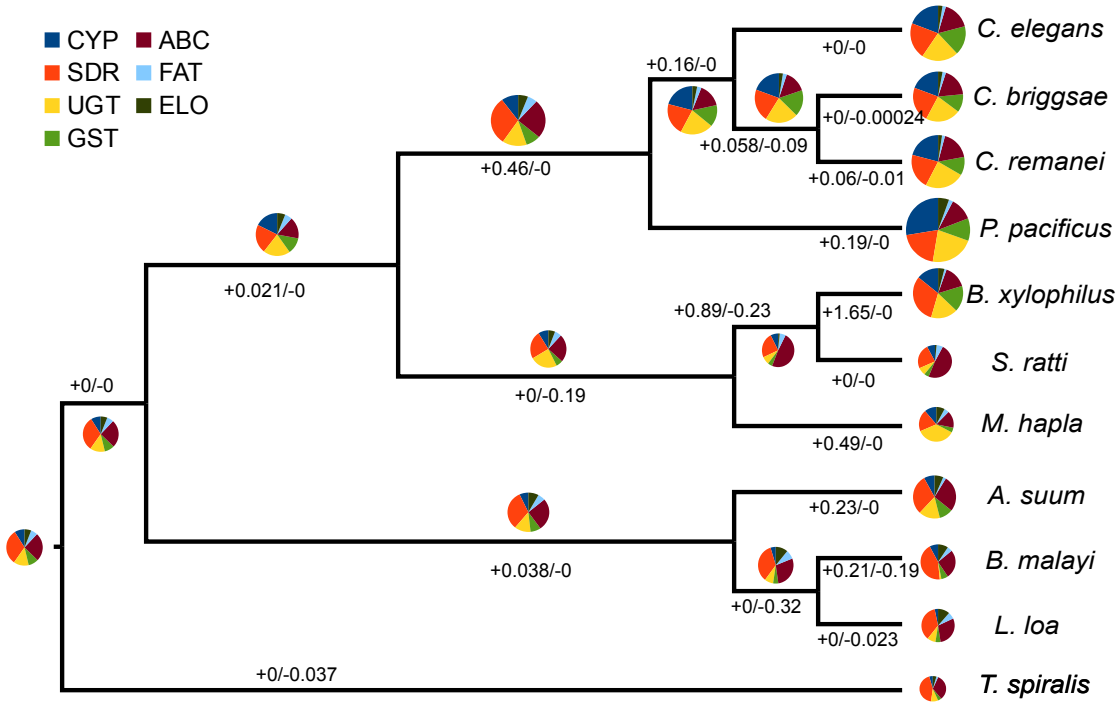


Figure 3.4: **Estimation of ancestral gene count for each family.** The ancestral proportion, global birth(+) and death(-) rates along all branches were estimated using the BDI-FR-ML Methods implemented in BadiRate [Librado et al., 2012] . The proportion of overall gene number in the investigated families are represented increasing size of pie-chart.

in gene counts in branches leading to clade IV and clade III, we found a strong increase in all clade V nematodes. Interestingly, the branch leading to *P. pacificus* shows the largest increase in global gene number. This analysis also shows that there is considerable variation in the rate of gene births and deaths across different gene families (Figure 3.4). In case of PUFA- metabolizing genes, a trend towards a significant drop in the nematodes gene number compared with their common ancestor was observed. In contrary, Xenobiotic-metabolizing genes show an increasing trend, with 27% of *P. pacificus* total genes belongs to CYPs family. In addition, UGT shows strong variation in gene number between species ranging from 2 % in *B. malayi* to 50% of in*M. hapla*. Together, this shows that the pattern of gene expansion looks random, with no obvious correlation between families with functional similarities (Figure 3.4).

3.2 Microevolutionary patterns of gene duplications in *Pristionchus pacificus*

3.2.1 Summary

The previous comparison of large gene families between *P. pacificus* and *C. elegans* that diverged more than 100 million years ago shows that gene duplication dominates the macroevolutionary

picture. The dynamic balance of gene gain and loss represents an important evolutionary process in the diversification species. However, to better understand how, existing deletions and duplications become fixed in populations, structural variations (SV) between different strains of the same species have to be investigated. By making use of already available genomic resequencing and transcriptomic data from two natural isolates of *P. pacificus* and a reference *P. pacificus* strain, the intra-species pattern of gene loss and gene gain has been investigated. To correctly interpret the genomic changes in an evolutionary sense, the analysis was restricted to regions of perfect synteny between *P. pacificus* and its sister species *P.exspectatus*. Using genomic resequencing data, deletions and duplications longer than 2kb were identified relative to the reference strain. Transcriptome analysis of deleted and duplicated genes in the syntenic region showed patterns contrasting the naive expectation that duplication leads to higher gene loss. While Loss of genes correlates with lack of expression, duplication does not show a trend towards an increase in gene expression. Differential gene expression analysis to find differences in gene expression levels between duplicated strain and reference strains shows that only 7 and 3 genes show double dosage after duplication. Detailed analysis of segregating sites that distinguish the duplicated copies, shows that in the majority of cases only one of the duplicated copies is expressed. Further analysis of deleted and duplicated genes, based on different orthology classes indicates strong evolutionary constraints acting to preserve synteny as even ¨transcriptionally silentd̈uplications are not tolerated in large parts of the genome. This further suggests and confirms previous observations of either neutral or deleterious nature of structural variants.

3.2.2 SVs are sparsely found in the genome

Based on significantly differential read coverage in resequencing data of two natural isolates RS5200 and RS5410 with respect to a reference strain PS312, we identified deletions and duplications [Xie and Tammi, 2009]. As differences in read coverages may result from reads that are too divergent to be mapped, we optimized the predictions of SVs using an empirical data set of hundreds of manually classified SV predictions and chose cutoffs that showed a good tradeoff between true and false positives (Figure 2.2B). For the strain RS5200, this approach resulted in 1621 and 609 predicted deletions and duplications, respectively, and 1642 and 565 predicted deletions and duplications for RS5410. Predicted SVs span a total range from two to 75kb. The portion of the *P. pacificus* genome that was predicted to be affected by deletions and duplications was 7.9Mb (4.6%) and 2.2Mb (1.3%) for RS5200 and 8.7Mb (5%) and 2.1Mb (1.2%) for RS5410. To rule out, that cnv-seq shows a tendency to predict deletions, rather than duplications, we tested RS5200 and RS5410 against each other showing very similar numbers of deletions and duplications, indicating that the higher number of predicted deletions is due to the fact, that we used PS312 as a control sample.

To find the genome-wide distribution and hotspots of SVs, we divided the *P. pacificus* assembly into non-overlapping windows of 100kb size and plotted the fraction of a given genomic window that was predicted as being deleted and duplicated, across the six *P. pacificus* chromosomes (Figure 3.5). Interestingly, we find two large regions spanning several hundred kilobases with extensive deletions for both strains on the X chromosome. These two blocks are located mostly outside of syntenic regions with the sister species P. exspectatus, potentially suggesting, that these regions might reflect recent lineage-specific SVs in the reference strain PS312. Predicted duplications show a more even distribution across the chromosomes with only minor clusters on chrIV, chrII, and chrX (Figure 3.5).

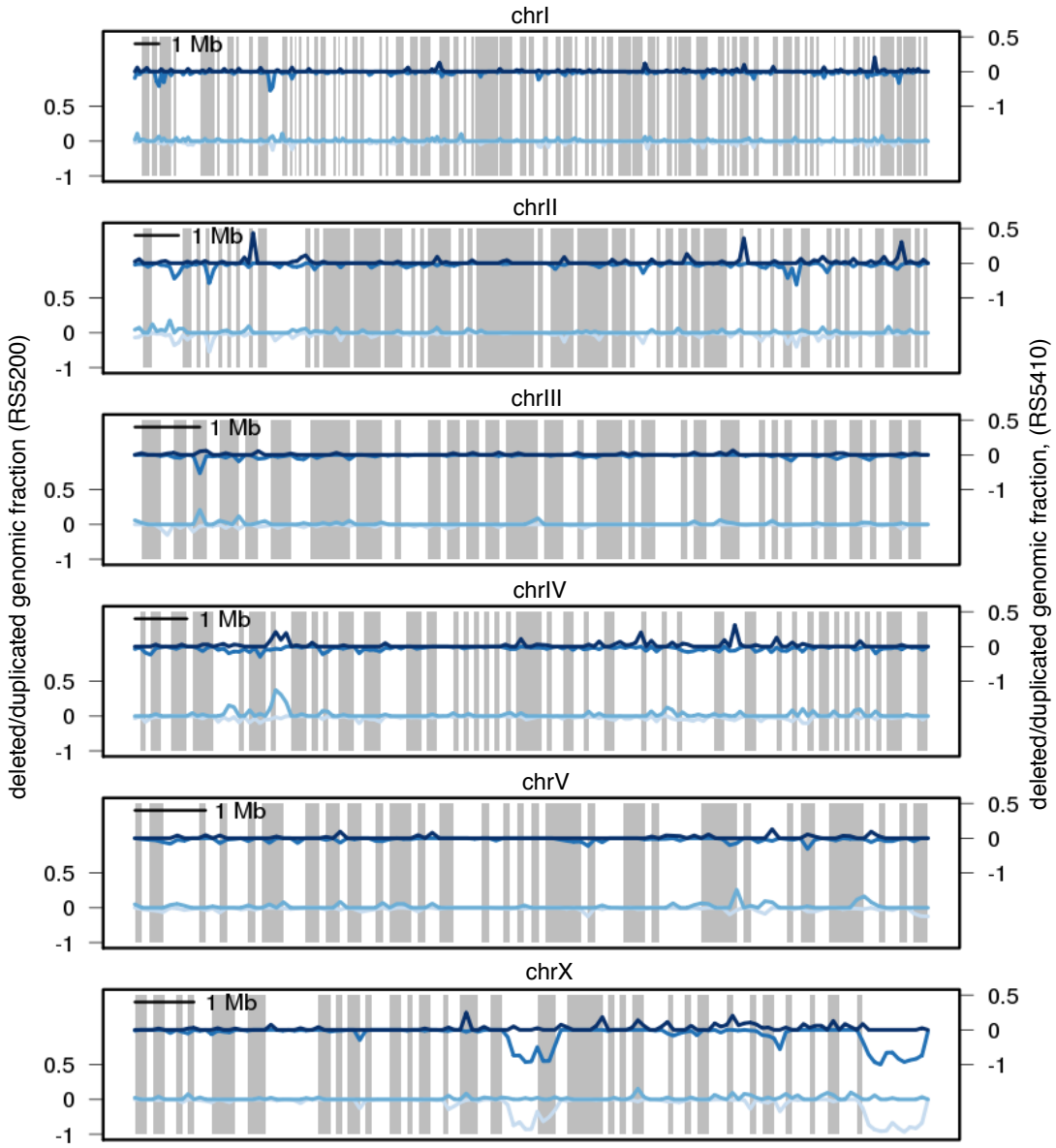


Figure 3.5: **Distribution of SVs across *P. pacificus* chromosomes.** The plot shows the fraction of genomic sequence within non-overlapping 100kb windows, predicted to be duplicated (positive values) and deleted (negative values) relative to the reference strain PS312. Gray boxes indicate conserved syntenic regions with the sister species P. exspectatus. While duplications exhibit a more even distribution, we identified two almost MB sized regions with high fraction of missing sequence on the X chromosome

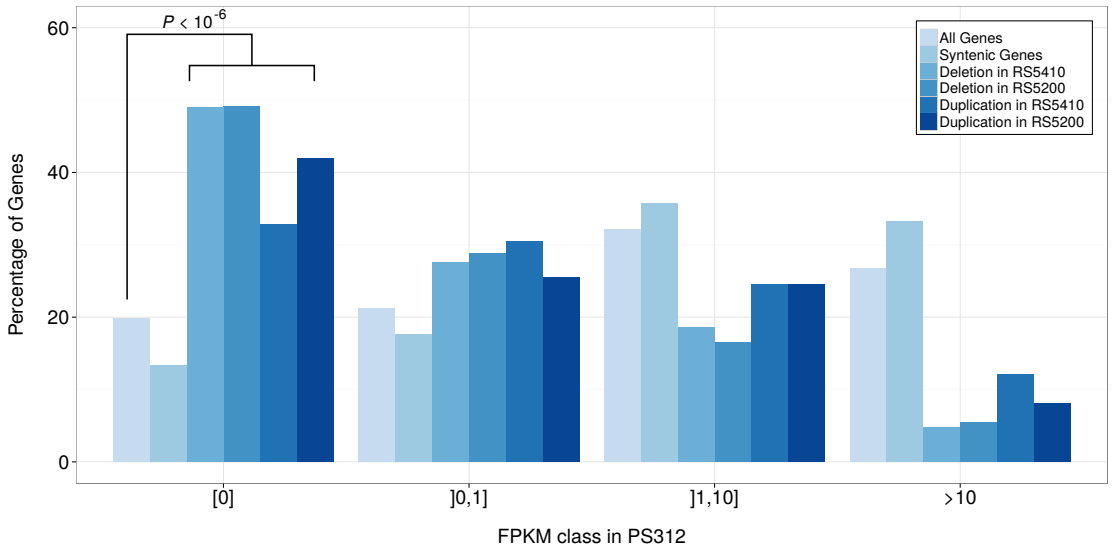


Figure 3.6: **Distribution of SVs in different expression class.** The plot shows the proportion of different categorical genes in different expression classes in the reference strain PS312. Gene categories include all genes, genes within syntenic regions and genes predicted to be deleted or duplicated. The FPKM value of zero represents lack of expression evidence and FPKM value > 10 represents robust expression.

3.2.3 SVs preferentially affect genes with low expression

Given that the breakpoints of SVs are not well defined and to account for false positive predictions, we considered only genes as being deleted or duplicated with respect to the reference strain PS312, if the complete gene from first to the last exon is covered by the SV. Following this definition, we find 1240 and 322 genes as being deleted and duplicated, respectively, for RS5200 and 1397 and 289 genes as being deleted and duplicated for RS5410. 65 of the duplicated genes and 732 of the deleted genes were predicted in both strains suggesting that these events predate the split of the two strains and would correspond to single evolutionary events.

To assess the effect of SVs on gene expression, we compared the gene expression levels of both deleted and duplicated genes in the reference strain PS312, where by definition all of them should be present as single copies. We found that genes affected by SVs in other strains show a tendency to have low or no expression in the reference strain. More precisely, more than 80% of genes deleted in both strains show expression levels less than 1 FPKM in the reference strain. This also applies to duplicated genes of which more than 65% have low expression levels (Figure 3.6). Compared with the proportion of lowly expressed genes in the entire *P. pacificus* gene set, the proportion of deleted and duplicated genes with low or no expression is significantly higher. This suggests that the SVs preferentially affect genes with low expression.

Interestingly, with respect to differential gene expression analysis, this finding indicates, that most gene loss and gain events will not result in significant differential expression calls because of the low statistical power to detect significant differences for genes with few read counts. To test this, we carried out differential expression analysis and found that only 25 and 24 (RS5410 and

RS5200 respectively) deleted and 7 and 3 (RS5410 and RS5200 respectively) duplicated genes are identified as being significantly differentially expressed (FDR corrected $p - value < 0.01$). Also due to the lower expression of affected genes, we do not expect SVs to dramatically affect global transcriptome profiles as is demonstrated in the strong correlation ($\rho > 0.89$, Spearman) between expression values [Ragsdale et al., 2013].

3.2.4 SVs as derived events in the natural isolates

We polarized SVs with the help of the genome of a closely related outgroup species to correctly interpret the identified effects of SVs on gene expression in an evolutionary sense. For Polarization, we made use of the recently assembled draft genome of the sister species *P. exspectatus* [Rödelsperger et al., 2014], which showed roughly 10% divergence to *P. pacificus* on the nucleotide level. We polarized the data by restricting the analysis to regions of perfectly collinear gene orders between the reference genome and the genome of the sister species. The identified perfectly collinear regions span roughly 74.1Mb (43%) of the *P. pacificus* genome assembly. All duplications and deletions within these regions were considered as derived events. This polarization reduced the number of genes to 288 deleted and 92 duplicated genes for RS5200 and 333 deleted and 112 duplicated genes for RS5410.

3.2.5 Deleted genes shows no expression

The distribution of expression levels in the reference strain and two natural isolates in defined expression classes were shown in Figure 3.7. The classes were defined based on the expression values into no evidence of expression ([0]), low expression ([0,1]), high expression ([1,10]) and robust expression ([> 10]).

In RS5200, 80% of deleted genes show no evidence of expression. This represents a highly significant trend towards loss of expression, as 65% of these genes show considerable expression in the reference strain ($P < 10^{-34}$, Fisher's exact test). In case of RS5410 as well, we observed a similar trend, significantly higher proportion of genes with no expression compared with reference strain ($P < 10^{-14}$, Fisher's exact test). Given the various levels of uncertainty such as false positive predictions and the imprecise boundaries of SVs, which may explain evidence of expression in genes that are predicted to be deleted, these results strongly support that a large fraction of our predictions are indeed correct and are also reflected in terms of gene expression levels. To test, whether also partial deletions result in loss of expression, we checked if deletions that affect at least half of a given gene showed similar levels of lack of expression and we found in both strains a significant increase of genes without expression ($P < 10^{-6}$, Fisher's exact test).

3.2.6 Duplicated genes are not correlated with increased expression

Along the same line, we also wanted to test the effect of gene duplication on expression levels. Naively, we would expect two patterns, a higher fraction of genes with expression evidence (FPKM > 0) and a trend towards doubled gene dosage in the strain carrying the duplication. In contrast to deletion, predicted duplicated genes do not show any trend towards an increase in gene expression. Comparing the proportion of duplicated genes between the reference strain and both natural isolates, we did not find a significantly reduced fraction of genes with no expression in both natural isolates. In order to test for differences in genes that are indeed expressed in the strain carrying

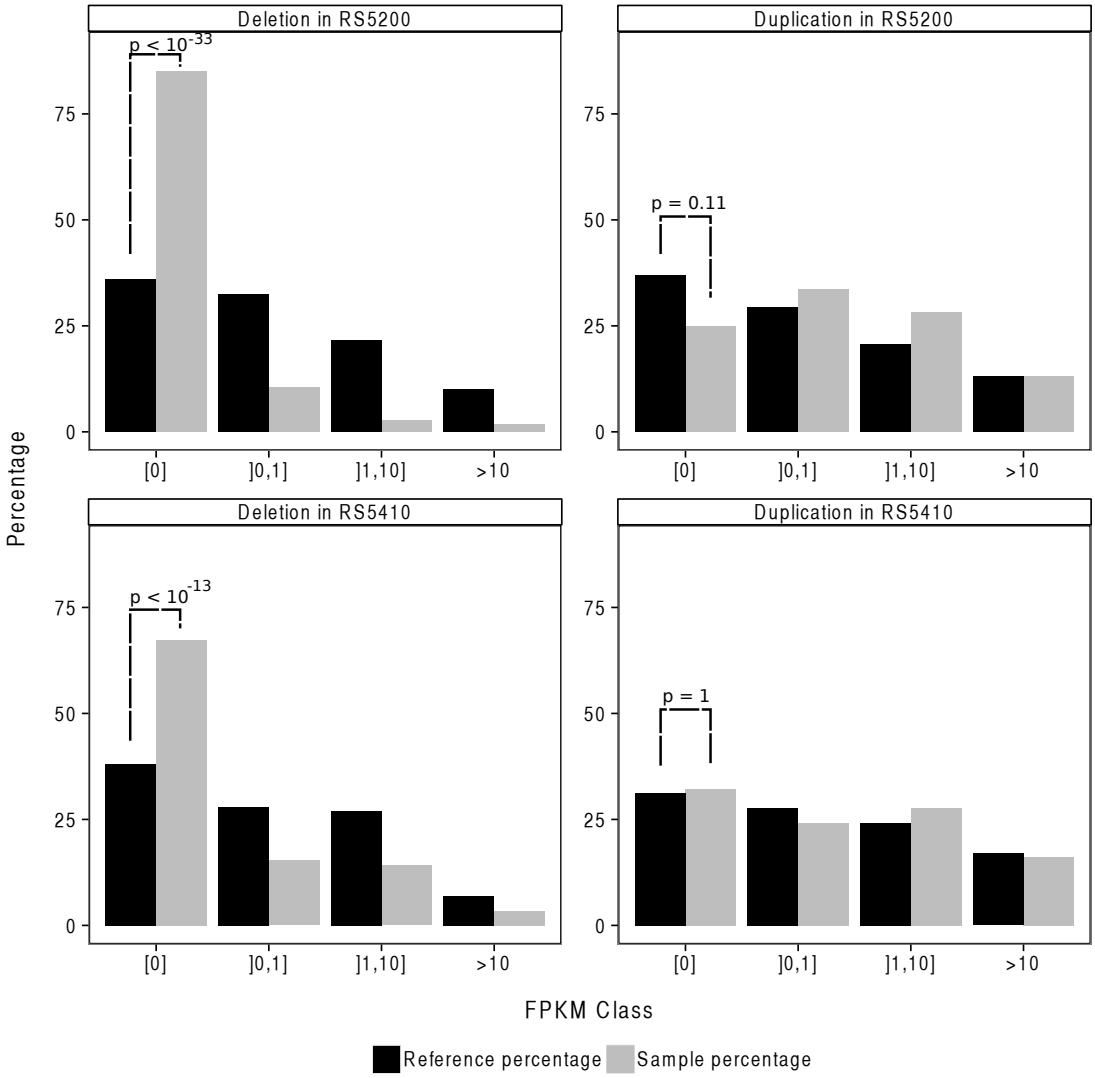


Figure 3.7: **Effect of duplications and deletions on gene expression levels** Distribution of genes predicted to be deleted and duplication in different expression classes.The black bar and the gray bar represents proportion of SV affected in genes in reference strain and the strain of interest respectively. Here we restricted the analysis to only those genes within perfect collinear regions between the reference genome and the sister species. While deletions lead to a strong increase in genes without expression, i.e. more than 80% (RS5200) and 65% (RS5410) of genes in predicted deletions show indeed no evidence of expression. In contrast, we do not see the opposite trend for duplicated genes.

the duplication as well as in the reference strain, we compared the fold changes of expressed genes (FPKM>0). We found that duplication does not lead towards increased fold change ($P = 0.35$ for RS5410 and $P = 0.39$ for RS5200, Wilcoxon-test).

This finding can be explained by the following three scenarios: i) The duplication does not encompass the whole gene including regulatory regions or the duplicated copy was inserted into a region where it is transcriptionally silent, ii) the higher dosage of the duplicated gene was compensated by mutations in cis-regulatory regions, iii) the duplicated gene is part of a highly self-regulatory network, that feeds back into its components and can thus compensate for the expression of the additional copy.

3.2.7 Biased expression of duplicated genes

In order to decide, which of the above-mentioned scenarios most likely explains our data, we first tested, whether both copies are indeed expressed. If indeed, both copies were expressed at similar levels, this would indicate that the expression level of both copies must somehow be downregulated by one of the two compensatory mechanisms (cis-regulatory mutations (ii) or feedback loops (iii)). To this end, we searched for segregating sites that potentially distinguish between the two copies. Such sites were identified previously as they appeared as apparent heterozygous variants in whole genome sequencing data [Rödelsperger et al., 2014]. *P. pacificus* is a diploid species, which does not rule out the possibility of truly heterozygous sites. However, all analyzed strains were inbred for at least ten generations [Rödelsperger et al., 2014] indicating that most of the apparent heterozygous variants derive from sites that segregate between copies of recently duplicated regions and for which resequencing reads are aligned to a single position in the reference genome assembly. Please note, that we do not have experimental proof that these sites are truly segregating between the two copies, but this interpretation is strongly supported by previous findings, that apparent heterozygous sites are clustered in regions of extremely high coverage, that their presence does not decrease by the degree of inbreeding, and that admixed strains do not show elevated levels of heterozygosity [Rödelsperger et al., 2014].

To test for the expression of both copies, we compared the ratios of segregating sites in the genomic resequencing and the transcriptome data. While we could only identify 45 informative sites in 21 genes with $\geq$ 10X coverage, in the genomic and transcriptome data for RS5410, we identified 199 sites in 18 genes, fulfilling these criteria for the RS5200 dataset. In the following, we will use the term 'alleles' to denote two non-identical nucleotides that distinguish the two copies. Figure 3.8 shows the distribution of minor allele frequencies for the genomic and transcriptomic data. In twelve cases, the difference between genomic and transcriptomic data proved to be statistically significant (P ¡ 0.05, Wilcoxon test with FDR correction). All these cases showed a bias of the minor allele not to be expressed. Similarly, in the majority of all tested genes, the same trend is apparent indicating that the new copy is either not complete or was inserted into a region where it is transcriptionally silent. We interpret the fact that few genes showed expression of both alleles as evidence that at least in some cases we were able to detect duplication events that produce a second functional copy. Intuitively, it is more likely that a deletion will have a functional effect on a gene rather than a duplication because as we see, a duplication event must encompass the whole gene including promoter sequence and the insertion site must be in the right genomic context. To rule out that the lack of the expected upregulation is just an effect of partial duplications, we generated a set of predicted tandem duplications using the software pindel [Ye et al., 2009], which allows identifying structural variations at nucleotide resolution. However, based on 152 tandemly

duplicated genes in RS5200 and 80 genes RS5410, we could neither detect a significant trend for more genes with expression evidence nor for higher fold changes when compared to unaffected genes.

Finally, we would like to point out that the lack of expression evidence for the second copy is not an ultimate proof for its non-functionality. After duplication, the fate of the new genes could be pseudogenization, neofunctionalization, or subfunctionalization [Ohno, 1970]. Gene expression domains can be highly stage-specific and in worms even cell-specific. Thus many functional genes with highly restricted expression patterns, which might be a result of subfunctionalization, are unlikely to be detected in RNA-seq data of pooled worms from various stages [Ragsdale et al., 2013]. Furthermore, given the overall lower expression level of genes that were affected by SVs (Figure 3.6), it might well be that expression level differences of just one order of magnitude would result in the lack of evidence for the second copy.

3.2.8 Evidence for negative selection and conservation of synteny

We next assessed the extent to which SVs affect gene classes as defined by different homology relationships with other nematodes (Figure 3.9). When compared to the overall distribution of gene classes, we find a strong depletion of deletions among single copy genes (one-to-one orthologs with *C. elegans* and genes with one-to-many relationships). On the contrary, multi-copy genes (many-to-X category, and orphans with paralogs) are found as being significantly enriched in SVs (Figure 3.9). The almost absence of one-to-one orthologs with *C. elegans* among SVs is expected, as the fact that these genes remained as single copies since the separation from their common ancestor, suggests a strong dosage sensitivity and consequently negative selection against SVs. Dosage sensitivity was proposed previously to cause selection against copy-number variations in the human genome [Schuster-Böckler et al., 2010]. Given that a depletion of SVs among certain genes has already been observed in the human genome [Schuster-Böckler et al., 2010] and that we would expect a strong signature of negative selection based on previous population genomic analysis [Rödelsperger et al., 2014], this finding strongly supports the validity of a large part of our SV predictions. Interestingly, Figure 3.9 shows that one-to-one orthologs are not only are depleted among deletions but also among duplications. Under the presumption, that most of the duplications are not functional, this raises the question why they are nevertheless selected against. We interpret this finding as indirect evidence for selection to preserve synteny as it was shown that duplications tend to be local [Katju and Lynch, 2003], leading to the possibility that insertion of duplicated sequences may interfere with long-range regulatory interactions or disrupt operon-like gene structures.

The absence of SVs in certain genomic regions could be explained if these genes were located in regions of peculiar properties such as low recombination rate. Since the available genetic linkage map of *P. pacificus* [Srinivasan et al., 2003] does not provide sufficient resolution to test this, we use nucleotide diversity as an indirect measure of recombination. We have previously shown that diversity is reduced in gene dense regions [Rödelsperger et al., 2014], which is compatible with a model of background selection, i.e. in regions with low recombination frequencies, neutral variation will be selected against if it is linked to deleterious sites, leading to a reduction of diversity. Comparing the presence of SVs with nucleotide diversity in the most deeply sampled clade from Rödelsperger et al. [Rödelsperger et al., 2014], we find that 100kb windows without SVs show a strong reduction in nucleotide diversity in all comparisons. More precisely, median diversity in regions without SVs is ten-fold lower than in regions with SVs, $P < 10^{-16}$, Wilcoxon rank-sum

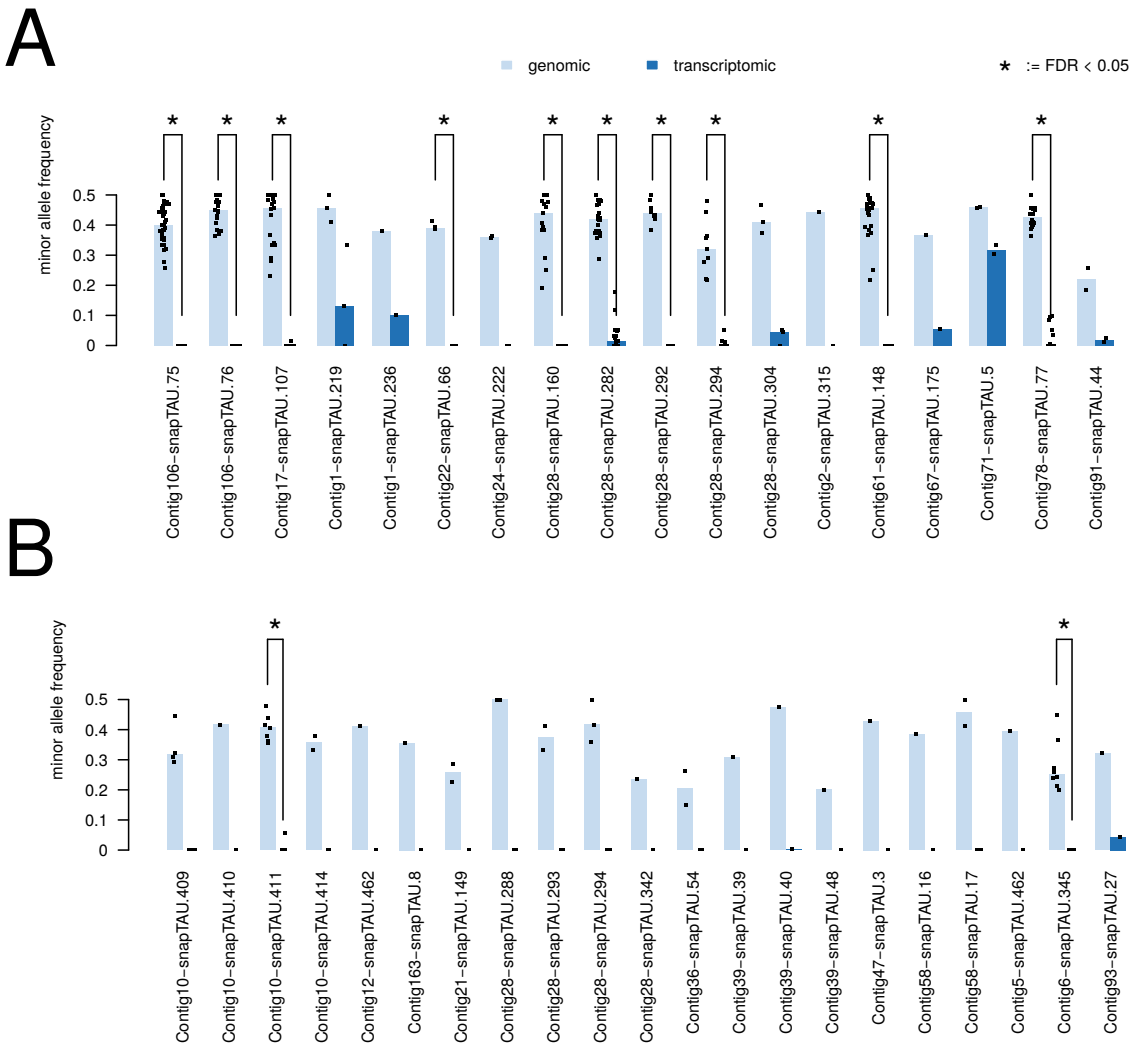


Figure 3.8: **Biased expression in the second copy of duplicated genes.** The plots show the frequency of reads carrying non-reference bases in the genomic and transcriptomic alignment. Here we only considered duplicated genes with putative segregating mutation between duplicates and with $\geq 10X$ coverage. The Panel A and B represent duplicated genes in RS5200 and RS5410 respectively. The Bar represents the median frequency of all informative sites per gene and dots represents frequency for each informative site. Among the analysed genes, most of the genes shows a strong difference in the frequency of genomic and transcriptomic data, which indicates that in these cases one of the duplicates is not expressed.

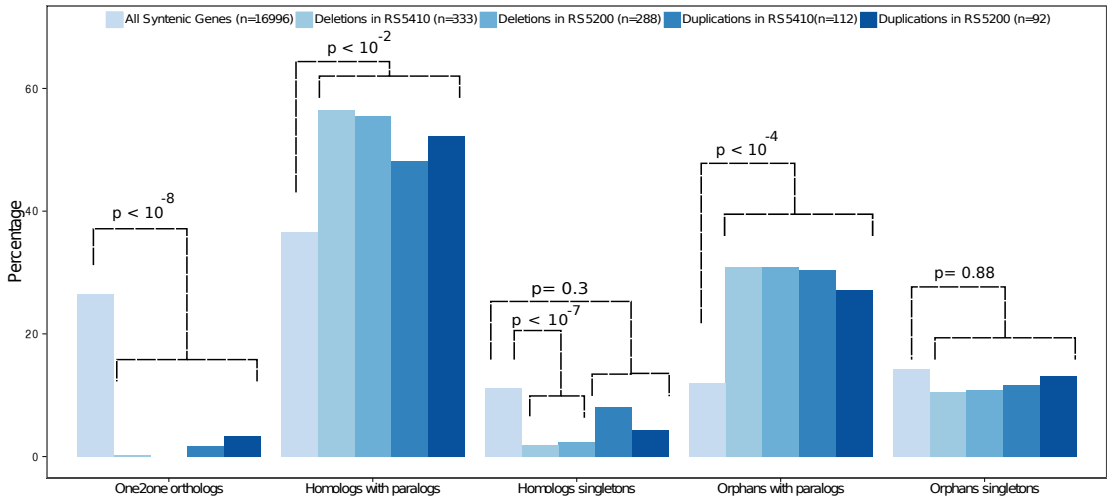


Figure 3.9: **Depletion of SVs among one-to-one othrologs.** The graph shows the distribution of deleted and duplicated genes in different gene classes. The genes were classified based on the presence of protein domain and homology relationship with other nematodes including *C. elegans*. We find strong depletion of both deleted and duplicated genes among one-to-one orthologs (one-to-one orthologs between *P. pacificus* and *C. elegans*). This suggests the action of purifying selection on the SV affected genes.

test). This further supports the action of negative selection in these regions and suggests that even neutral SVs might be removed from populations due to background selection.

3.3 Duplications of developmentally regulated genes in *Pristionchus pacificus*

3.3.1 Summary

The contrasting pattern of gene duplication at the macroevolution (cross-species comparison) and microevolution (within-species comparison) levels suggests different selection forces to dominate genomic signatures at different evolutionary time scales. Given the contrasting patterns between micro and macroevolution and that combining patterns of gene duplications with expression levels was quite insightful, we decided to the combine transcriptomic data from different developmental stages of *P. pacificus* with the duplication patterns that are found between *P. pacificus* and other nematodes including *C. elegans*. In this chapter, we compared the transcriptomes of different developmental stages of *P. pacificus* to investigate the developmental regulation and its association with gene duplication. Comparison of transcriptomes indicates that samples can be grouped into three stages, adult, dauer and early larval stage and more than 5000 *P. pacificus* genes are developmentally regulated. Additionally, we also found different gene families such as Heat shock proteins, collagens were enriched in among stage-specific expressed genes. To test the degree of conservation in developmental regulation, we performed a detailed phylogenetic analysis of, HSP20, and HSP70

using protein sequences from *P. pacificus* and *C. elegans* . This analysis suggests that developmentally regulated HSP genes have undergone lineage-specific duplications potentially resulting from selection to increase gene dosage in a developmental stage-specific manner.

3.3.2 Distinct transcriptome profiles of early larvae, dauer and adults

In order to investigate the developmental regulation of *P. pacificus* , we sequenced RNA-Seq libraries from 5 different developmental stages that comprise larval stages (J2 and J3), dauers and adults (J4 and adults). Using the Illumina platform for sequencing, we obtained between 13 million and 17 million reads per library and gene expression levels were estimated as fragments per kilobase per transcripts per million reads sequenced (FPKM). Principal component analysis was performed on the estimated gene expression levels, to get further insight into transcriptomes of the different developmental stages (Figure 3.10B). The result shows that 56% and 25% of global variability can be explained by the first two principal components, respectively. Another interesting observation from this analysis is the clustering of different samples into three distinct developmental stages: 1) early larval stage, 2) dauer larvae, and 3) late larvae and adult worms. Samples with the labels J2 and J3 were grouped in the early stage, while the adult stage contains a mix of samples that were labeled as J3, J4, and adult worms. The most parsimonious explanation for this observation is that our staging protocol resulted in an imperfect synchronization of worm cultures. The protocol we used was supposed to eliminate all hatched worms, but in our case it retains embryos, J1 and J2 larvae in *P. pacificus* . We attribute this to the inherent difference in the developmental stages of *C. elegans* and *P. pacificus* . In *P. pacificus* , the larvae hatch during the J2 stage, because the J1 stage is an embryonic molt within the egg. However, in *C. elegans* , the hatching takes places during the L1 (corresponding to the J1 in *P. pacificus*) stage [Hong and Sommer, 2006].

To support and validate the classification of transcriptome into three groups, we used three independent approaches. First, we calculated the number of differentially expressed genes in pairwise comparisons (Figure 3.10A) and used this number to cluster the samples hierarchically based on Euclidean distances (Figure 3.10C). The hierarchical clustering of samples also supports the classification based on PCA. Second, we ordered all transcriptomes using a PCA-based approach implemented in the software BLIND [Anavy et al., 2014] and found that the transcriptomes labeled as J4_1 and A1 were predicted as being from later stages than J3_1, J4_2 and A2. Finally, the classification of transcriptomes was further supported by the qRT-PCR experiments of six candidate genes (Appendix figure 6.6). All the three approaches suggest that the classification of developmental transcriptome into three groups is indeed correct.

3.3.3 Clusters of co-regulated /developmentally regulated genes

To define clusters of developmentally regulated genes, we used an unsupervised biclustering approach, as implemented in the software SAMBA [Tanay et al., 2002]. As a biclustering algorithm, SAMBA performs two-mode-way clustering, which performs simultaneous clustering of the genes and samples of an expression matrix to identify subsets of genes with similar expression profiles across a subset of samples. Using this approach and data of significant differential expression, we identified 29 partially overlapping biclusters (Figure figPpaDevTransD-F and Appendix figure 6.5). The number of genes in each bicluster varied from 50 to more than 1000 and in total the biclusters contains 5161 (17%) of predicted *P. pacificus* genes. The genes in the biclusters show distinct expression profiles across the stages. The distribution of expression levels of three biclusters was

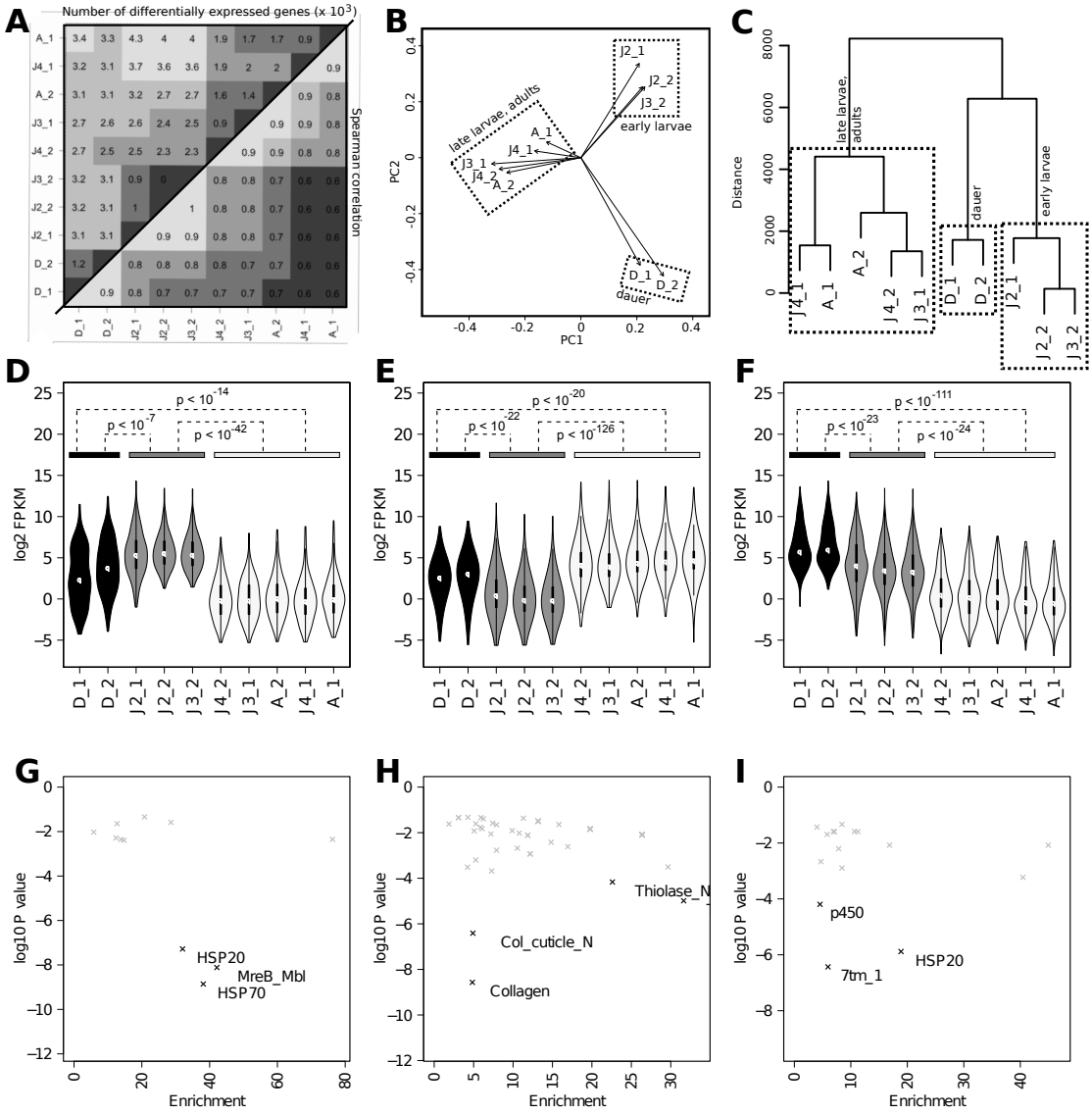


Figure 3.10: **Comparison and clustering of developmental transcriptome.** A) Spearman correlation of expression values(lower triangle) and number of differentially expressed genes (upper trianlge, number of genes in 10^3) across all pairwise comparison. B) Distribution of developmental transcriptome in the first two principal component(PC1 and PC2), indicates the samples are grouped into three clusters. C) Hierarchical clustering of transcriptomes based on the pairwise comparisons of all samples using Cuffdiff. D-F) Violin plot of expression values in all samples for genes clustered in bicluster 4 (panel D), 12 (E) and 24 (F) respectively. Color code indicates the rough grouping of samples into three developmental stages. The statistical significance of expression differences across stages is shown as the maximum p-value (Wilcoxon test) between any pairwise comparisons of samples across stages. G-I) Enrichment of Pfam domains in bicluster 4 (panelG), 12 (H) and 24 (I) respectively. The plots show the enrichment factor vs. the significance ($log10P$). The most significant protein families are highlighted in the individual plots.

shown in (Figure 3.10D-F), with genes in Bicluster 4, 12 and 24 exhibits higher expression levels in early larvae, dauer and adult samples respectively. Given the substantial expression variation exhibited by all of the 29 bicluster genes across 10 transcriptomes, we treated the bicluster genes as developmentally regulated genes.

3.3.4 Distinct gene families are overrepresented in stage-specific expression biclusters

Different gene families have been shown to play a vital role in developmental regulations in different organisms [Lander et al., 2010, Cao et al., 2013]. In order to check the contribution of different gene families to the developmental regulation in *P. pacificus* , we performed an overrepresentation analysis of predicted protein domains among developmentally regulated gene sets. The result of domain enrichment analysis shows that distinct gene families are enriched in biclusters. For example, gene families such as actin-like MreB proteins (PF06723), and heat shock proteins HSP20 (PF00011) and HSP70 (PF00012) were strongly enriched in Bicluster 4, which exhibits higher expression levels at the early larval samples (Figure 3.10F). In contrast, Thiolase and collagens were overrepresented in adult specific bicluster 12 (Figure 3.10H). In case of bicluster 24, which shows highest expression in dauer samples, gene families such as Cytochrome P540 (PF00067) and G-protein-coupled receptors (PF00108) were overrepresented. Interestingly, we found HSP20 proteins overrepresented in bicluster 4 also enriched in bicluster 24 (Figure 3.10I). This suggests the different members of a same gene family with divergent roles throughout the larval development of *P. pacificus* .

3.3.5 Evidence for lineage-specific duplication among developmentally regulated genes in HSP gene families

Next, we wanted to test whether the developmental regulation is conserved across species. One way to test the conservation is to check whether developmentally regulated genes have one-to-one orthologs in *C. elegans* . For this purpose, we reconstructed the phylogenetic trees of two gene families HSP70 and HSP20 using the protein sequences for *C. elegans* and *P. pacificus* (Figure 3.11 and 3.12). Members of both gene families show evidence for developmental regulation. Interestingly, in both families, we found that developmentally regulated genes have no one-to-one orthologs in *C. elegans* . In contrast, developmentally regulated genes are clustered together as subtrees in both gene families. The clustering of developmentally regulated genes indicates that they are paralogs and have generated by the gene duplication events in the *Pristionchus* lineage. The observed pattern lead us to hypothesize that the developmental regulation predates gene duplication (i.e the common ancestor of these paralogs was already developmentally regulated, before duplication events which give rise to the observed paralogs). We tested this by checking the expression profiles of members of individual subtrees in both HSP families (Figure 3.11B and 3.11C). We found that the paralogs have similar expression profiles and strong consistency across the ten transcriptomes. For example, all the paralogous genes in Figure 3.11C show higher expression in early larvae samples and strong consensus across all samples.

As discussed in the previous subsection, HSP20 family members were enriched in different biclusters with distinct expression profiles. This is further evident in the phylogenetic analysis of HSP20 family. We found distinct paralog groups of HSP20 having different expression profiles, where one group showed higher expression in dauer (Figure 3.12B) and another group in early larvae

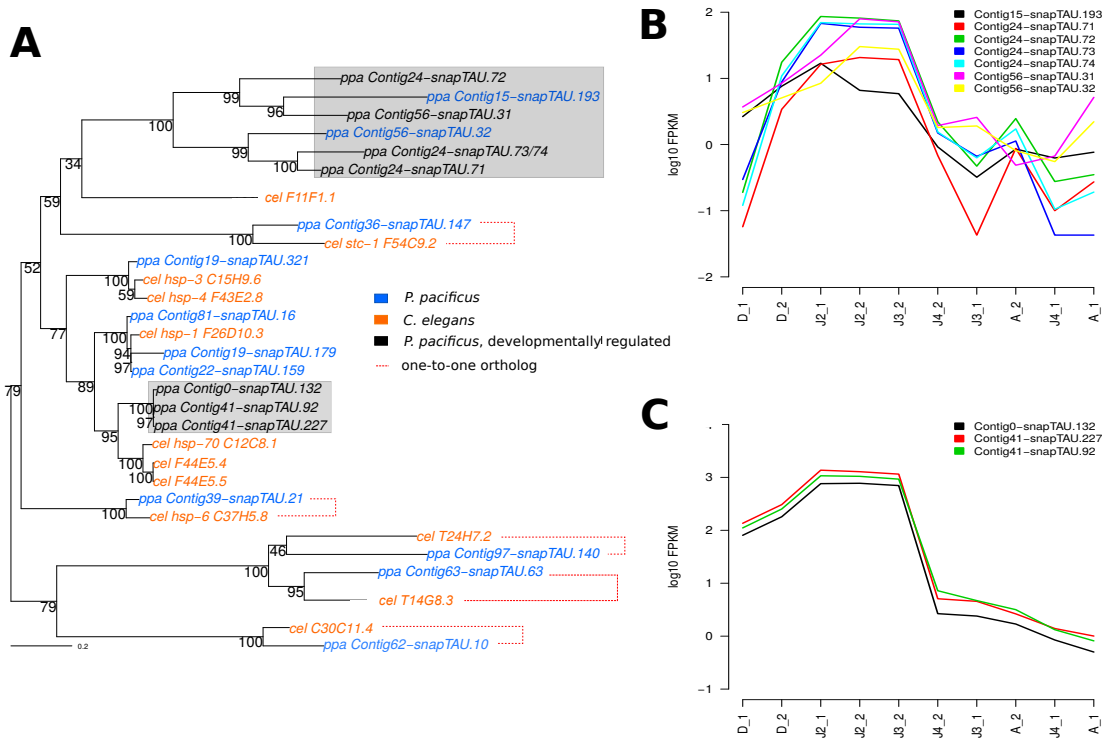


Figure 3.11: **Duplication and developmental regulation in the HSP70 gene family.** A) Phylogenetic trees of HSP70 reconstructed using the protein sequences from *C. elegans* and *P. pacificus*. The tree shows that clustering of developmentally regulated genes of *P. pacificus* are cluster together (gray box), indicating lineage-specific duplication. Panel B and C shows the expression value of developmentally regulated gene clusters in different transcriptomes. The developmentally regulated genes in HSP70 family shows high expression in early larvae samples

(Figure 3.12C). We also found a group with more divergent expression profiles among its members (Figure 3.12D). The pattern of divergent expression profiles exhibited by members of the same gene family can be explained by the mechanism of subfunctionalization, which indicates partition of ancestral properties among family members following gene duplication. However more experimental analysis is needed to allow a more robust investigation of this potential subfunctionalization, as the ancestral expression pattern is not known.

3.3.6 Majority of the developmentally regulated genes are duplicated genes

Detailed phylogenetic analysis of HSP families shows that developmentally regulated genes were product of gene duplication events. Together with expression data, we proposed that duplication is favored by selection to increase dosage of already developmentally regulated genes. In order to test whether the observed pattern represents a general trend in the evolution of developmental regulation in *P. pacificus*, we hypothesized that the developmentally regulated genes should be enriched with paralogous *P. pacificus* genes. To this end, we classified the *P. pacificus* genes into five homologous classes with respect to *C. elegans* and other related nematodes (See methods).

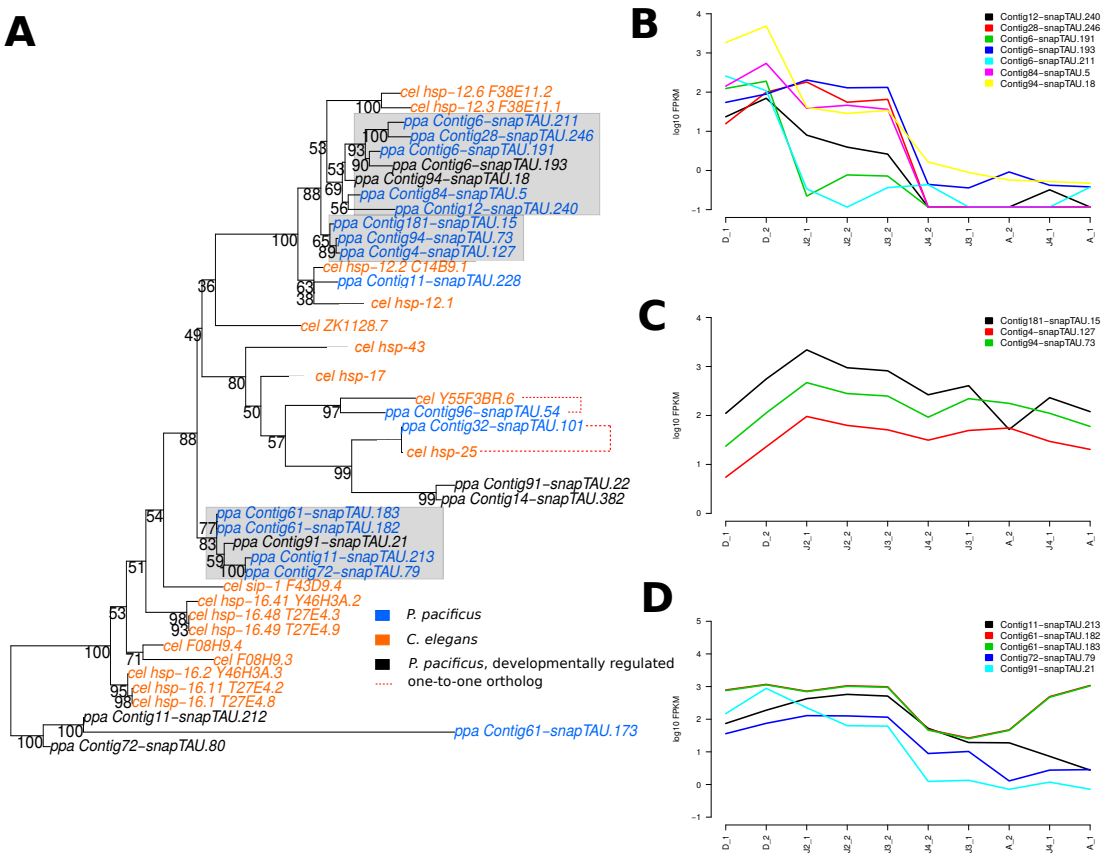


Figure 3.12:]

Duplication and developmental regulation in the HSP20 gene family. A) Phylogenetic trees of HSP20 reconstructed using the protein sequences from *C. elegans* and *P. pacificus*. B-D) Expression profiles of genes in paralogous groups.

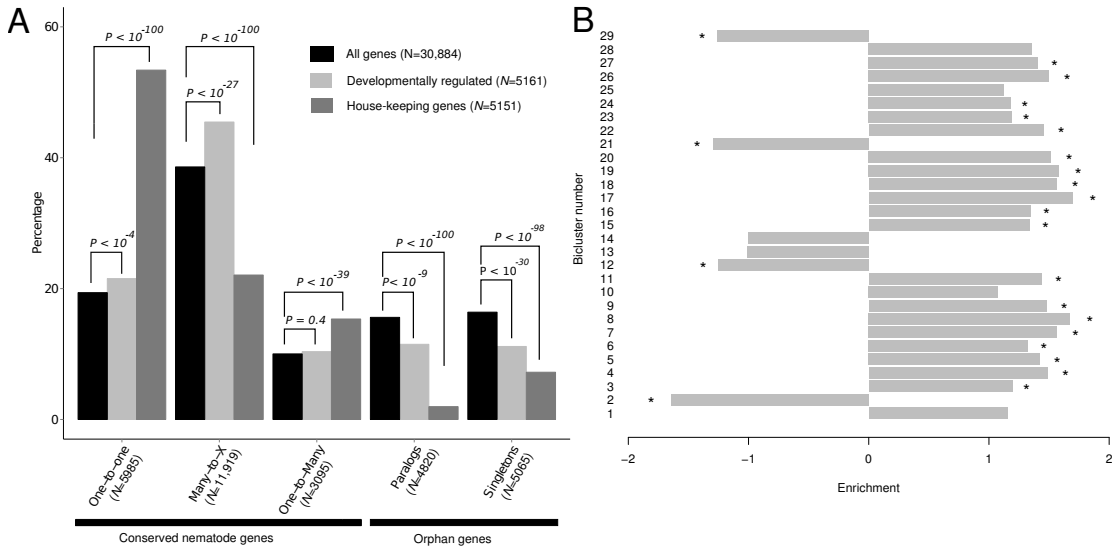


Figure 3.13: **Enrichment of duplicated genes among developmentally regulated genes** A) The proportion of all genes, developmentally regulated genes and house-keeping genes in different homology classes.The classes were defined by orthology relationship and expression pattern. Developmentally regulated genes (union of all bicluster genes) shows strong enrichment in conserved multicopy genes. B) Enrichment of multicopy genes among individual biclusters. Out of 29 biclusters, 19 show a significant enrichment in genes with paralogs.

Among other homologous classes, conserved multicopy genes (Many-to-X) represents *P. pacificus* genes with multiple paralogs within species, as well as homologs in other nematodes. This class includes genes with homologous relationship such as many-to-many, many-to-one, and many-to-zero with *C. elegans* and other nematodes (see methods). Next, we tested the enrichment of different homologous gene classes in three different datasets, entire *P. pacificus* gene set (All genes), developmentally regulated genes and housekeeping genes. As expected, conserved multicopy genes were strongly enriched in the developmentally regulated genes in accordance with our hypothesis (Figure 3.13A). More precisely, we found more than 50% of developmentally regulated genes to belong to the conserved multicopy gene class and it is significantly more than when compared with entire *P. pacificus* data set and housekeeping genes ($P < 10^{-27}$). While in other gene classes except for one-to-one orthologs, either all or housekeeping genes have a higher proportion than developmentally regulated gene sets. Even though multicopy genes dominate developmental regulation, we found a weaker but still significant enrichment of one-to-one orthologs (Figure 3.13A). This suggests that developmental regulation to a certain degree is conserved between *P. pacificus* and *C. elegans* . When individual biclusters were tested for the enrichment of homologous classes, we found the conserved multicopy genes to be significantly enriched in 19 out of 29 biclusters (Figure 3.13B). Taken together, these results indicate that developmentally regulated genes were subjected to lineage-specific duplications repeatedly within the *Pristionchus* lineage and that the developmental transcriptome is shaped by such ancient gene duplication events.

3.3.7 Functional characterization of developmental regulated gene clusters

We characterized the resulting gene sets by performing a Gene Ontology (GO) analysis based on *C. elegans* one-to-one orthologs using the David functional annotation webtool [Huang et al., 2009]. The dauer-specific bicluster 24 was most significantly enriched for G-protein coupled receptor protein signaling pathway (GO:0007186, 10-fold enriched, $P < 10^{-15}$) and neuropeptide signaling pathway (GO:0007218, 30-fold enriched, $P < 10^{-12}$). Other biclusters showed strong enrichment in biological processes such as molting cycle (GO:0042303), cell projection organization (GO:0030030), hedgehog receptor activity (GO:0008158), and chitin metabolic process (GO:0006030). In contrast, housekeeping genes only showed a strong overrepresentation of ribosomal proteins (Appendix table 6.1).

3.3.8 Comparison with previous expression-profiling studies

We compared our data set with three previous gene expression-profiling studies on *P. pacificus*. Appendix table 6.2 shows a summary of all biclusters of developmentally regulated genes, which showed a significant overlap ($P < 0.01$) with any of the previously identified gene sets [Sinha et al., 2012b, Rae et al., 2012, Sinha et al., 2012a]. When compared to the dauer vs. dauer exit experiment, five out of the six biclusters that show most significant enrichment with genes upregulated in dauers vs. dauer exit worms, also show trends for higher expression in dauers vs. adult worms and late larvae (Appendix figure 6.5). Similarly, the six biclusters that show most significant enrichment with genes downregulated in dauers vs. dauer exit worms also show trends for lower expression in dauers vs. adult worms and late larvae (Appendix figure 6.5). Although the data sets are not fully comparable because dauer-exit worms are not equivalent to adult worms, these findings further support that our expression measures based on RNA-seq experiments are largely robust and reproducible when compared to expression data obtained from microarrays [Sinha et al., 2012b].

In a previous study, Rae et al, [Rae et al., 2012] found that germline ablations in *P. pacificus* lead to increased longevity. In comparison with germline-ablated worms, the most significant association was a four-fold enrichment of house-keeping genes in genes that are downregulated upon germline ablation ($P < 10^{-300}$). We interpret this finding as evidence that general metabolic processes are slowed down in germline-ablated animals. On the contrary, most developmentally regulated clusters were found to be significantly depleted among genes that are downregulated upon germline ablation.

Next, we compared the identified gene sets with the transcriptional response of *P. pacificus* worms to four different pathogens (Xenorhabdus nematophila, Serratia marcescens, Staphylococcus aureus, Bacillus thuringiensis) [Sinha et al., 2012a]. Again, the most significant association was that housekeeping genes were significantly enriched in genes that are downregulated upon exposure to Xenorhabdus nematophila and Serratia marcescens. As these two pathogens kill most of *P. pacificus* worms within four days [Sinha et al., 2012a], we interpret these overlaps as a result of pathogenicity-associated necrosis, which leads to a breakdown of housekeeping functions in dying cells.

3.4 Comparative analysis of duplicated genes in three closely related *Pristionchus* species

The comparison of transcriptomic and genomic data from different *P. pacificus* strains and the comparison of selected gene families and the developmental transcriptome of *P. pacificus* with distantly related nematodes such as *C. elegans* reveal quite contradictory trends. More precisely, while at a microevolutionary level, duplications seem to be selected against or appear to be non-functional (the second copy is not expressed), the comparison with distantly related nematodes revealed large gene expansions that are stable over hundreds of million years. To resolve this apparently contradictory picture, we chose to investigate the patterns of duplication at an intermediate time scale. We therfore chose to to perform a detailed genome-wide comparative analysis of three closely related *Pristionchus* species; *P. pacificus*, *P. exspectatus*, and *P. arcanus* [Kanzaki et al., 2012]. A draft genome of *P. exspectatus* has been published previously [Rödelsperger et al., 2014], the draft genome of *P. arcanus* has been sequenced and assembled using similar methodologies (Prabh et al. unpublished data). We used *P. arcanus* as outgroup species to correctly interpret the change in gene expression levels in *P. pacificus* and *P. exspectatus*. As a first step, we identified genes belonging to different homology classes to distinguish conserved genes from orphan genes. Using the predicted protein coding genes of three species and OrthoMCL clustering algorithm, we identified 7958 one-to-one orthologs between 3 species (Figure 3.14). Additionally, we also identified 2372 gene clusters with single duplication events in one species and no duplication in other species $(1-1-2)$. More specifically, we found 232 and 1205 single duplication events in *P. pacificus* and *P. exspectatus*, respectively (The remaining 935 are specific to the outgroup species *P. arcanus*). This suggests that *P. exspectatus* genome has undergone more recent duplication events after the separation from *P. pacificus*. Other larger clusters, which represent different patterns of duplication, are heterogeneous and the number of clusters for each pattern is also quite small. As the small sample size of these clusters does not allow to infer statistically robust results, we did not investigate the effect of duplication on sequence and expression evolution in these heterogenous clusters any further.

3.4.1 Strong correlation in expression between genes in one-to-one orthologs

Given that a large number of genes are maintained as one-to-one orthologs, we wanted to test whether the corresponding expression levels are also comparable. To this end, we performed pairwise comparison of expression levels (FPKM) of genes in one-to-one orthologous clusters and calculated the correlation in expression levels across species. We found a strong positive correlation ($\rho \geq 0.67$) in expression levels across species among one-to-one orthologs (Figure 3.15). Furthermore, we also estimated the rate of evolution for each cluster by pairwise comparisons of one-to-one orthologs. We discarded clusters with $0.0001 \leq dS$ (synonymous substitution rate) ≥ 2 and $0.0001 \leq dN/dS(\omega) \geq 3$ and used mean omega value for each cluster. The median evolutionary rate of all one-to-one clusters is 0.2, which indicates an overall strong negative selection (Figure 3.17).

3.4.2 Majority of recently duplicated genes does not show increased gene dosage

The strong similarity in sequence and expression levels of one-to-one orthologs indicates strong conservation as a result of negative selection. So, we used the one-to-one orthologs as a baseline

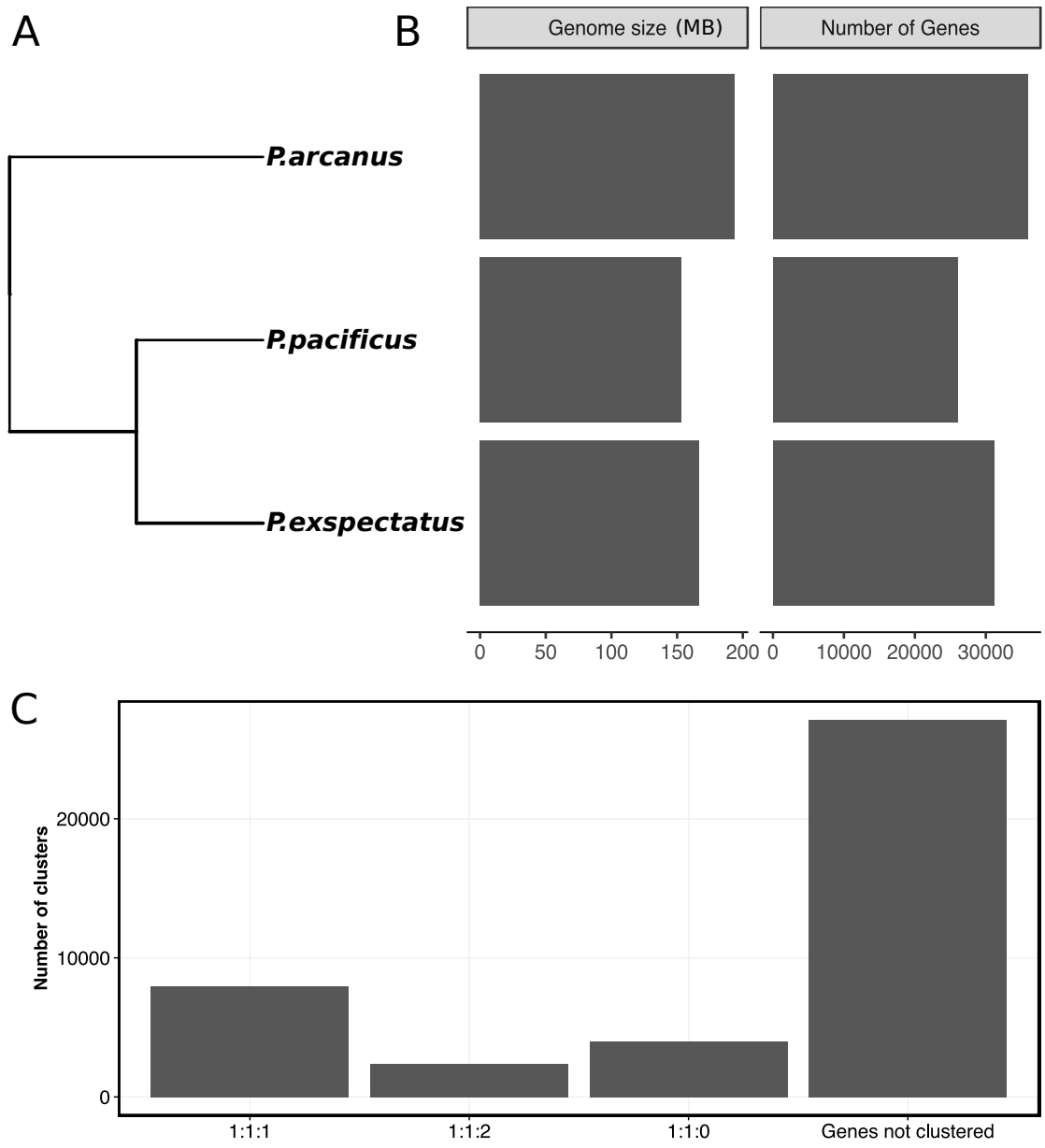


Figure 3.14: .

Schematic phylogeny of three *Pristionchus* species. B) represents the genome size (in MB) and the total number of annotated genes in each species.C) shows the number of one-to-one orthologs, total number of recent duplications, recent loss (1:1:0) and singletons (genes with no homologous sequences in other 2 speceis) in all three species.

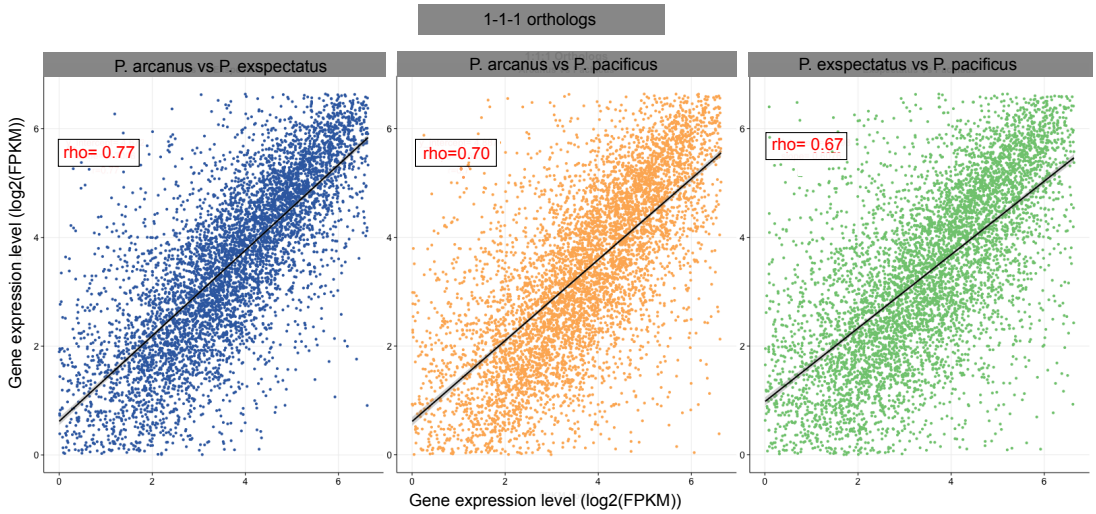


Figure 3.15: **Strong correlation in expression between one-to-one orthologs**. Each panel shows a pairwise comparison of log-transformed expression levels between different species with spearman correlation value (ρ or rho). The black line represents the correlation in expression profiles between species. The graph indicates strong correlation in expression levels across species.

reference for comparison with duplicated genes. Next, we wanted to test the effect of recent duplication on gene dosage, specifically $1-1-2$ orthoMCL cluster. $1-1-2$ cases provided an excellent basis to compare duplicated genes with non-duplicated orthologs in the same orthoMCL cluster. For instance, consider recent duplication resulting in two paralogous genes in *P.pacificus* and no duplication in the other two species. In this case, we can test the whether the increase in gene count leads to increase in gene expression in *P. pacificus* using *P. exspectatus* single copy gene as the baseline. Using this approach, we compared the summed expression levels of duplicated genes in *P. pacificus* with the expression level of its single copy ortholog in *P. exspectatus* (Figure 3.16 A). Surprisingly, we found no difference in the expression level between the duplicated and non-duplicated species ($P > 0.05$, Wilcoxon test) (Figure 3.16). This suggests that most duplication events do not increase gene expression levels, i.e. sum of expression levels of both copies. We further estimated the relative difference in the expression levels between the two duplicated copies to test if both copies are showing reduced expression. This detailed analysis of the expression levels of duplicated genes shows that in a majority of cases, one of the duplicated gene shows either really low or no expression evidence (Figure 3.16 B). These observations are completely consistent with the microevolutionary analysis and indicate that duplications that can be tolerated are most likely the ones, where the second copy is not or lowly expressed, where low expression can indicate expression in a limited number of cells or highly stage-specific expression. Additionally, we also estimated the evolutionary rate as measured in omega and compared the rate of evolution between duplicated and non-duplicated cluster members. More specifically, for each cluster, we estimated two omega values, one from the pairwise comparison of duplicated genes and another from the pairwise comparison of non-duplicated genes. Even though the median rate of evolution of duplicated genes pairs was significantly higher than the evolutionary rate of one-to-one orthologs and the baseline reference,

more than 75% of duplicated pairs have omega values are less than 0.5 (Figure 3.17). This indicates that despite a general relaxed constraint there is still negative selection acting against the change of the protein sequences in duplicated gene sequences. We also estimated branch-specific ω values to check if the duplicated gene with lower expression shows any difference in the sequence evolutionary rate in comparison the duplicated gene with higher expression levels. Even though branch-specific ω values were less than 1, we found no correlation between expression levels and ω values among young duplicates. Adding on the results from the previously mentioned microevolutionary analysis, the evidence of negative selection acting on both gene copies suggests that even though, one of the copies is only lowly or even not expressed at all under standard conditions, it still seems to be functional.

3.5 Duplication of gene families associated with parasitism in *Strongyloides papillosus*

3.5.1 Summary

Comparative genomics approaches provide unique opportunities to understand the evolutionary forces shaping the genomes different species. Using the comparative genomics and draft genome assemblies, Hunt et al [Hunt et al., 2016], investigated the genomic basis of parasitism in the Strongyloidies Spp and identified extreme duplication of Astacin and CAP gene families. In this work, we used comparative genomics approaches and developmental transcriptomics in *S. papillosus* to investigate the developmental regulation and evolutionary forces underlying the large expansion of Astacin and CAP gene families. To understand the developmental regulation, the transcriptomes of different developmental stages and sexes of *S. papillosus* were sequenced and compared with one another. The pairwise comparisons between the transcriptomes show that more than 73 % of predicted protein-coding genes were differentially expressed in at least one comparison. This indicates that majority of genes were developmentally regulated and were maintained as single copies in *S. papillosus* and *S. ratti*. Comparison with *S. ratti* gene expression profiles and protein sequences suggest conservation in developmental regulation at both sequence and expression levels. In accordance with the previous study [Hunt et al., 2016], we found that Astacin and CAP gene families were enriched in parasitic female comparisons. Detailed phylogenetic analysis of both families in combination with expression data shows that single expansion events early in the *Strongyloides* lineage result in most of the members. Interestingly, genes with similar expression profiles were clustered together as subtrees in both families and expression profiles vary between different subtrees in the same family. For example, in case of CAP family, all genes in one subtree show higher expression in infective larvae while genes in another subtree show higher expression in parasitic females. This pattern of divergent expression profiles by the members of the same gene family indicates that they have undergone subfunctionalization. Further, both gene families show more evidence for positive selection relative to genome-wide expectation. Taken together, this study provides the first insight into the developmental regulation of *S. papillosus* and shows that the patterns observed for parasitism associated gene families are shaped by the complex interplay between sequence and expression evolution.

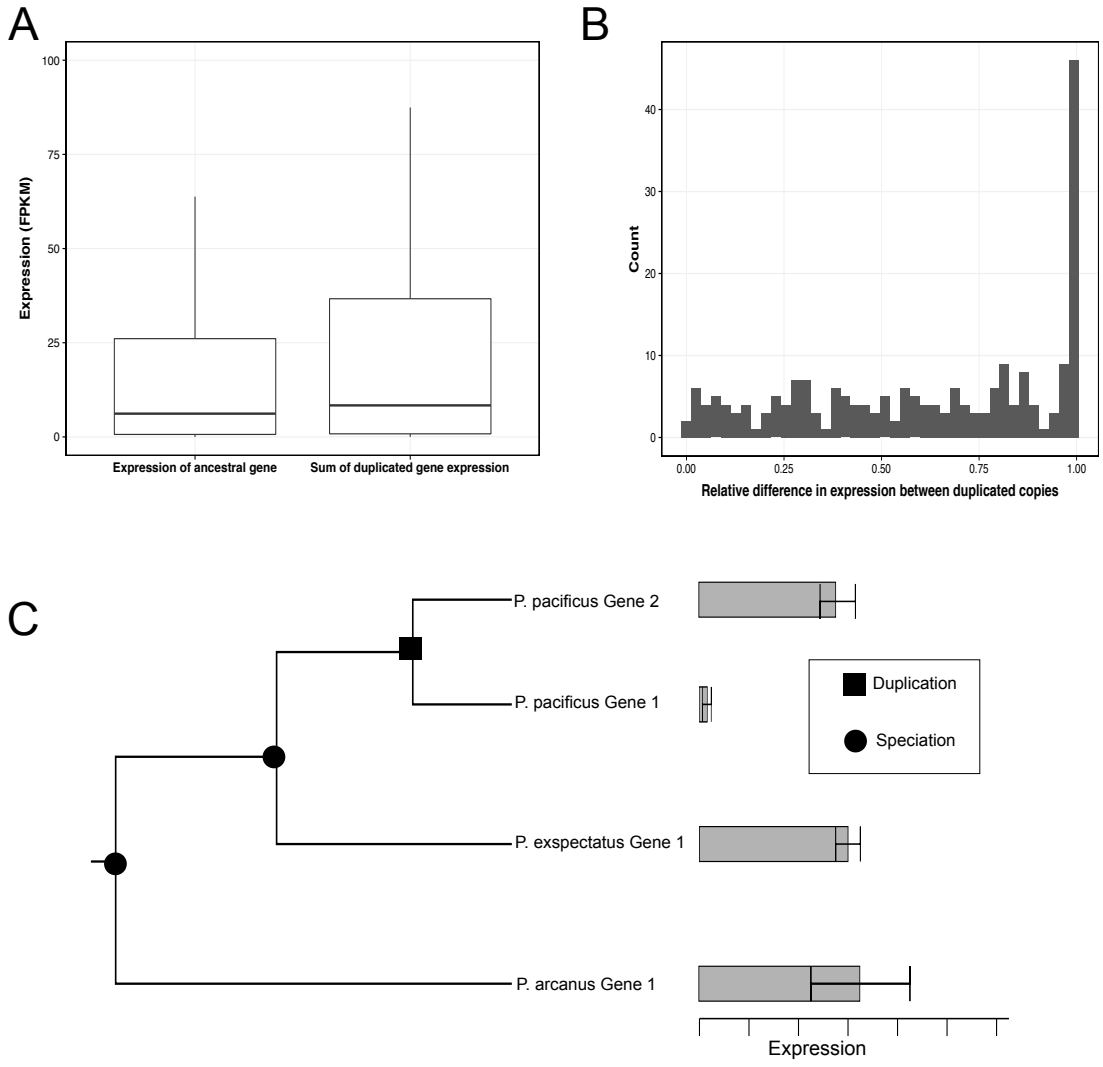


Figure 3.16: **Biased expression of recently duplicated genes in *P. pacificus*.** A) The box plot shows the distribution of median expression levels of ancestral or non-duplicated genes in comparison with the distribution of sum of expression of the duplicated copies. we found no statistically significant difference between the median of both distributions ($P > 0.05$, Wilcoxcon test). B) Histogram shows the distribution of relative difference in expression levels of the duplicated genes with 0 represents similar expression levels and 1 represents large difference in expression levels between the duplicated genes. C) Schematic gene tree of three *Pristionchus* species representing single duplication event in *P. pacificus* and expression values for each gene. We found that in the majority of 1-1-2 cases, one of the duplicated gene have similar expression value compared orthologs in other species and another duplicated gene have no or really low expression levels.

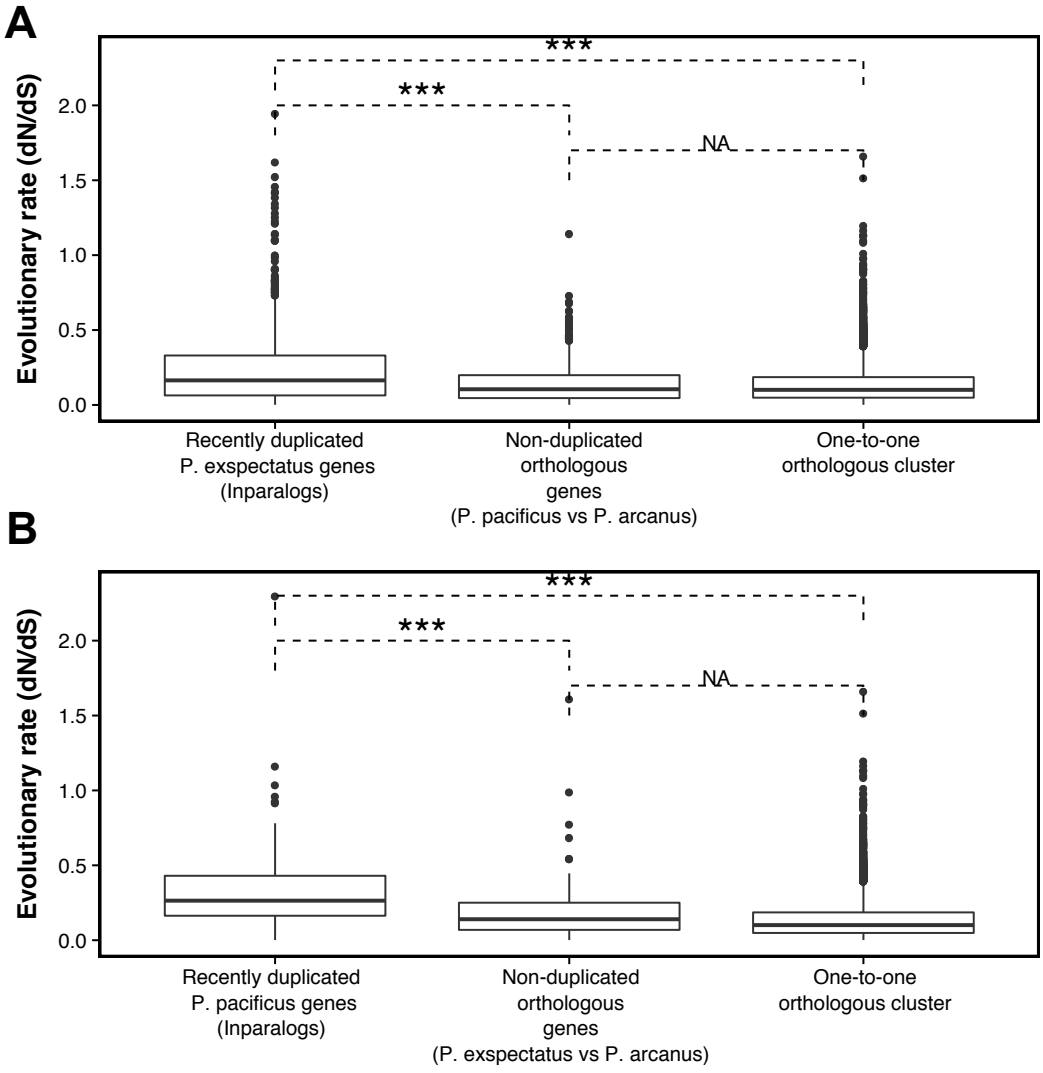


Figure 3.17: **Recent duplicates also show evidence for negative selection on protein sequences**. The graph shows the comparison of sequence evolutionary rate (ω) between duplicated pairs and non-duplicated orthologs (1-1-2 clusters). we also estimated ω value for 1-1-1 cases and used as a baseline. Each box-plot shows median, 25^{th} and 75^{th} percentile ω values. We estimated the statistical significance in median ω values of each group using Kruskal-Wallis test with posthoc nemenyi test. The star symbol indicates statistical significance with $p - value \leq 10^{-5}$ and 'NA' represent $p - value > 0.05$. Panel A and B shows comparisons of *P. exspectatus* and *P. pacificus* recent duplications. The comparison shows that the median ω value of non-duplicated homologs in 1-1-2 clusters is not significantly different from that of 1-1-1 clusters. Even though the median ω value of duplicated genes were generally higher that the non-duplicated homologs, more than 75% of duplicated pairs have $\omega < 0.5$, indicating strong negative selection.

3.5.2 *Strongyloides papillosus* as a model system

In this study, we focus on the nematode *Strongyloides papillosus*, a common gastrointestinal parasite of domestic and wild ruminants [Jäger et al., 2005]. *S. papillosus* has a unique and complex life cycle that consist of parasitic and free-living generations [Streit, 2017, Viney and Lok, 2015]. The adult parasites are only female and live in the mucosa of the small intestine where they lay embryonated eggs that pass with host feces. The parasitic female reproduces parthenogenetically (clonal reproduction). The progenies of parasitic females have two developmental choices: either to develop directly into an all-female infective third stage (iL3) (direct or asexual cycle) or to develop into facultative free-living males and females (indirect or sexual cycle). The free-living generations mate and undergo sexual reproduction in the environment and all their progeny develop into iL3. The life cycle of *S. papillosus* is illustrated in Figure 3.18A. The alternating life cycles between free-living and parasitic generations provide a unique opportunity to apply available genetic and molecular tools for a better understanding of the basic biology and evolution of parasitism in *Strongyloides* [Streit, 2017]. Hunt et al. [Hunt et al., 2016] reported draft genome sequences for four species of *Strongyloides* and transcriptomic comparisons between parasitic and free-living stages for *S. ratti* and S. stercoralis but not *S. papillosus*. To complement the work by Hunt et al. [Hunt et al., 2016]we have sequenced and characterized ten transcriptomes that were sampled throughout the development of *S. papillosus*.

3.5.3 Majority of *S. papillosus* genes are developmentally regulated

We sequenced ten transcriptomes of different developmental stages of *S. papillosus*, which includes, males and females, parasitic and non-parasitic, and larvae and adults (Table 3.1). The transcriptomes provide a unique opportunity to gain insight into the developmental process of *S. papillosus*. Principal component analysis of the estimated expression levels indicates that 55% of variation can be explained by the first two principal components (Figure 3.19A). The ten transcriptomes group into four clusters: young larvae (L1/L2), infective larvae (iL3), adult males and adult females (including parasitic and free-living). This indicates that variation between different developmental stages and sexes is considerably larger than variation between biological replicates (Figure 3.19A). Therefore, we conclude that the sequenced transcriptomes robustly capture gene expression profiles during the development of *S. papillosus*. By performing differential gene expression analysis in a pairwise manner, we found that the number of differentially expressed genes vary widely from 0.2% (parasitic L1/L2 vs. free-living L1/L2) to 45% (iL3 vs. parasitic female stage) (Figure 3.19B). Overall, 73% of *S. papillosus* genes are found as significantly differentially expressed ($FDR - correctedp - value < 0.05$) in at least one comparison. When focusing on the comparison between free-living and parasitic females, we found that only 10% of genes were differentially expressed and 4.4% (917) of genes are up-regulated in parasitic females. This relatively small set of infection-associated genes is consistent with previous data from Hunt et al. [Hunt et al., 2016] , which posited that 909 and 1188 genes were up-regulated in parasitic females in *S. ratti* and S. stercoralis respectively.

3.5.4 Developmental transcriptomes show high degree of conservation across *Strongyloides* species

Previously, we show from the analysis of the developmental transcriptome of *P. pacificus* that the developmentally regulated genes were enriched with multicopy genes arisen from ancient gene

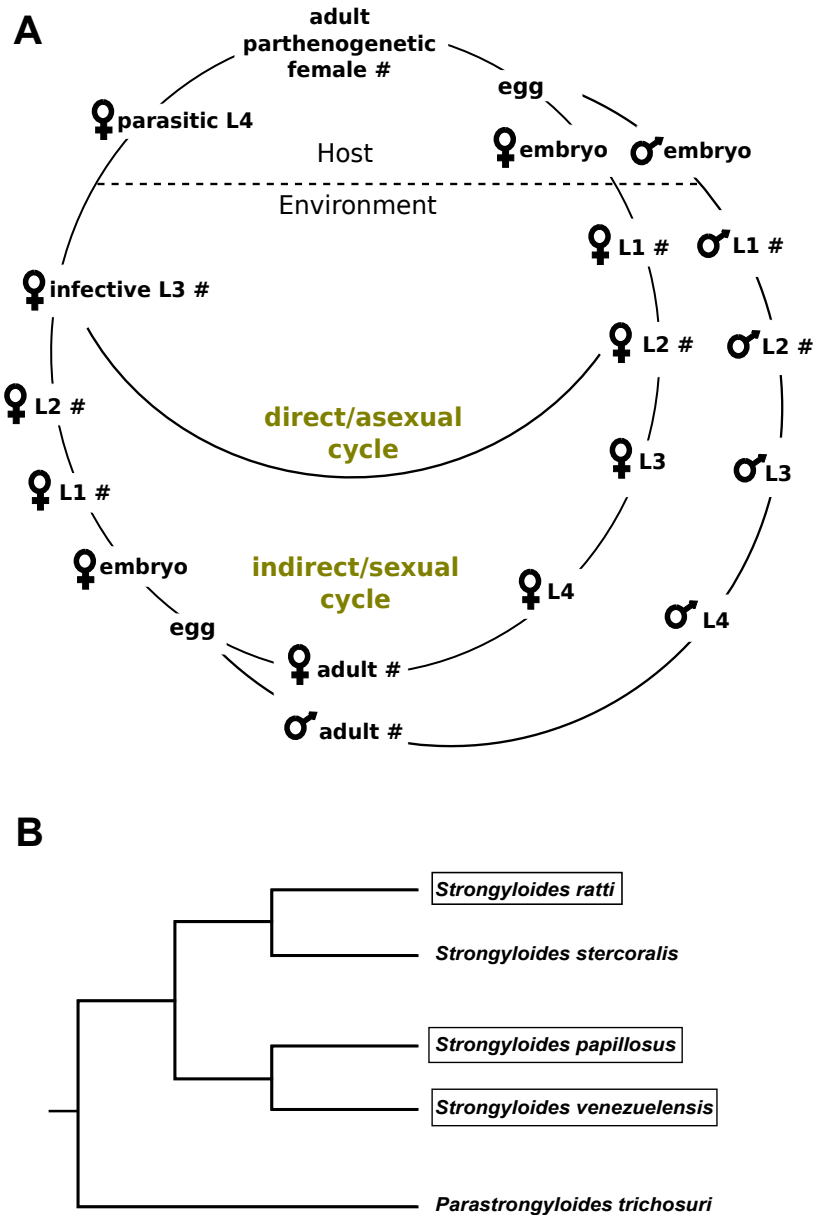


Figure 3.18: **The Life cycle of *S. papillosus*** A) The life cycle of S. papillosus consists of free-living and parasitic stages. Parasitic worms are female only. Within the host, parasitic females reproduce clonally. Their female offspring can either develop directly into infective larvae (iL3) or develop into facultative free-living females that reproduce sexually with males (indirect cycle). Their progenies develop into iL3. Stages for which transcriptomes were sequenced are labeled with #. B) The schematic tree shows the phylogenetic relationships between different Strongyloides species [Hunt et al., 2016]. The species that are considered for defining homology classes are highlighted by a box.

Developmental stages	Number of biological replicates	Sample name replicate 1	Sample name replicate 2	ENA Sample id replicate 1	ENA Sample id replicate 2
Parasitic females	1	Parasitic female		ERS1214240	
Free living Females	2	Free living female rep1	Free living female rep2	ERS1214234	ERS1214235
Free living males	2	Free living male rep1	Free living male rep2	ERS1214236	ERS1214237
Free living L1/L2 (Free-living mother derived L1/L2)	2	Free living L1/L2 rep1	Free living L1/L2 rep2	ERS1214232	ERS1214233
Infective larvae L3	2	iL3 rep1	iL3 rep2	ERS1214238	ERS1214239
Parasite L1/L2 (Parasitic mother derived L1/L2)	1	Parasitic L1/L2		ERS1214241	

Table 3.1: ***S.papillosus* sample information table.** General information about the sequenced developmental stages, number of biological replicates, sample names used in the main text and European nucleotide archive (ENA) accession id.

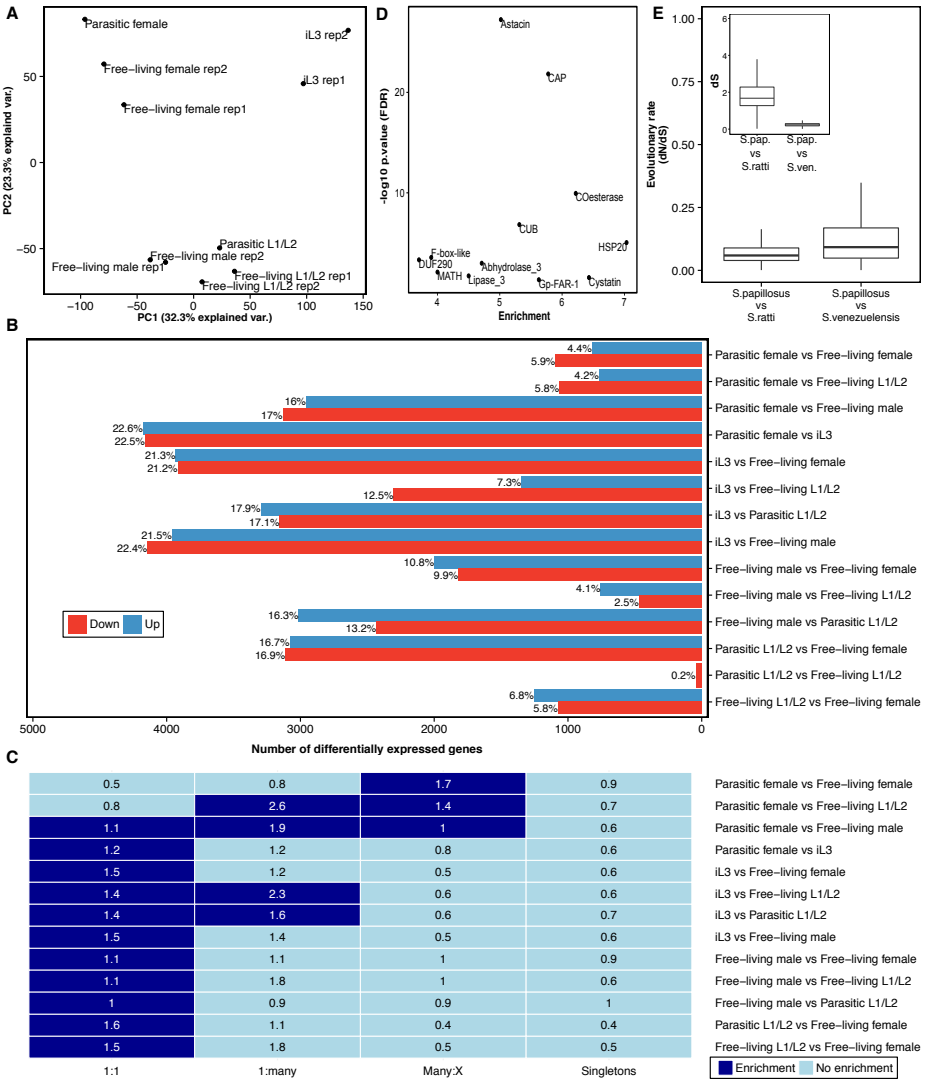


Figure 3.19: **Comparison of developmental transcriptome of *S.papillosus*** A) Principal component analysis of expression values shows a grouping of transcriptomes into distinct clusters that are defined by developmental stage and sex. B) Numbers of significantly differentially expressed genes across all pairwise comparison. C) Enrichment of different homology classes among differentially expressed genes with Dark blue color indicates significant enrichment (FDR corrected $p-value < 0.05$ and enrichment value > 1).The number inside each cell represents the enrichment value. The graph shows that the one-to-one orthologs are enriched in most of the comparison, indicating developmental regulation. D) Enrichment of protein domains (PFAM) among genes up-regulated in parasitic females in comparison with free-living females (FDR corrected $p-value < 0.05$ and enrichment value > 1). E) Box plot representation of the evolutionary rates of one-to-one orthologous genes as measured in $dN/dS(\omega)$ across two different species comparisons. This shows that one-to-one orthologous genes are under strong purifying selection. The inlay shows the age of one-to-one orthologous genes as measured in dS (synonymous substitutions rate). Median dS of S.papillosus-S.venezuelensis one-to-one orthologs are less than one, indicating that synonymous substitutions is not saturated.

duplication (This thesis chapter and [Baskaran et al., 2015]. We wanted to test whether we can observe a similar pattern in the comparison between *S. papillosus* and *S. ratti*. Thus, we first classified the differentially expressed genes into four different classes (Li 2003) based on their homology relationships with *S. ratti*: one-to-one orthologs (N=9302 genes) with one gene per species, 6428 *S. papillosus* genes with many-to-X relationships (many-to-many, many-to-one, many-to-zero), *S. papillosus* genes with one-to-many relationship in *S. ratti* (N=82), and *S. papillosus* singletons (N=2483) without any closely related gene either in *S. papillosus* or in *S. ratti*. Next, we tested whether genes that are identified as being differentially expressed in particular comparisons tend to be a result of duplication events since the separation from the *S. ratti* lineage. In contrast with previous results from the comparison between *P. pacificus* and *C. elegans*, which represents a much larger evolutionary distance [Baskaran et al., 2015], we found that in *Strongyloides* , the majority of gene sets identified by pairwise comparisons shows a significant enrichment of one-to-one orthologs (FDR corrected $p - value < 0.05$, Fisher's exact-test, Figure 3.19C). One notable exception is the genes that are differentially expressed between free-living females and parasitic females. This set is significantly enriched in genes that have undergone duplications in the *S. papillosus* lineage (many-to-X).

Given the substantial divergence between *S. papillosus* and *S. ratti*, (Figure 3.19E inset) and widespread evidence for negative selection ($\omega < 1$, Figure 3.19E), we further tested whether the conservation is also reflected at the level of expression. Visual inspection of phylogenetic trees (Figure 3.20 and Appendix Figure 6.9) in combination with expression data for two large gene families, GPCRs, and NGIC, shows that most genes in both gene families are indeed one-to-one orthologs and that expression profiles are indeed very similar. This implies that in addition to the high level of sequence conservation, as shown by the enrichment of one-to-one orthologs and a strong signature of negative selection ($\omega < 1$), the expression profiles between the two species are also highly conserved.

3.5.5 Enrichment of Astacin and CAP among genes with high expression in parasitic stages

To gain functional insight, we tested the enrichment of protein domains among the identified developmentally regulated genes. We found a large number of gene families enriched in different comparisons and some families in multiple comparisons. Protein kinases (PF00069), HSP20 (PF00011), Motile sperm domain containing proteins (PF00635), Collagens (PF01391), different subfamilies of GPCRs and a few other families are significantly enriched in more than five comparisons (Appendix Figures 6.7 and 6.7). We estimate that in total 221 families show enrichment in at least two comparisons. In accordance with the finding of Hunt et al. [Hunt et al., 2016] based on data from *S. ratti* and S. stercoralis, we found the Astacin (PF01400) and CAP (PF00188) gene families to be the most significantly enriched among genes up-regulated in parasitic females in comparison with free-living females ($P < 10^{-20}$, Fisher's exact-test, Figure 3.19D). Apart from the enrichment of members of these two gene families in the parasitic female stage, we also found a considerable number of Astacin and CAP genes up-regulated in iL3.

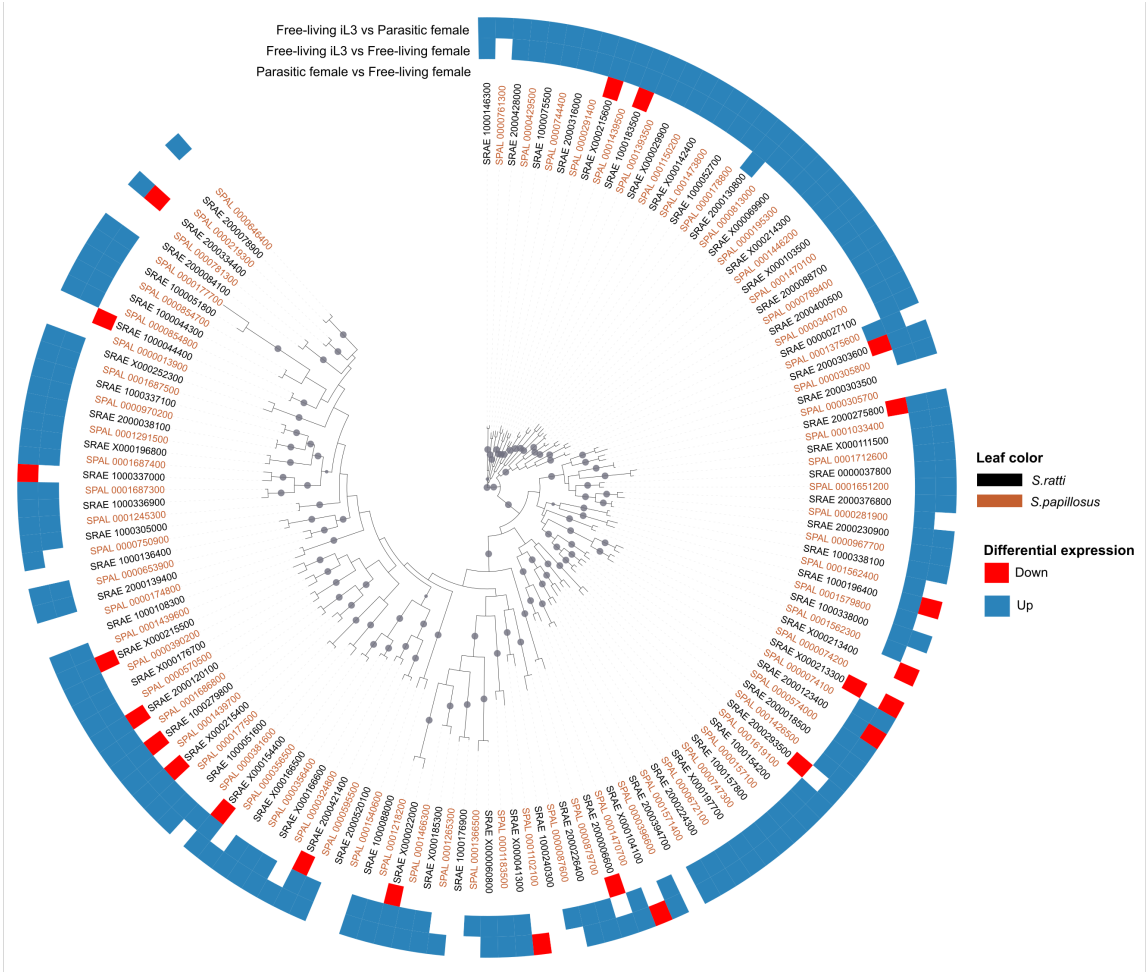


Figure 3.20: **Conservation of gene expression and protein sequence among NGIC gene familiy** Phylogenetic tree of NGIC family reconstructed using *S. papillosus* (Brown) and *S. ratti* (black) protein sequences. The heat map around the phylogenetic tree represents differential expression pattern of each gene in three different comparisons, where up-regulation, down-regulation and no change are shown in blue, red and white, respectively. This tree shows that a majority of one-to-one orthologs in NGIC gene family have similar expression patterns in *S. papillosus* and *S. ratti*. Dots indicate internal nodes with 100/100 bootstrap support.

3.5.6 Most members of Astacins and CAP families trace back to the single expansion events in the *Strongyloides* lineage

In order to investigate the underlying evolutionary history and to expand the previous findings of Hunt et al. [Hunt et al., 2016], we reconstructed phylogenetic trees for both families and integrated the newly acquired *S. papillosus* expression data (Figures 3.21 and 3.22, respectively). *C. elegans* and Drospohila Spp CAP and Astacin protein sequences were used as outgroups to determine the evolutionary pattern of ancestral genes. Both phylogenies show that most genes fall into *S. papillosus* specific subtrees. Similarly, *C. elegans* paralogs also cluster in lineage-specific subtrees, indicating that there is a trend for duplication of these families even in non-parasitic nematode lineages. Interestingly, the few *S. papillosus* Astacin genes that appear to have one-to-one orthologs in *C. elegans* (nas-36, dpy-31, nas-37, nas-33, nas-30, toh-1, hch-1 in Figure 3.22) show quite unique expression profiles, suggesting that these profiles are tightly regulated and are highly dosage sensitive (Figure 3.22). While the bootstrap support for a single expansion of CAP genes giving rise to almost all extant members of this family is weak (52/100), there is one branch with a support of 96/100 replicates that contains 97 (59%) of CAP genes (Figure 3.21). For Astacins, we indeed found one highly supported branch that gave rise to 193 (97%) of all genes in this gene family (Figure 3.22).

3.5.7 Astacin and CAP subtree genes shows two distinct expression profiles

Integration of *S. papillosus* expression data allows us to perform a detailed visual inspection of the differential expression patterns across both families. Using this approach, we found the above-mentioned finding that most genes either show high expression in parasitic females or in iL3. These two patterns dominate the general picture of expression profiles and comparison with the gene tree reveals a strong phylogenetic signature in the expression profiles, i.e. most iL3 specific genes cluster in one subtree while genes with high expression in parasitic females cluster in a different subtree. Even though we are far from understanding the roles of Astacins and CAP genes in *Strongyloides* parasites, our analysis strongly indicates the fact that mutually exclusive groups of Astacins and CAP genes play a role at the stage where parasites infect the host and at the stage where adult parasites proliferate within the host. This clearly demonstrates the mechanism of subfunctionalization in both gene families.

3.5.8 Strong signature of positive selection in CAP and Astacin gene families

Finally, we wanted to test whether the adaptation to novel hosts is paralleled by increased evidence for positive selection. As the divergence between *S. papillosus* and *S. ratti* indicates that synonymous substitutions are already saturated (more than one substitution is expected per synonymous site, Figure 3.19E inset), we identified orthologous clusters with the more closely related S. venezuelensis [Hunt et al., 2016] and screened all orthologous clusters with more than two genes for positive selection. More precisely, we performed a likelihood ratio test to decide whether a model including a number of sites with $\omega > 1$ explains the alignments better than models with either only negative or neutral selection (see Methods). Overall, we found only 28 (67%) and 14 (58%) of orthologous clusters in Astacin and CAP families, respectively, with evidence for positive selection. This represents a strong enrichment when compared to all other clusters tested (N=943), of which only 21% showed evidence for positive selection (FDR corrected $p-value < 0.05$, Fisher's exact test).

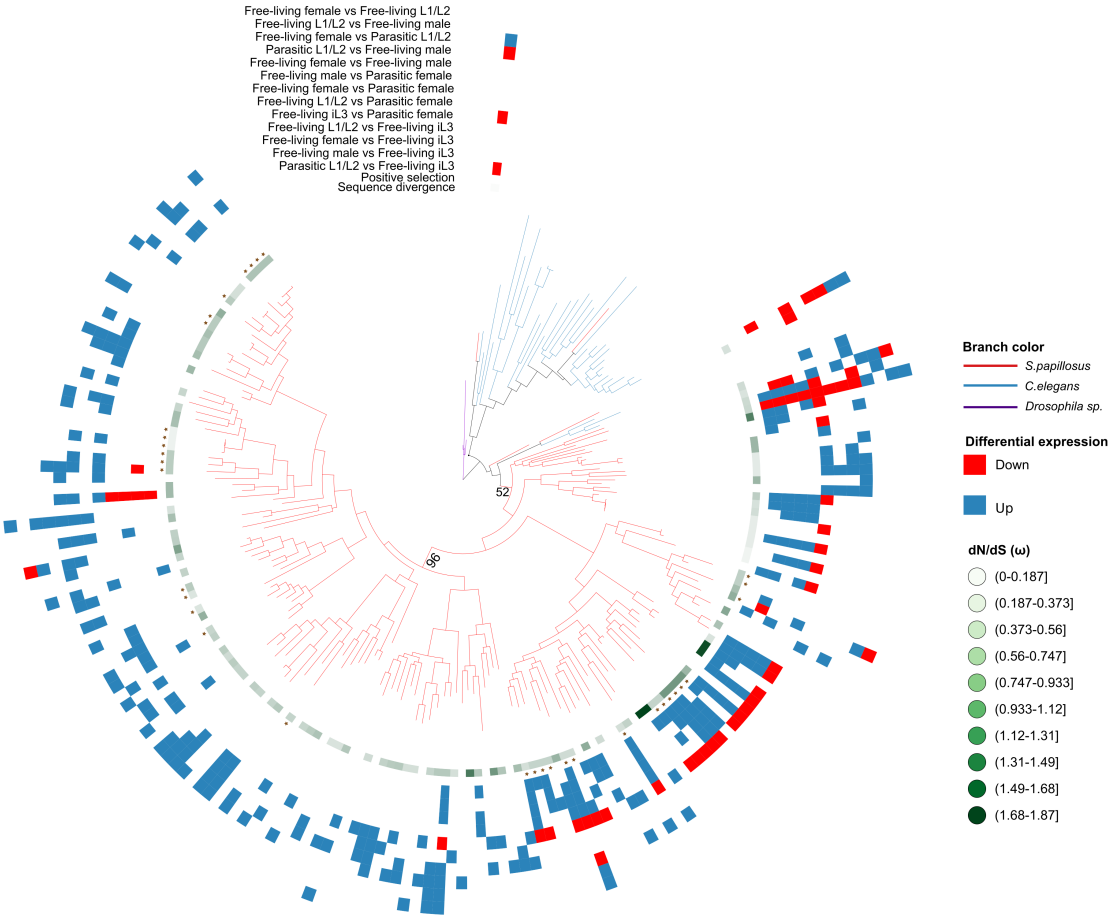


Figure 3.21: **Gene duplication in CAP gene family.** Phylogenetic tree reconstructed using CAP amino acid sequence from *S. papillosus* (red branches), *C. elegans* (blue branches) and *Drosophila Sp.* (purple branches). The tree shows lineage-specific expansion of CAP genes in the *Strongyloides* lineage. Due to poor visibility, only bootstrap values that are relevant for our analysis are shown. The color patterns and symbols for each *S. papillosus* gene were generated by testing sequence evolution and differential expression. The green gradient shows the sequence evolutionary rate as measured in $dN/dS(\omega)$. The star symbol represents the evidence for positive selection in the gene sets. The heatmap of blue, red and white shows the differential expression pattern for each *S. papillosus* gene based on the pairwise comparison of all sequenced developmental stages. Up-regulation, down-regulation and no change in gene expression are shown in blue, red and white, respectively.

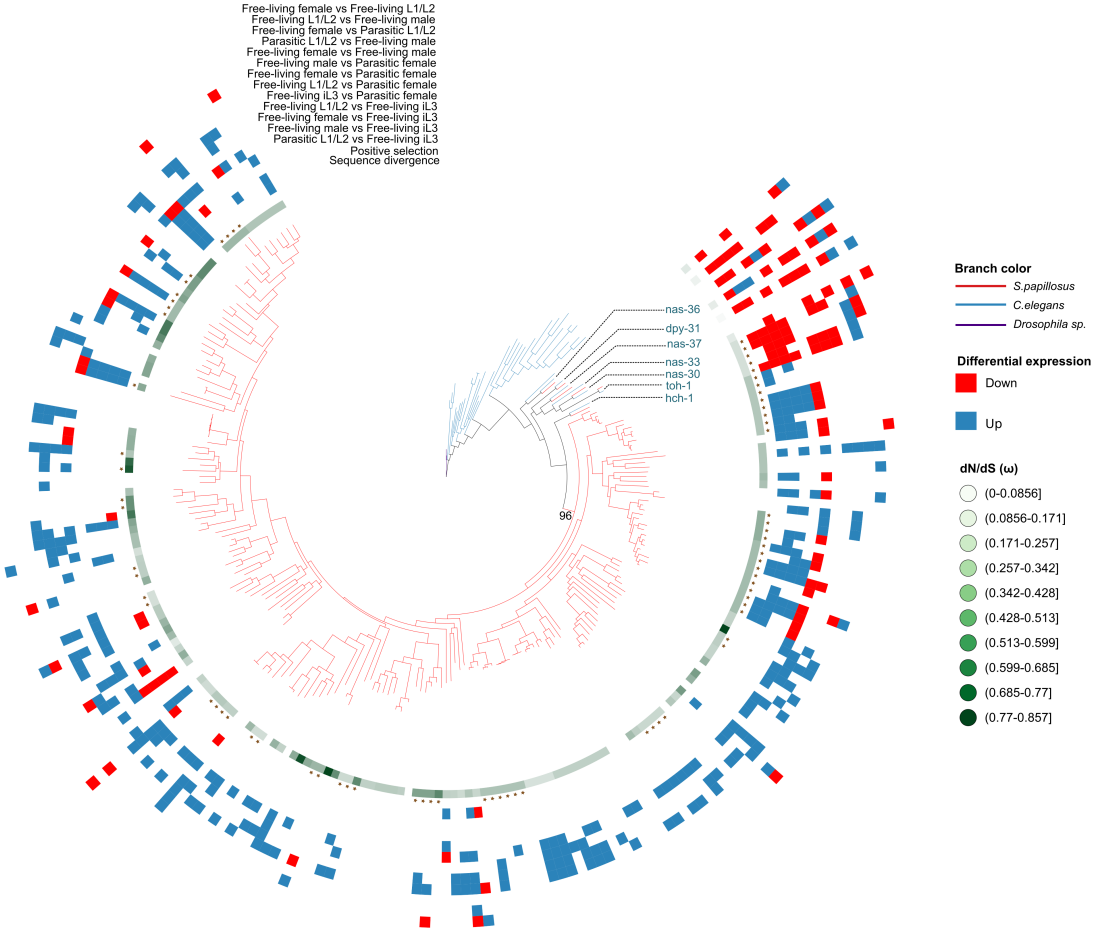


Figure 3.22: **Gene duplication in Astacin gene family.** Phylogenetic tree of Astacins reconstructed using protein sequences from *S. papillosus* (red branches), *C. elegans* (blue branches) and *Drosophila Sp.* (purple branches). The tree shows lineage-specific expansion of CAP genes in the *Strongyloides* lineage. The color pattern and symbols for each *S. papillosus* gene were generated by testing sequence evolution and differential expression. The green gradient shows the sequence evolutionary rate as measured in $dN/dS(\omega)$. The star symbol represents the evidence for positive selection in the gene sets. The heatmap of blue, red and white shows the differential expression pattern for each *S. papillosus* gene based on the pairwise comparison of all sequenced developmental stages. Up-regulation, down-regulation and no change in gene expression are shown in blue, red and white, respectively.

Chapter 4

Discussions

This chapter contains content from the following publications. The copyright holder has granted the re-use permission.

Markov, G. V., Baskaran, P., and Sommer, R. J. (2015). The Same or Not the Same: Lineage-Specic Gene Expansions and Homology Relationships in Multigene Families in Nematodes. Journal of Molecular Evolution, 80(1):18-36.

Baskaran, P. and Rödelsperger, C. (2015). Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus. PloS one, 10(6) e0131136.

Baskaran, P., Rödelsperger, C., Prabh, N., Serobyan, V., Markov, G.V., Hirsekorn, A., and Dieterich, C. (2015). Ancient gene duplications have shaped developmental stage-specic expression in Pristionchus pacificus. BMC Evol Biol, 15(1).

Baskaran,P., Jaleta, T., Streit, A. and Rödelsperger C. (2017). Duplications and Positive Selection Drive the Evolution of Parasitism-Associated Gene Families in the Nematode Stongyloides papillosus. Genome Biol Evo,l 9(3): 790-801.

In this work, we systematically studied the dynamics of evolutionary forces acting on duplicated genes and its impact on gene expression levels in different nematode species.

4.1 Widespread evidence for gene duplication in nematode multigene families

In the first part (Chapter 3.1), we compared the protein sequences of seven gene families belonging to 11 nematode species to investigate the orthology relationships among enzymes involved in major pathways. By manually curating these protein sequences we discarded false positive predictions and ensured robustness of the datasets for detailed analysis. By performing phylogenetic analyses on these gene families we show that only a small proportion of genes were maintained as one-to-one orthologs between *P. pacificus* and *C. elegans*. Specifically, less than 8% of *P. pacificus* and 12 % *C. elegans* genes analyzed were one-to-one orthologs. This indicates that majority of the genes involved in the detoxification of xenobiotic and PUFA synthesis pathways are derived from lineage-specific processes. The proportion of one-to-one orthologs observed in our sampling dataset is smaller than previous comparisons based on entire sequence data of *P. pacificus* (29,201 genes) and *C. elegans* ($\approx$ 20,000 genes), where more than 16% of genes in both species were maintained

as one-to-one orthologs [Dieterich et al., 2008]. Additionally, we also a found large variation in the number of genes involved in the xenobiotic metabolism than genes involved in the citrate cycle, which is consistent with the previous observation [Dieterich et al., 2008].

While out of the seven families analyzed, six families show evidence for large-scale amplification events in a lineage-specific manner; the ABC transporter family is an exception with a high degree of conservation between *C. elegans* and *P. pacificus*. Also, families involved in PUFA synthesis show a unique pattern, with strong conservation in Caenorhabditis genus and lineage-specific expansion in Pristionchus genus. Additionally, pristionchus genus specific large-scale expansions were observed in UGT family, where 34 and 54 genes were generated from two amplification events. Our finding of an overall increase in gene count on the branch leading to *P. pacificus* is consistent with the observation that in general gene families show higher duplication rate in *P. pacificus* than in *C. elegans*. One of the major evolutionary implications of this study is that it raises questions about the relationship between lineage-specific amplification and functional specificity. It is assumed that enzyme coding genes that are involved in important pathways are preserved as one-to-one orthologs between species and this helps to maintain functional conservation. However, our findings shows a completely opposite trend, strong depletion of one-to-one orthologs as a result of large-scale lineage-specific expansions among gene families involved in detoxification and PUFA synthesis pathways. This strongly suggests that during the course of nematode evolution, the functional specificity of many enzymes has changed.

Consistent with our observation, the unusual amount of convergence, redundancy and promiscuity among many enzyme coding genes have been reported previously by various studies. The elongation of fatty acids by the member of both elongase and reductase gene families represents a classic instance of functional convergence between different enzymes [Watts, 2009]. Similarly, promiscuous nature of enzymes involved in PUFA metabolism and *C. elegans* xenobiotic enzyme F54F3.4 has also been reported earlier [Watts, 2009, Kisiela et al., 2011]. Taken together, the large-scale lineage-specific amplifications observed in the analyzed multigene families might be the result of relaxed evolutionary constraint by functional convergence, promiscuity, and redundancy of the genes involved in the metabolic pathways. The observation of distinct rate of duplication in different gene families also suggests the relaxation of functional constraints differs between families. In general, the observed pattern of amplification and lack of one-to-one orthologs shows that the evolution of most of the analyzed gene families was shaped by massive expansion events.

4.2 Intra-species comparison within *P. pacificus* shows that most duplications are either deleterious or neutral

The question how genes are gained and lost during evolution is central to understanding the diversity across the animal kingdom. To complement previous comparisons of nematode gene repertoires at cross-species levels [Dieterich et al., 2008, Markov et al., 2015], we have investigated intra-species patterns of gene loss and gain in the nematode *P. pacificus* (Chapter 3.2). Such comparisons spanning different time periods have the potential to illuminate different aspects of evolutionary forces acting on the *P. pacificus* genome. Along these lines, it has been previously shown that mutations that occurred during one hundred generations under laboratory conditions show characteristics of neutral variation [Weller et al., 2014], whereas even in the most closely related natural isolates, a strong evidence for purifying selection was observed [Rödelsperger et al., 2014]. At the level of gene gain and loss, previous comparative genomic studies have highlighted massive gene con-

traction and expansion events that were hypothesized to reflect adaptation to new environments and lifestyles such as the necromenic association of Pristionchus nematodes with scarab beetles [Dieterich et al., 2008, Herrmann et al., 2006]. However, at a population level, the prevailing pattern seems to be, that most deletions and duplications are either neutral or deleterious. Evidence for negative selection as shown in the strong depletion of SVs in large parts of the genome is consistent with previous evidence for purifying selection as obtained from single nucleotide variation data, such as the loss of nonsynonymous diversity over time [Rödelsperger et al., 2014].

Despite the fact that copy number variations and SVs have extensively been studied in various species [Maydan et al., 2007, Maydan et al., 2010, Vergara et al., 2014, Schuster-Böckler et al., 2010, Zhou et al., 2011, Handsaker et al., 2015], only a few studies have combined SVs with expression data [Schuster-Böckler et al., 2010, Zhou et al., 2011, Handsaker et al., 2015]. Interestingly, we find that while deletions show a strong impact on gene expression, duplications have virtually no effect on the transcriptomic level. Based on the fact that most of the genes, for which different gene copies could be distinguished based on segregating sites, showed a strong allelic bias in their expression, we hypothesize that the missing signal is due to either incomplete duplicates or duplicate insertions into transcriptionally silenced genomic regions. Further analysis of tandem duplications at nucleotide level resolution showed that even in cases where we can be sure that the duplication encompasses the first and last exon, we do not find the expected effect on expression data. This is in contrast to a previous study in human [Handsaker et al., 2015], which showed a strong effect of duplications on gene expression levels. However, we would like to point out, that although in many cases, gene expression levels scaled linearly with copy number, only a few cases truly showed that duplication doubled the gene dosage [Schuster-Böckler et al., 2010, Handsaker et al., 2015]. Furthermore, genetic diversity in *P. pacificus* is roughly one order of magnitude higher than in humans, suggesting that at the evolutionary distances between our *P. pacificus* strains, selection might have had more time to purge functional duplications that have slightly deleterious effects. Similarly, a recent gene duplication in *P. pacificus* which was identified by our depth of coverage approach showed significant differential expression and was present only in the reference (PS312) and in the most closely related strain [Mayer et al., 2015]. However, much more genomic and transcriptome data is needed to study the connection between age and effect of duplications on expression.

Intuitively, deletions are more likely to affect genes than duplications, because even a partial deletion may disrupt gene structure whereas a partial duplication most likely results in a non-functional gene fragment. Consistently, it has previously been shown that median size of recent duplications in *C. elegans* is shorter than average gene length (2.5kb) [Katju and Lynch, 2003], further supporting the finding of abundant non-functional gene copies. One major drawback of our study is that using gene expression level to evaluate functionality is naive.

The fact that they were generated and maintained may suggest that some of the duplications might be beneficial for certain strains. However, in the absence of true functional data for most of the genes in *P. pacificus*, the use of gene expression as a first proxy for functionality revealed very interesting trends that might also be reflected at the functional level. Despite the fact that most novel gene copies are not expressed at a comparable level to their ancestral copy, an open question remains, why there seems to be strong selection against these apparent nonfunctional duplications in large parts of the genome. We can only speculate, that insertions of even non-functional duplicated sequences may interfere with long-range regulatory interactions and may disrupt gene and operon-like structures. Previous analysis in *C. elegans* has shown that recent duplications tend to be locally and show a trend towards dispersal across the genomes with increasing age [Lynch and Katju, 2004,

Katju and Lynch, 2003]. Thus, the local nature of duplications suggests that in regions where synteny is important, even transcriptionally silent duplications may have deleterious effects.

4.3 Evidence for selection for higher gene dosage in developmentally regulated genes of *P. pacificus*

The pattern observed at the microevolutionary levels suggests selection acting against changes in the copy number and consecutively against the variation in expression levels of SV affected genes. In contrast, at the level of macroevolution, we observed massive expansion events of multigene families of *P. pacificus* and *C. elegans*. Given this, we wondered whether the impact on gene expression also varies at different evolutionary time scales. We tested this hypothesis by comparing the transcriptomes of distinct developmental stages of *P. pacificus* to identify and reconstruct the evolutionary history of developmentally regulated genes. By sequencing the transcriptome of different developmental stages of *P. pacificus*, this study (Chapter 3.3) provides the first expression profiling data for early larval stages of *P. pacificus*. Using a Unsupervised biclustering approach on expression data we detected 29 biclusters, which show developmental-stage specific regulation throughout the time course (Figure 3.10D-F, Additional file 2: Figure S1). A complementary set of 5151 (17 %) potential house-keeping genes were identified that showed robust expression in all samples and that did not reveal any signal for significant differential expression in any pairwise comparison.

In this study (Chapter 3.3), we characterized the pattern of conservation and divergence within a broader evolutionary time scale comparison. In two gene families (HSP20 and HSP70), where we performed a detailed analysis combining phylogenetic reconstruction and expression analysis, we did not find any developmentally regulated gene with a one-to-one ortholog in *C. elegans*. In contrast, we observed a tendency of developmentally regulated genes to occur in *P. pacificus*-specific subtrees suggesting that they had undergone duplication events after the separation from the *C. elegans* lineage. Even paralogous genes that were not captured as developmentally regulated due to missing significance in the differential expression analysis showed very similar expression profiles as the genes that were indeed identified to be developmentally regulated.

We examined these findings on a genome-wide scale by testing the over-representation of different homology classes in developmentally regulated gene set (Figure 3.13). Consistent with the case study of heat shock proteins, we found that the strongest enrichment of putative paralogs, i.e. conserved multicopy genes (many-to-X category in Figure 3.13) among developmentally regulated gene set. More precisely we found that 19 out of 29 biclusters showed a significant enrichment for conserved multicopy genes. This indicates that the developmental transcriptome of *P. pacificus* is shaped by ancient gene duplication events. Such ancient duplications may represent a plausible evolutionary mechanism to increase the dosage of developmentally regulated genes. The implicated positive selection on gene dosage as it had already been suggested by Ohno [Ohno, 1970], is often neglected in discussions of models for gene duplication [Rogozin et al., 2014, Ohno, 1970, Lynch and Katju, 2004]. Instead, more focus is given to mechanisms that may give rise to neofunctionalization and subfunctionalization within a gene family. Such trends may be supported by the finding of dauer-specific and early larval-specific paralogous clusters in one gene family (Figure 3.12B,C), as well as divergent expression profiles of genes within one paralogous cluster (Figure 3.12C). However, further experimental and computational work is needed to allow a more detailed characterization of patterns of subfunctionalization in *P. pacificus*.

Our findings reveal trends that likely evolved in the range of hundreds of millions of years and suggest that evolutionary patterns that can be detected at this level may be completely different from patterns at the microevolutionary level.

4.4 The study of parasitic nematodes reveals strong support for the importance of gene duplication in adapting to new environments

The above studies highlight the different aspects of evolutionary forces acting on the *P. pacificus* genomes at the micro and macroevolutionary level. This intrigued us to test what happens at the intermediate level by comparing closely related species. We compared *P.pacificus* with two closely related pristionchus species, *P.exspectatus* and *P.arcanus* to investigate the impact of gene duplication on expression levels (Chapter 3.4). More specifically we were interested in cases with single duplication event in *P.pacificus* and no duplication in other two species. We found combined expression levels of recently duplicated genes in *P.pacificus* are not significantly different from the expression levels of outgroup species. Additional analysis of these recent duplicates shows a trend similar to the microevolutionary trend. Therefore we switched to a system, where we have good candidate gene families that are likely involved in the adaptation to new environments.

To explore the evolutionary forces at the intermediate levle, we compared the developmental transcriptome of parasitic nematode *S. papillosus* with its closely related species, *S. ratti* and *S.venezuelensis* (Chapter 3.5). This first characterization of developmental transcriptomes from *S. papillosus* showed that the majority of *S. papillosus* genes are developmentally regulated, and, in addition, most of them have one-to-one orthologs in *S. ratti*. Furthermore, the new transcriptomic data allowed us to gain some insights into the evolution of gene expression in the Strongyloides lineage, which has not been the focus of the previous study by Hunt et al. [Hunt et al., 2016]. Our comparative transcriptomic analysis shows remarkably similar expression between orthologous gene pairs (Figure 3.20). Given the substantial divergence between the species, this indicates a high degree of evolutionary constraint acting at sequence and expression level. This high degree of conservation is in contrast to our previous study comparing the developmental transcriptome of *P. pacificus* with *C. elegans*, in which most developmentally regulated genes are derived from ancient gene duplications [Baskaran et al., 2015]. However, much more similar to the microevolutionary picture of *P. pacificus* [Baskaran and Rödelsperger, 2015].

One notable exception to the general widespread constraints on sequence and expression levels, is the enrichment of duplicated genes in the comparison between free-living females and parasitic females. This comparison if of utmost important as it may reveals genes responsible for the parasitic activities of *S. papillosus*. The enrichment of duplicated genes in this comparison may represent the adaptation to discrete and continuous changes in the environment, such as host switches and adaptation of the host immune response. Our speculation is further supported by the evidence for positive selection in Astacins and CAP genes, which are most significantly enriched in the comparison between free-living females and parasitic females. Together with initial evidence from experimental evolution studies supporting a strong adaptive potential of gene duplications in nematodes [Farslow et al., 2015], our results suggest that despite its generally conservative nature, evolution can be, under certain circumstances, extremely fast. If gene duplications have an adaptive potential, e.g. increase the ability of *Strongyloides* nematodes to infect their host, it is much

more likely that this will increase the probability of their retention. Thus, it could be possible that the strongly divergent patterns which dominate comparisons of distantly related species, such as between *P. pacificus* and *C. elegans* [Baskaran et al., 2015] or between *S. stercoralis* and *C. elegans* [Stoltzfus et al., 2012], reflect the evolutionary history of multiple adaptive events. The alternative scenario that most of these changes can be explained by neutral events (e.g. genetic drift and demographic effects) appears less likely, as the overwhelming picture of several recent studies of genome evolution in nematodes seems to be that selection generally acts against changes in protein-coding regions [Rödelsperger et al., 2014, Baskaran and Rödelsperger, 2015].

Overall, our studies captured the dynamics of evolutionary forces acting on nematode genomes at different time scales and provide the first insight into the developmental regulation of parasitic and non-parasitic nematodes. The prevalence of negative selection on duplicated genes observed at the microevolutionary levels can be associated with the time since duplication. Since the *P. pacificus* strains have diverged from each other for over 1 million generations we can speculate that these predicted duplicated copies are in their early days and might not have embedded into the gene-regulatory networks. This could explain the biased expression of duplicated copies, as the new copy might not have been integrated into the expression network. Expression of both copies of some duplicated genes suggests that the copies might be using the regulatory elements of parent gene. In contrast, most of the duplicated genes observed at the macroevolutionary level and in the comparison of closely related species might be already integrated into the biological networks, as they are stable in the genome for a substantial time period.

Chapter 5

Conclusion

This thesis investigated the evolutionary forces which act on duplicated genes and thus determine their retention or loss. By investigating the protein sequence and expression profiles of duplicated genes in nematodes at different evolutionary time scales, we found that different selection forces shape the evolution of duplicated genes and gene families. Initial models proposed by Ohno, Lynch and others provided a first insight into how the evolutionary fate of duplicated genes are decided [Ohno, 1970, Lynch and Force, 2000, Lynch et al., 2001, Kaessmann, 2010]. Even though such models are quite informative in understanding the underlying forces acting on duplicated genes, they failed to captures some important aspects. For instance, Ohno's neofunctionalization model assumes one copy acquires new function, while the other maintains the ancestral function. In this case, it is not clear how natural selection distinguishes two identical copies and select one for a new function [Bergthorsson et al., 2007]. In this thesis, evolutionary comparisons at various time-scales revealed quite heterogeneous patterns. At population level, and comparison between closely related species, selection against duplications seems to be the dominating process. The few duplications that are tolerated, appeared at the first glance to be non-functional, as one of the copies seems not be expressed. This indicates that in contrast to Ohno's assumption of two initially identical copies, the two duplicates are unequal at birth. Surprisingly, even if one of the duplicated copies shows very little evidence of expression, our data strongly suggest that they are still subject to purifying selection and thus are functional. This purifying selection may be relaxed in individual cases and may in the long run lead to accumulation of mutations that change their function. However, these individual events are unable to explain the pulses of duplications (multiple rounds of duplications which increase the overall gene dosage) that are observed in the comparison with *C. elegans*. We therefore assumed, that constantly changing environments can lead to strong selection for higher gene dosage as well as drive functional divergence in order to adapt to novel challenges. This has been demonstrated in the example of duplications in the Astacin and CAP gene families in *Strongylodies* nematodes to overcome host-specific immune response or increase the virulence. Such rare but strong adaptive events may eventually explain the patterns observed in comparisons of long evolutionary distances. Together, these findings show that the retention of gene duplicates cannot explain by a single process. More likely, their retention is a result of a complex interplay between multiple different processes, such as increase of gene expression, generation of novel copies that can be the target of positive selection at amino acid sequences, and a combination of such subfunctionalization/neofunctionalization at sequence and expression level. This highlights the need to choose a right ecological context in order to better understand patterns of genome

evolution.

Chapter 6

Appendix

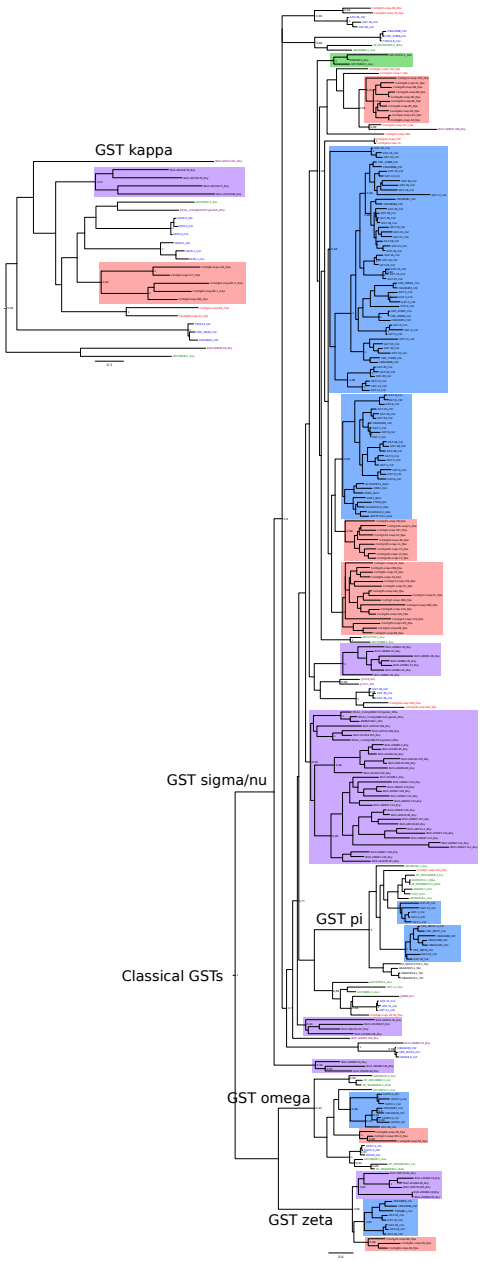


Figure 6.1: **Detailed phylogeny and gene duplication in nematode GST family.** Maximum-likelihood trees of nematode GST (kappa and classical GST) , an extended version of Figure 3,1. Different colors represent different nematode species included in this study (see Figure 3.1 for details). Only lineage-specific expansion events that give rise to at least three paralogs were highlighted. we found three expansion events in *P.pacificus* resulting in more than three classical GST genes.

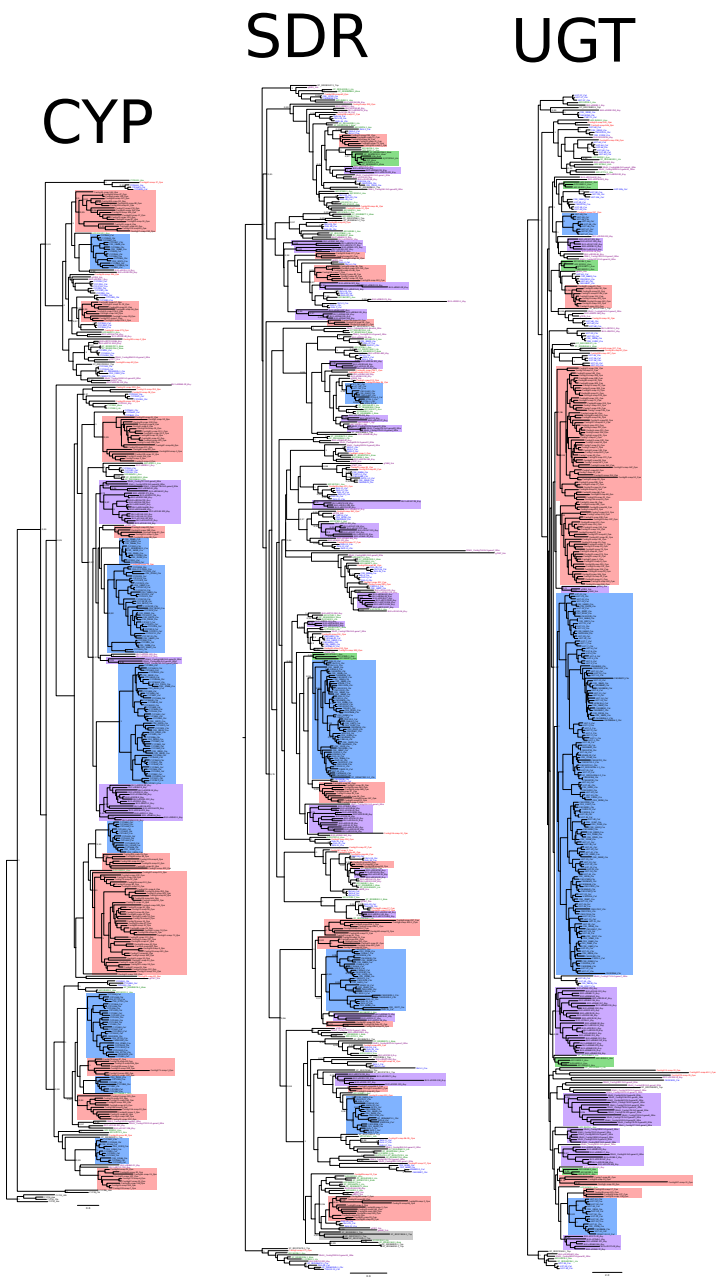


Figure 6.2: **Maximum-likelihood phylogeny of CYP, UGT and SDR families.** Extensive phlyogenetic trees of nematode CYP, SDR and UGT reconstructed using Maximum-likelihood approach based on LG substitution matrix. Different colors represent different nematode species included in this study (see Figure 3.1 for details). Only lineage-specific expansion events that give rise to at least three paralogs were highlighted. Values in each node represent statistical support based on likelihood ratio tests, with only values above 0.97 are considered as reliable.

Figure 6.3: **Gene duplication in elongase families.** Maximum-likelihood phylogenetic tree of nematode elongase familiy shows three lineage-specific expansion of elongases in *P.pacificus* (highlighted in red). We found only two *P.pacificus* genes have one-to-one orthologs in *C.elegans*.

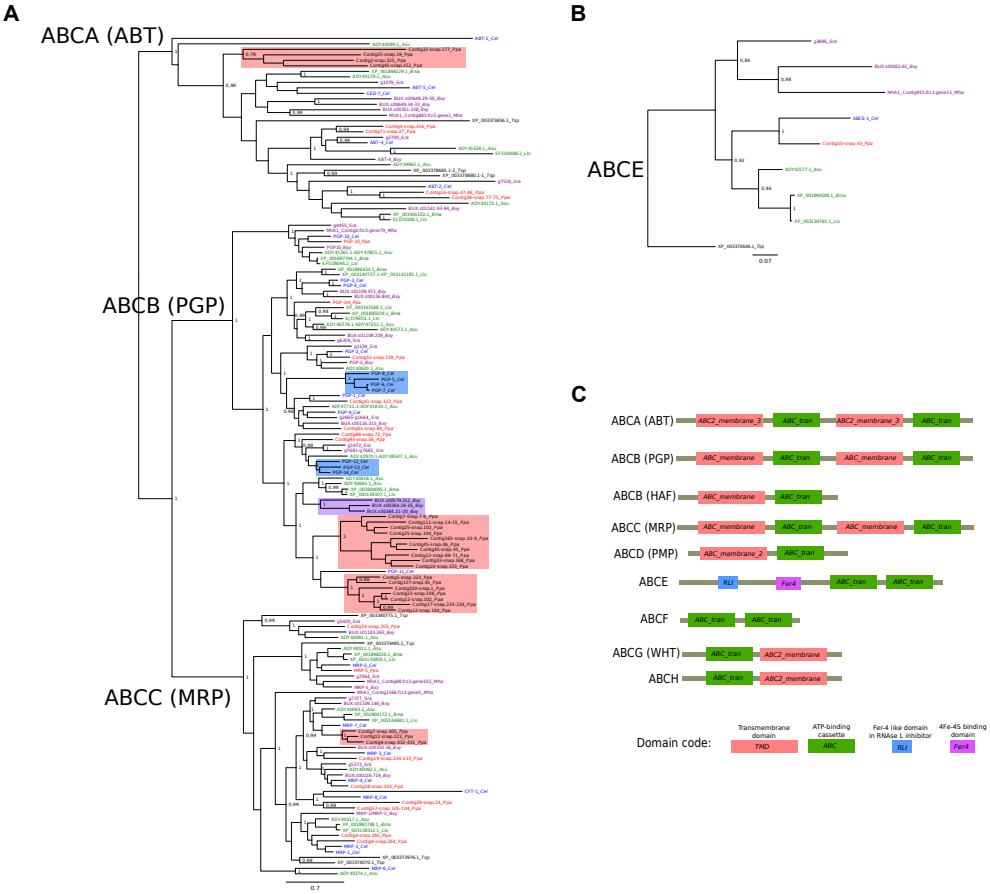
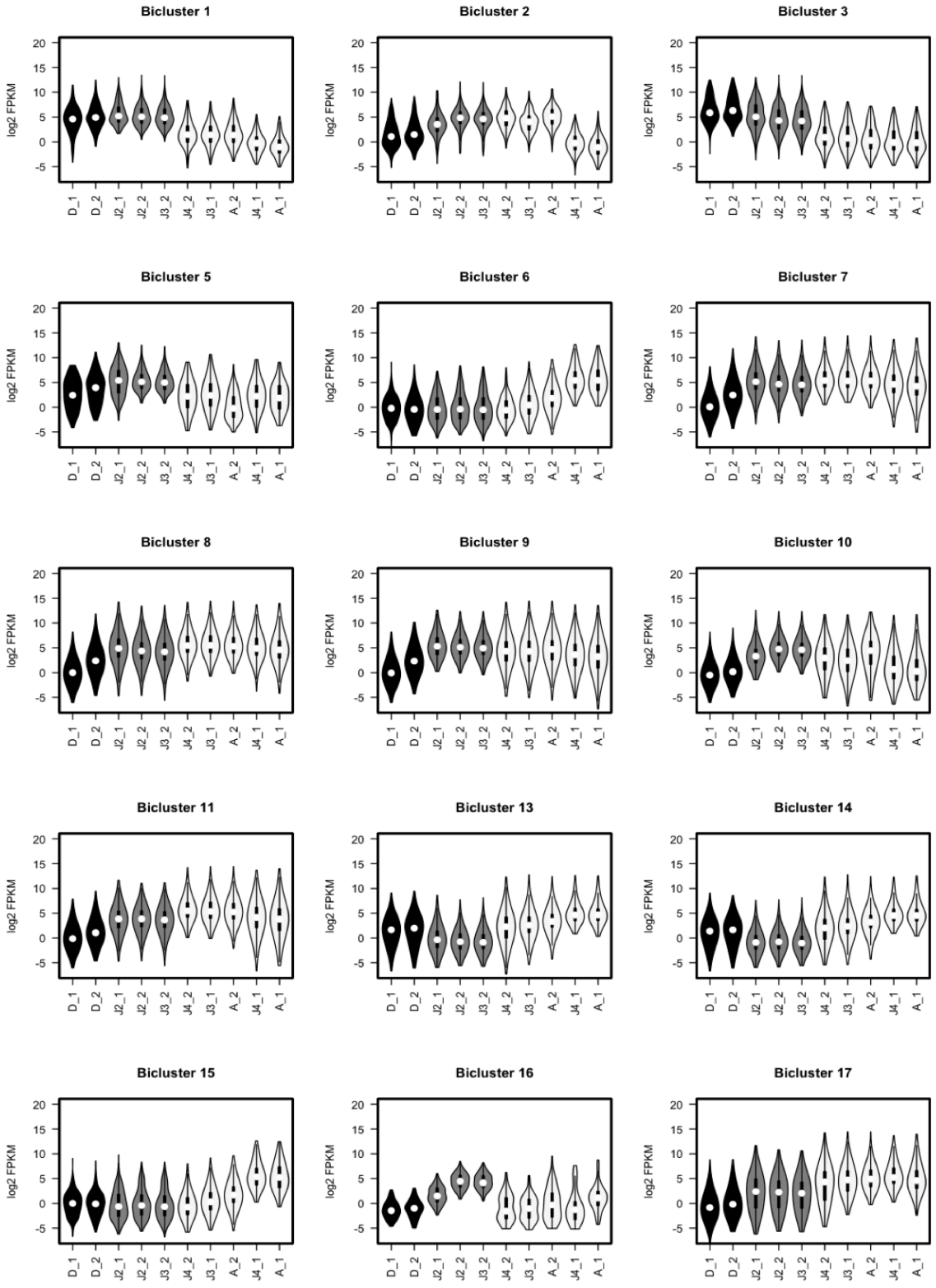


Figure 6.4: **Detailed phylogenetic tree and gene duplication in ABC families.** Panel A and B shows the maximum-likelihood phylogenetic tree of nematode ABC sub families (ABCA. ABCB, ABCC and ABCE) using LG substitution matrix. Panel C shows schematic representation of domain architecture of different ABC sub families. Phylogenetic trees of rest of the ABC subfamilies are included in the main text. Even though ABC subfamilies show strong conservation across different nematodes, specifically between *P. pacificus* and *C. elegans*, ABCB, ABCC and ABCA subfamilies have underground at least one expansion event *Pristionchus* lineage.


Bicluster 1
Bicluster 2
Bicluster 3
Bicluster 5
Bicluster 6
Bicluster 7
Bicluster 8
Bicluster 9
Bicluster 10
Bicluster 11
Bicluster 13
Bicluster 14
Bicluster 15
Bicluster 16
Bicluster 17
log2 FPKM
20
15
10
5
0
-5
D_1
D_2
J2_1
J2_2
J3_2
J4_2
J3_1
A_2
J4_1
A_1

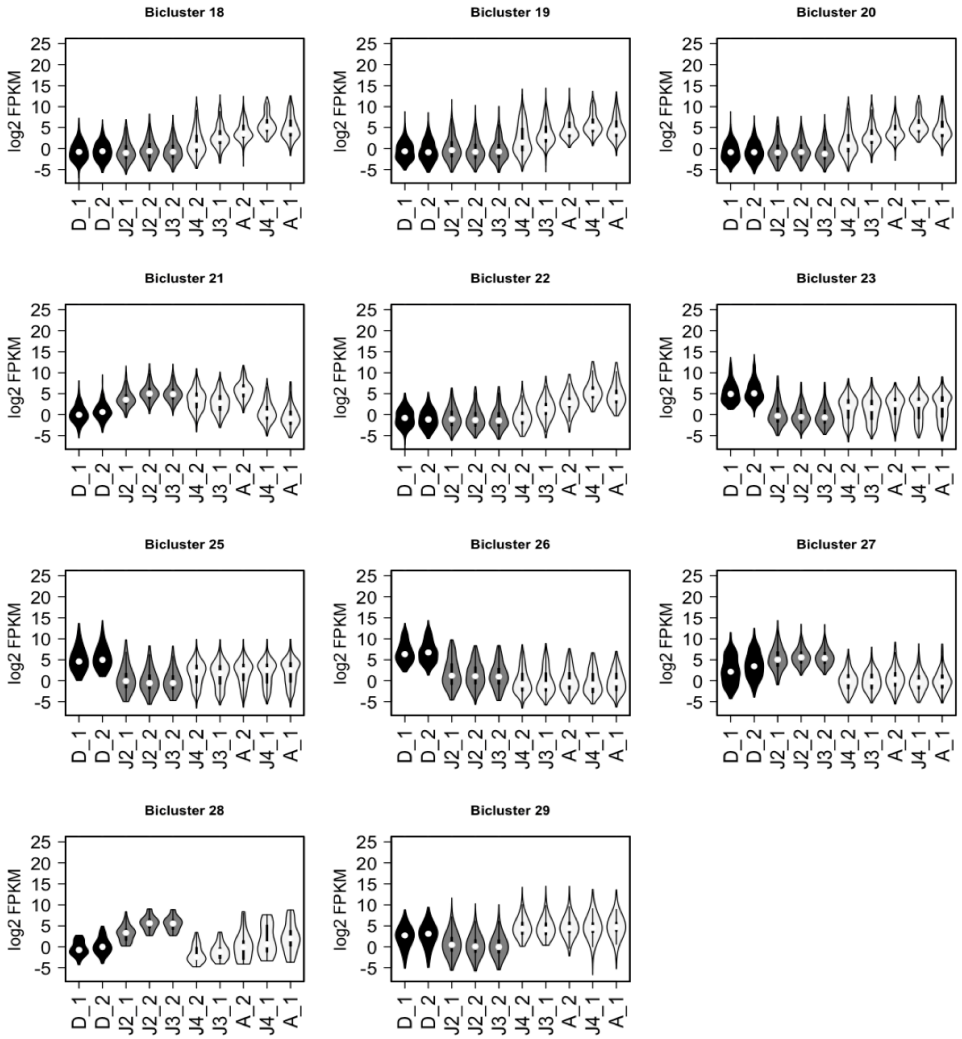


Figure 6.5: **Expression levels of bicluster genes across *P.pacificus* developmental transcriptomes.** Distribution of expression levels for all genes in each bicluster across all ten samples. The samples were grouped based on the expression levels using different computational methods and confirmed using qPCR. Black, gray and white color represent dauer larva, early larvae and adults samples respectively. Significant differential expression was observed for all biclusters in at least one comparison, which supports the interpretation that all identified biclusters represent genes that are developmentally regulated.

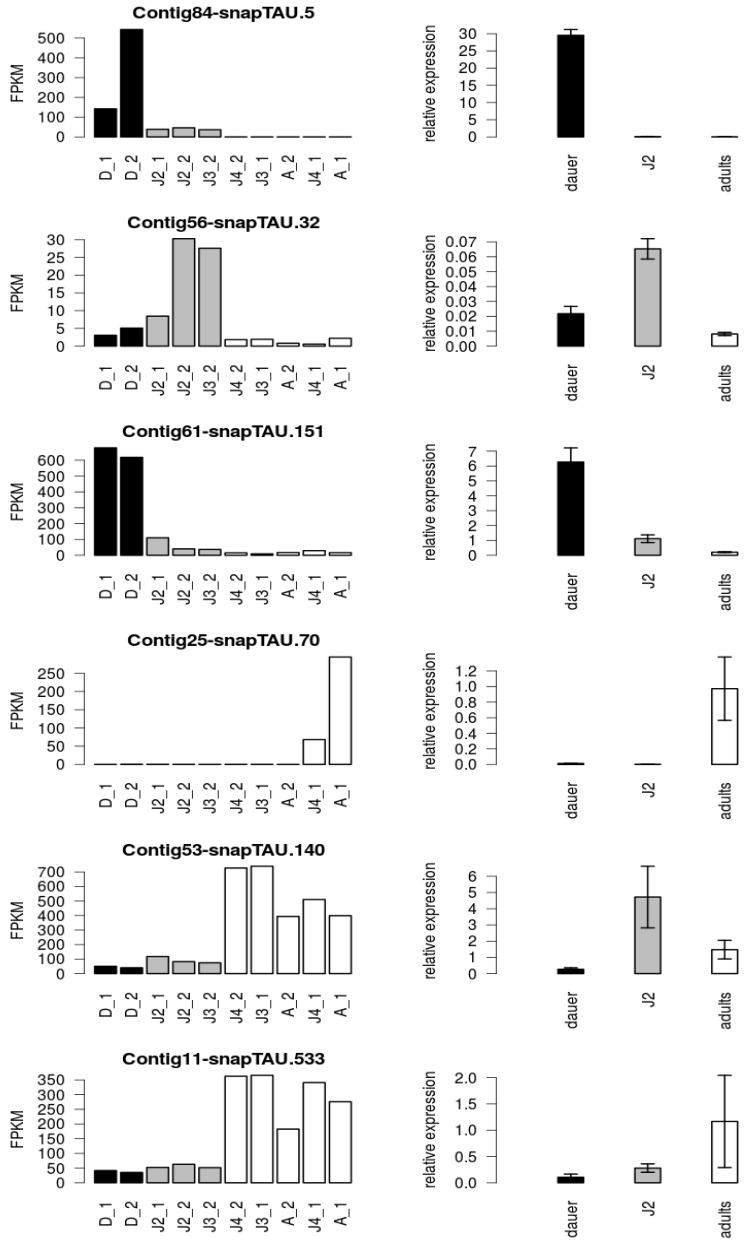


Figure 6.6: **qRT-PCR validation of developmental regulation.** Comparison of expression values of six candidate genes obtained by RNA-seq and qRT-PCR. The left column shows RNA-seq expression values (measured in FPKM) of candidate genes across all samples. The right column shows the results of qRT-PCR as relative expression (Height of the bar represent mean and error bar represent standard deviation of four biological replicates). Black, gray and white color represent dauer larva, early larvae and adults samples respectively. All the genes shows very consistent pattern of expression across the samples between both methods, except Contig53-snapTAU.140 gene.

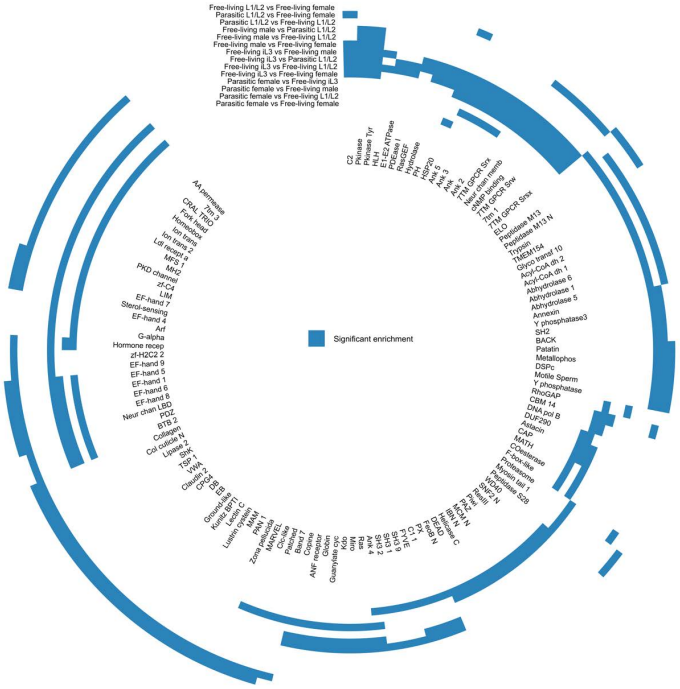


Figure 6.7: **PFAM domain enrichment among up-regulated genes.** The Heatmap shows the enrichment of different PFAM domains among genes up-regulated in different comparisons of developmental stages. Blue box indicates significant over-representation in a comparison and white box no significance.

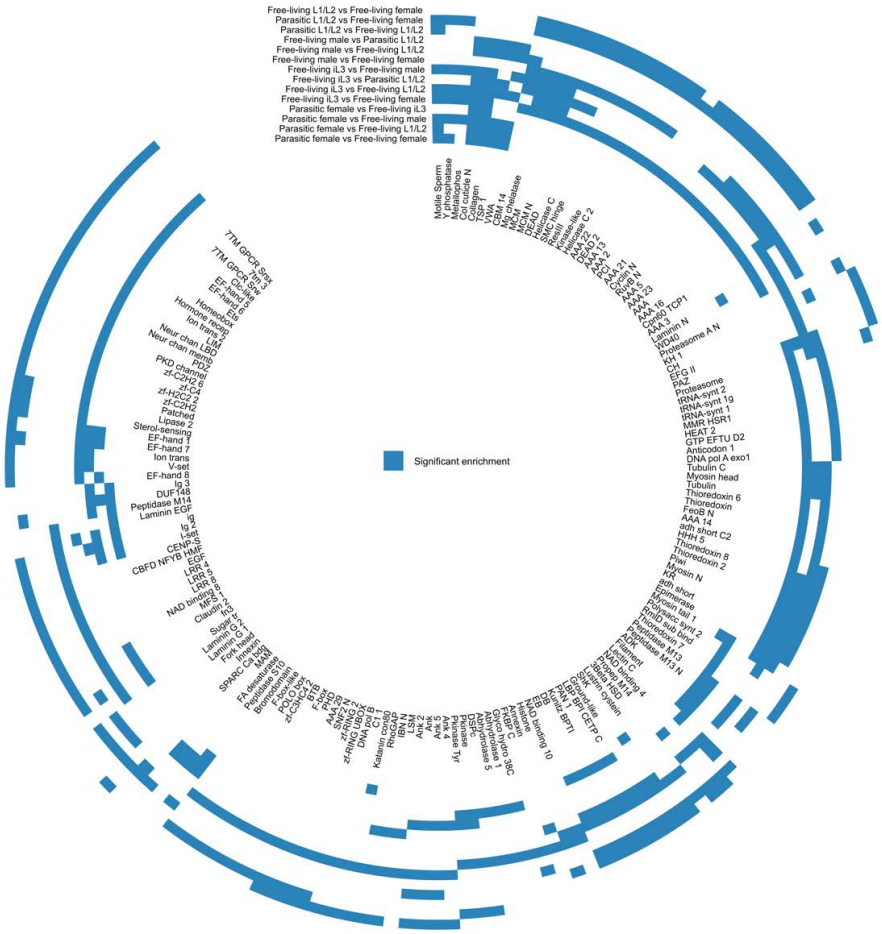


Figure 6.8: **PFAM domain enrichment among down-regulated genes.** The Heatmap shows the enrichment of different PFAM domains among genes down-regulated in different comparisons of developmental stages. Blue box indicates significant over-representation in a comparison and white box no significance.

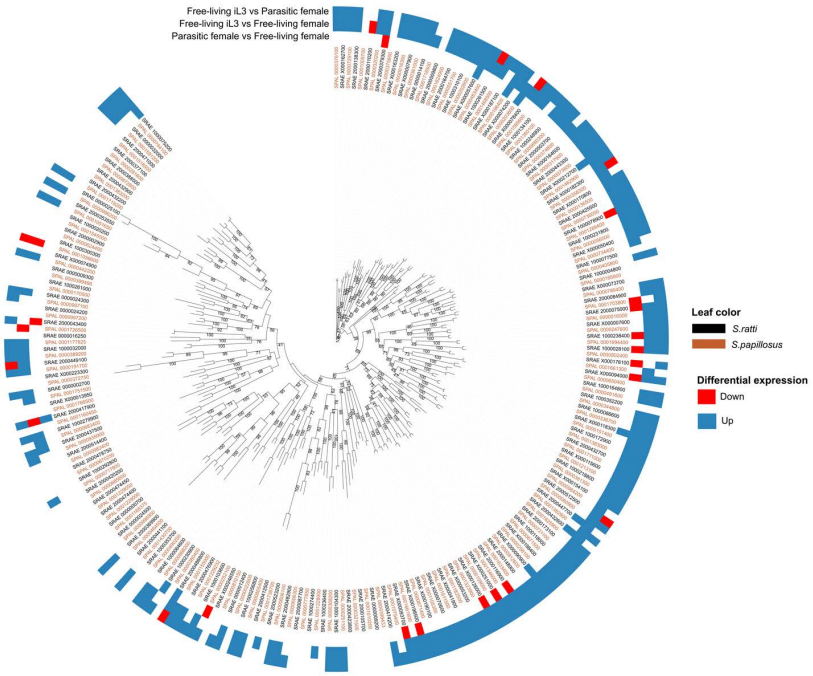


Figure 6.9: **Gene expression and sequence evolution in GPCR gene family.** Phylogenetic tree of GPCR genes from S. papillosus (green) and S. ratti (black). Genes without one-to-one orthology between S. papillosus and S. ratti were removed for the sake of simplicity of the visualization. The heatmap around the tree indicates differential expression patterns of each gene in three different comparisons. Up-regulation, down-regulation and no change in gene expression are shown in blue, red and white, respectively.

Bicluster	GO_term	Description	Enrichment	FDR
1	GO:0030030	cell projection organization	10.7	0.008
2	GO:0042303	molting cycle	6.1	1.20E-08
2	GO:0018996	molting cycle- collagen and cuticulin-based cuticle	6.1	1.20E-08
2	GO:0018988	molting cycle- protein-based cuticle	6.1	1.20E-08
2	GO:0042302	structural constituent of cuticle	23.7	1.18E-05
2	GO:0008158	hedgehog receptor activity	16.2	2.8E-04
3	GO:0007218	neuropeptide signaling pathway	41.4	1.15E-14
3	GO:0007186	G-protein coupled receptor protein signaling pathway	10.1	5.61E-12
3	GO:0007166	cell surface receptor linked signal transduction	7.6	1.20E-09
7	GO:0044421	extracellular region part	15	4.85E-06
7	GO:0005578	proteinaceous extracellular matrix	15	4.18E-05
7	GO:0031012	extracellular matrix	14.4	5.75E-05
7	GO:0044420	extracellular matrix part	26.9	8.2E-04
7	GO:0005604	basement membrane	26.9	8.2E-04
7	GO:0005576	extracellular region	5.3	0.002
8	GO:0005578	proteinaceous extracellular matrix	21.4	1.80E-06
8	GO:0031012	extracellular matrix	20.7	2.50E-06
8	GO:0044421	extracellular region part	19.3	4.61E-06
8	GO:0005576	extracellular region	6.4	0.002
8	GO:0005604	basement membrane	32.2	0.007
8	GO:0044420	extracellular matrix part	32.2	0.007
10	GO:0018988	molting cycle- protein-based cuticle	6.2	8.84E-09
10	GO:0018996	molting cycle- collagen and cuticulin-based cuticle	6.2	8.84E-09
10	GO:0042303	molting cycle	6.2	8.84E-09
10	GO:0008158	hedgehog receptor activity	16.9	2.0E-04
10	GO:0005576	extracellular region	4.9	2.5E-04
21	GO:0018988	molting cycle- protein-based cuticle	6.5	1.01E-08
21	GO:0018996	molting cycle- collagen and cuticulin-based cuticle	6.5	1.01E-08
21	GO:0042303	molting cycle	6.5	1.01E-08
21	GO:0008158	hedgehog receptor activity	19.4	7.33E-05
21	GO:0006030	chitin metabolic process	26	0.002
21	GO:0006022	aminoglycan metabolic process	24	0.003
21	GO:0005576	extracellular region	4.6	0.003
24	GO:0007186	G-protein coupled receptor protein signaling pathway	9.5	9.07E-16
24	GO:0007218	neuropeptide signaling pathway	29.8	9.99E-13
24	GO:0007166	cell surface receptor linked signal transduction	7.1	1.29E-12
Housekeeping	GO:0030529	ribonucleoprotein complex	2.1	9.85E-32
Housekeeping	GO:0005840	ribosome	2.1	4.07E-24
Housekeeping	GO:0033279	ribosomal subunit	2.3	0.006

Table 6.1: **Enrichment of gene ontology terms among developmentally regulated genes.** Overrepresentation of GO terms was done by borrowing annotations from C. elegans. For this purpose, C. elegans one-to-one orthologs of bicluster and housekeeping genes were tested for enrichment using the David functional annotation webtool using the total set of one-to-one orthologs as background. As some biclusters had only a few genes with one-to-one orthologs in C. elegans, only gene sets with at least 90 one-to-one orthologs were tested. Only GO terms with a fold enrichment greater or equal than 2 and an FDR corrected p-value below 0.01 are shown.

Bicluster	Dauer Vs Dauer_exit				Germline ablation				Response to *Xenorhabdus nematophila*			
	Up (n= 3545)		Down (n= 1394)		Up (n= 994)		Down (n=2391)		Up (n= 848)		Down (n=4921)	
	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue
1	2.9	4.4E-28	-	-	2.6	3.8E-06	-	-	6.1	1.7E-28	-	-
3	5.1	1.6E-93	-	-	3.5	3.3E-10	-	-	5.3	3.5E-20	-	-
4	2.5	2.3E-14	-	-	4.5	2.6E-14	-	-	7.7	1.1E-33	-	-
5	-	-	2.7	0.002	3.3	9.8E-04	-	-	8.7	2.7E-16	-	-
6	1.5	1.2E-04	-	-	5.1	1.3E-36	2.1	5.9E-11	-	-	2.3	2.2E-28
7	-	-	9.6	3.4E-272	2	8.5E-06	-	-	-	-	3.7	1.3E-146
8	-	-	10	3.7E-266	2.3	7.7E-09	-	-	-	-	4	2.0E-170
9	-	-	7.9	4.9E-197	2	1.6E-06	-	-	2.3	8.3E-09	2.8	2.6E-75
10	-	-	2.4	3.0E-12	-	-	-	-	-	-	1.3	0.005
11	-	-	8.5	2.5E-188	1.9	4.9E-05	-	-	-	-	3.5	1.8E-118
12	-	-	2.6	1.5E-17	1.9	5.3E-05	1.4	0.001	-	-	1.8	7.6E-17
13	-	-	1.4	8.55382E-04	4.3	2.7E-59	1.5	1.1E-07	1.9	2.4E-06	1.7	7.6E-20
14	-	-	-	-	4.5	4.5E-58	1.6	3.4E-09	1.9	2.3E-06	1.7	2.2E-17
15	-	-	-	-	5.4	2.8E-41	1.8	1.3E-07	-	-	2.1	4.3E-20
17	-	-	6.2	2.9E-86	7.3	2.9E-77	-	-	3.3	1.6E-14	3.1	5.7E-68
18	-	-	-	-	11.7	1.9E-102	-	-	5.3	3.2E-23	1.7	1.3E-06
19	-	-	-	-	10.6	4.4E-90	-	-	5.6	1.4E-26	1.6	1.2E-05
20	-	-	-	-	10.9	5.3E-102	-	-	5.4	1.3E-26	1.5	3.3E-05
22	-	-	-	-	13.3	6.9E-181	-	-	4.7	1.1E-25	-	-
23	3.9	4.8E-72	-	-	-	-	-	-	-	-	-	-
24	4.9	4.8E-124	-	-	2.9	3.5E-09	-	-	4.1	5.5E-18	-	-
25	3.7	1.7E-79	-	-	-	-	-	-	-	-	-	-
26	5.5	2.6E-69	-	-	2.6	5.2E-04	-	-	-	-	-	-
27	2.4	4.7E-13	-	-	4	9.19E-12	-	-	7.4	8.49E-33	-	-
29	-	-	3	3.1E-23	-	-	1.6	3.3E-05	-	-	2	7.2E-23

Bicluster	Response *to Serratia marcescens*				Response to *Staphylococcus aureus*				Response to *Bacillus thuringiensis*			
	Up (n= 192)		Down (n=1006)		Up (n= 178)		Down (n= 140)		Up (n= 156)		Down (n=61)	
	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue	Enrichment	pvalue
1	5.5	2.1E-06	-	-	15	2.1E-26	-	-	6.3	1.7E-06	-	-
3	6.1	8.5E-07	-	-	13.1	7.1E-20	-	-	5	2.3E-04	-	-
4	9.5	1.6E-11	-	-	19.9	3.8E-31	-	-	11.7	6.6E-13	-	-
5	6.4	0.003	-	-	12.1	1.9E-06	-	-	7.9	0.002	15.2	0.001
6	-	-	4.3	4.7E-27	-	-	6.8	6.8E-10	-	-	-	-
7	5.2	2.8E-12	3.5	2.1E-26	5	1.1E-10	6.6	7.1E-14	7.6	3.0E-19	8.5	7.05E-10
8	5.2	1.8E-11	4.4	1.4E-39	5.4	2.1E-11	10.6	1.9E-27	6.4	1.5E-13	9.9	1.94E-11
9	5.2	8.7E-13	2.1	4.3E-08	5.6	1.3E-13	3.6	3.9E-05	9	4.0E-26	6.5	8.60E-07
10	-	-	-	-	-	-	4.4	9.1E-06	-	-	5.4	2.9E-04
11	3.7	3.8E-06	4.1	3.2E-32	2.6	0.003	12	6.9E-32	4.9	3.7E-08	18	2.1 E-26
12	-	-	3.9	1.2E-30	-	-	5.1	1.8E-08	-	-	-	-
13	-	-	3.5	2.4E-40	-	-	4.9	2.5E-12	-	-	-	-
14	-	-	3.4	2.8E-33	-	-	4.9	4.1E-11	-	-	-	-
15	-	-	3.8	1.7E-21	-	-	5.8	5.6E-08	-	-	-	-
17	2.9	0.001627387	4.4	2.6E-31	-	-	13.7	2.4E-32	3.2	0.001	14.1	2.1 E-15
18	-	-	2.7	4.7E-07	-	-	6.6	9.4E-07	-	-	-	-
19	-	-	2.6	1.0E-06	-	-	7	1.7E-07	-	-	-	-
20	-	-	2.5	1.6E-06	-	-	5.4	1.7E-05	-	-	-	-
21	-	-	-	-	-	-	4.5	0.001	-	-	-	-
24	3.9	1.5E-04	-	-	7.6	2.8E-12	-	-	-	-	-	-
27	7.5	2.6E-08	-	-	18.6	3.1E-29	-	-	10.6	1.52E-11	-	-
29	-	-	3.8	1.3E-27	-	-	5.4	6.2E-09	-	-	-	-

Table 6.2: **Comparisons with previous *P. pacificus* transcriptome profiles** Comparisons with previous P.pacificus trancriptome profiles were done to test for enrichment of biclustered genes in various experimental studies. The data used in this comparisons were differentially expressed gene lists from studies on dauer vs dauer exit worms, germline ablated worms and worms exposed to four different bacterial pathogens. P-values were computed using Fisher?s exact test and FDR for multiple testing corrections. Only biclusters with enrichment value greater than 1 and p-value less than 0.01 are shown.

Chapter 7

Contributions

A notable amount of ideas for this work was developed in discussion with Dr. Christian Rödelsperger (CR) and Prof. Ralf Sommer (RJ). Most single ideas are not traceable to one person as they developed over time, but it is noteworthy that this dissertation is not conceivable without their contributions.

CHAPTER 2, 3 and 4

This work is part of the manuscripts that have been published in Journal of Molecular Evolution [Markov et al., 2015], BMC Evolutionary Biology [Baskaran et al., 2015], PLoS ONE [Baskaran and Rödelsperger, and Genome Biology and Evolution [Baskaran et al., 2017]. Permission to reproduce the text and figures has been obtained from the publishers. These published work were carried out in collaboration with Christian Rödelsperger, Teggen Jaleta (TJ), Gabriel Markov (GM), Neel Prabh (NP), Antje Hirsekorn (AH), Christoph Dieterich (CD), Adrian Streit (AS) and Ralf Sommer. CR, AS, CD and RJ conceived and designed the projects. I and CR carried out the genomics, transcriptomics data analysis. I and GM carried out evolutionary analysis, and interpretation. TJ, NP and AH performed the wet lab experiments. I, CR, TJ, GM wrote the manuscripts. I, CR and GM created the figures.

Bibliography

[Abad et al., 2008] Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V. C., Caillaud, M.-C., Coutinho, P. M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J. V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T. R., Markov, G. V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T. J., Blaxter, M., Bleve-Zacheo, T., Davis, E. L., Ewbank, J. J., Favery, B., Grenier, E., Henrissat, B., Jones, J. T., Laudet, V., Maule, A. G., Quesneville, H., Rosso, M.-N., Schiex, T., Smant, G., Weissenbach, J., and Wincker, P. (2008). Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nature biotechnology*, 26(8):909–915.

[Abzhanov et al., 2004] Abzhanov, A., Protas, M., Grant, B. R., Grant, P. R., and Tabin, C. J. (2004). Bmp4 and Morphological Variation of Beaks in Darwin's Finches. *Science*, 305(5689).

[Adler et al., 2014] Adler, M., Anjum, M., Berg, O., Andersson, D. I., and Sandegren, L. (2014). High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Mol Biol Evol*, 31.

[Alexeyenko et al., 2006] Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. In *Bioinformatics*, volume 22, pages 9–15.

[Altenhoff et al., 2012] Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Computational Biology*, 8(5):e1002514.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.

[Anavy et al., 2014] Anavy, L., Levin, M., Khair, S., Nakanishi, N., Fernandez-Valverde, S. L., Degnan, B. M., and Yanai, I. (2014). BLIND ordering of large-scale transcriptomic developmental timecourses. *Development*, 141(5).

[Asojo et al., 2007] Asojo, O. A., Homma, K., Sedlacek, M., Ngamelue, M., Goud, G. N., Zhan, B., Deumic, V., Asojo, O., and Hotez, P. J. (2007). X-ray structures of Na-GST-1 and Na-GST-2 two glutathione s-transferase from the human hookworm Necator americanus. *BMC Structural Biology*, 7(1):42.

[Assis and Bachtrog, 2013] Assis, R. and Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila. *Proceedings of the National Academy of Sciences*, 110(43):17409–17414.

[Bailey and Eichler, 2006] Bailey, J. a. and Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews. Genetics*, 7(7):552–64.

[Barrière and Félix, 2006] Barrière, A. and Félix, M.-A. (2006). Isolation of C. elegans and related nematodes. *WormBook : the online review of C. elegans biology*, pages 1–9.

[Baskaran et al., 2017] Baskaran, P., Jaleta, T. G., Streit, A., and Rödelsperger, C. (2017). Duplications and Positive Selection Drive the Evolution of Parasitism-Associated Gene Families in the Nematode Strongyloides papillosus. *Genome Biology and Evolution*, 9(3):790–801.

[Baskaran and Rödelsperger, 2015] Baskaran, P. and Rödelsperger, C. (2015). Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus. *PloS one*, 10(6):e0131136.

[Baskaran et al., 2015] Baskaran, P., Rödelsperger, C., Prabh, N., Serobyan, V., Markov, G. V., Hirsekorn, A., and Dieterich, C. (2015). Ancient gene duplications have shaped developmental stage-specific expression in Pristionchus pacificus. *BMC Evol Biol*, 15(1).

[Bedford and Hartl, 2009] Bedford, T. and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences*, 106(4):1133–1138.

[Bento et al., 2010] Bento, G., Ogawa, A., and Sommer, R. J. (2010). Co-option of the hormone-signalling module dafachronic acidDAF-12 in nematode evolution. *Nature*, 466(7305):494–497.

[Bergthorsson et al., 2007] Bergthorsson, U., Andersson, D. I., and Roth, J. R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A*, 104.

[Blaxter and Koutsovoulos, 2015] Blaxter, M. and Koutsovoulos, G. (2015). The evolution of parasitism in Nematoda. *Parasitology*, (Suppl 1):S26–39.

[Blaxter, 2003] Blaxter, M. L. (2003). Nematoda: genes, genomes and the evolution of parasitism. *Advances in parasitology*, 54:101–95.

[Blaxter et al., 1998] Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T., and Thomas, W. K. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature*, 392(6671):71–75.

[Borchert et al., 2010] Borchert, N., Dieterich, C., Krug, K., Schütz, W., Jung, S., Nordheim, A., Sommer, R. J., and Macek, B. (2010). Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome Research*, 20(6):837–846.

[Bose et al., 2014] Bose, N., Meyer, J. M., Yim, J. J., Mayer, M. G., Markov, G. V., Ogawa, A., Schroeder, F. C., and Sommer, R. J. (2014). Natural variation in dauer pheromone production and sensing supports intraspecific competition in nematodes. *Current Biology*, 24(13):1536–1541.

[Brown et al., 2008] Brown, C. M., Reisfeld, B., and Mayeno, A. N. (2008). Cytochromes P450: a structure-based summary of biotransformations using representative substrates. *Drug Metabolism Reviews*, 40(sup3):1–288.

[Cao et al., 2013] Cao, H., Shockey, J. M., Klasson, K. T., Chapital, D. C., Mason, C. B., and Scheffler, B. E. (2013). Developmental Regulation of Diacylglycerol Acyltransferase Family Gene Expression in Tung Tree Tissues. *PLoS ONE*, 8(10):e76946.

[Capella-Gutiérrez et al., 2009] Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15):1972–1973.

[Casewell et al., 2011] Casewell, N. R., Wagstaff, S. C., Harrison, R. A., Renjifo, C., and Wüster, W. (2011). Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Molecular Biology and Evolution*, 28(9):2637–2649.

[Chen and Zhang, 2012] Chen, X. and Zhang, J. (2012). The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. *PLoS Computational Biology*, 8(11):e1002784.

[Coghlan, 2005] Coghlan, A. (2005). Nematode genome evolution. *WormBook : the online review of C. elegans biology*, pages 1–15.

[Conant and Wolfe, 2008] Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature reviews. Genetics*, 9(12):938–50.

[Corsi et al., 2015] Corsi, A. K., Wightman, B., and Chalfie, M. (2015). A transparent window into biology: A primer on Caenorhabditis elegans. *Genetics*, 200(2):387–407.

[Darriba et al., 2011] Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165.

[De Ley, 2006] De Ley, P. (2006). A quick tour of nematode diversity and the backbone of nematode phylogeny. *WormBook*.

[Deng et al., 2010] Deng, C., Cheng, C.-H. C., Ye, H., He, X., and Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences*, 107(50):21593–21598.

[Denver et al., 2011] Denver, D., Clark, K., and Raboin, M. (2011). Reproductive mode evolution in nematodes: Insights from molecular phylogenies and recently discovered species. *Molecular Phylogenetics and Evolution*, 61(2):584–592.

[DH and Fitch, 2005] DH, F. and Fitch, D. H. A. (2005). Evolution: An ecological context for C. elegans. *Curr Biol*, 15(17).

[Dieterich et al., 2008] Dieterich, C., Clifton, S. W., Schuster, L. N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte, H., Yang, S.-P., Wilson, R. K., and Sommer, R. J. (2008). The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nature Genetics*, 40(10):1193–1198.

[Dieterich and Sommer, 2009] Dieterich, C. and Sommer, R. J. (2009). How to become a parasite - lessons from the genomes of nematodes. *Trends in genetics : TIG*, 25(5):203–9.

[Dobin et al., 2013] 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(1):15–21.

[Dulai et al., 1999] Dulai, K. S., von Dornum, M., Mollon, J. D., and Hunt, D. M. (1999). The Evolution of Trichromatic Color Vision by Opsin Gene Duplication in New World and Old World Primates. *Genome Research*, 9(7):629–638.

[Durbin et al., 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis*. Cambridge University Press.

[Eichler, 2001] Eichler, E. E. (2001). Recent Duplication and the Dynamic Mutation of the Human Genome. *Trends in Genetics*, 17(11):661–669.

[Emanuel and Shaikh, 2001] Emanuel, B. S. and Shaikh, T. H. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat Rev Genet*, 2(10):791–800.

[Escriva et al., 2006] Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., and Laudet, V. (2006). Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS genetics*, 2(7):e102.

[Farslow et al., 2015] Farslow, J. C., Lipinski, K. J., Packard, L. B., Edgley, M. L., Taylor, J., Flibotte, S., Moerman, D. G., Katju, V., and Bergthorsson, U. (2015). Rapid Increase in frequency of gene copy-number variants during experimental evolution in Caenorhabditis elegans. *BMC Genomics*, 16(1):1044.

[Fierst et al., 2015] Fierst, J. L., Willis, J. H., Thomas, C. G., Wang, W., Reynolds, R. M., Ahearne, T. E., Cutter, A. D., and Phillips, P. C. (2015). Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. *PLoS genetics*, 11(6):e1005323.

[Fitch, 1970] Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2):99–113.

[Force et al., 1999] Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-L. L., Postlethwait, J., Allendorf, F. W., Utter, F. M., May, B. P., Gerhart, J., Kirschner, M., Graf, J. D., Kobel, H. R., Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A., Sidow, A., Jowett, T., Mancera, M., Amores, A., Yan, Y.-L. L., Kimura, M., Lewis, W. H., Lynch, M., Walsh, J. B., Ohno, S., and Raff, R. A. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45.

[Ghedin et al., 2007] Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L., Guiliano, D. B., Miranda-Saavedra, D., Angiuoli, S. V., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N. M., Wortman, J. R., Feldblyum, T., Tallon, L., Schatz, M., Shumway, M., Koo, H., Salzberg, S. L., Schobel, S., Pertea, M., Pop, M., White, O., Barton, G. J., Carlow, C. K. S., Crawford, M. J., Daub, J., Dimmic, M. W., Estes, C. F., Foster, J. M., Ganatra, M., Gregory, W. F., Johnson, N. M., Jin, J., Komuniecki, R., Korf, I., Kumar, S., Laney, S.,

Li, B.-W., Li, W., Lindblom, T. H., Lustigman, S., Ma, D., Maina, C. V., Martin, D. M. A., McCarter, J. P., McReynolds, L., Mitreva, M., Nutman, T. B., Parkinson, J., Peregrín-Alvarez, J. M., Poole, C., Ren, Q., Saunders, L., Sluder, A. E., Smith, K., Stanke, M., Unnasch, T. R., Ware, J., Wei, A. D., Weil, G., Williams, D. J., Zhang, Y., Williams, S. A., Fraser-Liggett, C., Slatko, B., Blaxter, M. L., and Scott, A. L. (2007). Draft genome of the filarial nematode parasite Brugia malayi. *Science (New York, N.Y.)*, 317(5845):1756–60.

[Gouy et al., 2010] Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2):221–224.

[Handsaker et al., 2015] Handsaker, R. E., Doren, V. V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., and McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3):296–303.

[Herrmann et al., 2006] Herrmann, M., Mayer, W. E., and Sommer, R. J. (2006). Nematodes of the genus Pristionchus are closely associated with scarab beetles and the Colorado potato beetle in Western Europe. *Zoology*, 109(2):96–108.

[Hillier et al., 2005] Hillier, L. W., Coulson, A., Murray, J. I., Bao, Z., Sulston, J. E., and Waterston, R. H. (2005). Genomics in C. elegans: So many genes, such a little worm.

[Hong and Sommer, 2006] Hong, R. L. and Sommer, R. J. (2006). Pristionchus pacificus: A well-rounded nematode.

[Huang et al., 2009] Huang, D. W., Lempicki, R. a., and Sherman, B. T. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

[Hughes and Kaufman, 2002] Hughes, C. L. and Kaufman, T. C. (2002). Hox genes and the evolution of the arthropod body plan. *Evolution & Development*, 4:459–499.

[Huminiecki et al., 2009] Huminiecki, L., Goldovsky, L., Freilich, S., Moustakas, A., Ouzounis, C., and Heldin, C.-H. (2009). Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom. *BMC evolutionary biology*, 9:28.

[Hunt et al., 2016] Hunt, V. L., Tsai, I. J., Coghlan, A., Reid, A. J., Holroyd, N., Foth, B. J., Tracey, A., Cotton, J. A., Stanley, E. J., Beasley, H., Bennett, H. M., Brooks, K., Harsha, B., Kajitani, R., Kulkarni, A., Harbecke, D., Nagayasu, E., Nichol, S., Ogura, Y., Quail, M. A., Randle, N., Xia, D., Brattig, N. W., Soblik, H., Ribeiro, D. M., Sanchez-Flores, A., Hayashi, T., Itoh, T., Denver, D. R., Grant, W., Stoltzfus, J. D., Lok, J. B., Murayama, H., Wastling, J., Streit, A., Kikuchi, T., Viney, M., and Berriman, M. (2016). The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nature Genetics*, 48(3):299–307.

[Hurles, 2004] Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. *PLoS biology*, 2(7):E206.

[Innan and Kondrashov, 2010] Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews. Genetics*, 11(2):97–108.

[Jäger et al., 2005] Jäger, M., Gauly, M., Bauer, C., Failing, K., Erhardt, G., and Zahner, H. (2005). Endoparasites in calves of beef cattle herds: management systems dependent and genetic influences. *Vet Parasitol*, 131(3-4):173–191.

[Johnson et al., 2010] Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, 11:431.

[Kaessmann, 2010] Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes.

[Kaletta and Hengartner, 2006] Kaletta, T. and Hengartner, M. O. (2006). Finding function in novel targets: C. elegans as a model organism. *Nature reviews. Drug discovery*, 5(5):387–98.

[Kanzaki et al., 2012] Kanzaki, N., Ragsdale, E. J., Herrmann, M., Mayer, W. E., and Sommer, R. J. (2012). Description of Three Pristionchus Species (Nematoda: Diplogastridae) from Japan that Form a Cryptic Species Complex with the Model Organism P. pacificus. *Zoological Science*, 29(6):403–417.

[Katju and Bergthorsson, 2013] Katju, V. and Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet*, 4.

[Katju and Lynch, 2003] Katju, V. and Lynch, M. (2003). The structure and early evolution of recently arisen gene duplication in the Caenorhabditis elegans genome. *Genetics*, 165.

[Khaitovich et al., 2004] Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Pääbo, S. (2004). A Neutral Model of Transcriptome Evolution. *PLoS Biology*, 2(5):e132.

[Kieninger et al., 2016] Kieninger, M., Ivers, N., Rödelsperger, C., Markov, G., Sommer, R., and Ragsdale, E. (2016). The Nuclear Hormone Receptor NHR-40 Acts Downstream of the Sulfatase EUD-1 as Part of a Developmental Plasticity Switch in Pristionchus. *Current Biology*, 26(16):2174–2179.

[Kim et al., 2013] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36.

[Kimura, 1983] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

[Kisiela et al., 2011] Kisiela, M., El-Hawari, Y., Martin, H.-J., and Maser, E. (2011). Bioinformatic and biochemical characterization of DCXR and DHRS2/4 from Caenorhabditis elegans. *Chemico-Biological Interactions*, 191(1):75–82.

[Kondrashov et al., 2004] Kondrashov, F. A., Koonin, E. V., Fisher, R., Fisher, R., Wright, S., Orr, A., Kacser, H., Burns, J., Cornish-Bowden, A., Hurst, L., Randerson, J., Fisher, E., Scambler, P., Strachan, T., Read, A., Veitia, R., Papp, B., et Al., Consortium, G. O., Steinmetz, L., et Al., Jimenez-Sanchez, G., et Al., Haldane, J., Fisher, R., Ohno, S., Lynch, M., Force, A., Moore, R., Purugganan, M., Koonin, E., et Al., Wheeler, D., et Al., Boeckmann, B., et Al., Goffeau, A., and et Al. (2004). A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in genetics : TIG*, 20(7):287–90.

[Koonin and Galperin, 2003] Koonin, E. V. and Galperin, M. Y. (2003). *Sequence - Evolution - Function*. Springer New York.

[LaMunyon and Ward, 1999] LaMunyon, C. W. and Ward, S. (1999). Evolution of sperm size in nematodes: sperm competition favours larger sperm. *Proceedings. Biological sciences*, 266(1416):263–7.

[Lander et al., 2010] Lander, N., Bernal, C., Diez, N., Añez, N., Docampo, R., and Ramírez, J. L. (2010). Localization and developmental regulation of a dispersed gene family 1 protein in Trypanosoma cruzi. *Infection and immunity*, 78(1):231–40.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359.

[Letunic and Bork, 2011] Letunic, I. and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, 39(suppl):W475—-W478.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Li, 2003] Li, L. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9):2178–2189.

[Librado et al., 2012] Librado, P., Vieira, F. G., and Rozas, J. (2012). BadiRate: Estimating family turnover rates by likelihood-based methods. *Bioinformatics*, 28(2):279–281.

[Lindblom and Dodd, 2006] Lindblom, T. H. and Dodd, A. K. (2006). Xenobiotic detoxification in the nematodeCaenorhabditis elegans. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305A(9):720–730.

[Lunter and Goodson, 2011] Lunter, G. and Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939.

[Lynch and Conery, 2000] Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*, 290(5494):1151–5.

[Lynch and Force, 2000] Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473.

[Lynch and Katju, 2004] Lynch, M. and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates.

[Lynch et al., 2001] Lynch, M., O'Hely, M., Walsh, B., and Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804.

[Makova and Li, 2003] Makova, K. D. and Li, W. H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Research*, 13(7):1638–1645.

[Markov et al., 2015] Markov, G. V., Baskaran, P., and Sommer, R. J. (2015). The Same or Not the Same: Lineage-Specific Gene Expansions and Homology Relationships in Multigene Families in Nematodes. *Journal of Molecular Evolution*, 80(1):18–36.

[Markov et al., 2016] Markov, G. V., Meyer, J. M., Panda, O., Artyukhin, A. B., Claaßen, M., Witte, H., Schroeder, F. C., and Sommer, R. J. (2016). Functional Conservation and Divergence of daf-22 Paralogs in Pristionchus pacificus Dauer Development. *Molecular biology and evolution*, 33(10):2506–14.

[Maydan et al., 2007] Maydan, J. S., Flibotte, S., Edgley, M. L., Lau, J., Selzer, R. R., Richmond, T. A., Pofahl, N. J., Thomas, J. H., and Moerman, D. G. (2007). Efficient high-resolution deletion discovery in Caenorhabditis elegans by array comparative genomic hybridization. *Genome Res*, 17(3):337–347.

[Maydan et al., 2010] Maydan, J. S., Lorch, A., Edgley, M. L., Flibotte, S., and Moerman, D. G. (2010). Copy number variation in the genomes of twelve natural isolates of Caenorhabditis elegans. *BMC Genomics*, 11.

[Mayer et al., 2015] Mayer, M. G. M. M. G., Rödelsperger, C., Witte, H., Riebesell, M., and Sommer, R. J. (2015). The Orphan Gene dauerless Regulates Dauer Development and Intraspecific Competition in Nematodes by Copy Number Variation. *PLoS genetics*, 11(6):e1005146.

[Mayer et al., 2011] Mayer, W. E., Schuster, L. N., Bartelmes, G., Dieterich, C., and Sommer, R. J. (2011). Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. *BMC evolutionary biology*, 11(1):13.

[McGaughran et al., 2013] McGaughran, A., Morgan, K., and Sommer, R. J. (2013). Natural variation in chemosensation: lessons from an island nematode. *Ecology and evolution*, 3(16):5209–24.

[McGaughran et al., 2016] McGaughran, A., Rödelsperger, C., Grimm, D. G., Meyer, J. M., Moreno, E., Morgan, K., Leaver, M., Serobyan, V., Rakitsch, B., Borgwardt, K. M., and Sommer, R. J. (2016). Genomic Profiles of Diversification and Genotype-Phenotype Association in Island Nematode Lineages. *Molecular biology and evolution*, 33(9):2257–72.

[McGaughran and Sommer, 2014] McGaughran, A. and Sommer, R. J. (2014). Natural variation in cold tolerance in the nematode Pristionchus pacificus: the role of genotype and environment. *Biology open*, 3(9):832–8.

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

[Meyer et al., 2016] Meyer, J. M., Markov, G. V., Baskaran, P., Herrmann, M., Sommer, R. J., and Rödelsperger, C. (2016). Draft Genome of the Scarab Beetle Oryctes borbonicus on La Réunion Island. *Genome biology and evolution*, 8(7):2093–105.

[Mitreva et al., 2011] Mitreva, M., Jasmer, D. P., Zarlenga, D. S., Wang, Z., Abubucker, S., Martin, J., Taylor, C. M., Yin, Y., Fulton, L., Minx, P., Yang, S.-P., Warren, W. C., Fulton, R. S., Bhonagiri, V., Zhang, X., Hallsworth-Pepin, K., Clifton, S. W., McCarter, J. P., Appleton, J., Mardis, E. R., and Wilson, R. K. (2011). The draft genome of the parasitic nematode Trichinella spiralis. *Nature Genetics*, 43(3):228–235.

[Moreno et al., 2016] Moreno, E., McGaughran, A., Rödelsperger, C., Zimmer, M., and Sommer, R. J. (2016). Oxygen-induced social behaviours in Pristionchus pacificus have a distinct evolutionary history and genetic regulation from Caenorhabditis elegans. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1825).

[Nehrt et al., 2011] Nehrt, N. L., Clark, W. T., Radivojac, P., and Hahn, M. W. (2011). Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Computational Biology*, 7(6):e1002073.

[Nielsen, 2005] Nielsen, R. (2005). *Statistical Methods in Molecular Evolution*. Statistics for Biology and Health. Springer New York, New York, NY.

[Ohno, 1970] Ohno, S. (1970). *Evolution by gene duplication*. Springer Verlag, Berlin.

[Opperman et al., 2008] Opperman, C. H., Bird, D. M., Williamson, V. M., Rokhsar, D. S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T. D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B. R., Thomas, V. P., and Windham, E. (2008). Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39):14802–7.

[Paradis et al., 2004] Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

[Parker et al., 2004] Parker, J. A., Holbert, S., Lambert, E., Abderrahmane, S., and Néri, C. (2004). Genetic and pharmacological suppression of polyglutamine-dependent neuronal dysfunction in Caenorhabditis elegans. *Journal of molecular neuroscience : MN*, 23(1-2):61–8.

[Perbandt et al., 2005] Perbandt, M., Hoppner, J., Betzel, C., Walter, R. D., and Liebau, E. (2005). Structure of the Major Cytosolic Glutathione S-Transferase from the Parasitic Nematode Onchocerca volvulus. *Journal of Biological Chemistry*, 280(13):12630–12636.

[Perry et al., 2007] Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., and Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256–60.

[Prabh and Rödelsperger, 2016] Prabh, N. and Rödelsperger, C. (2016). Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics*, 17(1):226.

[Prince and Pickett, 2002] Prince, V. E. and Pickett, F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics*, 3(11):827–837.

[Qian and Zhang, 2008] Qian, W. and Zhang, J. (2008). Gene dosage and gene duplicability. *Genetics*, 179(4):2319–2324.

[R Core Team, 2014] R Core Team (2014). R: A Language and Environment for Statistical Computing.

[Rae et al., 2012] Rae, R., Sinha, A., and Sommer, R. J. (2012). Genome-Wide Analysis of Germline Signaling Genes Regulating Longevity and Innate Immunity in the Nematode Pristionchus pacificus. *PLoS Pathogens*, 8(8).

[Ragsdale et al., 2013] Ragsdale, E., Müller, M., Rödelsperger, C., and Sommer, R. (2013). A Developmental Switch Coupled to the Evolution of Plasticity Acts through a Sulfatase. *Cell*, 155(4):922–933.

[Ragsdale and Ivers, 2016] Ragsdale, E. J. and Ivers, N. A. (2016). Specialization of a polyphenism switch gene following serial duplications in Pristionchus nematodes. *Evolution*, 70(9):2155–2166.

[Remm et al., 2001] Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052.

[Rödelsperger and Dieterich, 2010] Rödelsperger, C. and Dieterich, C. (2010). CYNTENATOR: Progressive gene order alignment of 17 vertebrate genomes. *PLoS ONE*, 5(1).

[Rödelsperger et al., 2014] Rödelsperger, C., Neher, R. A., Weller, A. M., Eberhardt, G., Witte, H., Mayer, W. E., Dieterich, C., and Sommer, R. J. (2014). Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data. *Genetics*, 196(4):1153–1165.

[Rödelsperger and Sommer, 2011] Rödelsperger, C. and Sommer, R. J. (2011). Computational archaeology of the Pristionchus pacificus genome reveals evidence of horizontal gene transfers from insects. *BMC evolutionary biology*, 11(1):239.

[Rödelsperger et al., 2013] Rödelsperger, C., Streit, A., Sommer, R. J., Rödelsperger, C., Streit, A., and Sommer, R. J. (2013). Structure, Function and Evolution of The Nematode Genome. In *eLS*. John Wiley & Sons, Ltd, Chichester, UK.

[Rogozin, 2014] Rogozin, I. B. (2014). Complexity of gene expression evolution after duplication: protein dosage rebalancing. *Genetics research international*, 2014:516508.

[Rogozin et al., 2014] Rogozin, I. B., Managadze, D., Shabalina, S. A., and Koonin, E. V. (2014). Gene Family Level Comparative Analysis of Gene Expression in Mammals Validates the Ortholog Conjecture. *Genome Biology and Evolution*, 6(4):754–762.

[Rohlfs et al., 2014] Rohlfs, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular biology and evolution*, 31(1):201–11.

[Sanghvi et al., 2016] Sanghvi, G. V., Baskaran, P., Röseler, W., Sieriebriennikov, B., Rödelsperger, C., and Sommer, R. J. (2016). Life History Responses and Gene Expression Profiles of the Nematode Pristionchus pacificus Cultured on Cryptococcus Yeasts. *PLOS ONE*, 11(10):e0164881.

[Schliep, 2010] Schliep, K. P. (2010). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

[Schuster-Böckler et al., 2010] Schuster-Böckler, B., Conrad, D., and Bateman, A. (2010). Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions. *PLoS ONE*, 5(3):e9474.

[Shaye and Greenwald, 2011] Shaye, D. D. and Greenwald, I. (2011). OrthoList: A Compendium of C. elegans Genes with Human Orthologs. *PLoS ONE*, 6(5):e20085.

[Sheehan et al., 2001] Sheehan, D., Meade, G., Foley, V. M., and Dowd, C. A. (2001). Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *The Biochemical journal*, 360(Pt 1):1–16.

[Sheps et al., 2004] Sheps, J. A., Ralph, S., Zhao, Z., Baillie, D. L., and Ling, V. (2004). The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. *Genome Biology*, 5(3):R15.

[Sievers et al., 2014] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. (2014). Fast scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539.

[Sinha et al., 2012a] Sinha, A., Rae, R., Iatsenko, I., and Sommer, R. J. (2012a). System Wide Analysis of the Evolution of Innate Immunity in the Nematode Model Species Caenorhabditis elegans and Pristionchus pacificus. *PLoS ONE*, 7(9).

[Sinha et al., 2012b] Sinha, A., Sommer, R. J., and Dieterich, C. (2012b). Divergent gene expression in the conserved dauer stage of the nematodes Pristionchus pacificus and Caenorhabditis elegans. *BMC genomics*, 13(1):254.

[Sommer, 2006] Sommer, R. J. (2006). Pristionchus pacificus. *WormBook*, pages 1–8.

[Sommer and McGaughran, 2013] Sommer, R. J. and McGaughran, A. (2013). The nematode Pristionchus pacificus as a model system for integrative studies in evolutionary biology. *Molecular Ecology*, 22(9):2380–2393.

[Sommer and Streit, 2011] Sommer, R. J. and Streit, A. (2011). Comparative Genetics and Genomics of Nematodes: Genome Structure, Development, and Lifestyle. *Annual Review of Genetics*, 45(1):1–20.

[Soshnikova et al., 2013] Soshnikova, N., Dewaele, R., Janvier, P., Krumlauf, R., and Duboule, D. (2013). Duplications of hox gene clusters and the emergence of vertebrates. *Developmental Biology*, 378(2):194–199.

[Srinivasan et al., 2003] Srinivasan, J., Sinz, W., Jesse, T., Wiggers-Perebolte, L., Jansen, K., Buntjer, J., Van Der Meulen, M., and Sommer, R. J. (2003). An integrated physical and genetic map of the nematode Pristionchus pacificus. *Molecular Genetics and Genomics*, 269(5):715–722.

[Stamatakis, 2014] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

[Stein et al., 2003] Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D. H. A., Fulton, L. A., Fulton, R. E., Griffiths-Jones, S., Harris, T. W., Hillier, L. W., Kamath, R., Kuwabara, P. E., Mardis, E. R., Marra, M. A., Miner, T. L., Minx, P., Mullikin, J. C., Plumb, R. W., Rogers, J., Schein, J. E., Sohrmann, M., Spieth, J., Stajich, J. E., Wei, C., Willey, D., Wilson, R. K., Durbin, R., and Waterston, R. H. (2003). The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics. *PLoS Biology*, 1(2):e45.

[Stoltzfus et al., 2012] Stoltzfus, J. D., Minot, S., Berriman, M., Nolan, T. J., and Lok, J. B. (2012). RNAseq analysis of the parasitic nematode Strongyloides stercoralis reveals divergent regulation of canonical dauer pathways. *PLoS neglected tropical diseases*, 6(10):e1854.

[Streit, 2017] Streit, A. (2017). Genetics: modes of reproduction and genetic analysis. *Parasitology*, pages 1–11.

[Studer and Robinson-Rechavi, 2009] Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5):210–216.

[Sugino et al., 2006] Sugino, R. R. P., Innan, H., Ohta, T., Li, W.-H., Gao, L.-z., Innan, H., Sugino, R. R. P., Innan, H., Innan, H., Teshima, K., Innan, H., Ohno, S., Ohta, T., Hurst, L., Smith, N., Brown, D., et Al., Ghaemmaghami, S., et Al., Sharp, P., Li, W.-H., Papp, B., et Al., Kondrashov, F., Koonin, E., Davis, J., Petrov, D., Steinmetz, L., et Al., Innan, H., Wolfe, K., Shields, D., Kellis, M., et Al., Dietrich, F., and et Al. (2006). Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends in genetics : TIG*, 22(12):642–4.

[Suyama et al., 2006] Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(WEB. SERV. ISS.).

[Tanay et al., 2002] Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S136–S144.

[Teshima and Innan, 2008] Teshima, K. M. and Innan, H. (2008). Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*, 178(3):1385–98.

[The C. elegans Sequencing Consortium et al., 1998] The C. elegans Sequencing Consortium, Equence, C. E. S., Iology, T. O. B., The, C., Consortium, S., and Consortium, T. C. e. S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396):2012–2018.

[Thorvaldsdóttir et al., 2013] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–92.

[Trapnell et al., 2012] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–578.

[Uebbing et al., 2016] Uebbing, S., Künstner, A., Mäkinen, H., Backström, N., Bolivar, P., Burri, R., Dutoit, L., Mugal, C. F., Nater, A., Aken, B., Flicek, P., Martin, F. J., Searle, S. M. J., and Ellegren, H. (2016). Divergence in gene expression within and between two closely related flycatcher species. *Molecular Ecology*, 25(9):2015–2028.

[Ugot et al., 2001] Ugot, J.-p. H., Aujard, P. B., and Orand, S. M. (2001). Biodiversity in helminths and nematodes as a eld of study: an overview. *Nematology*, 3(3):199–208.

[van Hoof, 2005] van Hoof, A. (2005). Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics*, 171(4):1455–61.

[van Rossum et al., 2004] van Rossum, A. J., Jefferies, J. R., Rijsewijk, F. A. M., LaCourse, E. J., Teesdale-Spittle, P., Barrett, J., Tait, A., and Brophy, P. M. (2004). Binding of hematin by a new class of glutathione transferase from the blood-feeding parasitic nematode Haemonchus contortus. *Infection and immunity*, 72(5):2780–90.

[Veitia, 2005] Veitia, R. A. (2005). Gene dosage balance: deletions, duplications and dominance. *Trends in Genetics*, 21(1):33–35.

[Vergara et al., 2014] Vergara, I. a., Tarailo-Graovac, M., Frech, C., Wang, J., Qin, Z., Zhang, T., She, R., Chu, J. S. C., Wang, K., and Chen, N. (2014). Genome-wide variations in a natural isolate of the nematode Caenorhabditis elegans. *BMC genomics*, 15(1):255.

[Viney and Lok, 2015] Viney, M. E. and Lok, J. B. (2015). The biology of Strongyloides spp. *WormBook*, pages 1–17.

[Wagner et al., 2003] Wagner, G. P., Amemiya, C., and Ruddle, F. (2003). Hox cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25):14603–6.

[Ward and Moreno-Hagelsieb, 2014] Ward, N. and Moreno-Hagelsieb, G. (2014). Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss? *PLoS ONE*, 9(7):e101850.

[Watts, 2009] Watts, J. L. (2009). Fat synthesis and adiposity regulation in Caenorhabditis elegans.

[Weadick and Sommer, 2016] Weadick, C. J. and Sommer, R. J. (2016). Mating System Transitions Drive Life Span Evolution in Pristionchus Nematodes. *The American naturalist*, 187(4):517–31.

[Weller et al., 2014] Weller, A. M., Rödelsperger, C., Eberhardt, G., Molnar, R. I., and Sommer, R. J. (2014). Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode Pristionchus pacificus. *Genetics*, 196(4):1145–1152.

[Wilecki et al., 2015] Wilecki, M., Lightfoot, J. W., Susoy, V., and Sommer, R. J. (2015). Predatory feeding behaviour in Pristionchus nematodes is dependent on phenotypic plasticity and induced by serotonin. *Journal of Experimental Biology*, 218(9).

[Xie and Tammi, 2009] Xie, C. and Tammi, M. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10(1):80.

[Yang, 2007] Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

[Ye et al., 2009] Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871.

[Zhang, 2003] Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298.

[Zhao et al., 2013] Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14(Suppl 11):S1.

[Zhou et al., 2011] Zhou, J., Lemos, B., Dopman, E. B., and Hartl, D. L. (2011). Copy-number variation: The balance between gene dosage and expression in Drosophila melanogaster. *Genome Biology and Evolution*, 3(1):1014–1024.