Phylogenomic analysis of *Pristionchus* nematodes with the focus on orphan genes

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

vorgelegt von
Neel Duti Prabh
aus Sitamarhi, Bihar, Indien

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	18.06.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Daniel Huson
2. Berichterstatter:	Prof. Dr. Ralf J. Sommer
3. Berichterstatter:	Prof. Dr. Erich Bornberg-Bauer

Acknowledgements

First, I want to thank Dr. Christian Rödelsperger for guiding me throughout the course of my doctoral research and bearing with my constant chatter. His mentoring allowed me to maintain poise during the difficult periods and I have learned much from him. I thank Prof. Dr. Ralf J. Sommer for giving me the opportunity to conduct my doctoral research under his aegis and for pushing me beyond the limits of what I thought could be achieved. I thank Prof. Dr. Daniel Huson and Prof. Dr. Richard Neher for being part of my thesis advisory committee and suggesting timely course-corrections to achieve my research goal. I also thank the student and faculty members of the IMPRS “Molecules to organisms” graduate program and Dr. Adrian Streit, their questions and inputs were invaluable. I thank Dr. Sarah Danes, the IMPRS Ph.D. program coordinator, for her constant support.

Although I joined the SommerLab as a bioinformatician, a large part of my doctoral research has been spent working in the wet lab. In this regard, I like to especially thank two people, first, Dr. Vahan Serboyan for being my mentor in the wet lab and displaying confidence in me, and second, Hanh Witte for supporting me all the time. Here, I also thank our lab technicians Waltraud Röseler, Gabi Eberhardt, Heike Haussmann and Dr. Christa Lanz for their kind help. My lack of experience in the wet lab was covered through constant interactions with lab-members including Dr. James Lightfoot, Dr. Cameron Weadick, Dr. Vladislav Susoy, Dr. Jan Falke, Dr. Martin Wilecki, Dr. Michael Werner, Dr. Eduardo Moreno, Dr. Anja Holz, and Dr. Arpita Kulkarni. Here, I also thanks Dr. Amit Sinha for helping me make the decision to venture in the wet lab. I was also fortunate to get advice from Dr. Praveen Baskaran and Dr. Gabriel Markov to mitigate any problems that I faced in the dry lab. Dr. Baskaran was especially supportive of me in the dry-lab, and his inputs have allowed timely completion of my projects. I also thank Metta Riebesell, Tobias Loschko, Sandra Mäck, Aida Kalac, Nermin Akduman, Mohnnad Dardiry, Shuai Sun, Sara Wighard, Devansh Sharma, and all the other members of the SommerLab for their inputs and help.

Apart from their direct scientific contributions, I was also able to enjoy the company of my lab-mates socially. Dr. Gaurav Sanghvi, Dr. Suryesh Namdeo, and Dr. Praveen Baskaran were not only my lab mates but also my housemates. It was their presence that has ensured that I could enjoy a healthy social life and feel at home away from home. I also thank Dr. Gauri Tendulkar for being an ideal housemate and a great calming influence.

Bogdan Sieriebriennikov is the last lab mate that I would like to thank, but among all my lab mates he has had the most impact on my doctoral research. I have thoroughly enjoyed our conversations both scientific and otherwise. Although we rarely agreed, still our interaction has been a great learning experience for me. On the personal front, I feel privileged to have a friend like him. I also thank Joachim Sieler, whose support was cherished both on the professional and the personal front.

I thank my family members for their constant support. My parents have always motivated me to deliver my best efforts and this has allowed me to immerse myself deep in my work. I thank Dr. Vikrant Singh Rajpoot, Dr. Pankaj Kumar, and Dr. Nikhil Singh. This brings me to my wife Nidhi, whom I cannot thank enough for being a partner in all aspects and taking over all my responsibilities whenever needed.

Finally, I thank the German taxpayer.

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, November 2018

Neel Duti Prabh

Table of Contents

Zusammenfasssung

Der Großteil der biologischen Funktionen innerhalb einer Zelle wird von Proteinen ausgeübt. Daher, ist es unumgänglich, dass die Gene, die diese Proteine kodieren, konserviert bleiben um die ordnungsgemäße Faltung ihrer Primärsequenz zu gewährleisten. Sequenzhomologie zwischen Genen verschiedener Arten erlaubt uns, Rückschlüsse über deren gemeinsame Herkunft zu ziehen. Komparative Genomanalysen, die sich mit dem Vergleich von DNA- und Proteinsequenzen beschäftigen, haben jedoch eine Vielzahl von protein-kodierenden Genen, sogennante Waisengene, hervorgebracht, die keine Sequenzhomologie außerhalb einer Gruppe nahe verwandter Arten aufweisen. In meiner Doktorarbeit habe ich den Fadenwurm *Pristionchus pacificus* als Modellsystem benutzt, um folgende Fragestellungen zu untersuchen: Sind Waisengene echt? Wenn ja, wie alt sind sie, wie evolvieren sie und was ist ihr Ursprung? Der Reichtum an beschriebenen *Pristionchus* Arten hat mir dabei ermöglicht, basierend auf neuesten Sequenziermethoden, einen Datensatz aus zehn Genomen zu erstellen, der maximal vergleichbar ist und dem eine leiterartige Phylogenie zugrunde liegt. Basierend auf Selektionsanalysen konnte ich die protein-kodierende Natur der meisten Waisengene nachweisen, da Selektion in ihnen gegen einen Austausch von Aminosäuren wirkt. Dabei korreliert die Stärke der Selektion mit dem Alter der Gene, was zeigt, dass Waisengene schneller evolvieren als konservierte Gene. Schließlich habe ich über den Vergleich nahe verwandter Genome unterschiedliche Mechanismen für die Enstehung von Waisengenen aufdeckt. Dies zeigt, dass Waisengene sowohl durch Divergenz von Genfragmenten entstehen können, als auch komplett *de novo* aus nicht-kodierenden Sequenzen. Zusammenfassend deuten die Ergebnisse meiner Arbeit daraufhin, dass Waisengene von hoher biologischer Bedeutung sind und deshalb auf keinen Fall vernachlässigt werden dürfen.

Summary

Proteins perform the bulk of the activity inside each living cell. Thus, it is important that a gene coding for a given protein remains conserved to maintain the proper folding of the primary amino acid sequence. Sequence homology between genes from different species allows us to trace the shared ancestry of the individual genes and species. However, the field of comparative genomics, which deals with sequence comparison, is filled with protein-coding genes that lack detectable sequence homology outside a species or a group of closely related species, such genes are classified as ‘Orphan genes’. During my doctoral research, I have tried to answer the following questions: Are *Pristionchus pacificus* orphan genes real or not? If yes, how old are these genes and how do they originate? I verified the protein-coding nature of orphan genes by estimating the selection pressure on their primary amino acid sequence. These findings indicate that the majority of orphan genes are under strong selection against non-synonymous amino acid changes and hence are real protein-coding genes. Further, by sequencing the genomes of six *Pristionchus* and two non-*Pristionchus* Diplogastrid species, I have generated a phylogenomic dataset with an underlying ladder-like structure around *P. pacificus*. This has allowed me to uncover the dynamics that shape the evolution of young and old gene families. Further, by demonstrating the diverse gene origin mechanisms, I have also determined that both sequence divergence and *de novo* gene creation contribute to the emergence of novel genes in *Pristionchus* nematodes. My results indicate that the genes without homology are biologically important and must not be ignored.

List of publications

1. **N Prabh**, W Roeseler, H Witte, G Eberhardt, R J Sommer, C Rödelsperger: *Deep taxon sampling reveals the evolutionary dynamics of novel gene families in the Pristionchus genome*. Genome Research 2018

2. M S Werner, B Sieriebriennikov, **N Prabh**, T Loschko, C Lanz, R J Sommer: *Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation.* Genome Research 2018

3. E Moreno, M Lenuzzi, C Rödelsperger, **N Prabh**, H Witte, W Roeseler, M Riebesell, R J Sommer: *DAF-19/RFX controls ciliogenesis and influences oxygen-induced social behaviours in Pristionchus pacificus.* Evolution & Development 2018

4. C Rödelsperger, W Roeseler, **N Prabh**, K Yosida, C Weiler, M Hermann, R J Sommer: *Phylotranscriptomics of Pristionchus nematodes reveals parallel gene loss in six hermaphroditic lineages*. Current Biology 2018

5. B Sieriebriennikov, **N Prabh***, M Dardiry*, H Witte, W Roeseler, M Kieninger, C Rödelsperger, R J Sommer: *A Developmental Switch Generating Phenotypic Plasticity Is Part of a Conserved Multi-gene Locus*. Cell Reports 2018 * Equal contribution

6. C Rödelsperger, J M Meyer, **N Prabh**, C Lanz, F Bemm, R J Sommer: *Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model Organism Pristionchus pacificus*. Cell Reports 2017

7. **N Prabh**, C Rödelsperger: *Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs?*. BMC Bioinformatics 2016

8. P Baskaran, C Rödelsperger, **N Prabh**, V Serobyan, G V Markov, A Hirsekorn, C Dieterich: *Ancient gene duplications have shaped developmental stage-specific expression in Pristionchus pacificus*. BMC Evolutionary Biology 2015

List of figures

List of tables

List of abbreviations

BUSCO	:	Benchmarking Universal Single-Copy Orthologs
cDNA	:	Complementary deoxyribonucleic acid
Chr	:	Chromosome
d_N	:	rates of non-synonymous changes
DNA	:	Deoxyribonucleic acid
d_S	:	rates of synonymous changes
FDR	:	False discovery rate
FPKM	:	fragments per kilo base of transcript per million mapped reads
GTR	:	General time reversible
K	:	Kilo, 1000
kb	:	Kilo base
LRT	:	Likelihood ratio test
Mb	:	Mega base
N50	:	Measure of genome assembly quality
ORF	:	Open-reading frame
PCR	:	Polymerase chain reaction
RNA-seq	:	Ribonucleic acid sequencing
RT-PCR	:	Reverse transcription polymerase chain reaction
SSOG	:	Species-specific orphan genes
TROG	:	Taxonomically-restricted orphan genes

Chapter 1: Introduction

Proteins are made from chains of amino acid residues. They are the workhorses that perform the majority of cellular activity. Since the information required to make proteins is embedded in the genome of an organism, the fidelity of DNA replication is of paramount importance. The genome holds the instructions for both the primary amino acids sequence of the proteins, which ultimately folds into the native protein structure, and the spatiotemporal regulatory information for their production. The primary sequence of proteins can be deciphered from the nucleotide sequence of the genome. This facilitates prediction of the protein-coding complement of an organism by recognizing the stretches of DNA, called open-reading frames (ORFs), that can be translated into proteins. Nevertheless, every sequence that has an ORF is not recognized as a bonafide protein. This raises the question, what is a bonafide protein?

Functional annotation through experimental verification is considered the ultimate proof for the protein-coding nature of a gene. However, only a small fraction of conserved protein-coding genes have been functionally validated. Still most conserved genes are considered bonafide proteins, because an establishment of clear homology with experimentally annotated proteins is considered as a reliable evidence for the protein-coding nature of most genes given that they have a complete ORF. It is at this stage that the orphan genes, which lack detectable homology with other known genes and hence fail to qualify as conserved genes, are not considered as bonafide proteins unless experimentally verified. Functional annotation at the protein level can establish an orphan gene as a real protein-coding gene. However, the absence of experimental data casts a proverbial shadow of doubt over the protein-coding nature of such genes. This was indeed the case when the first contiguous stretch of a eukaryotic DNA, the Chr III of *Saccharomyces* cerevisiae, was sequenced and it was reported that more than half of all the predicted ORFs lacked any function [1]. Although at that time these genes were not referred to as orphan genes, in principle they fell under the orphan gene category. My doctoral research has been aimed at finding evidence for protein-coding nature of *P. pacificus* orphan genes, characterizing their evolutionary trajectories, and illustrating mechanisms behind their origination. Thus, for the rest of this dissertation I will focus mainly on protein-coding genes and use the term gene to refer to protein-coding genes unless mentioned otherwise.

Orphan genes

Orphan genes were introduced into the scientific lexicon by Dujon and Casari et al. [2,3]. They coined the term ‘Sequence orphans’ while defining the yeast proteins that lacked homologs in other species. Although what Dujon referred to as the “The mystery of the orphan genes” was already being discussed as early as 1992, when the yeast chromosome III was sequenced [1], still the term ‘orphan’ was brought to the fore when the entire yeast genome was sequenced and analysed. Subsequently such genes were also identified in microbes where they are referred to as ‘ORFans’ [4]. In 2001, Schmid and Aquadro carried out a comprehensive analysis of orphan genes from *Drosophilla* [5]. Based on the existing data on rapidly diverging genes of *Drosophila* [6], Schmid and Aquadro concluded that orphan genes are mainly constituted by rapidly diverging genes and annotation artifacts. However, careful functional examination of a mouse orphan gene revealed for the first time that a functional gene can emerge *de novo* from the non-coding DNA [7]. Since then, the *de novo* emergence of orphan genes has been reported in several species including humans [8–10].

What is an orphan gene?

Orphan genes are the set of genes and gene predictions that are made conspicuous by their lack of homology with an established set of conserved genes. Thus, in a given bioinformatic framework, any gene that is not recognized as a conserved gene gets classified as an orphan gene. These genes have also been known as sequence orphans, young, pioneer, or novel genes [2,11,12]. Generally, the homology of conserved genes is established through blast based similarity searches against known genes from other species. The lack of detectable homology leads to classification of a gene as an orphan gene, which also means that some genes can be identified either as an orphan or conserved gene depending on the sensitivity of the homology detection protocol.

In the past two decades multiple studies have investigated orphan genes and this heightened interest in orphan genes is directly associated with the increase in the amount of whole genome sequencing data [13]. As the cost of genome sequencing has come down, more genomes are getting sequenced by the day. Predicting the entire protein-coding complement of an organism is one of the main reasons motivating most genome sequencing project. Thus, draft genome assembly is immediately followed by annotation of protein-coding genes and this leads to identification of orphan genes. Orphan genes are found in the genomes of each newly sequenced species [11]. This has led to the point that the total number of orphan genes has by far exceeded the total number of the known gene families [12]. However, annotation artifacts can also inflate the number of orphan genes in a genome [14]. Thus, in order to reduce the number of annotation artifacts, advanced gene annotation algorithms can construct their gene models based on the transcriptome evidence from the given species and protein homology evidence from other species [15]. Nevertheless, multiple ORFs lacking homologs in other species are predicted to code for proteins and hence get designated as orphan genes.

Homology inference does not establish biological function

The legibility of amino acid sequence allows homology inference among the protein-coding genes of various organisms. This also permits mapping of the functional attributes of a well-annotated gene on to its homolog from a relatively less studied neighbor. However, the extent of functional conservation of homologous proteins between two species may depend on their phylogenetic distance [16,17]. Moreover, the ortholog conjecture suggests that the orthologous genes share greater functional similarity than the paralogs [18], which raises a question about the exact functional role of the paralogs. Nevertheless, homology detection can reliably establish the shared ancestry between genes. Thus, once a gene from a non-model species is identified as a clear homolog of a functionally annotated gene from a model species, it is accepted as a bonafide protein even in the absence of experimental evidence. Conversely, lack of homology prohibits any readymade functional inference for orphan genes and raises doubts about the reliability of their gene models, unless they are clearly supported by functional data. Although it has been known for decades that not all conserved genes exhibit phenotypes that can be readily assessed through genetic screens [19], still the absence of experimental data greatly undermines the functional relevance of an orphan gene and raises doubt over its protein-coding nature.

The abundance of orphan genes

Most genomes contain 10-30% orphan genes and this number tends to be on the higher side in species from phylogenetically isolated taxa [11,12]. *Pristionchus pacificus,* the focal organism of my doctoral research, is placed within one such taxon [20]. *P. pacificus* belongs to the Diplogastridae family of nematodes [21]. The closest neighbors of these nematodes with sequenced genomes are from the Rhabditidae family, which includes *C.*

elegans [22]. These two families diverged from each other around 60-90 million years ago [23,24]. Thus, it did not come as a total surprise that only 20% of the *P. pacificus* predicted proteins share 1:1 orthology with *C. elegans* proteins, while nearly 1/3rd of all the genes predicted in *P. pacificus* are orphans [20]. Khalturin et al. have suggested that the homologs for many orphan genes will be found as genome data from closely related species becomes available [11]. If homologs of an orphan gene are found only within a closely related group of species belonging to a particular taxon, then such genes are labelled as lineage-specific or taxonomically-restricted orphan genes (Fig. 1.1). The number of taxonomically-restricted orphan genes is expected to increase with the depth of phylogenetic sampling within the taxon [11]. Moreover, finding homologs of orphan genes in closely related species increases the confidence in the reliability of these gene models because protein sequence homology of otherwise non-coding regions are unlikely to be maintained across species boundary [25]. Thus, conservation of orphan genes among closely related species points towards some functional role that might only be significant within the given taxon [11].

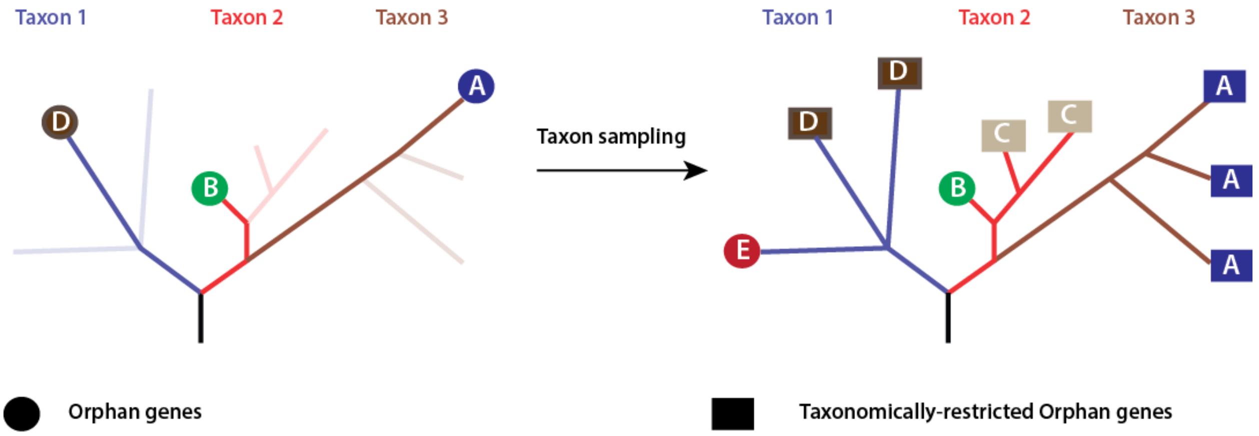


Figure 1.1: Taxonomically-restricted orphan genes. Three taxa each with three species are shown. In the beginning, genome of only one species is available from each taxon (shown as bold lines). Their orphan genes A, B, and D are represented as circles. Increased genome sampling for each taxa shows that some orphan genes A, C, and D are found in neighboring species and hence get classified as Taxonomically-restricted orphan genes, represented as rectangles. Also new orphan gene E is found. Figure adapted from Khalturin et al. 2009 [11].

Orphan genes play important role in lineage-specific adaptations

Recent reports have indicated that orphan genes are involved in stress response, adaptation to fluctuating environment, lineage-specific adaptations and many other biological processes [11,26–29]. Taxonomically-restricted orphan genes, found only in a particular taxon, have been shown to play an important role in lineage specific adaptation and can act as a drivers of phenotypic novelty [30]. In mammals, orphan genes have been shown to contribute to lineage specific traits such as milk production, immune response, and reproduction [31,32]. They are also identified as facilitator of major evolutionary innovations in other lineages [33,34].

Mechanisms of gene origin

Several models for the emergence of orphan genes have been suggested, but horizontal gene transfer, duplication-divergence, and *de novo* models remain the most widely accepted [2,11,12]. Gene duplication followed by sustained sequence divergence may cause a distant paralog to appear as an orphan gene (Fig. 1.2a). The *de novo* emergence of an open reading frame from a previously non-coding region has also been demonstrated as a possible mechanism facilitating the appearance of orphan genes (Fig. 1.2b). However, investigating whether an orphan gene fits one of these models is a difficult proposition and requires both exhaustive computational and manual analysis of individual cases.

Duplication-divergence

Gene duplication

Gene duplication has been considered as the sole mechanism for new gene origin for the better part of the last century. Although Susumu Ohno's acclaimed book "Evolution by Gene Duplication", published in 1970, firmly placed gene duplication as the singular mechanism behind new gene origin, the discussion over the role of gene duplication in shaping the organismal evolution had already begun during the early part of the 20th century. The duplication of genetic material was first recognized by Kuwada, in 1911, as a chromosome duplication event in maize [35]. Soon chromosome copy number variations were reported in other plant varieties and species [36]. In 1918, Calvin Bridges suggested that the duplications could explain the increase in the length of *Drosophila* chromosomes carrying identical genes, which could mutate separately and diversify [37]. It is important to note that this suggestion already carried the conceptual seed for the later named "duplication-divergence" mechanism, even though at that time the proper concept of a gene was not yet formalized. In 1932, Haldane proposed the idea that duplication may be favorable as it opens up the possibility of altering the duplicated gene without being disadvantageous to the organism and multi-copy genes would be less susceptible to harmful mutations [38]. The conceptual underpinning of the “sub-functionalization” mechanism was put forth by Serebrovsky in 1938 [39]. He concluded that “This principle of loss of duplicate functions by one of the homologues in the process of genic evolution . . . should result in a specialization of genes, when each then fulfills only one function which is strictly limited and important for the life of the organism” [39,40].

By the early 1940s, links between gene duplication and organismal complexity were already being stipulated [41,42]. In 1947, Metz argued that “New elements must be added. Otherwise we would have to assume that the primordial amoeba was endowed with all the germinal components now present throughout the wide range of its descendants, from protozoa to man” [43]. The fomenting of ideas on gene-duplication in the 1940s was nicely capped by Stephens in his paper “Possible Significance of Duplication in Evolution” as he questioned the role of the accumulation of slow allelic mutation in shaping the organismal evolution and proposed that evolutionary progress could only be achieved by increasing the number of genetic loci, either by *“de novo”* synthesis of new loci from non-genic source or by duplication of existing genetic loci [44]. Thus, Stephens had not only expanded on the existing ideas of duplication-divergence but he also explicitly mentioned *de novo* gene creation as an alternative mode of new gene origin. However, he considered the origin of new genetic loci through duplication as a well-established phenomenon and postulated that finding evidence for *de novo* gene origin will be too difficult. Indeed, the rapid accumulation of evidence supporting the role of gene duplication made further exploration of *de novo* gene evolution an unproductive endeavor in the second half of the 20th century.

The copious ideas about gene duplication, including subfunctionalization and neofunctionalization, in the preceding decades had poised the field for the publication of Ohno's book in 1970. In this book, he mainly summarized multiple pieces of evidence gathered from various studies, including his own previous work, and strongly argued in favor of gene duplication as the only plausible mechanism for new gene origin: “The creation of metazoans, vertebrates and finally mammals from unicellular organisms would have been quite impossible, for such big leaps in evolution required the creation of new gene loci with previously nonexistent functions. Only the cistron which became redundant was able to escape from the relentless pressure of natural selection, and by escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus” [45].

Divergence after duplication

In 1935, while investigating duplication and insertion of short chromosomal fragments in fruit fly [46], Muller proposed that “Following such duplication, it is to be expected that the redundant loci will come to have divergent mutations established in them in the course of evolution, and so gradually will become more differentiated, until they can finally be regarded as quite non-homologous genes”. This statement by Muller highlights the principle behind the duplication-divergence scenario of the orphan gene origin. In other words, this mechanism posits that a new gene that is created through duplication is not restrained by a strong selective constraints and thus goes through a phase of sequence divergence through which all traces of homology, with the other copy, are lost (Fig. 1.2a). This loss of discernable homology leads to classification of the duplicated gene as an orphan [2,11,12].

Limited horizontal gene transfer in *P. pacificus*

The exchange of genes among different evolutionary lineages is known as horizontal gene transfer, it can be considered as an extended duplication mechanism involving transfer of the duplicated copy to another species [12]. Although horizontal gene transfer frequently occurs in prokaryotes [47], so far it is known to play a limited role in the metazoan genome evolution [48,49]. The gain of orphan genes in *P. pacificus* nematodes through horizontal gene transfer has been investigated in two separate studies [22,50]. Both reports suggest that horizontal gene transfer makes a negligible contribution to the overall number of *P. pacificus* genes, hence obviating the need for further exploration of this mechanism in the evolution of *P. pacificus* genes and genome.

***De novo* gene origin**

Both duplication-divergence and horizontal gene transfer involve reuse of existing proteins to give rise to new protein-coding genes. However, the *de novo* gene origin mechanism suggests that new proteins can arise from ancestrally non-coding sequences (Fig. 1.2b). These non-coding sequences can either be from non-genic loci or from an alternate reading frame of existing genic loci [44,51]. However, in the second half of the 20th century the *de novo* gene origin mechanism was mostly ignored as a plausible process of novel gene formation. Ohno was not the only influential scientist of the 20th century who favored duplication-divergence as the sole mechanism behind new gene origin. Francois Jacob in his 1977 paper titled “Evolution and Tinkering” claimed that “Evolution does not produce novelties from scratch….. The probability that a functional protein would appear *de novo* by random association of amino acids is practically zero . . . . creation of entirely new nucleotide sequences could not be of any importance in the production of new information" [52]. The most potent line of argument against *de novo* gene origin was the lack of experimental evidence that clearly supported origin of novel genes from ancestrally non-coding regions. Nonetheless in 2008, the first demonstration of a functional protein arising from previously non-coding RNA was done in yeast [53]. Stephens had envisaged in 1951 that the investigation of *de novo* origin of new genes will be a difficult proposition and this remains true till date. However, now we know that the probability of functional protein arising out of random non-coding genomic region is non-zero. In fact, the first clear piece of evidence supporting *de novo* gene origin was put forward in 1984 by none other than Ohno himself [54]. While analyzing an enzyme that gave Flavobacteria the ability to degrade nylon, Ohno found that this protein is a result of an evolutionary innovation that allows it to be coded from the genomic locus of a previously existed protein but in an alternative reading frame. As the alternative reading frame was previously non-coding, this makes the nylon degrading enzyme a *de novo* protein. However, Ohno chose not to pursue further exploration into the *de novo* origin of such genes.

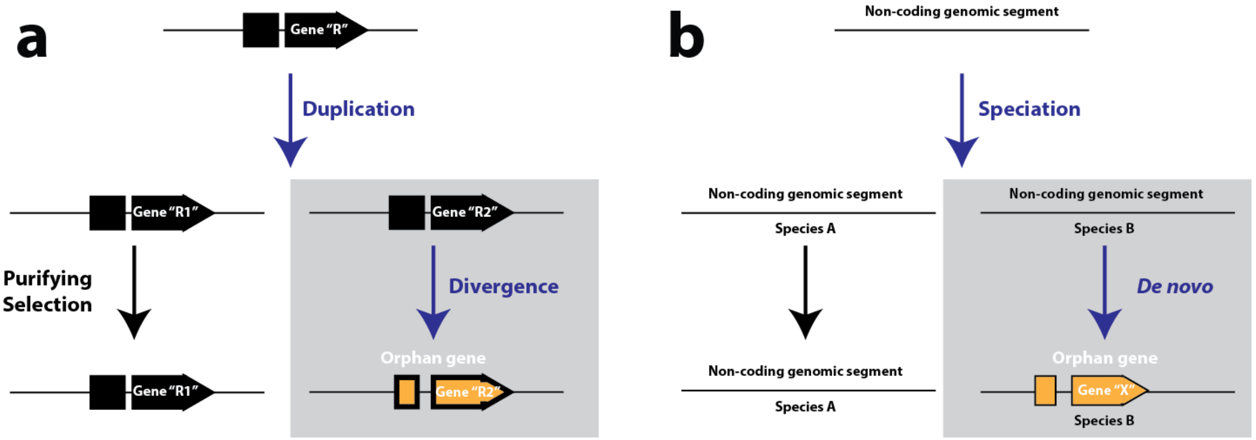


Figure 1.2: Gene origin. a. Duplication-Divergence mechanism: First a gene gets duplicated and then one of the duplicates starts to diverge while the other copy remains under purifying selection. Persistence duplication can remove the signatures of homology and thus the gene gets classified as an orphan gene, marked in yellow. b. *De novo* gene origin: A new gene arises *de novo* in species B from ancestrally non-coding sequence. This is established by identifying the corresponding non-coding genomic segment in a sister species A.

Overprinting

New protein-coding genes can evolve in two ways. First, by *de novo* evolution of an open reading frame within previously untranslated stretches of the genome, either from intronic or intergenic regions. Second, by utilizing an alternate reading frame of a previously existing gene to gain a novel ORF. Actualization of more than one ORFs from the same locus is known as overprinting [51]. The nylon degrading enzyme discussed in the previous section would be a perfect example of overprinting if its ancestral reading frame was also maintained [54]. Overprinting can result from *de novo* opening of an alternate reading frame, which overlaps with the ancestral gene. As the new ORF comes from a previously non-coding reading frame the gene is considered to posit *de novo* origin [51,54]. Kesse and Gibbs analyzed several loci in viruses that present two ORFs and concluded that most loci had one ORF that belonged to old genes, while the other ORF was new and appeared *de novo* [55]. Over the last two decades, several studies have identified candidate loci for overprinting in eukaryotic genomes [56–64]. Thus, *de novo* formation of genes from alternate reading frame of ancestrally coding transcripts has also been established as a verified mechanism of orphan gene origin.

De novo gene origin at ancestrally non-coding loci

Generation of new genes through overprinting involves utilization of previously coding genomic loci in an alternate fashion, however recent studies have shown that new genes can also arise from ancestrally non-coding segments of the genome [10,39]. These non-coding segments can either be intergenic (between two old genes) or intronic. *De novo* gene formation from such loci involves two steps, one of which is to gain an ORF and the other is to gain transcriptional and translational regulation [12]. Recently, Carvunis et al. have proposed that the *de novo* birth of genes from non-genic sequences takes place through an intermediate but reversible proto-gene stage [26].They suggested that the proto-genes first get transcribed and then the non-coding transcripts gains an ORF. Although the reverse order of these events could not be ruled out, as their model was based on the reported translation of non-coding transcripts throughout the genome, this pervasive translation could already provide raw material for natural selection [65].

The selective constraint on most *de novo* genes is likely to be extremely weak and hence from the entire set of *de novo* genes present in a genome only a small fraction will be retained in long-term. Thus, the complement of *de novo* genes shows poor overlap even between closely related taxa. As a result, majority of *de novo* genes identified till date are restricted to a particular lineage or species. This also introduces difficulty in the verification of a *de novo* candidate because the criterion requires identification of the homologous non-coding segment in a related species, and due to lack of strong selective constraints such regions are rarely found in the neighboring species [25]. Hence, even though *de novo* genes have been discovered in many eukaryotic lineages including yeasts, animals, and plants, each individual analysis has verified only a small number of candidates [7,8,10,26,53,66–74]. Further, to my knowledge, *de novo* gene origin has never been reported in nematodes.

Pristionchus pacificus

Nematodes are one of the most species-rich taxa with around 30,000 described species, but their actual number is estimated to be over a million [75–77]. Although *Pristionchus pacificus* is among the most extensively studied nematodes [78], unlike *Caenorhabditis elegans*, it has not yet been established as a model organism for widespread scientific research. This is also indicated by the observation that out of the 28 nematode genomes available on the Wormbase (ftp://ftp.wormbase.org/pub/wormbase/species, release WS254) eight belong to *Caenorhabditis* genus and only two belong to *Pristionchus* genus. During the course of my doctoral research, I have made an effort to generate comparable genome data within the *Pristionchus* genus. This will provide genomic resources to expedite comparative genomic analysis of *Pristionchus* nematodes and will also help in establishing *Pristionchus pacificus* as a major model organism.

Pristionchus pacificus is an emerging model organism

P. pacificus is a model organism that has been used to do comparative evo-devo studies with *C. elegans* and other nematodes [79]. In recent years, it has also been used for in-depth analysis into phenotypic plasticity using its mouth-form dimorphism [80]. Our lab has established several genetic and molecular tools including CRISPR making the species genetically amenable [81]. Moreover, it has a chromosome scale genome assembly, which is one of the best-assembled nematode genomes [82]. Thus, the expertise available within the lab enables us to carry forward genetics, reverse genetics, and comparative genomics analysis to answer many interesting biological questions. Furthermore, *P. pacificus* is known to be associated with scarab beetles in the wild [83]. This association has allowed our lab to sample more than 40 culturable *Pristionchus* species and over one thousand of *P. pacificus* strains from different parts of the world and it is evident that we have yet not reached the species saturation levels within the genus [84].

Phylogenomics

The dense phylogenetic sampling within the *Pristionchus* genus provides an opportunity to employ the comparative method for conducting a phylogenomic analysis of *Pristionchus* nematodes [85]. Phylogenomics is a combination of genome analysis and evolutionary studies that involves the study of genomes under a given evolutionary framework [86]. Originally, phylogenomics was employed to predict the function of a new protein through common ancestry [87,88]. However, as more genome data became available, a new approach was used to study the traits restricted to the particular taxon of the phylogenetic tree. This aspect of phylogenomics is especially relevant to the study of orphan genes and has allowed a better understanding of the lineage-restricted genes based on the width of their distribution [89–92]. Mainly, there are two requirements for a comprehensive phylogenomic analysis. First, an accurate species tree that helps with the selection of species that are best placed to study a particular question. Second, the generation of comparable genome data for the selected species. Susoy et al. generated a robust molecular phylogeny for the *Pristionchus* genus [84]. This has enabled me to carry on an in-depth phylogenomic analysis of *Pristionchus* nematodes (Fig. 1.3).

Aims of the thesis

My doctoral thesis has been motivated by the large number of orphan genes that are found in *P. pacificus*. Although in the recent years several studies have investigated orphan genes in various organisms, at the inception of my doctoral research none of the reported studies had systematically investigated the protein-coding nature of orphan genes. Thus the first question that I have tried to answer in this thesis is: Are orphan genes real protein-coding genes or merely prediction artifacts? In order to answer this question, I established a method that could distinguish between real protein-coding genes and annotation artifacts. In chapter three of this thesis, I discuss this method and the associated results.

The next question raised in this thesis is based on the proposed abundance of orphan genes in phylogenetically isolated species. The hypothesis proposed by Khalturin et al. suggests that as we sequence more genomes around an otherwise isolated species, the homologs of more orphan genes are found and they can be reliably placed in taxonomically restricted gene families. Thus, the second question I have tried to answer in this thesis is: Can deep taxon sampling help us in furthering our understanding of the evolutionary dynamics of novel gene families? In order to answer this question, I created a dataset of assembled genomes of different *Pristionchus* and diplogastrid species, which allowed me to carry a phylogenomic analysis of *Pristionchus* nematodes. In chapter four, I explain the methods used to ensure comparability of different genome assemblies and the first insights that I gained using these assemblies.

The ladder-like phylogeny of assembled genomes created around *P. pacificus* allowed me to investigate origin of its orphan genes. In chapter five, I address the questions about the age of orphan genes, their rate of emergence, and the ratio of Species-specific and taxonomically-restricted orphan genes. Finally, in chapter five, I try to answer the question: What are the mechanisms that lead to creation of novel genes?

Main results

- Majority of orphans are real genes. This is based on selection analysis, which has predictive power to identify real genes even in the absence of other lines of evidences.
- Deep taxon sampling allows age estimation of *Pristionchus* gene families.
- Young genes are frequently lost and remain under relaxed selective constraint.
- Old genes are concentrated at chromosome centers but are generally under strong selective pressure irrespective of their location.
- A steady rate of gene birth is observed along the *pacificus* lineage.
- Novel genes can arise from existing genes either through divergence or ORF switching.
- Novel nematode genes can also arise de novo from ancestrally non-coding regions.

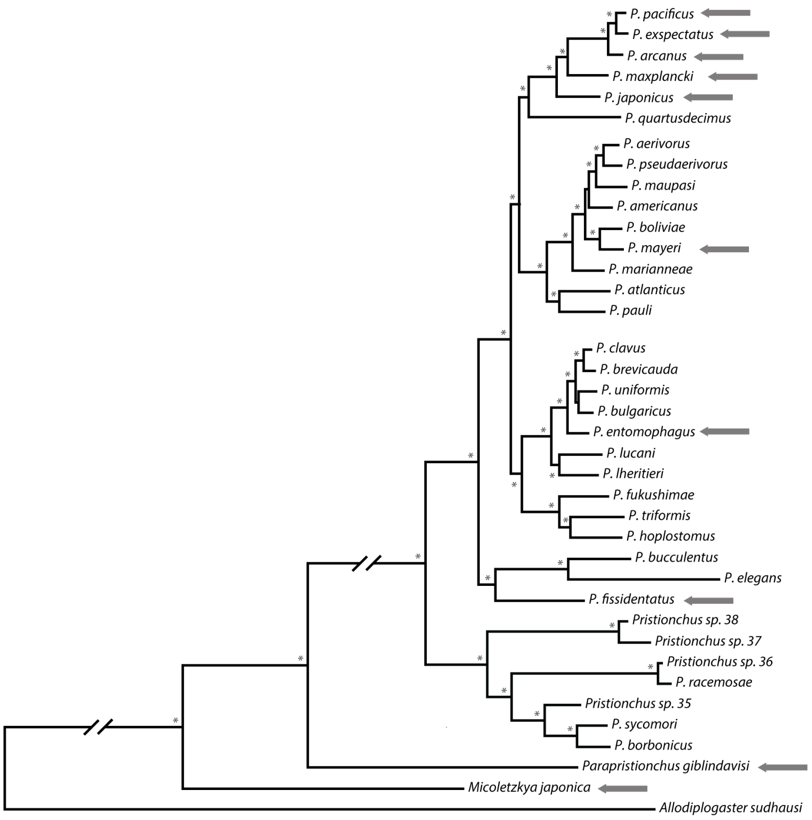
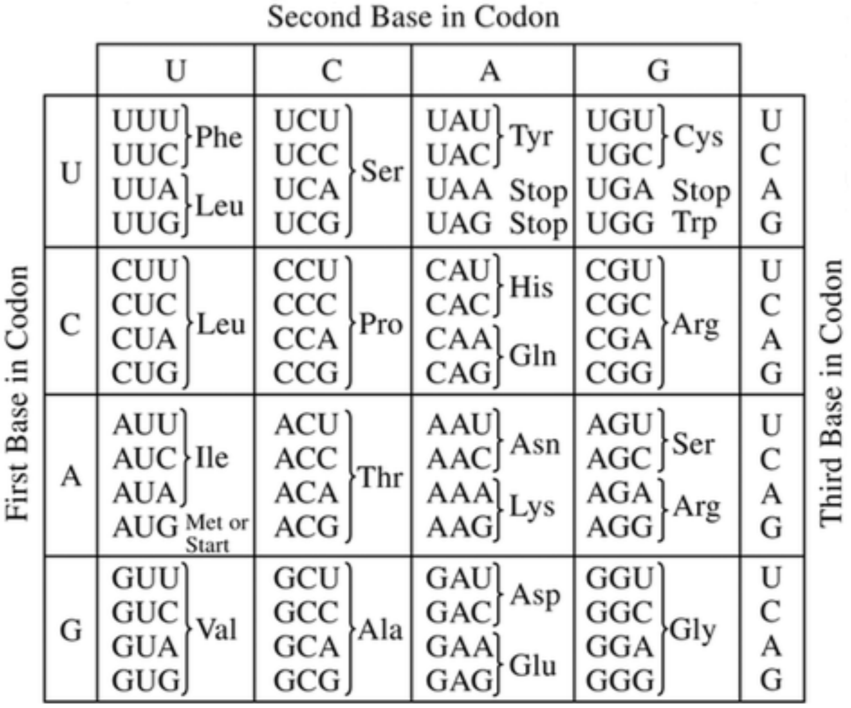


Figure 1.3: **Phylogeny of *Pristionchus* species** inferred from 18S and 28S rRNA genes and 27 ribosomal protein genes by Susoy et al. 2016 [84]. Asterisk represent node support of 100% posterior probability. Arrows represent species analysed in this study.

Chapter 2: Background

Selection analysis based on codon substitution

Due to the degeneracy of codons, many substitutions at the nucleotide sequence level do not get reflected at the amino acid level and are hence referred to as ‘synonymous or silent substitutions’, which can be easily distinguished from ‘non-synonymous or replacement substitution’ that lead to substitutions at the amino acid level (Fig. 2.1). Since natural selection mainly acts at the protein level, the strength of selection acting on the synonymous and non-synonymous mutations vastly differs and they also get fixed at different rates. Hence, the comparison between these two substitution rates is considered a reliable method to uncover the effect of natural selection at the protein level [93–96]. Such comparisons do not require a prior knowledge of the species divergence times or absolute substitution rates.



First Base in Codon	Second Base in Codon								Third Base in Codon
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met or Start	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 2.1: Codon table. Overview of the codons and the three letter codes for their corresponding amino acid residues.

Generally, distances for both synonymous and non-synonymous substitutions are calculated and then get defined as the number of synonymous substitutions per site (d_S) and the number of non-synonymous substitutions per non-synonymous site (d_N), respectively. Although it is possible to employ the heuristic counting method for d_S and d_N estimation, in this thesis, I have used the maximum likelihood method for two main reasons [97]. First, the maximum likelihood method is simple and unlike the counting method does not require the explicit estimation of different transition and transversion rates or catering for unequal codon frequencies. Second, the maximum likelihood method can also accommodate more realistic models of codon substitution, such as general time reversible or GTR model [98], which is not possible under the counting methods.

I have estimated the d_S and d_N values with the codeml suite of the PAML software, which also gives a value for their ratio called ω (d_N/d_S) that measures the strength of selection acting at the protein level [99,100]. If the d_S is greater than d_N, then the ω value will be less than one and the protein is considered to be under purifying or negative selection. If the d_N is greater than d_S, then the ω value will be more than one and the protein is considered to be under adaptive or positive selection. However, if a protein is not under selection and evolves neutrally then both rates should be equal and the ω value will be equal to one. Considering that the annotation artifacts or pseudogenes should not be under purifying selection at the protein level, my null hypothesis is that the ω value for such genes should not show significant deviation from one (see methods section from chapter 3), which means that they should portray neutral evolution. Thus, a statistically significant ω value of less than one indicates that the gene is under purifying selection to preserve its protein sequence and hence unlikely to be an annotation artifact.

Gene structure annotation

Although the drastic decrease in the sequencing cost brought on by the next-generation sequencing protocols has greatly expedited the rate at which new genomes are being sequenced, these protocols generally do not match the genome contiguity of the previously used shotgun sequencing approach. The fragmented genomes, as well as the lack of pre-existing high-quality gene models for most organisms, complicate the task of reliable gene structure prediction. Prediction of accurate protein-coding gene models remains one of the most important steps of genome sequencing projects because any error introduced at this level will persist in subsequent analysis and will cast doubt over inferences made regarding gene family evolution.

Gene structure prediction of phylogenetically isolated organisms involves the use of *ab initio* gene prediction software. Some software, such as SNAP and AUGUSTUS [101,102], can utilize existing transcriptome for initial training. Employment of a prior training step is known to yield more reliable gene models [103,104]. Further, it is also possible to merge gene predictions from several methods and use additional evidence to select one or more predicted gene models from a particular locus. In this thesis, I have used Maker2, which is one such 'chooser' pipeline, that can select the most representative prediction based on evidence, allows multiple isoform predictions from the same loci, and can also add UTRs by inferring the RNA-seq data [15]. Most importantly, Maker2 also allows the simultaneous use of protein sequences from closely related species and the transcriptome assembly of the given species to improve the accuracy of the gene predictions. This function is especially useful for large taxon genome projects that need to annotate several genomes at once because it allows splice-aware mapping of pre-existing proteins from closely related species and facilitates reliable gene structure prediction, even if only limited transcriptome evidence exists for most species within the taxon. However, the major drawbacks of the Maker2 pipeline include generation of a large number of temporary files and long running time, especially if multiple iterations are needed to run, even after massive parallelization. Although these issues can impede the genome projects, still parallel annotation of several related genomes through the Maker2 pipeline yields accurate and highly comparable protein predictions (see BUSCO columns in Table 4.1). This tends to limit technical artifacts and allows more reliable inferences regarding the gene and gene family evolution.

Chapter 3: Are orphan genes real?

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

> Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016;17: 226.

Author	Author Position	Scientific ideas %	Data generation %	Analysis and interpretation %	Paper writing
N.P.	1	50	100	75	60
C.R.	2	50	0	25	40
Title of the paper:	Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics.				
Status:	Published				

In this chapter, I have examined the protein-coding nature of orphan genes in *P. pacificus* using three criteria. First, I assessed whether they are expressed. Second, I searched for direct support towards the protein-coding nature of an orphan gene by finding a match in available proteomic data. Third, in the absence of peptide data I used negative selection that constrains the rate of non-synonymous amino acid substitutions as an indirect evidence for the protein-coding nature. Orphan genes that get expressed but do not fulfill any of the other two criteria were considered as candidates for non-coding RNAs. The results posit that between 39–77% of orphan genes are protein-coding, indicating that orphan genes play a significant role in the biology of *P. pacificus*. Application of this methodology on other taxonomically under-sampled groups will further support the veracity of orphan genes and their biological significance.

Results

More than 80% of orphan genes are transcribed

I used a previously published dataset of orphan genes (N = 9,885), conserved genes (N = 20,999), and 14 RNA-seq experiments in *P. pacificus* to assess the transcription of both orphan and conserved genes [105]. This dataset was also used to examine the correspondence between the total number of expressed genes and the number of included

RNA-seq samples (Fig. 3.1a). First, I defined two different thresholds on the magnitude of expression, which was measured as fragments per kilobase transcript per million fragments (FPKM) sequenced. While a value of FPKM > =10 indicates robust expression, any value of FPKM > =1 can still be accepted as a reliable evidence of expression because many functionally annotated genes display FPKM values well below 10 [106,107]. Even though conserved genes show higher levels of both expressed (*N* = 18997,~90%) and robustly expressed genes (*N* = 14010, ~67%), still in this analysis 7997 (81%) of orphan genes are expressed in at least one of the samples (Fig. 3.1a).

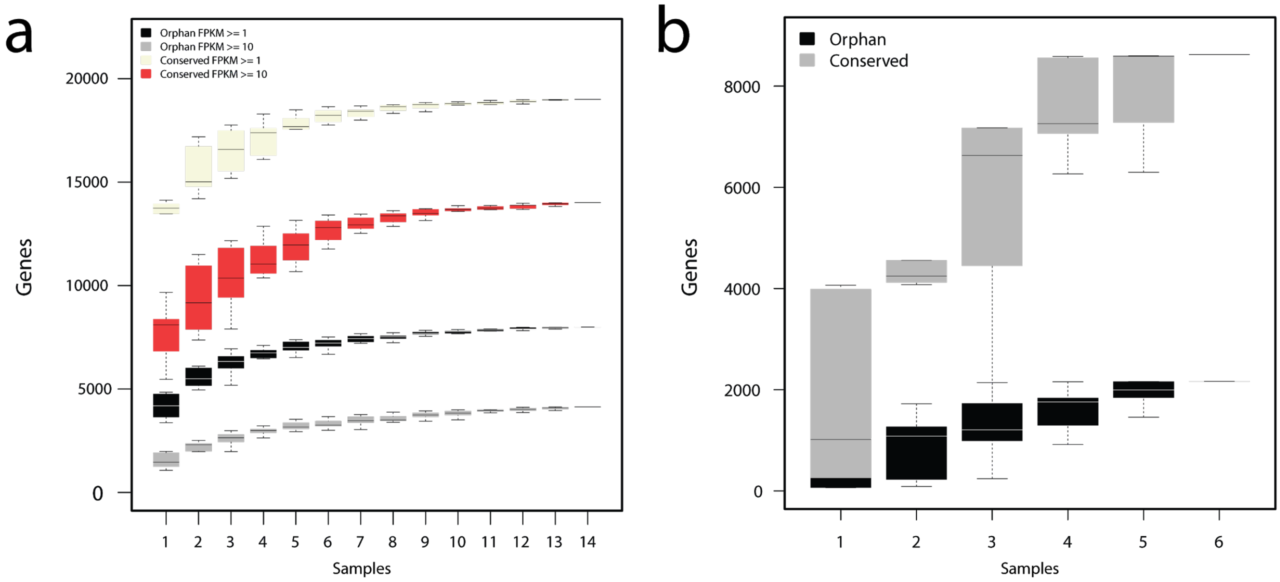


Figure 3.1: Transcription and differential expression of orphan genes. a) RNA sequencing data from 14 experiments are used to determine the number of both orphan and conserved genes getting expressed over the expression threshold of FPKM 1 and 10. b) Differential expression pattern of both orphan and conserved genes from 6 Microarray based differential expression analysis experiments.

Orphan genes are integrated into gene-regulatory networks

Apart from using transcription as a necessary criterion for a real gene, I also quantified the genes that are differentially expressed under variable conditions such as distinct developmental stages or exposure to different pathogens. The rationale behind this analysis is the assumption that an orphan gene that gets differentially expressed under varying conditions is most likely integrated into an existing gene-regulatory network. Therefore, differential expression further supports the protein-coding nature of an orphan candidate. On the other hand, other orphan genes that posit constitutively low or negligible expression levels may not have had enough time to be integrated into any existing regulatory network. Further, this also suggests that most older genes must have already integrated into a gene network and they will posit a stronger differential expression pattern compared to new genes. Consistent with this assumption, previous research from our group has shown that conserved genes are significantly enriched among the genes that are developmentally regulated [105]. However, by analyzing previously conducted transcriptome profiling studies (*N* = 6 gene sets) [108,109], I find that along with 8623 (41%) conserved genes even 2165 (22%) *P. pacificus* orphan genes are differentially expressed. This implies that these orphan genes must have persisted for a considerable amount of evolutionary time within the genome for them to be integrated into a regulatory circuit.

Only 4% of orphan genes shows direct evidence for translation

Next, I assessed the direct evidence for the translation of genes by testing for matches in available proteomics data [20,110]. Employing 100% sequence identity over the full peptide length as a criterion to search through the peptide data (*N* = 51,224 peptide sequences), I found peptide evidence for 428 (4%) orphan genes. This number is considerably low with respect to the 5177 (25%) conserved genes that have peptide evidence, but it is compatible with a previous study that reported the depletion of orphan genes in transcriptome and peptide data [110]. However, it is also clear that three-quarters of conserved genes also lack peptide evidence, thus suggesting that the absence of peptide evidence is not a sufficient criterion to distinguish real protein-coding genes from potential artifacts or non-coding RNAs. Hence, I employ genomic resources to obtain indirect evidence for protein-coding genes, i.e. selection against non-synonymous substitutions.

Comparative genomics of orphan genes

As most orphan genes are not present in peptide data, I use selection against non-synonymous substitutions in protein-coding genes as an indirect evidence for translation. However, to estimate this evolutionary constraint, which is called negative or purifying selection, at least two sequences are needed. This is problematic for orphan genes since they do not have homologs in other species. However, this problem can be overcome by doing comparative analysis with the genome data from a closely related species. To this end, I use *P. exspectatus*, the recently sequenced sister species of *P. pacificus*. The *P. pacificus* and *P. exspectatus* genomes show roughly 10% sequence divergence in alignable regions [105,111]. For further analysis, all genes from both the species were segregated into 14,656 different orthologous clusters using OrthoMCL [112]. Fig. 3.2a depicts the distribution of these clusters into six different categories based on the number of genes of each species present in the clusters. Evidently, the majority of clusters contain only two members, i.e. one from each species. After removing *P. exspectatus* specific clusters, hybrid clusters, and the short peptide containing clusters or poorly aligning clusters (see Methods), I was left with 10,327 clean clusters containing one or more *P. pacificus* conserved genes and 3,273 clusters carrying one or more *P. pacificus* orphan genes. These two cluster sets contain 3,891 orphan and 13,103 conserved genes from *P. pacificus.* Both the conserved and the orphan gene clusters were subdivided into two datasets. The first dataset made by the clusters that contain at least one gene from both the species was called 'orthologous clusters'. The second dataset was called 'paralogous clusters' and consisted of all the clusters containing more than one *P. pacificus*, however, all *P. exspectatus* genes were removed from this dataset. When combined together both the datasets represent all *P. pacificus* genes from the 10,327 clean cluster except the singleton genes, which are present only as single copy genes in *P. pacificus* and lack a corresponding homolog in *P. exspectatus.* In order to include *P. pacificus* singletons and study selection at a closer timescale, I created a third dataset called 'clade A1-A2 orthologs' ($N = 30{,}884$). This dataset was employed to compare the divergence of orthologous gene pairs (cluster size = 2) across two geographically isolated *P. pacificus* lineages [111].

Both orphan and conserved genes are under negative selection

I used the ratio of synonymous to non-synonymous rate of amino acid substitution (d_N/d_S), also called omega (ω), as the measure of selective pressure for each cluster of the three above-mentioned datasets using the PAML suit [99]. An ω value equal to 1 indicates neutral evolution, while $\omega < 1$ can be interpreted as evidence for negative selection. The estimation of ω value of orthologous clusters showed that the conserved gene clusters are under relatively stronger negative selection than the orphan gene clusters (Fig. 3.2b). Nonetheless, the majority of orphan gene clusters were also under negative selection. The results from similar analysis done on the paralogous dataset were also comparable, given that only 11% of conserved clusters and 15% of orphan clusters were present in this dataset (Fig. 3.2c). The results for clade A1-A2 dataset also maintains the trend of showing stronger selection on conserved genes. Nevertheless, the orphan genes were also shown to be under robust negative selection (Fig. 3.2d).

Ortholog, paralog, and intra-species comparisons complement each other

Similar to the expression data analysis, I defined both a liberal and a conservative criterion to estimate the number of genes under negative selection. First, I arbitrarily chose ω < 0.6 as the liberal cutoff for a cluster to be considered as evolving under negative selection. Further, the conservative criterion required that the negative-selection model posits a statistically significant difference when compared with a neutral model. Thus, for each cluster the PAML was run twice, once allowing a single but free to change ω value for the entire cluster tree (alternate model: HA) and then for the second tree ω was fixed at 1 for the entire cluster tree (null model: H0). For each cluster, the likelihood ratio test was conducted with one degree of freedom and at P-value < 0.05 (FDR adjusted) for each cluster to determine the significance of the alternate model. The combination of ω < 1 and FDR adjusted P-value < 0.05 was used as the second more conservative criterion. It is to be noted here, that the lack of statistical significance does not exclude the presence of evolutionary constraint. In many cases, the lack of statistical significance manifests from the low statistical power of comparison between small proteins or due to the little divergence between the sequences.

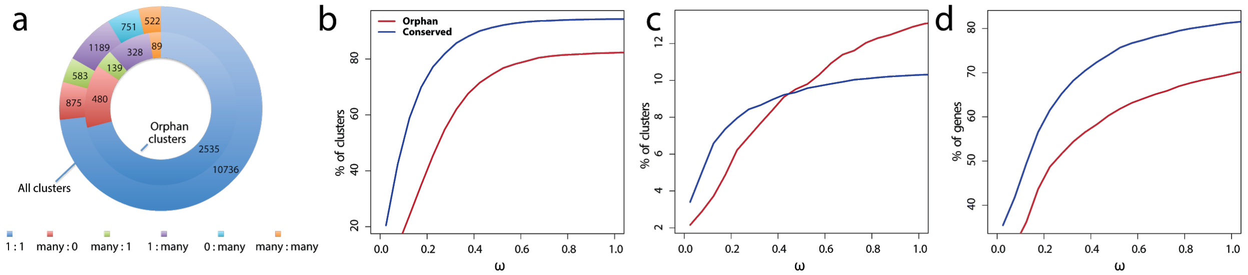


Figure 3.2: Orphan genes are under strong negative selection. a) Distribution of gene clusters based on the number of homologs between *Pristionchus pacificus* and *Pristionchus exspectatus*. b) Comparison of the variation in the proportion of Orphan gene clusters and conserved gene cluster (Y-axis) under given ω value in the *P. pacificus – P. exspectatus* orthologous clusters dataset. c) Comparison of the variation in the proportion of Orphan gene clusters and conserved gene cluster (Y-axis) under given ω value in *P. pacificus* paralogous clusters dataset. d) Comparison of the variation in the proportion of Orphan gene clusters and conserved gene cluster (Y-axis) under given ω value in clade A1 – clade A2 paralogs dataset.

Employing both of the above-mentioned criteria I compared orphan candidates from all three datasets, which were found to be under negative selection. Using the cutoff of ω value lower than 0.6, I identified 7545 (76%) orphan genes under negative selection in at least one of the three datasets (Fig. 3.3a). Here, the largest contribution came from the intra-species comparison. However, since the evolutionary distances in the intra-species comparison are rather small, random fluctuations in the number of non-synonymous and synonymous substitutions can also generate ω values below 0.6. Given that the drive for positive selection is unlikely to cause substitutions along complete genes, ω > 1 are generally due to such random fluctuations; based on this I consider that a subset of the identified orphan genes with ω < 0.6 can be due to random noise. Therefore, I emphasize that the liberal cutoff of ω < 0.6 has to be regarded as the upper limit of the estimated number of negatively selected orphan genes.

Using the likelihood ratio test I identified 3818 (39%) orphan genes that show statistically significant negative selection in at least one of the three datasets (Fig. 3.3b). Moreover, out of 3273 orthologous orphan clusters 2899 clusters (89%) posit significantly better goodness of fit for negative selection compared with the neutral evolution model. The corresponding figure for the conserved clusters is 9838 (95%) out of 10,327. In the paralogous clusters dataset, 297 (57%) out of 514 orphan clusters and 997 (82%) out of 1222 conserved clusters posited significant goodness of fit for negative selection. In clade A1-A2 dataset, only 769 (8%) out of 9885 orphan gene clusters and 4767 (23%) out of 20,999 conserved gene clusters showed significant goodness of fit supporting negative selection, suggesting that the divergence between two *P. pacificus* lineages is in general not sufficient to gather robust evidence for negative selection at the single gene level.

In order to further evaluate both the liberal and the conservative criteria, I compared the expression evidence available for the candidates identified under each criterion (Fig. 3.3c). This demonstrated that a cutoff of ω < 0.6 on one hand captures almost all examples of significant negative selection, while on the other hand it also minimizes the fraction of candidate orphan genes that lack expression evidence. I have also observed that changing this cutoff value to 0.5 or 0.7 exceedingly impairs this balance. In summary, my analysis suggests that a large fraction (39–76%) of orphan genes are under strong negative selection.

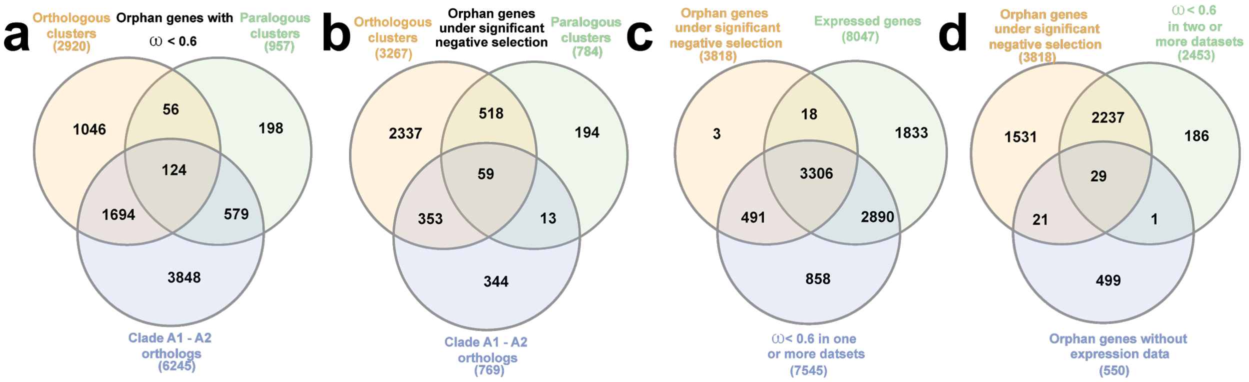


Figure 3.3: Complementary of different datasets. a) Venn diagram for three different datasets using a definition of < 0.6. b) Venn diagram for three different datasets using a definition of ω < 1 and P<0.05. c) Comparison of different thresholds to define negative selection. d) Overlap between orphan genes that lack any expression data, are under significant negative selection and also have ω value less than 0.6 in at least two of the three datasets.

The evolutionary constraint can predict gene expression

The previous result confirmed that a considerable number of orphan genes are under negative selection. Next, I combined the expression data with the results of the selection analysis to screen for orphan genes that show evidence for negative selection but lack expression evidence. To this end, the list of 3818 orphan genes showing statistically significant negative selection was intersected with two other lists: first, the list of orphan genes without expression data (*N* = 550, FPKM = 0 in all 14 RNA-seq experiments) and second, the list of genes that had ω < 0.6 in at least two out of the three datasets (Fig. 3.3c). In total, 29 genes were found to be present in all three lists, out of which eleven genes having more than 2 exons in their predicted open reading frame were chosen as the candidates for validation by RT-PCR using primers overlapping predicted exon-exon junctions. PCR products for three out of the 11 candidate genes were obtained and then sequenced, but only when PCR was done using cDNA and not genomic DNA as the template (Fig. 3.4a). Sequencing of the PCR products resulted in expressed sequence tags that matched the gene predictions (Fig. 3.4b). This result demonstrated that even in the absence of evidence from 14 RNA-seq samples, RT-PCR was able to detect expression evidence of orphan genes in standard mixed-stage worm cultures. As expression profiles can be highly stage and tissue-specific, I speculate that sequencing of more stage and tissue-specific RNA samples can validate predicted gene structures of many orphan genes that are not supported by transcriptome evidence generated thus far.

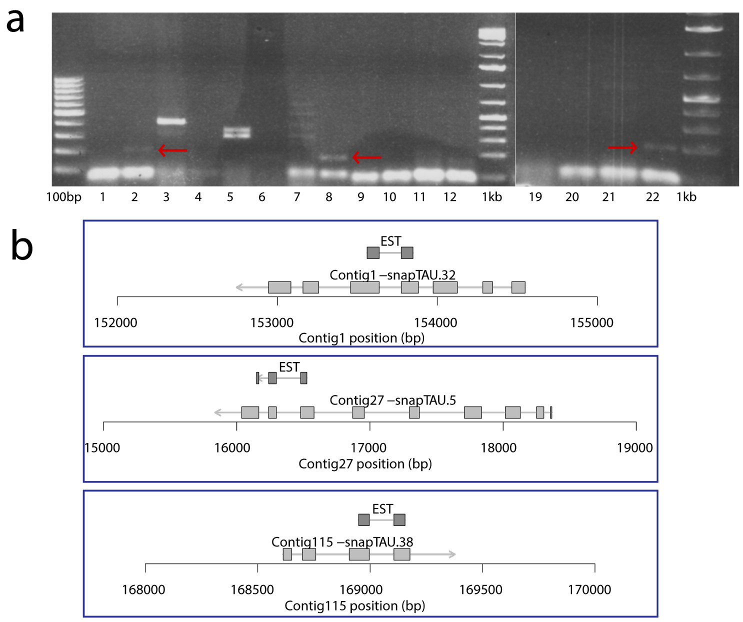


Figure 3.4: Validation of orphan genes. a) PCR validation experiments for eleven candidate orphan genes. Genomic DNA (odd numbers) and cDNA (even numbers) was amplified using the same primer pairs. In three cases, we obtained bands in the expected size range. b) Sequencing of amplification products resulted in ESTs that confirmed the gene structure.

Differences among various gene classes

Based on the available expression data and estimates for negative selection, I divided all *P. pacificus* genes into four distinct classes. The first class was made of 497 potential prediction artifacts or pseudogenes (orphan genes with FPKM values below or equal to one and not under strong negative selection, i.e. $\omega < 0.6$ in any analyses). The second class contained 837 candidates for non-coding RNAs (FPKM greater than 10 in at least one RNA-seq dataset, but not under strong negative selection). The conserved genes and the orphans that either showed statistically significant negative selection or were found in the peptide data constituted the third and the fourth classes respectively. I compared four gene features (Transcript length, number of exons, GC content, and contig size percentile) across all four gene classes (Fig. 3.5). While the conserved and the negatively selected orphan genes tend to be longer and have more exons than the potential prediction artifacts/pseudogenes and the non-coding RNA candidates ($P < 0.001$, Wilcoxon ranksum test, Fig. 3.5a-b), at the level of GC content and the contig size percentile no obvious differences were detected (Fig. 3.5c-d). Given that the fragmented assemblies were identified as a potential source of prediction artifacts, it was interesting to observe that in all the four gene classes nearly 90% of genes reside within the top 1% of largest contigs. There were no significant trends towards aggregation of potential artifacts on the smaller contigs. Thus, partial gene models resulting from fragmented contigs are an unlikely source for the majority of prediction artifacts or pseudogene candidates.

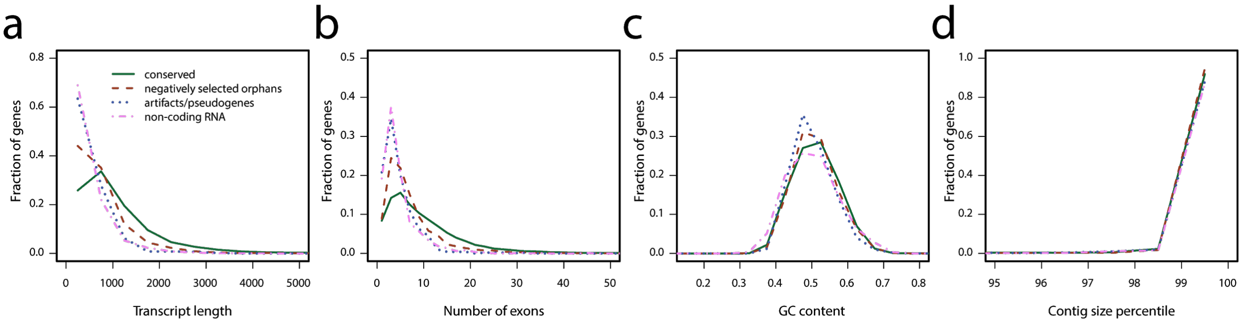


Figure 3.5: Differences between various gene classes. a) Comparison of transcript length for potential prediction artifacts/pseudogenes, non-coding RNA candidates, negatively selected orphan genes, and conserved genes. The y-axis denotes the fraction of genes at a given transcript length. b) Comparison of number of exons. c) GC content distribution for all four gene classes. d) Distribution of contig size percentiles among all four classes. The top 1% of largest contigs harbors roughly 90% of genes for all four gene classes.

Conclusion

Based on these results, I make two main conclusions. First, majority of *P. pacificus* orphan genes are real protein-coding genes. This conclusion is mainly supported by the intra-species selection analysis, however, both the selection analysis with the sister species and transcriptome evidence point towards the same general direction. Nevertheless, having based my main conclusion on the selection analysis, I also wanted to check if selection analysis results also have predictive power on the protein-coding nature of genes. Based on the analysis of candidate genes shown to be under strong negative selection at the protein level, but lacking transcriptome evidence, I conclude that the selection analysis can itself predict protein-coding nature of genes, which then can be verified with further experimental analysis.

Methods

Orphan definition and homology clustering

Orphan genes were defined previously using blastp comparisons against 14 non-diplogastrid nematode outgroup genomes [105]. This procedure identified orphan genes that were restricted to the family Diplogastridae. In total, 20,999 genes were identified as conserved genes and remaining 9,885 genes were classified as orphan genes. I used the OrthoMCl software [112], with default parameters, to create homologous protein clusters between *P. pacificus* (version TAU) and *P. exspectatus* (version SNAP2012) genes. This allowed me to estimate selection pressure at an interspecies (orthologous clusters) as well as intra-species level (clusters with paralogs). The clusters were divided into orphan and conserved categories. I identified 253 ‘hybrid clusters’ carrying both orphan and conserved genes of *P. pacificus*. Closer analysis of these clusters showed that events such as gene fusion, gene split, pseudogenization, and variation in divergence rates lead to mixing of orphan and conserved genes in a single cluster. Therefore, I decided to exclude these hybrid clusters from further analysis.

Transcription and translation analysis

Expression evidence for the predicted genes was assessed using previously generated RNA-seq [105] and microarray data [108]. The correspondence between the total number of genes with expression evidence and the number of analyzed RNA-seq samples was assessed using random permutations of RNA-seq samples (*N* = 14, Fig. 3.1a) and differentially expressed gene sets (*N* = 6, Fig. 3.1b).

Mass spectrometry data from earlier experiments was analyzed to gather evidence for translation [20,110]. The peptide sequences generated from the mass spectrometry experiments were compared with the predicted protein database using blastp. A predicted gene was accepted as a translated gene, only if it had a blastp hit matching entire peptide length with 100% identity.

Divergence estimation between different P. pacificus lineages

In order to study selection at a microevolutionary time-scale, I used whole genome resequencing data from two deeply sampled *P. pacificus* clades [111], clade A1 from Asia (*N* = 15 strains), clade A2 from Southern and Central America and the Indian ocean (*N* = 16 strains). These two clades are geographically isolated and show approximately 1% genome-wide divergence. I extracted fixed differences to the reference genome (*N* = 485,795 for clade A1 and *N* = 618,650 for clade A2), which were covered in all sequenced strains per clade (by at least two reads with a samtools genotype quality above 20) and mapped them onto the reference genome.

Estimation of selection pressure

For each cluster, the predicted proteins were aligned using MUSCLE [113] and the resulting protein alignments were converted into codon alignments using PAL2NAL [114]. A minimum cutoff of 150 bases codon alignment was set to avoid bias introduced in the analysis by poor alignment or short peptides, as a result of which 231 clusters were removed from further analysis. RAXML was employed to create the phylogenetic tree for each cluster containing three or more proteins, for clusters smaller in size an unrooted tree was created. For each cluster, the Codeml package of the PAML suite [99] was run twice: first to obtain a single omega (ω) value for the entire tree (alternate model, HA) along with its associated maximum likelihood score, and then again to obtain the maximum likelihood score for the model with a fixed ω value of one for the entire tree (null model, H0).

In order to test the statistical significance of the estimated ω value, I performed Likelihood ratio tests (LRT) for each cluster using the maximum likelihood score generated from both the runs. Considering that my null model is the model where ω is fixed at one and my alternate model is where a single ω value is freely estimated for the entire cluster tree, the required assumption for a LRT i.e. that two models are nested is readily fulfilled. The test statistic is double the difference in log-likelihood (lnL) scores for the two models (LRT statistic = 2 (lnL HA - lnL H0)). For large samples, LRT statistic follows a chi-square distribution with degrees of freedom that is the difference in the number of freely estimated parameters between the alternate and null models. Here the only parameter that differs between the models is ω and thus the degree of freedom is 1. Hence the P-value of the LRT statistic was calculated from a chi-square distribution with a degree of freedom of 1 and then adjusted for false discovery rate (FDR). The alternate model was considered significantly better than the null model if the LRT statistic P-value (FDR adjusted) was less than 0.05.

Worm collection, DNA extraction, and RNA preparation

P. pacificus worms were grown on nematode growth media agar plate and the *Escherichia coli* strain OP50 was used as food source [115]. Adult worms were washed from the plate using M9 buffer and frozen at −80 °C with Trizol for RNA preparation and without Trizol for DNA extraction.DNA extraction was performed using Sigma GenElute Mammalian Genomic DNA Miniprep kit (Catalog number G1N70-1KT) as per manufacturer’s instructions.

For RNA preparation, worm pellets frozen with Trizol were put through three freeze/thaw cycles in liquid nitrogen, followed by vigorous vortexing and 10 min incubation at room temperature. After centrifugation at 14000 rpm for 10 min at 4 °C, 100 µl chloroform was mixed with the supernatant, vortexed and incubated for 5 min at room temperature. The samples were again centrifuged at 14000 rpm for 15 min at 4 °C and the upper phase containing the RNA was transferred into a new tube. RNA precipitation was carried out by adding 0.5 µl glycogen blue and 250 µl isopropanol and then incubating at −20 °C for a few hours. The precipitated RNA was centrifuged at 14000 rpm for 10 min at 4 °C and then the pellet was washed with ethanol. Genomic DNA was digested using Promega RQ1 DNAse (Catalog number M6101). The dried pelleted RNA was resuspended in 20 µl TE-buffer and incubated at 60 °C for 10 min to dissolve [107]. Invitrogen SuperScript® II Reverse Transcriptase kit (Catalog number 18064–014) was used to reverse transcribe cDNA as per manufacturer’s instructions.

Candidate validation

PCR primers pairs ranging between 24 to 26 nucleotides in length were designed for all the 11 candidates (Supplemental Table S3.1) and then ordered to Eurofins for synthesis. PCR reaction was carried out using these primers against both *P. pacificus* genomic DNA and cDNA. The PCR program was 5 min at 94 °C, then 35 cycles of 3 steps - 94 °C for 30 s, 60 °C for 30 s and 72 °C for 30 s and then 72 °C for 10 min. The PCR products were run on 1.5% agarose gel for 60 mins to check their size range. Finally, the PCR products were sequenced using BigDye v3.1 Cycle Sequencing Kit (Catalog number 4337457) from the Thermo Fischer Scientific as per instruction. The sequences derived from sequencing reaction were manually aligned to corresponding candidate regions.

Chapter 4: Evolutionary dynamics of novel gene families

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

Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in nematodes. Genome Res. 2018; doi:10.1101/gr.234971.118

Author	Author Position	Scientific ideas %	Data generation %	Analysis and interpretation %	Paper writing
N.P.	1	60	70	75	70
W.R	2	0	10	0	0
H.W.	3	0	9	0	0
G.E.	4	0	9	0	0
R.J.S.	5	10	2	0	10
C.R.	6	30	0	25	20
Title of the paper:	Deep taxon sampling reveals the evolutionary dynamics of novel gene families in nematodes.				
Status:	Published				

Although in the previous chapter I established that the majority of orphan genes are real protein-coding genes, still all orphan genes with homologs in sister species were bracketed in a single age group and hence were only estimated to be at least as old as the last common ancestor of *P. pacificus* and *P. exspectatus*. Thus, in order to overcome the lack of dense genomic sampling around *P. pacificus*, I decided to sequence the genomes of six other *Pristionchus* species and two non-*Pristionchus* diplogastrid species. In this chapter, I set out to illustrate generic features of *Pristionchus* genome evolution by first creating a dataset of comparable genomes and then analyzing the evolutionary dynamics of *Pristionchus* gene families. The results suggest that old genes are robustly expressed, show lower substitution rates, and remain concentrated at chromosome centers. On other hand, novel genes are mainly present at chromosome arms, show higher substitution rates, get expressed at a lower level and show higher propensity towards gene loss.

Results

Genome assemblies of 10 diplogastrid nematodes as a platform for comparative phylogenomics

To understand the dynamics of gene gain and loss within the *Pristionchus* lineage, I sequenced eight new diplogastrid genomes to complement the two existing draft genomes of the sister species *P. pacificus* [82] and *P. exspectatus* [111]. Specifically, I sequenced genomes of three gonochoristic species (*P. arcanus, P. maxplancki* and *P. japonicus*) and three hermaphroditic species (*P. mayeri, P. entomophagus* and *P. fissidentatus*) of the genus *Pristionchus* along with the gonochoristic *Micoletzkya japonica* and *Parapristionchus giblindavisi* [84,116]. Additionally, I resequenced the genome of the hermaphroditic *P. pacificus* on Illumina platform to increase comparability (see Methods for details). Each species was carefully chosen to generate a deep taxon sampling within the *Pristionchus* genus based on our current knowledge of the molecular phylogeny [84]. Furter, the two non-*Pristionchus* species were selected as outgroup (Fig. 4.1a, Fig. 1.3). The genome sizes of *Pristionchus* species in the scaffolded assemblies varied between 151 Mb and 297 Mb (Table 4.1). Studies conducted in *Caenorhabditis* nematodes have reported that hermaphroditic mode of reproduction can result in reduced genome size [117–120]. In this study, gonochoristic species do not generally have larger genomes than hermaphroditic species. However, while comparing the *P. pacificus* with its two closest relatives, *P. exspectatus* and *P. arcanus*, I found that the trend for smaller genomes in hermaphrodites holds true (Fig. 4.1a, Table 4.1).

To assess the quality of the genome assemblies, I compared measures of contiguity (N50: numbers of scaffolds), completeness (BUSCO: percentage of raw reads reads represented in the assembly) and correctness (paired ends in proper orientation and ambiguous fraction). The largest differences were caused due to the switching of assembly strategy during the course of this study. Particularly, the older and more aggressive ALLPATHS-LG assembly strategy, which was based on an initial assembly of overlapping read pairs, generated substantially fewer contigs at the cost of higher levels of ambiguous base calls [121]. The more recent approach, implemented through the DISCOVAR *de novo* software (https://software.broadinstitute.org/software/discovar), yields an initial assembly based on a PCR free library. These assemblies tend to have higher number of contigs, but also considerably reduced levels of ambiguity (Table 4.1). Nonetheless, these differences between assembly strategies do not have an obvious impact on the N50, BUSCO, or any of the other assembly quality measures. Therefore, I conclude that all our assemblies are of comparable quality.

Table 4.1: **Summary of basic assembly features.** Genome size denotes the range between assembled and scaffolded genomes.

Species	Genome size (Mb)	Number of scaffolds	N50 (kb)	Assembler	Depth	Fraction of mapped reads (%)	Read pairs in correct orientation (%)	Ambiguous fraction	BUSCO (%)
P. pacificus	143-151	33,047	438	DISCOVAR	73 X	92-93	94	1.1×10^{-4}	87
P. exspectatus	167-178	4,412	142	ALLPATHS-LG	97 X	90	95-96	1.5×10^{-3}	91
P. arcanus	195-203	4,263	271	ALLPATHS-LG	72 X	96-97	96-97	1.3×10^{-3}	92
P. maxplancki	222-266	69,506	309	DISCOVAR	50 X	95	95	3.9×10^{-4}	91
P. japonicus	199-223	33,291	448	DISCOVAR	49 X	96	96	1.6×10^{-4}	90
P. mayeri	267-297	84,599	235	DISCOVAR	32 X	95	93	1.5×10^{-4}	87
P. entomophagus	242-264	72,722	369	DISCOVAR	36 X	97	97	1.0×10^{-4}	87
P. fissidentatus	233-247	56,870	443	DISCOVAR	39 X	98	94	1.1×10^{-4}	90
P. giblindavisi	178-201	7,303	112	ALLPATHS-LG	50X	94-95	81-92	1.3×10^{-3}	79
M. japonica	180-202	137,965	189	DISCOVAR	61 X	97	87	4.8×10^{-4}	87

Single evolutionary events can explain the majority of gene families

The ladder-like phylogenetic tree (Fig. 4.1a) first allowed the tracking of the phylogenetic origin of genes on nine ancestral nodes and then the assignment of genes into Age classes. I predicted gene annotations based on protein homology and RNA-seq data for all 10 species (Supplemental Table S4.1) and created orthologous gene clusters with OrthAgogue [122] (Fig. 4.1b). OrthAgogue is a faster re-implementation of the extensively used OrthoMCL pipeline [112]. In total, 38,639 clusters with two or more genes were generated, they contain 68-81% of genes in a given genome (Fig. 4.1c). More than 5000 clusters contained at least one gene from each species and hence their origin could be mapped back to the last common ancestor of all 10 diplogastrid nematodes (Fig. 4.1d). Such clusters were marked as 'Age class ix' in our analysis (Fig. 4.1b). Clusters that were only missing *M. japonica* genes, but had at least one gene from the other nine species, were designated as 'Age class viii' (Fig. 4.1d). It is important to note that such clusters either represent an *M. japonica*-specific loss or a taxon-restricted gain. Further, multiple clusters were restricted within a monophyletic sublineage and were marked as 'Age class vii

- i' (Fig. 4.1d). Thus, the lower the Age class, the more recent was the origin of the genes in it.

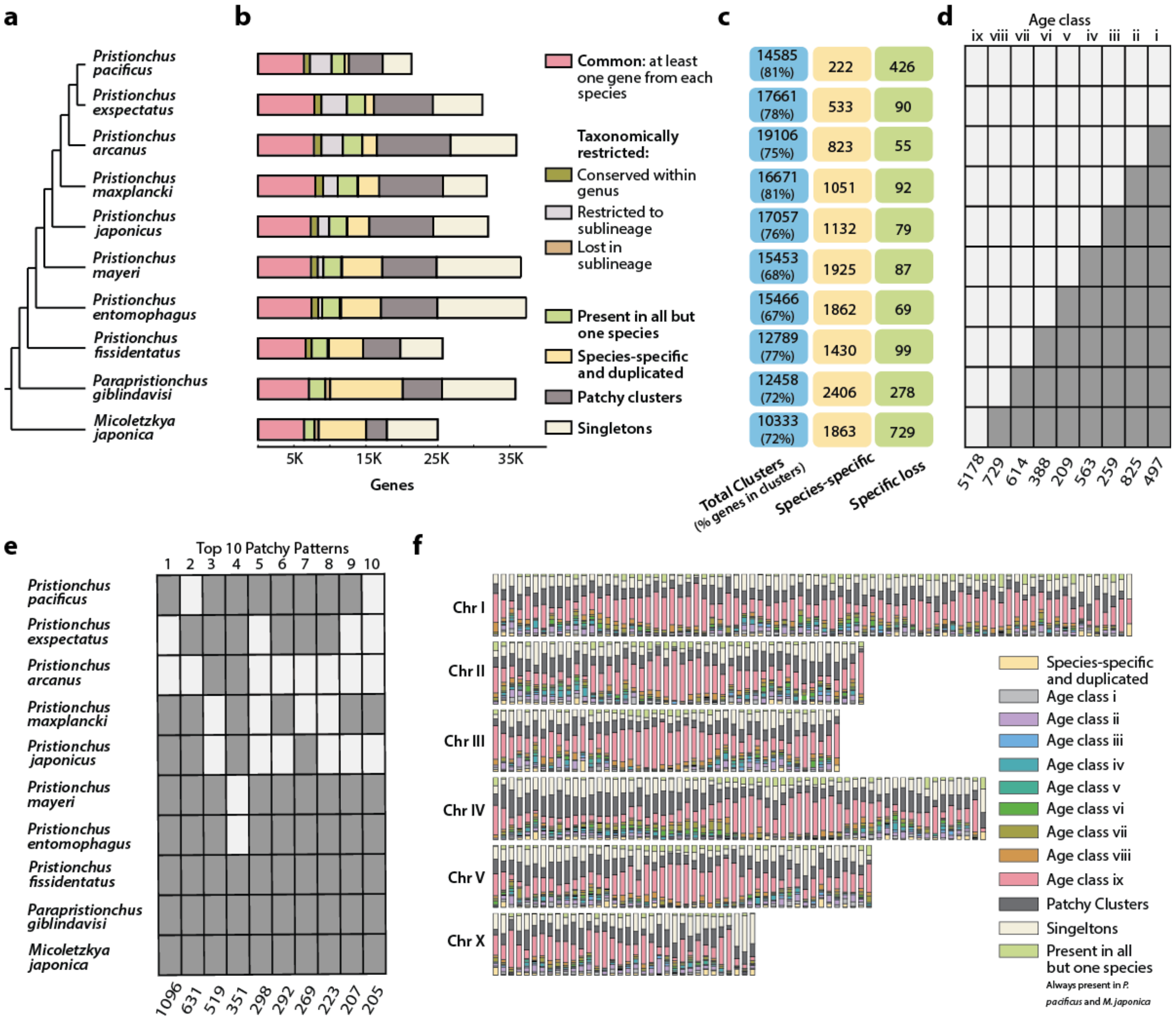


Figure 4.1: Gene classes of *Pristionchus* nematodes and their distribution on *P. pacificus* chromosomes. (a) Overview of phylogenetic relationship among the 10 diplogastrid species. (b) Distribution of genes within orthology classes across the 10 diplogastrid genomes. (c) Numbers of total clusters per species and the percentage of all genes within these clusters, followed by the number of Species-specific clusters, and clusters that have been exclusively lost in the given species. (d) Graphical representation of the Age classes, light rectangle indicates presence of a gene family in the given species and dark rectangle indicates absence of this gene family. The roman numerals at the top of the box indicate the relative age of the Age class. (e) Top 10 species distribution patterns in Patchy clusters. (f) Distribution of all orthology classes in non-overlapping 500 Mb windows across chromosomes suggests that older genes are overrepresented at the chromosome centers. Chromosome II, III, IV, and V have their centers at the middle, Chromosome I has two chromosome centers and Chromosome X has no obvious center.

Next, I identified clusters in which the species distribution of the genes could be parsimoniously explained by gene loss restricted to a sublineage ('Lost in sublineage', Fig. 4.1b). There were multiple clusters having at least one gene from all but one species. I recognized such clusters as 'Species-specific loss' (Fig. 4.1c). Finally, there were many clusters with genes only from a single species, such clusters were classified as 'Species-specific' clusters. They were constituted by Species-specific genes that got duplicated and thus form entire clusters made of paralogs (Fig. 4.1c). Consistent with the phylogeny that governs our study design [84], longer branches between the extant taxa and the more ancestral inner nodes show elevated levels of Species-specific duplications and gene losses. As it can be difficult to distinguish true losses from missing evidence [123,124], the numbers of Species-specific losses within most of the *Pristionchus* species seem rather stable and show substantial increase only in the two outgroups (Fig. 4.1c).

In addition to the cluster categories described thus far, I was left with the genes from every species that were not present in gene clusters. Such genes were classified as 'Singletons'. Although I suspect that some Singletons can be gene annotation artifacts, the results from previous chapter suggests that the majority of Singletons are real protein-coding genes. However, the lack of homologous sequences prohibits any type of selection analysis. Therefore, I focused on orthologous gene clusters with members from at least two species. Taken together, the analysis of these cluster showed that up to 67% of *P. pacificus* gene families can be parsimoniously explained by singular evolutionary events such as a gene gain or a gene loss.

Young gene families are frequently lost

Although the orthologous clusters can be explained by a single evolutionary event, still, 38% of all clusters showed uneven distribution patterns and could only be explained by taking more than one evolutionary event into account. Such clusters were labelled as 'Patchy clusters'. Further analysis of the most common species distribution patterns among the Patchy clusters revealed that most of the top 10 patterns can be explained by just two evolutionary events, *i.e.* a gain at some internal node within the *Pristionchus* genus, followed by a loss either at one of the derived internal nodes or in an extant species (Fig. 4.1e). Specifically, nine out of the 10 most abundant Patchy cluster patterns were not older than the common ancestor of *P. pacificus* and *P. japonicus*. This indicates that younger gene families are more susceptible to gene loss. Further, I found the most abundant Patchy clusters could not distinguish the two different modes of reproduction. Thus, I conclude that the majority of observed genomic changes are better explained by phylogeny.

A chromosome-scale assembly of the *P. pacificus* genome [82] gave me the opportunity to map the genes from different categories onto the six chromosomes. I created non-overlapping windows of 500 kb on each chromosome and calculated the proportion of genes falling into different categories or Age classes within a given window (Fig. 4.1f). The majority of chromosomes showed enrichment for old cluster categories , Age class ix, at the chromosome centers. Note that chromosome centers are not centromeres, as nematodes have holocentric chromosomes [125]. Instead, chromosome centers were defined based on genomic signatures such as high gene density, low repeat content, and reduced nucleotide diversity [82]. Thus, *P. pacificus* Chromosome I seems to have two center-like regions. The finding that the Patchy clusters are preferentially situated at chromosome arms is also consistent with the fact that they mostly represent young gene families that have been secondarily lost in one of the species [126–128].

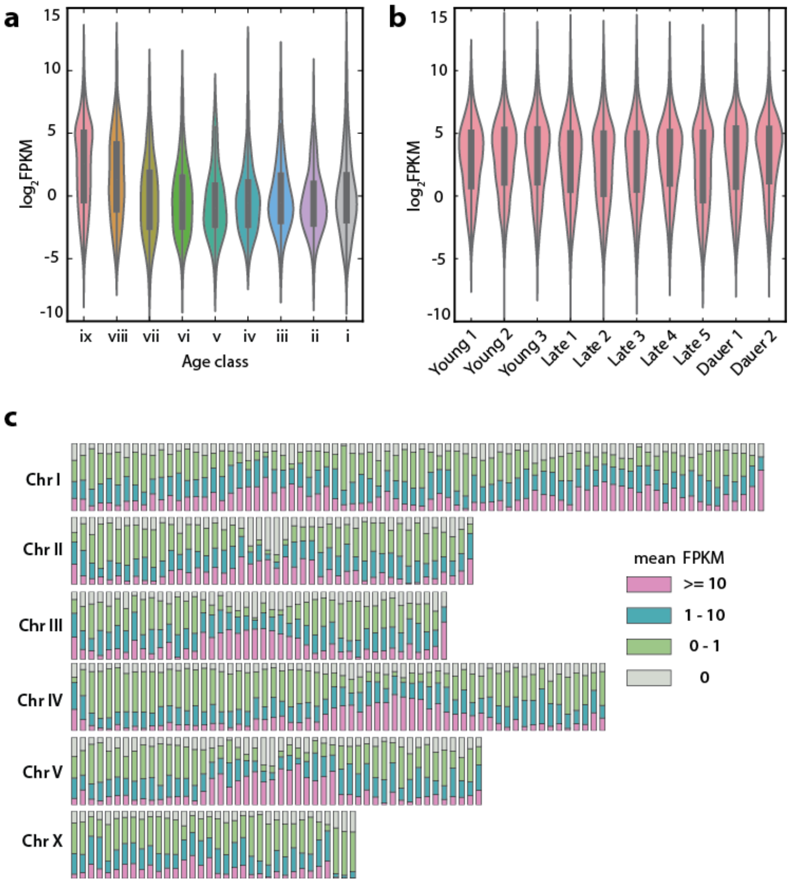


Figure 4.2: Expression increases over time. (a) Expression values for *P. pacificus* genes from different Age classes in an RNA-seq dataset of late larvae and adults (Late 1) indicate that older Age classes are expressed at higher levels. (b) Age class ix genes are expressed at a constitutively high level in all 10 developmental transcriptomes. (c) Distribution of expression classes across the *P. pacificus* chromosomes.

Expression levels increase with gene age

To study the evolution of gene expression over time, I compared different Age classes and their gene expression profiles from several developmental stages of *P. pacificus* [105]. I observed that in all samples, the younger Age classes (Age class i-vii) were expressed at a lower level than the old Age classes (Age class ix and viii), this suggests that the expression levels increase with gene age (Fig. 4.2a, Supplemental Fig. S4.1). Also, Age class ix genes were expressed at relatively high levels in all samples (Fig. 4.2b). Although the correlation between the Age classes and their expression levels was relatively weak (Spearman's rho=0.33, P-value < 2^{-64}, Fig. 4.2a), this was improved by calculating the mean expression value of all genes in all samples (Spearman's rho=0.46, P-value < 2^{-64}). While mapping gene expression levels along 500 kb non-overlapping windows on each chromosome, I observed that the genes with highest expression values (mean FPKM >= 10) are also enriched at the chromosome centers (Fig. 4.2c). Incidentally, at the chromosome centers some windows also had the highest concentration of genes without any expression evidence (mean FPKM = 0), this is most likely due to the presence of older genes with precise spatio-temporarily regulated expression. In summary, the analysis of expression data suggests that young genes generally have either low or spatio-temporarily regulated expression and that their expression tend to rise or become broader over time.

All Age classes are evolutionarily constrained

Next, I investigated the selective pressure acting on the different Age classes. To this end, I estimated the rates of non-synonymous changes (d_N), synonymous changes (d_S) and ω (d_N/d_S) for 1:1 orthologs of *P. pacificus* and every other species. The rates of synonymous changes (d_S) obtained from the pairwise species comparisons were used as a proxy for divergence time and the peak of the d_S distribution was consistent with the species phylogeny (Fig. 4.3a). The two closest neighbor of *P. pacificus*, i.e. *P. exspectatus* and *P. arcanus*, showed d_S peaks at 0.2 and 0.5 substitutions per site. The ω distributions exhibited that all Age classes are under evolutionary constraint (Fig. 4.3b). Additionally, the ω distributions also reflected the species phylogeny, indicating that the older species pairs were under stronger selection (Fig. 4.3b). However, it should be noted that the patterns of ω distribution may also reflect the fact that longer time intervals facilitate the removal of deleterious alleles [111,129,130]. Therefore, I decided to narrow my focus on a fixed evolutionary distance by only considering *P. pacificus* - *P. exspectatus* pairwise dataset for further analysis.

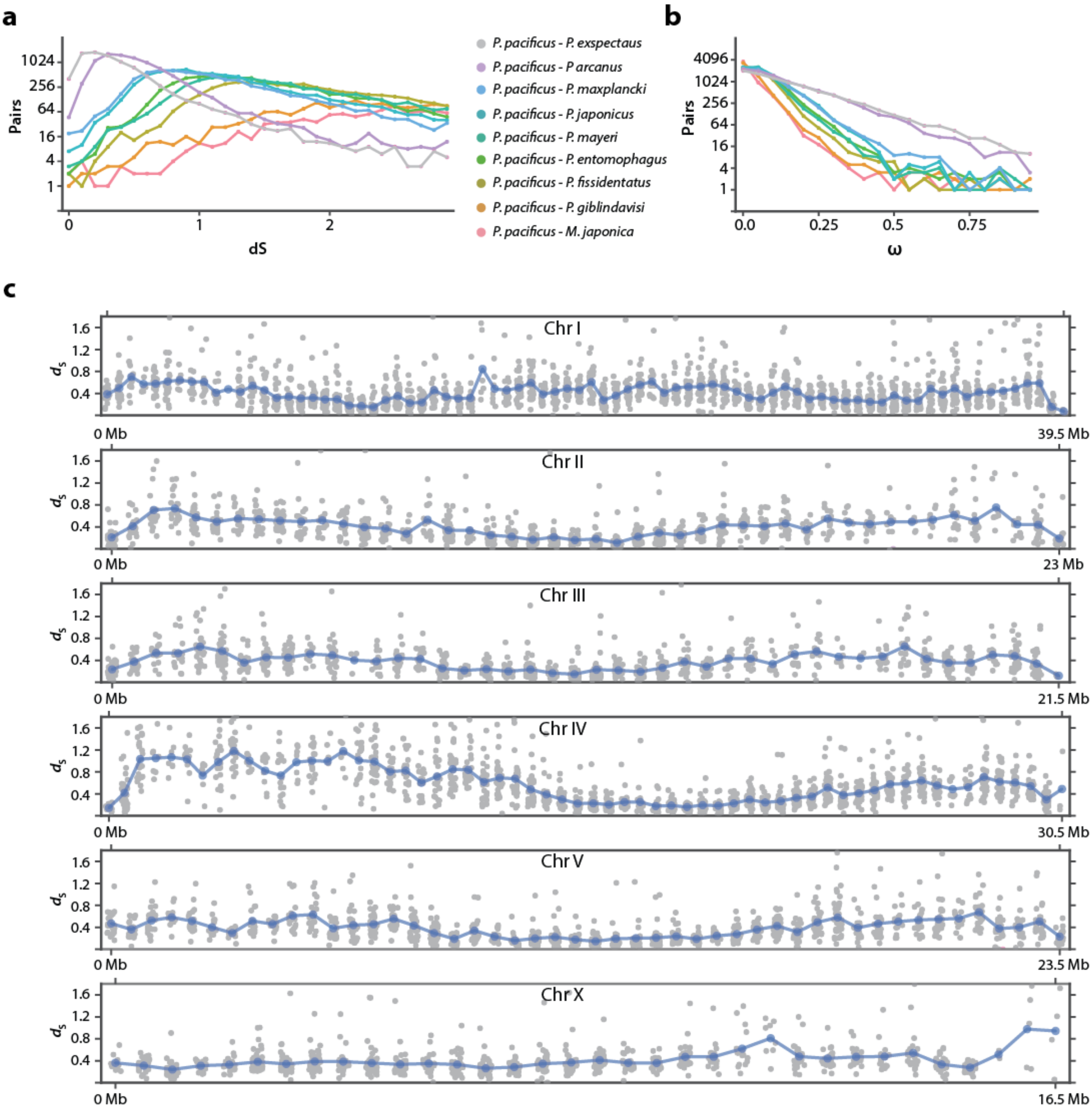


Figure 4.3: Divergence estimates across different time-scales and their chromosomal distribution. (a & b) Pairwise d_S (a) and ω (b) distribution between *P. pacificus* and all other species support the underlying species phylogeny [84]. (c) d_S value of each 1:1 ortholog between *P. pacificus* and *P. exspectatus* were mapped on the *P. pacificus* chromosomes with a running mean for each window (in blue).

Divergence profiles reflect fast evolving chromosome arms and stable centers

A previous report from our lab has concluded that nucleotide diversity is non-uniformly distributed across the *P. pacificus* chromosomes [82]. This suggests that d_S may also vary between different chromosomal locations. To investigate d_S variation across the chromosome, I plotted the d_S values for all pairwise *P. pacificus* - *P. exspectatus* comparisons for 500 kb non-overlapping windows and the running mean for each window (Fig. 4.3c). Median d_S between *P. pacificus* and *P. exspectatus* was 0.33 (IQR= 0.21-0.51), which would roughly correspond to a divergence time of 1-5 mya [131]. Similar to the nucleotide diversity profile across the chromosomes [82], I observed that the d_S values are lower at the chromosome centers and are higher at the arms. Previously, while analyzing evolutionary rates in *Arabidopsis*, Yang and Gaut proposed three non-exclusive processes to explain similar variation in divergence rates, which are codon bias, a non-uniformly distributed mutation rate, and population genetic processes such as background selection [132]. I ruled out codon bias and mutation rate as the main processes behind this variation, as mutation accumulation line experiments in *P. pacificus* and some other nematodes could not provide evidence supporting mutation rate biases [133,134], and the positive correlation between d_N and d_S (Spearman's rho = 0.63 with a P-value < 2^{-64}) restricts the role of codon bias, thus only leaving background selection as a plausible explanation. Further, since the spatial aggregation of Age classes coincided with the d_S distribution and previous analysis of evolutionary constraint have reported old genes to be under stronger purifying selection (Fig. 4.3b), I hypothesized that differences in distribution of Age classes can give the impression of faster evolving chromosome arms and slower evolving chromosome centers.

Young genes evolve more rapidly

Finally, I wanted to examine if chromosomal location determines the level of divergence via background selection. Therefore, I tested whether the degree of selective pressure differs between Age classes. To this end, I decided to look at the ω distribution of different Age classes by segregating them into two d_S ranges (0 - 0.4 and 0.4 - 0.8). While the lower d_S range largely represented chromosome centers, the upper range mainly captured genes at chromosome arms (Fig. 4.3c). In both categories, I observed that the old Age classes were under strong purifying selection (Spearman's rho = 0.56, P-value < 2^{-64}, Fig. 4.4a-b). Although the segregation of d_S corrected for synonymous divergence, I also compared ω distribution for different Age classes along the chromosomes. For this, I divided the Age classes into 'young' (Age class i-viii) and 'old' (Age class ix), and then plotted their ω distribution along 5 Mb non-overlapping window (Fig. 4.4c-d). Again, I observed that the old genes remained under strong purifying selection, while the young genes evolved more rapidly, demonstrating that it is the different proportion of Age classes within a given chromosomal region that explains the non-uniform divergence along chromosomes (Fig. 4.3c).

Further, I quantified the significance of the comparisons of d_N, d_S, and ω along the chromosomes (Supplemental Fig. S4.2). These comparisons were generally statistically significant, supporting the idea that selection can act on individual genes. I conclude that at large evolutionary distances, as the separation among different *Pristionchus* species, the major determinant of the strength of evolutionary constraint acting on most genes are the genes themselves.

Conclusion

Based on the results discussed in this chapter I make four main conclusions. First, the ladder-like phylogeny of the 10 diplogastrid nematode genomes allowed me to trace the evolutionary history of the majority of *P. pacificus* genes including orphan genes, which did not show any homology outside the diplogastrid family [135]. The availability of a chromosome-scale assembly enabled me to map the *P. pacificus* gene predictions to chromosomes based on their respective Age classes [82], portraying that old genes are concentrated at the chromosome centers. This is consistent with the general tendency of new genes to cluster in certain chromosomal segments, which has been associated with other features such as late replication timing and transposons [126,136,137].

Second, young gene families are lost more frequently than the old gene families. This conclusion was based on the distribution of the top 10 patchy clusters, which is skewed towards the genes that originated within the *pacificus* clade but have been lost in one or more species.

Third, the results show that older Age class genes are either more broadly or more highly expressed compared with genes from younger Age classes. This trend persists at every life stage that I looked at, suggesting that in general gene expression increases or becomes broader with time. This finding is also consistent with previous studies in animals and plants [137–139].

Fourth, although the chromosome arms and centers show distinct levels of divergence, this pattern is mainly created by differences composition of young and old Age classes, which themselves show different levels of evolutionary constraint. Additionally, in agreement with prior studies [137,140,141], younger Age classes evolve more rapidly than older Age classes. This suggests that at evolutionary time-scales, such as the separation among different *Pristionchus* species, selection tends to act on individual genes independent of their chromosomal location.

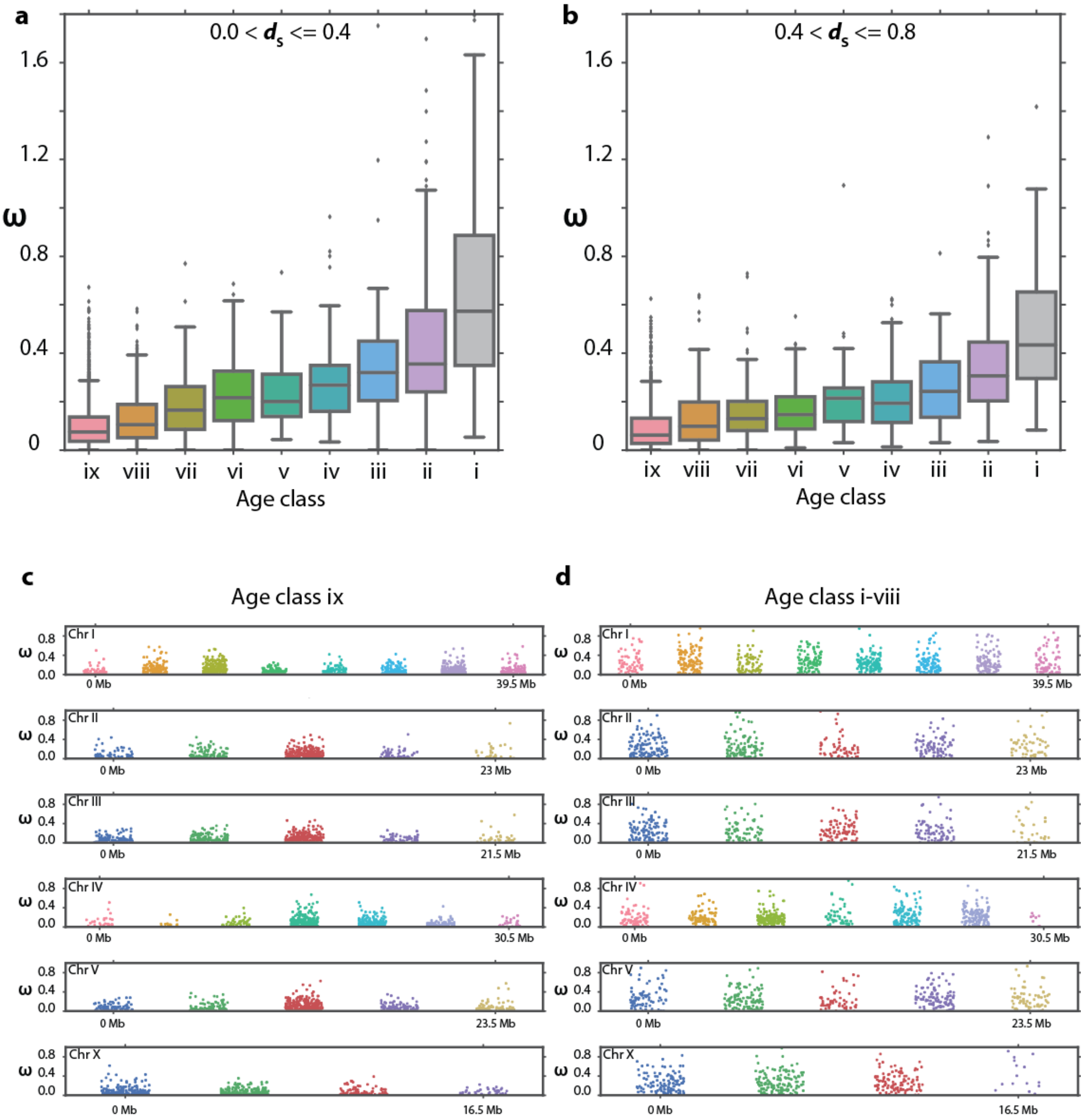


Figure 4.4: Young genes evolve more rapidly. (a & b) ω values decrease with age in both the d_S ranges indicating that young genes evolve rapidly and become more constrained over time. The ω values of 1:1 orthologs between *P. pacificus* and *P. exspectatus* of Age class ix (c) and Age class i-viii (d) in 5 Mb windows show that young genes are less constrained irrespective of the chromosomal location. For comparison, in both panel c and d, corresponding windows on each chromosome have the same color.

Methods

DNA extraction, sequencing, assembly, and scaffolding

All nematodes worms were grown on nematode growth medium (NGM) plates and gonochoristic species were inbred for 10 generations before DNA extraction. I rinsed the plates with M9 buffer and collected worm pellets by centrifugation at 1300 rpm for 3 minutes at 4°C. Then I followed the method specified by Rödelsperger et al. for DNA extraction [82]. Overlapping and mate pair illumina libraries for *P. arcanus* and *P. giblindavisi* were generated based on the protocol specified by Rödelsperger et al [111,121]. For the other species, PCR free libraries were generated using TruSeq DNA PCR-Free Library Prep kit following the manufacturer's protocol and the sequencing was done on Illumina MiSeq. These species included *P. pacificus* itself, which I chose to resequence and assemble *de novo* in order to make the datasets more comparable. Initial assemblies were created with the DISCOVAR *de novo* assembler (version r52488, https://software.broadinstitute.org/software/discovar). I checked for *E. coli* contamination by blastn search against in-house and NCBI *E. coli* genomes and removed dubious contigs after manual inspection. Final scaffolding was done with SSPACE_Basic_v2.0 [142] using mate pair libraries of sizes 1.5, 3, 5, and 8 kb, which were generated with Nextera Mate Pair Sample Preparation Kit.

Assembly evaluation

To evaluate the completeness of final assemblies, I calculated the fraction of raw reads represented in each final assembly. This was done by aligning reads from individual libraries using BWA (version 0.7.12-r1039) and stampy (version v1.0.21 r1713), and extracting the fraction of aligned reads from the SAMtools flagstat program (version 0.1.19-96b5f2294a) output [143–145]. The SAMtools flagstat output also provided information on the fraction of correctly oriented paired-ends reads that can be interpreted as a measure of correctness. Additionally, based on the realignments, I defined the ambiguous fraction as the fraction of genome assembly with heterozygous variant calls [111]. Finally, I employed the universal single-copy orthologs benchmarking (BUSCO, version 3.0.1) approach as an extra measure for assembly completeness [146]. Based on the qualification of the BUSCO gene set to be conserved as single copy >90% of genomes, the effective maximum expected score should be slightly above 90% and is attained for the *P. arcanus* (Table 4.1) genome as well as the earlier published *P. pacificus* genome [82].

RNA extraction, sequencing and assembly

Worm pellets for all species were gathered by the above mentioned methods and were resuspended in 10 volumes of TRIzol. RNA extraction was done using Direct-zol RNA miniprep kit (Zymo research) and library preparation was done with Illumina TruSeq RNA Library Prep Kit v2. These libraries were sequenced on Illumina HiSeq 3000. I assembled the transcriptome with 'trinityrnaseq-2.2.0' [147]. For *P. pacificus*, I also generated a strand-specific transcriptome assembly using previously published RNA-seq data [148,149].

Gene annotation

I employed both AUGUSTUS (3.2.2) and SNAP within the Maker2 (v2.31.8) pipeline for protein-coding gene prediction [15,101,102]. Three iterations of this pipeline were run, in the first run the gene finders were trained on the given transcriptome assembly. In the second run, I generated gene models that were at least partially supported by the given transcriptome (AED_threshold < 1). In the final run, along with all the evidence used in the second run, I additionally used gene models resulting from the second run of all other species as protein homology data. Moreover, in the final run, I retained predicted gene models without any transcriptome or homology evidence (AED_threshold ≤ 1). For runs 2 and 3, I used minimum contig length of 2 kb (min_contig=2000). PFAM domains were annotated with InterProScan-5.19-58.0 [150]. In order to visualize the genomic features distribution across chromosomes, *P. pacificus* protein annotations were mapped on to the El Paco assembly of *P. pacificus* using the exonerate protein2genome program (version 2.2.0) [151].

Orthology clustering and inference of gene gain and loss

I ran pairwise blastp (E-value < 10^{-5}) between all species pairs in our analysis and created orthologous gene clusters with OrthAgogue and MCL (both programs were run with default settings) [122,152]. Based on the presence and absence of genes from different species, each cluster was segregated into different categories. Based on maximum parsimony, clusters were classified into Age classes, each of which corresponds to a single origin at an internal branch of the phylogeny (Fig. 4.1a).

Expression analysis

I mapped stage-specific transcriptome data, from 10 samples, generated by Baskaran et al. to the *P. pacificus* genome using TopHat2 [105,153]. Then, I computed the expression values for the *P. pacificus* annotations in each sample with Cufflinks 2.2.1 [154]. Expression patterns, mean expression values, and mapping of mean expression pattern on chromosome were generated with custom Python scripts.

Estimation of evolutionary constraints

Pairwise 1:1 orthologs between *P. pacificus* and other species were extracted by selecting only those clusters that have only one gene each from *P. pacificus* and the other species. I aligned 1:1 orthologs with MUSCLE [113] and converted protein alignments into codon alignments using PAL2NAL [113,114]. These codon alignments were passed on to PAML to compute the rate of synonymous (d_S) and non-synonymous (d_N) substitution, and ω (d_N/d_S) values [99].

Chapter 5: Illustrating diverse mechanisms of orphan gene emergence

This chapter contains content from the following manuscript.
Prabh N, Rödelsperger C. The diversity of orphan gene origin - illustrated.

Author	Author Position	Scientific ideas %	Data generation %	Analysis and interpretation %	Paper writing
N.P.	1	70	NA	75	80
C.R.	2	30	NA	25	20
Title of the paper:	The diversity of orphan gene origin - illustrated.				
Status:	Submitted				

In the last chapter, I employed deep taxon sampling of nematode genomes to probe the evolutionary dynamics of novel genes. In this chapter, I focus on Species-specific orphan genes (SSOGs) and show that they account for roughly 10% of all genes in each Pristionchus spcies irrespective of the sampling depth. Phylostratigraphic analysis indicates an exceptionally high number of SSOGs, which could be explained by the presence of a rapidly evolving gene pool or by a constant fraction of annotation artifacts. Based on sequence searches in other closely related genomes, I found traces of homology for 61% of *P. pacificus* SSOGs, which allowed me to gain mechanistic insights into their emergence. Manual inspection of high-confidence SSOGs revealed heterogeneous divergence mechanisms including chimeric origin, alternative reading frame usage, and gene splitting with subsequent exon gain. In addition, I present two cases of complete *de novo* origination from non-coding regions, which represents the first report of *de novo* genes in nematodes.

Results

Roughly 10% of all genes are Species-specific irrespective of the sampling depth

To investigate how variable is the fraction of orphan genes across the diplogastrid genomes, I applied a three-step filtering procedure to define a robust orphan gene set for the eight *Pristionchus* and two outgroup species (Fig. 5.1a). Given that all 10 genomes have around one third of genes classified as orphan, I next explored how conserved these genes are within the family Diplogastridae. First, I identified orphan genes that have a blastp hit (E-value < 10^{-3}) in at least one other diplogastrid. If an orphan gene fulfilled this criterion, it was labelled as 'Taxonomically-restricted orphan gene' or 'TROG', otherwise it was classified as 'Species-specific orphan gene' or 'SSOG'. According to this classification, more than 70% of all orphan genes are classified as TROGs for every *Pristionchus* species except for *P. fissidentatus* (Fig. 5.1b, Supplemental Fig. S5.1). Thus, approximately 10% of all predicted genes in different *Pristionchus* species lack any homology at the protein level with any other species and fall in the SSOGs category. This lack of phylogenetic signal is unexpected, since the taxonomic sampling is much deeper around the focal species *P. pacificus* (Fig. 5.1a) and encompasses the two sister species, *P. exspectatus* and *P. arcanus*, that can still form viable but sterile hybrids with *P. pacificus* [155]. Hence I naively anticipated that this should result in a much lower fraction of SSOGs in the focal species and its close neighbors. While I cannot rule out that a constant fraction of erroneous gene annotations partially contributes to this pattern, however, these results are consistent with the idea that novel genes are frequently generated as a result of pervasive transcription but rarely reach fixation and are rapidly lost [156].

SSOGs make the most gene rich phylostratum

To gain more detailed insights into the age distribution of *P. pacificus* orphan genes, I separated them into different phylostrata that can be mapped to the most recent common ancestors of *P. pacificus* and the other diplogastrid species (Fig. 5.1c). Based on the parsimonious assumption that the breadth of a gene's phylogenetic distribution is an indicator of its age, a gene that is shared with a distantly related species is expected to be older than a gene that is only shared with a close neighbor. Thus, each orphan gene was placed into the phylostratum that points to the most recent common ancestor of *P. pacificus* and its most distantly related species that has a homolog of this gene [157]. Additionally, the *P. pacificus* SSOGs were placed in the 'Phylostratum 0'. I found that the number of genes in a given phylostratum correlates with the amount of divergence between its ancestor to the next extant species (Spearman's rho = 0.6, Fig. 5.1d), suggesting a constant rate of fixation of novel genes within the *Pristionchus* lineage. I have excluded phylostrata 8 and 9 from this analysis, as I have previously shown that novel gene families have a high propensity of being lost, which could lead to an underestimation of the number of genes gained at these ancient splits. 'Phylostratum 0' constitutes an exception as it has by far the highest number of genes among all 10 phylostrata, yet the length of the terminal branch leading towards *P. pacificus*, which depicts this phylostratum, is relatively short (Fig. 5.1a, 1d). Given that the number of genes in this phylostratum is exceptionally high, this further supports that most gene-like sequences are not long-lived enough to survive a speciation event.

Most *P. pacificus* SSOGs have traces of homology in closely related genomes

The deep taxon sampling around the focal species *P. pacificus* allowed me to screen for traces of homology of *P. pacificus* SSOGs in the genomes of sister species, which could be indicative of their mechanism of origin. To this end, I performed various blast searches against the annotated transcripts, genome assembly, and transcriptome assembly (Fig. 5.1e). While tblastn searches against the genome assembly of other species may identify homologous non-coding regions of *de novo* candidates, I additionally performed a blastn search against the annotated transcripts to screen for potential cases of ORF switching, and a blastn search against the transcriptome assembly to assess the degree of missing homology due to assembly gaps. 504 (32%) of *P. pacificus* SSOGs show blast hits in all three target database types, which after closer investigation was seen to be largely due to overlapping gene structures such as 3' UTR overlap of neighboring genes [149,158]. Another 479 (31%) of *P. pacificus* SSOGs did not show hits in any of the databases and were labeled 'Untraceable'. Among the remaining SSOGs, I found only 29 (2%) with a hit in the transcriptome assembly but not in the genome or the annotated transcripts. The fraction of putative assembly gap genes is constantly low for all the genomes supporting their comparable quality [23]. In total, 1082 (61%) of *P. pacificus* SSOGs exhibit detectable traces of homology in the genomes of other closely related species, demonstrating that the taxon sampling of the phylogenomic dataset is sufficient to study the mechanisms of origin for the most *P. pacificus* SSOGs in greater detail.

Identification of a high-confidence candidate set for origin analysis

Given that more than a thousand *P. pacificus* SSOGs have traces of homology in closely related sister species and the gene structures of orphan genes in general are poorly supported by expression evidence [135], I first needed to define a high-confidence candidate set of SSOGs that could be used for detailed gene origin analysis. I only considered SSOGs with more than one annotated exon, because I hypothesized that this additional layer of regulated expression involving the proper splicing of the transcripts would yield a more likely protein-coding gene candidate with confirmed regulated expression as opposed to pervasive transcription of non-genic sequences [159]. Additionally, the splice sites can be informative to better predict the correct orientation of the gene, which is essential to elucidate their origin. I manually inspected all *P. pacificus* SSOGs except the Untraceable genes, in total 1082 candidate loci, to find gene structures that are fully confirmed by raw read alignments of existing RNA-seq datasets and I insisted on finding a minimum of two raw RNA-seq reads that aligned with each coding exon and two spliced reads that span such exons. Eventually, I established 29 SSOGs with fully confirmed gene structures (Fig. 1e) which formed the high-confidence candidate set. Based on my investigation, I provide examples for six plausible mechanisms that explain the origin of SSOGs including two examples of *de novo* genes. Most of the remaining high-confidence candidates are either instances of the proposed mechanisms or their origin cannot be unambiguously concluded.

Divergence by recycling of ancestrally protein-coding fragments

The first mechanism alludes to chimeric gene formation resulting in an SSOG with two exons that are derived by partial duplication from distinct source genes. The paralogous exons from both the ancestral source genes are duplicated and then inserted in close proximity to facilitate the formation of a novel ORF (Fig. 5.2a). Considering that such genes can be created by minimal contribution from existing genes, local alignment based tools may fail to detect the homology of these genes with their paralogous exons. For example, PS312-mkr-S378-0.29-mRNA-1 is a *P. pacificus* SSOG with two exons. Its first exon has 100% protein identity with an exon from a *P. exspectatus* gluthatione peroxidase gene, while its second exon shows partial identity with an exon of another conserved gene (exspectatus-mkr-S_440-0.48-mRNA-1, Fig. 5.2b). Orthologs of both *P. exspectatus* genes are maintained in *P. pacificus*. Based on blastp comparison, neither of the two exons of the candidate gene match their paralogous exons from the two *P. exspectatus* genes. This suggests that even if high percentage-identity is retained between paralogous exons, chimeric genes can lack detectable blast homology with related genic parts from other species leading to their classification as SSOG.

The second mechanism of SSOG creation is based on splitting of an ancestral gene (Fig. 5.2c). After the split, either both or one of the fragments can diverge from the ancestral sequence and can also acquire new exons. If the length of the ancestral exon or exons in a given split gene is small, a moderate level of divergence can result in a failure to detect homologous sequences. Based on synteny information and spliced alignment, I mapped the first exon of this gene to the first exon of a conserved gene (exspectatus-mkr-S_158-0.48-mRNA-1) in *P. exspectatus*, another *P. pacificus* gene is homologous to the remaining exons of the *P. exspectatus* gene. Upon manual inspection, I found that the first exon of the *P. pacificus* SSOG has acquired an insertion that shifted its reading frame and renders protein homology undetectable. Although some of the N-terminal residues are identical to the *P. exspectatus* protein (Fig. 2e), the remaining residues from the first exon of the candidate gene were found to be derived from other reading frames of the orthologous *P. exspectatus* exon. Hence, it is clear that the predicted ORF from the first exon of the candidate gene is mainly derived from the non-ancestral reading frame. Moreover, the initial segment, which partially retains the ancestral ORF, is not large enough to facilitate homology detection. Ancestry of the second exon of the *P. pacificus* SSOG could not be established even after manual inspection. This suggests that the second exon has been acquired *de novo*. Thus, origin of the candidate gene can be attributed to gene split, partial ORF shift, and *de novo* acquisition of a new exon.

New gene creation through alternative reading frame usage

So far, I have discussed two mechanisms of new gene creation that require deviation from an existing gene structure but maintain the ancestral reading frame either fully or partially. Here I discuss a third mechanism that involves strand switching, which results in a completely new ORF (Fig. 5.3a). The *P. pacificus* SSOG PS312-mkr-S198-1.6-mRNA-1, which has two coding exons, is an example of such a mechanism. In *P. pacificus*, this gene is placed within an intron of a conserved gene (Supplemental Fig. S5.2). This intron is 2.1 kb long in *P. pacificus*. The corresponding intron of the *P. exspectatus* ortholog is 1.4 kb long and shows no homology to the candidate SSOG at the nucleotide or the protein level (Supplemental Fig. S5.2). Spliced alignment of the candidate SSOG on to the *P. exspectatus* genome did not generate any match. Thus, I performed a blastn match against the *P. exspectatus* genome at a relaxed threshold of E-value < 10 (Fig. 5.3b). The resulting aligned genomic section was traced to a single exon of exspectatus-mkr-S_1052-0.1-mRNA-1 gene whereby the candidate has some sequence identity with a reading frame from the reverse strand of the *P. exspectatus* gene (Fig. 5.3b-c).

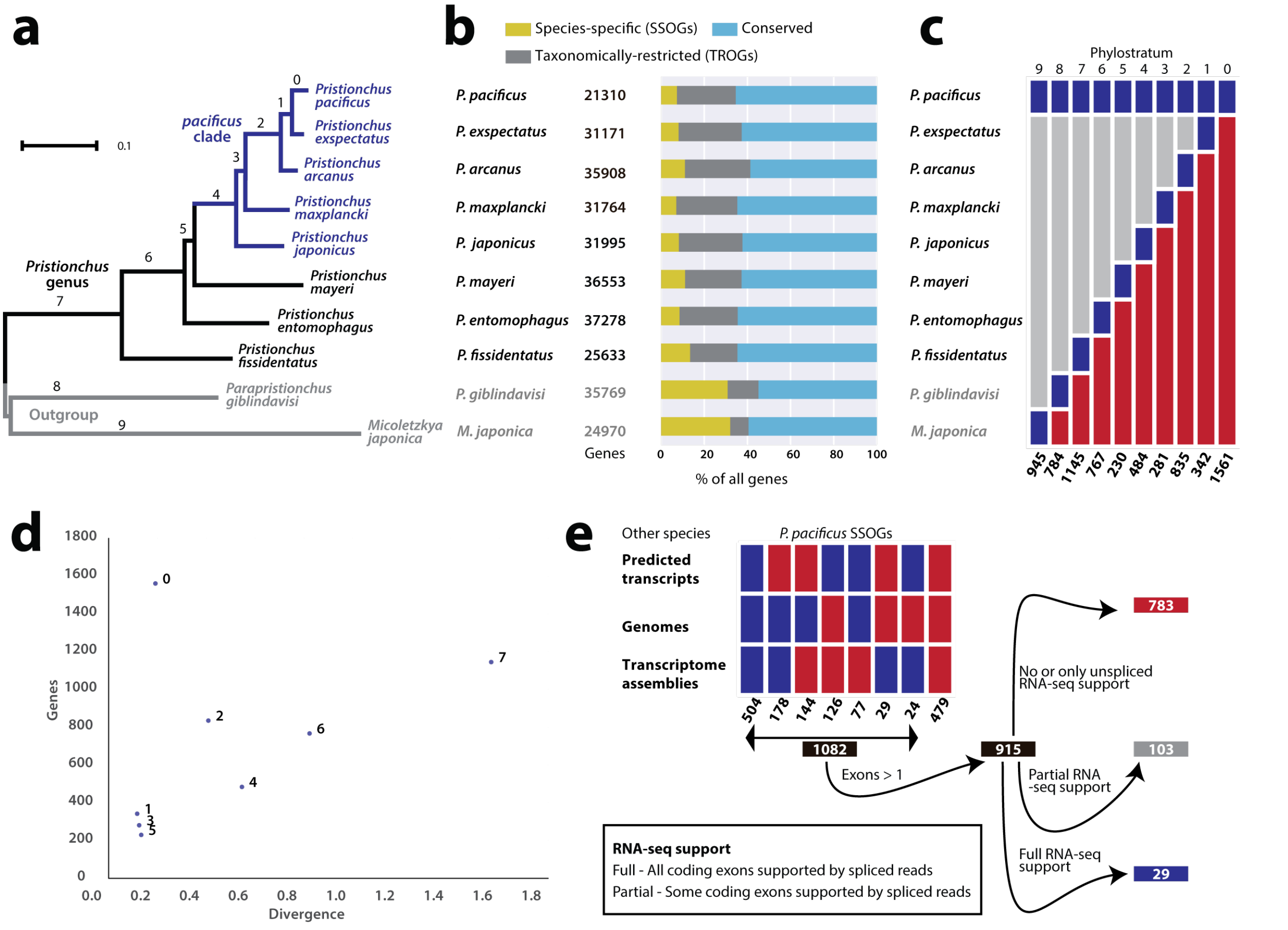


Figure 5.1: Fraction of SSOGs is consistent within the *Pristionchus* genus irrespective of divergence time (a) The maximum-likelihood phylogenetic tree of the species analysed in this study, adapted from Rödelsperger et al. 2018 [160]. Branch lengths denote the number of amino acid substitutions per site. The numbers correspond to the phylostrata from panel c. **(b)** The horizontal stacked bars show the fractions of conserved genes, TROGs, and SSOGs. **(c)** The 10 phylostrata depict the origin of *P. pacificus* orphan genes along the diplogasrid lineage. Blue boxes indicate presence of *P. pacificus* orphan genes and the most distant diplogastrid species that has homologs of these gene, red bars indicate absence of homologs, and grey bars indicated homologs may or may not be present. The number of *P. pacificus* orphan genes in each phylostratum are at the bottom. **(d)** On x-axis I have the divergence estimate for each phylostratum and on y-axis I have the number of genes in them. Phylostratum 0 is the clear outlier that has the highest number of genes within little divergence time. **(e)** The heatmap shows traces of homology for *P. pacificus* in predicted transcripts, genomic, and transcriptomic data of other species. The rectangles indicate whether traces of homology were found (blue) or not (red). Manual inspection of *P. pacificus* RNA-seq data resulted in a high-confidence dataset of 29 *P. pacificus* SSOGs which were taken as the starting point for origin analysis.

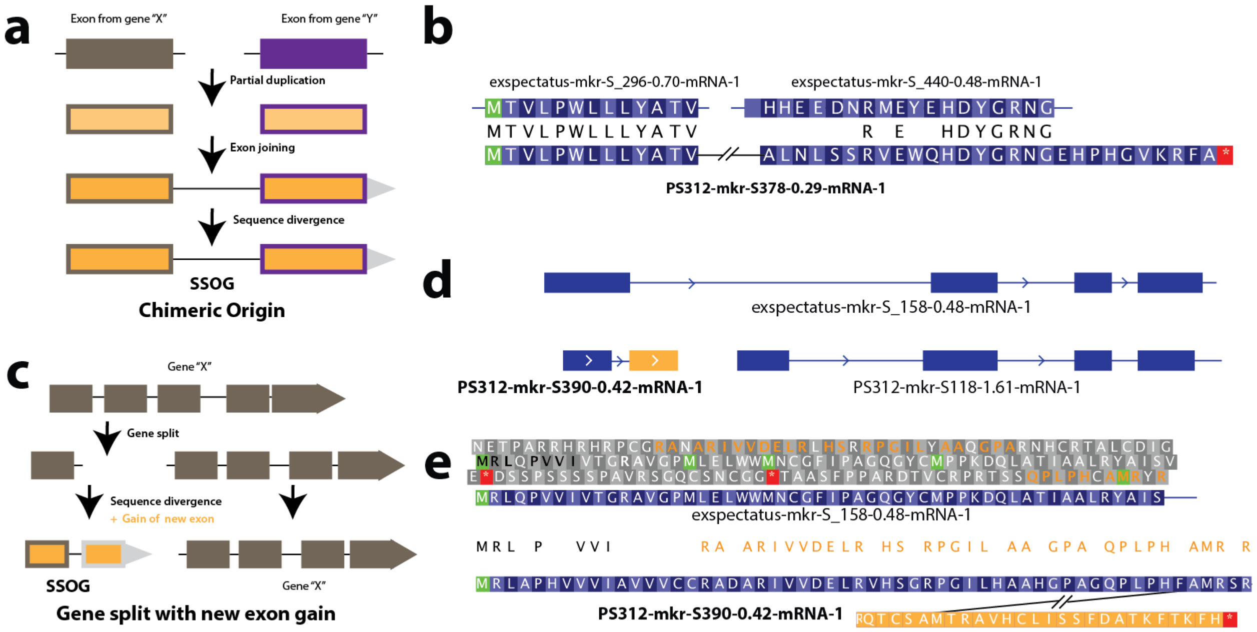

a
Exon from gene "X"
Exon from gene "Y"
Partial duplication
Exon joining
Sequence divergence
SSOG
Chimeric Origin
b
exspectatus-mkr-S_296-0.70-mRNA-1
exspectatus-mkr-S_440-0.48-mRNA-1
MTVLPWLLLYATV
HHEEDNRMEYEHDYGRNG
MTVLPWLLLYATV
R E HDYGRNG
MTVLPWLLLYATV
ALNLSSRVEWQHDYGRNGEHPHGVKRFA*
PS312-mkr-S378-0.29-mRNA-1
c
Gene "X"
Gene split
Sequence divergence
+ Gain of new exon
SSOG
Gene "X"
Gene split with new exon gain
d
exspectatus-mkr-S_158-0.48-mRNA-1
PS312-mkr-S390-0.42-mRNA-1
PS312-mkr-S118-1.61-mRNA-1
e
NETPARRHRHRPCGRANARIVVDELRLHSRRPGILYAAQGPARNHCRTALCDIG
MRLQPVVIVTGRAVGPMLELWWMNCGFIPAGQGYCMPPKDQLATIAALRYAISV
E*DSSPSSSSPAVRSGQCSNCGG*TAASFPPARDTVCRPRTSSQPLPHCAMRYR
MRLQPVVIVTGRAVGPMLELWWMNCGFIPAGQGYCMPPKDQLATIAALRYAIS
exspectatus-mkr-S_158-0.48-mRNA-1
MRL P VVI RA ARIVVDELR HS RPGIL AA GPA QPLPH AMR R
MRLAPHVVVIAVVVCCRADARIVVDELRVHSGRPGILHAAHGPAGQPLPHFAMRSR
PS312-mkr-S390-0.42-mRNA-1
RQTCSAMTRAVHCLISSFDATKFTKFH*

Figure 5.2: Sequence divergence and ORF shift erode evidence of homology. (a) The schematic overview shows an example of an SSOG with chimeric origin. Two exons gained from partial duplication of two distinct genes are joined together and with time sequence divergence occurs. Thus, traces of sequence homology with the original exons become hard to detect and such genes get classified as an SSOG. **(b)** This example shows a *P. pacificus* SSOG (PS312-mkr-S378-0.29-mRNA-1) of chimeric origin and its alignments with parts of two conserved *P. exspectatus* genes. Identical amino acid residues are labelled in black between the *P. pacificus* and *P. exspectatus* exons. Even though the first exon is 100% identical with its homolog, the stretch of alignment is not long enough to be detected by blastp at the stipulated E-value cutoff. **(c)** Schematic overview of a gene split with subsequent exon gain which results in an SSOG **(d)** The *P. pacificus* SSOG PS312-mkr-S390-0.42-mRNA-1 is homologous to the first exon of a conserved *P. exspectatus* gene. The neighboring gene shows homology with the remaining exons, indicating that the SSOG is derived from a gene split event. **(e)** The alignment of the *P. pacificus* SSOG with *P. exspectatus* is spread over multiple reading frames. Amino acid identity between the predicted reading frame of both the proteins are marked in black and those from the other reading frame of the exspectatus gene are marked in saffron. The residues corresponding to the *P. pacificus* SSOG in different reading frames of the *P. exspectatus* sequence are also labelled in black.

Although the identity between the candidate and its putative homologous exon from *P. exspectatus* is not substantial (tblastn E-value = 2.37), based on this alignment I propose that the candidate SSOG gene and the homologous exon of the *P. exspectatus* gene, share a common ancestry. I propose that the exon was duplicated in the lineage leading to *P. pacificus*, was split by gaining an intron, and switched strand to gain a novel ORF. The evidence for strand switching comes from strand-specific RNA-seq data (Supplemental Fig. S5.2).

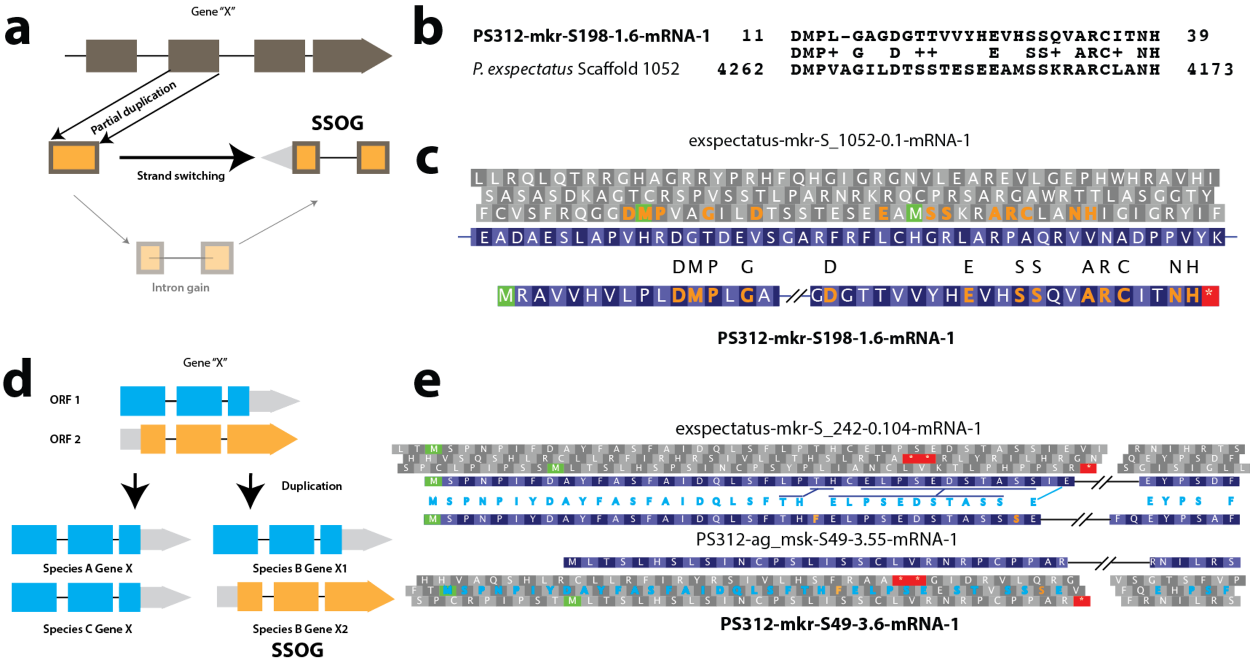

a
Gene "X"
Partial duplication
SSOG
Strand switching
Intron gain
b
PS312-mkr-S198-1.6-mRNA-1 11 DMPL-GAGDGTTVVYHEVHSSQVARCITNH 39
DMP+ G D ++ E SS+ ARC+ NH
P. exspectatus Scaffold 1052 4262 DMPVAGILDTSSTESEEAMSSKRARCLANH 4173
c
exspectatus-mkr-S_1052-0.1-mRNA-1
LLRQLQTRRGHAGRRYPRHFQHGIGRGNVLEAREVLGEPHWHRAVHI
SASASDKAGTCRSPVSSTLPARNRKRQCPRSARGAWRTTLASGGTY
FCVSFRQGGDMPVAGILDTSSTESEEAMSSKRARCLANHIGIGRYIF
EADAESLAPVHRDGTDEVSGARFRFLCHGRLARPAQRVVNADPPVYK
DMP G D E SS ARC NH
MRAVVHVLPLDMPLGA GDGTTVVYHEVHSSQVARCITNH*
PS312-mkr-S198-1.6-mRNA-1
d
Gene "X"
ORF 1
ORF 2
Duplication
Species A Gene X
Species B Gene X1
Species C Gene X
Species B Gene X2
SSOG
e
exspectatus-mkr-S_242-0.104-mRNA-1
MSPNPIYDAYFASFAIDQLSFTH ELPSEDSTASS E EYPS F
PS312-ag_msk-S49-3.55-mRNA-1
PS312-mkr-S49-3.6-mRNA-1

Figure 5.3: Switching to an alternate reading frame gives rise to SSOGs. (a) Partial duplication in combination with an intron gain can allow opening of a new reading frame from the opposite strand. **(b)** *The P. pacificus* SSOG*,* PS312-mkr-S198-1.6-mRNA-1, is an example of both strand switching and exon splitting. Here I show amino acid identity and similarity between my candidate SSOG with the translation from the *P. exspectatus* genome. **(c)** This is a two exon gene, and both the exons share a remote homology with opposite strand of one single exon of a *P. exspectatus* gene at the aligned locus. The identical amino acid residues between the *P. pacificus* SSOG and its corresponding *P. exspectatus* ORF are marked in saffron. **(d)** The schematic overview illustrates a case of actualization of an alternative reading frame by duplication. Overprinting describes a gene with two alternate ORFs. Gene prediction tools generally do not annotate alternate overlapping ORFs from the same strand. However, duplication might generate gene copies where the alternative ORF will be annotated. Nevertheless, in species with single copy of this gene only one ORF gets predicted and due to lack of protein homolog in other species the alternate ORF will be categorized as SSOG. **(e)** PS312-mkr-S49-3.6-mRNA-1 is a four exon, HT category SSOG. Its *P. exspectatus* homolog is predicted from the same strand but in a different reading frame. Both genes maintain both ORFs. I found a *P. pacificus* gene, PS312-ag_msk-S49-3.55-mRNA-1, which is predicted in the *P. exspectatus* ORF and their identical amino acid residues are marked in turquoise between their exons and also in corresponding reading frame of my candidate SSOG. Comparison of this reading frame between the two *P. pacificus* genes shows two residues, in saffron, that are uniquely found in these genes. This indicates that SSOGs can be generated by prediction of an alternate ORF.

The fourth mechanism deals with genes that can have more than one overlapping ORFs. This phenomenon is known as overprinting and has been reported in several studies [51,54–60,62,64]. Generally, gene prediction tools only annotate non-overlapping ORFs from the same strand of the DNA. However, if an ancestral gene with two overlapping same strand ORFs gets duplicated in a lineage, one of the duplicates can switch to the less common ORF (Fig. 5.3d). This will lead to classification of the duplicated gene as a SSOG, as the corresponding ORF has not been annotated in any other species. I found that the *P. pacificus* SSOG PS312-mkr-S49-3.6-mRNA-1 is one candidate for such a scenario as it lacks protein homologs with any other species. Interestingly, this gene has a paralog, PS312-ag_msk-S49-3.55-mRNA-1, at the predicted transcript level (blastn E-value = 0.00, identity = 92.34%). However, the protein predicted from the candidate SSOG is in a reading frame that differs from that of the transcript level paralog. I found that both ORFs are available to both paralogs. The predicted ORF of the paralog is conserved within the genus and has its orthologous ORF in *P. exspectatus* (Fig. 5.3e). Selection analysis indicates that the predicted *P. pacificus* ORF shows an ω value of 1.6 whereas the ancestral ORF shows evidence of negative selection (ω = 0.38), suggesting that the predicted *P. pacificus* ORF is an annotation artifact. However, in the absence of conclusive evidence, I cannot completely reject the predicted reading frame. Therefore, it is plausible that gene duplication allows actualization of such alternative ORFs.

Heuristic failures in homology detection contribute to classification as SSOGs

The fifth mechanism of SSOG formation specifically deals with the fact that blast programs implement a heuristic approach to find sequence matches and that these programs are typically run with default settings. It is obvious that lowering thresholds (e.g. E-value) or switching to a more sensitive alignment approach (e.g. exonerate) facilitates the identification of homologous sequences for a number of *P. pacificus* SSOGs that were missed by blast programs. This has been illustrated by the identification of homologous regions for the previously described divergence cases (Fig. 5.2b, 3b). During my investigation of high-confidence SSOG candidates, I encountered two repeat rich SSOGs, PS312-mkr-S142-0.63-mRNA-1 and PS312-mkr-S81-0.14-mRNA-1, where more detailed investigation of the syntenic region facilitated the identification of a homologous region in the *P. exspectatus* genome. However, it appears that even when blast's repeat filtering is switched off, it fails to detect homology due to the combination of a small non-repetitive match and indels as well as substitutions in the repeat-rich region (Fig. 5.4b). Even though I cannot be sure how specific this behaviour is to repeat-rich genes, these two examples, together with the previous examples illustrate how the failure of any heuristic approach to detect homology will inevitably lead to the classification of certain genes with homologs as SSOGs.

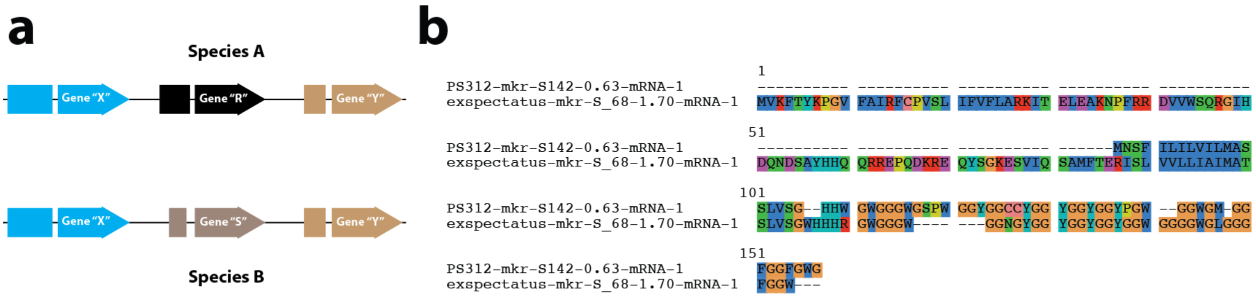

a
Species A
Gene "X"
Gene "R"
Gene "Y"
Gene "X"
Gene "S"
Gene "Y"
Species B
b
1
PS312-mkr-S142-0.63-mRNA-1 ---------- ---------- ---------- ---------- ----------
exspectatus-mkr-S_68-1.70-mRNA-1 MVKFTYKPGV FAIRFCPVSL IFVFLARKIT ELEAKNPFRR DVVWSQRGIH
51
PS312-mkr-S142-0.63-mRNA-1 ---------- ---------- ---------- ------MNSF ILILVILMAS
exspectatus-mkr-S_68-1.70-mRNA-1 DQNDSAYHHQ QRREPQDKRE QYSGKESVIQ SAMFTERISL VVLLIAIMAT
101
PS312-mkr-S142-0.63-mRNA-1 SLVSG--HHW GWGGGWGSPW GGYGGCCYGG YGGYGGYPGW --GGWGM-GG
exspectatus-mkr-S_68-1.70-mRNA-1 SLVSGWHHHR GWGGGW---- ---GGNGYGG YGGYGGYGGW GGGGWGLGGG
151
PS312-mkr-S142-0.63-mRNA-1 FGGFGWG
exspectatus-mkr-S_68-1.70-mRNA-1 FGGW---

Figure 5.4: Failures in homology detection lead to classification as SSOGs. (a) The *P. pacificus* SSOG PS312-mkr-S142-0.63-mRNA-1 is found in a conserved syntenic region with the *P. exspectatus* TROG exspectatus-mkr-S_68-1.70-mRNA-1. Both proteins are 'GGX' repeat rich proteins and share a small non-repetitive part, but blastp failed to identify both proteins as homologous.

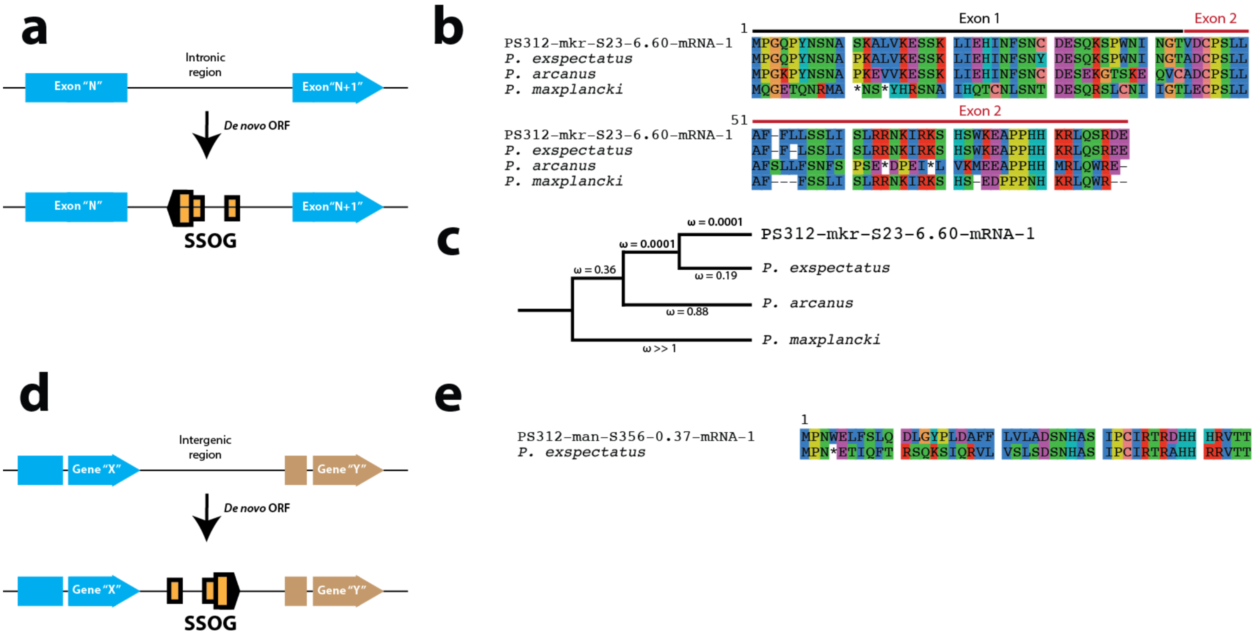

a
Intronic region
Exon "N"
Exon "N+1"
De novo ORF
Exon "N"
Exon "N+1"
SSOG
b
Exon 1
Exon 2
1
PS312-mkr-S23-6.60-mRNA-1 MPGQPYNSNA SKALVKESSK LIEHINFSNC DESQKSPWNI NGTVDCPSLL
P. exspectatus MPGQPYNSNA PKALVKESSK LIEHINFSNY DESQKSPWNI NGTADCPSLL
P. arcanus MPGKPYNSNA PKEVVKESSK LIEHINFSNC DESEKGTSKE QVCADCPSLL
P. maxplancki MQGETQNRMA *NS*YHRSNA IHQTCNLSNT DESQRSLCNI IGTLECPSLL
Exon 2
51
PS312-mkr-S23-6.60-mRNA-1 AF-FLLSSLI SLRRNKIRKS HSWKEAPPHH KRLQSRDE
P. exspectatus AF-F-LSSLI SLRRNKIRKS HSWKEAPPHH KRLQSREE
P. arcanus AFSLLFSNFS PSE*DPEI*L VKMEEAPPHH MRLQWRE-
P. maxplancki AF---FSSLI SLRRNKIRKS HS-EDPPPNH KRLQWR--
c
ω = 0.0001
ω = 0.0001
ω = 0.36
ω = 0.19
ω = 0.88
ω >> 1
PS312-mkr-S23-6.60-mRNA-1
P. exspectatus
P. arcanus
P. maxplancki
d
Intergenic region
Gene "X"
Gene "Y"
De novo ORF
Gene "X"
Gene "Y"
SSOG
e
1
PS312-man-S356-0.37-mRNA-1 MPNWELFSLQ DLGYPLDAFF LVLADSNHAS IPCIRTRDHH HRVTT
P. exspectatus MPN*ETIQFT RSQKSIQRVL VSLSDSNHAS IPCIRTRAHH RRVTT

Figure 5.5: *De novo* gene birth. (a) A *de novo* gene can originate as an antisense transcript in the intron of another gene. *De novo* creation of such an ORF can be verified by finding the corresponding intron in a related species that lacks this ORF. **(b)** PS312-mkr-S23-6.60-mRNA-1 is two exon *P. pacificus* gene that is located in an intron of another *P. pacificus* host gene. Based on the identification of orthologous intron of the host gene in other species, I have created an alignment of my candidate and translation of its corresponding reading frame from other species. It is clear that the same ORF also exists in *P. exspectatus*. However, *P. arcanus* has two stop codons (*) in the middle of the 2nd exon and *P. maxplancki* has two stop codons in the 1st exon itself. **(c)** Selection analysis done on the alignment from panel b, indicates that the predicted ORF has been under strong selection towards the *P. pacificus* lineage. This trend may have started from the common ancestor of *P. pacificus, P. exspectatus* and *P. arcanus*. **(d)** A *de novo* gene can originate from ancestrally intergenic region. **(e)** PS312-man-S356-0.37-mRNA-1 gene contains a single coding exon and its homologous reading frame in *P. exspectatus* is found at a non-transcribed intergeneic location and has an early stop codon (*). This gene does not show sequence homology with any other species but *P. exspectatus*, and hence has most likely emerged in *P. pacificus* lineage post speciation.

Evidence for *de novo* genes in *P. pacificus*

All the five mechanisms described in the previous sections portray how new genes can be created from old genes. Some of these mechanisms may involve the acquisition of new exons from non-coding segments of the genome, but they mainly contain exons that are derived from ancestrally coding segments of the genome. PS312-mkr-S23-6.60-mRNA-1 is a *P. pacificus* SSOG with two coding exons, placed within a single intron of a the *P. pacificus* homolog of *C. elegans* C27F2.7 (Supplemental Fig. S5.3).

The intronic location of my candidate SSOG within a conserved gene helped me in identifying the orthologous genomic locations in other *Pristionchus* species. Based on the spliced alignment of my candidate on the genomes of other species I was able to extract the orthologous sequences from *P. exspectatus, P. arcanus* and *P. maxplancki* (Fig. 5.5b). No transcriptional evidence for the genomic regions corresponding to their extracted ORFs was found in *P. exspectatus, P. arcanus* and *P. maxplancki* (Supplemental Fig. S5.3). Nevertheless, the length of the *P. exspectatus* ORF matches that of the *P. pacificus* prediction. Additionally, the *P. arcanus* ORF aligns well with the *P. pacificus* ORF but contains two stop codons in the middle of the second exon. Furthermore, the sequence extracted from *P. maxplancki* has stop codons at the 11th and 14th position, and no Methionine thereafter to make an abridged ORF. This suggests that the ORF at this locus was engendered in the common ancestor of *P. pacificus*, *P. exspectatus*, and *P. arcanus*. Moreover, the lack of ORF in *P. maxplancki* and alignable region in other species confirms the *de novo* origin of this gene. To further support the protein-coding nature of my *de novo* candidate, I carried out selection analysis on the predicted ORF of *P. pacificus* and the protein translation from the other species. In this analysis, I allowed each branch of the tree to have an independent ω value. Here, the branches leading from the common ancestor of *P. pacificus*, *P. exspectatus* and *P. arcanus*, towards the *P. pacificus* lineage are under extremely strong negative selection (Fig. 5.5c). This suggests that since its emergence, the *de novo* gene has been maintained as a protein-coding gene in the lineage leading to *P. pacificus*.

Our second *de novo* candidate PS312-man-S356-0.37-mRNA-1 is a two exon gene with its entire coding sequence in the 2nd exon. Since the candidate could be mapped on to the genomes of none of the other species but *P. exspectatus*, I was only able to extract the orthologous *P. exspectatus* sequence from a conserved syntenic region (Supplemental Fig. S5.4). Nevertheless, the absence of transcription in *P. exspectatus* and the presence of a stop codon at the 4th position of the extracted *P. exspectatus* sequence confirms the non-genic and non-transcribed status of the *P. exspectatus* sequence. Thus, I conclude that the *P. pacificus* SSOG PS312-man-S356-0.37-mRNA-1 is a *de novo* gene that has emerged from a previously non-coding intergenic region in the *P. pacificus* lineage.

Conclusion

To my knowledge, the analysis done in this chapter is the first of its kind in nematodes and allows me to make three major conclusions. First, the number of SSOGs is exceptionally high. Given the dense taxonomic distribution of the *pacificus* clade species, i.e. *P. pacificus, P. exspectatus, P. arcanus, P. maxplancki* and *P. japonicus*, in my analysis, I expected these species to be depleted in SSOGs. However, my results posited comparable fractions of SSOGs across the *Pristionchus* genus, which does not correspond with the difference in their relative divergence (Fig. 5.1a-b). The lack of correspondence between divergence time and number of SSOGs is also observed in the phylostratigraphic analysis.

Second, the high phylogenetic resolution of my dataset allowed me to find homologous traces for 1082 (61%) SSOGs of *P. pacificus* in the genomes of closely related sister species. This demonstrates the usability of my phylogenomic dataset to study the origin of *P. pacificus* SSOGs.

Third, both divergence of existing genic segments and *de novo* creation of new genic segments contribute towards birth of SSOGs. Based on manual inspection of the 29 fully supported candidate SSOGs, I found evidence for six mechanisms that potentially explain their origin. I demonstrate that the first five mechanisms involve recycling of ancestrally protein-coding gene segments to engender new genes. The final mechanism is illustrated by two *de novo* genes that have fully emerged from ancestrally non-coding regions

Methods

Identification of orphan genes

The genome, protein and transcript data of 24 non-diplogastrid nematodes were obtained from Wormbase (WormBase web site, http://www.wormbase.org, release WS254, date 7/18/16). The phylogenomic dataset for the 10 diplogastrid nematodes was gathered from my previous publication [23] and is available at http://www.pristionchus.org/download. All the Uniprot knowledgebase taxonomic divisions SwissProt data was downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/. The invertebrate taxon contained a single *Pristionchus* species gene, Q9NHZ4, which was removed from further analysis.

We first identified all conserved genes for the 10 diplogastrid nematodes using the following approach:

1. Classify all genes that have blastp match (E-value $\leq 10^{-3}$) with any non-diplogastrid nematode protein as 'Conserved genes'. For the remaining genes go to step 2.
2. Classify all genes that have a tblastn match (E-value $\leq 10^{-5}$) in any non-diplogastrid nematode genome as 'Conserved genes'. For the remaining genes go to step 3.
3. Classify all genes that have blastp match ($E \leq 10^{-3}$) with any protein from any Uniprot knowledgebase taxonomic divisions as 'Conserved genes'. The proteins classified as conserved genes at this step are candidates for horizontal gene transfer.

The remaining genes were classified as 'Orphan genes'. All blast runs were conducted, with version 2.6.0+, under default parameters (including no filtering of low complexity regions by SEG) unless mentioned otherwise.

Classification of orphan genes

The availability of 10 diplogastrid genomes provided me with the opportunity to further investigate the *Pristionchus* orphan genes. my first aim was to identify the orphan genes that have a homolog in at least one other diplogastrid species. Thus, for each species I selected the subset of orphan genes that have blastp match (E-value $\leq 10^{-3}$) with at least one other diplogastrid species. This subset of orphan genes was classified as TROGs. The remaining orphan genes were classified as SSOGs, as they did not show blastp match with any other species. It is important to note here that for the identification of TROGs I have only used protein homology. I did not employ tblastn against genomes to avoid detection of pseudogenes or non-coding genomic regions as protein homologs. Further, since a ladder-like species phylogeny exists around the focal species *P. pacificus* (Fig. 5.1a) [160], I decided to trace the origin of *P. pacificus* TROGs and SSOGs on this phylogeny. For this, I employed the phylostratigraphy approach [157]. This approach is based on finding the oldest ancestral node of a given phylogenetic tree where the founding member of a gene family can be traced back to. Thus, I divided the diplogastrid family tree into nine phylostrata. ‘Phylostratum 1’ corresponds to the most recent common ancestor of *P. pacificus* and *P. exspectatus*. Additionally, I created ‘Phylostratum 0’ that includes *P. pacificus* SSOGs and hence is the youngest phylostratum.

Mapping of gene models on the genome of other species

The synteny relation between genes from *P. pacificus* and the other species was derived using CYNTENATOR [161]. Pairwise blastp results for each species pair and two files containing genomic location of genes in both the species, were provided as input to the software. The output file contained a list of genes from both species within the syntenic blocks. Spliced alignment of gene models from one species to the genome of another species was done by employing the protein2genome model of the Exonerate tool [151].

Gene structure validation

One of the main aims of this study was to elucidate the mechanistic details underpinning the birth of new genes. However, even with my structured approach of dividing the orphan genes into several categories and subcategories, I was unable to put forward a clear hypothesis on this matter. Thus, I decided to create a set of most reliable candidate genes to better understand the processes that foster new *P. pacificus* genes. For this, I limited myself to the *P. pacificus* SSOGs with confirmed gene structure. The validation of predicted gene structure was done by visual inspection, in IGV [162], of raw RNA-seq data aligned with the *P. pacificus* genome [23,163]. I used TopHat v2.1.1 and STAR version 020201 for aligning the raw reads to genome [153,164]. Single exon genes were filtered out. Only multi-exon genes with minimum two spliced RNA-seq reads aligning to all coding exons and minimum two spliced reads straddling such exons, were assigned ‘fully confirmed gene structure’ status. If, only few, but not all exons of a gene satisfiedthese criteria, then it was assigned ‘partially confirmed gene structure’ status. For overlapping genes from opposite strands, strandedness of strand-specific RNA-seq data was used as an additional confirmation step.

Selection analysis

For selection analysis of the SSOG candidates, their orthologous reading frames (including in-frame stop codons) from sister species were extracted and manually adjusted. Protein alignment of the candidate and its corresponding reading frames from one or more sister species was done using MUSCLE and visualisation was done with SeaView [113,165]. The protein alignment was converted to codon with PAL2NAL [114]. Selection analysis was done with codeml suite of PAML [99]. Species tree was passed as gene tree to PAML. If the corresponding homologous region from only one sister species was included in the analysis I generated a single ω value for the entire tree, else I generated independent ω values for each branch of the tree. The statistical significance of the resulting ω values was calculated using the likelihood ratio test at the P-value threshold of 0.05. Only the statistically significant results were reported.

Chapter 6: Conclusion and Discussion

During the course of my doctoral research, I have attempted to bring the comparative method to the genomics of *Pristionchus* nematodes [85]. To my knowledge, this is the first phylogenomic study of nematodes, where 10 species of a family were chosen to create a ladder-like phylogeny so that *P. pacificus,* the focal species, invariably remains under a monophyletic clade. Additionally, whole genome sequencing and annotation of each species were done within a single lab, ensuring that the resultant genomes and gene annotations are highly comparable. Analysis of the resultant dataset allowed me to comprehensively address the questions that I set out to answer at the beginning of my doctoral research. In the following sections, I discuss the results presented in chapter three, four, and five.

Orphan genes are real

Although all genomes are reported to have orphan genes [11], a comprehensive test for protein-coding nature of such genes was unavailable at the inception of my doctoral research. In chapter three, I presented a bioinformatic pipeline that systematically searches for evidence supporting the protein-coding nature of each orphan candidate. Relying on selection analysis, which shows that 76% of orphan genes are under negative selection, I concluded that the majority of *P. pacificus* orphan genes are not annotation artifacts but real protein-coding genes. Besides, this finding is in agreement with several other studies from primates, fish, and insects [140,166–170]. It is important to point out that most technical artifacts such as incorrect alignments, pseudogenes, and mispredictions will bias signal towards neutral selection. Hence, further error correction will only increase the fraction of orphan genes with reliable protein-coding evidence.

The protein-coding nature of orphan genes was further supported by the orthologous clustering data generated in chapter four, which covers 81% of *P. pacificus* genes, and the classification of 79% of *P. pacificus* orphan genes as TROGs in chapter five. Given that *P. pacificus* and its closest neighbor diverged 1-5 million years ago, the identification of an orthologous gene in a sister species is in itself a strong evidence for selection acting to preserve the protein-coding nature of an orphan gene, because in the absence of purifying selection most genes will fail to maintain an ORF and thus will quickly get pseudogenized.

Deep taxon sampling is pivotal for uncovering the evolutionary dynamics of genes and gene families

After establishing the protein-coding nature of orphan genes through multiple lines of corroborating evidence, I tried to investigate the evolutionary dynamics that shape the genomes of *Pristionchus* nematodes. The generation of a ladder-like phylogeny around *P. pacificus* proved to be instrumental in this regard. The usefulness of the deep taxon sampling, with emphasis on creating a highly comparable dataset, was made obvious by the observation that single evolutionary events could explain the origin of up to 67% of *P. pacificus* gene families, which could be segregated into distinct Age classes. Here, I chose to identify the Age classes based on orthologous clustering rather than adopting a phylostratigraphy approach [157], because the clustering methods are able to break large gene families into small but densely connected clusters and hence can split the clusters that arose by recent duplication. Since my aim was to study the evolutionary processes that act on young genes irrespective of their origin, I intentionally included recently duplicated genes which are known to follow distinct evolutionary trajectories [141,171–175].

Further, the origin of most of the remaining gene families, which belong to the patchy clusters category, can be explained with just two evolutionary events. Additionally, this result also revealed that the young gene families are especially prone to gene loss and thus portray a patchy distribution pattern. To this end, the pairwise selection analysis showed that the young genes are evolutionarily less constrained and this increases their susceptibility towards deleterious substitutions, which can then lead to gene loss. Conversely, old genes were observed to be under strong purifying selection and thus are less likely to be lost.

The classification of genes into Age classes also revealed that old genes are enriched at the autosome centers, which also portray low substitution rates. However, any suspicion over a link between chromosomal location and evolutionary rates was put to rest by comparing the ω distribution of both young and old genes in non-overlapping windows along each chromosome. The results clearly showed that selection acts on individual genes irrespective of their chromosomal location and the low substitution rates observed at the autosome centers are an artifact of the enrichment of old genes at these locations. A similar pattern was also observed with expression data, as old genes were shown to be highly expressed and the autosome centers retained the highest expression values.

Phylostratigraphy reveals that the SSOGs make the most gene rich phylostratum

Although the orthologous clustering approach yielded deep insights into the evolutionary dynamics of *Pristionchus* gene families, in order to trace the evolutionary origin of *P. pacificus* orphan genes I decided to use the phylostratigraphy approach [157]. This method allowed me to estimate the number of genes that originated at each internal node on the *Pristionchus* phylogeny. The resulting data presented a steady rate of gene gain within the *Pristionchus* genus as the number of genes in each phylostratum and its estimated divergence was shown to be correlated. However, the phylostratum corresponding to the *P. pacificus* SSOGs is an exception, as it is by far the most abundant phylostratum but also happens to be one of the least divergent. This raises the question whether the abundance of SSOGs is due to erroneous gene annotations and missing gene models in sister species, or if it actually represents a rapidly evolving gene pool with high rates of gene gain and loss.

Given the widespread evidence of pervasive transcription [159], it is plausible that the novel gene-like sequences arise much faster than they become fixed in a population. Apart from suggesting that novel genes can result from pervasive transcription and translation [26,176], some independent studies in insects, mammals, and nematodes indicate that many novel genes are rapidly lost [23,69,140,156]. In my opinion, both processes, rapid emergence of gene-like sequences and erroneous annotations, substantially contribute to the abundance of SSOGs. However, since SSOGs are poorly supported by transcriptome data, future functional genomic or population-scale studies will be needed to conclusively distinguish annotation artifacts from real genes.

Both sequence divergence and *de novo* gene origin contribute to the emergence of the orphan genes

To prioritize the list of candidate SSOGs that was to be manually investigated, I defined high-confidence candidates based on gene structure and transcriptional evidence. Exclusion of single exon genes was one of the major contributors towards the acceptance of fewer genes as high-confidence candidates. Although *de novo* origin of single exon genes has been reported in both animal and plant species [69,177,178], due to limited proteome and transcriptome data for *P. pacificus* orphan genes [135], in this study, I emphasized on the alignment of spliced raw RNA-seq reads as the primary requirement to confirm regulated expression of my candidates [179]. Under the assumption that *de novo* genes typically arise as single exon genes with subsequent intron gain, the analyzed high-confidence SSOGs might represent an older age class of more established genes that in addition may also display a different composition of origin mechanisms.

The protein homology of SSOGs with their ancestrally coding gene segments was rendered undetectable mainly due to two reasons. First, when the alignable region is small, chimeric, non-contiguous, or repeat-rich, then divergence decreases the sensitivity of homology detection. Formation of novel genes due to rapid evolution of repeat-rich low complexity sequences and chimeric genes is well documented in the literature [180–186]. In chapter five, I have shown examples of both chimeric and repeat-rich genes being classified as SSOGs. Second, when the ancestral ORF gets fully or partially replaced by an alternate ORF, then the process of reading frame shift eliminates sequence homology at the protein level [187]. Maintenance of both the ancestral and alternate ORFs within the same gene is called 'overprinting' [51,55]. In chapter five, I have shown two cases of potential frame shifts, the first case shows the replacement of the ancestral reading frame, while the second case could represent an example for overprinting. My candidate gene from the first case, PS312-mkr-S390-0.42-mRNA-1, can also be taken as an example of a mix between divergence and *de novo* formation, as the alternate ORF is largely non-existent in other species. The formation of this gene results from several steps, which include splitting of the ancestral gene, sequence divergence, reading frame shift and *de novo* acquisition of a new exon. Thus, I argue that this gene is a product of 'mixed origin mechanism', as both divergence and *de novo* origin mechanisms have contributed to its birth. Given the limited number of cases that could be analyzed at this level of detail, I cannot comment on the extent to which the proposed mechanisms contribute to the emergence of novel genes and in future more comprehensive studies will be needed to quantify their contributions.

The analysis of two *de novo* origin genes in chapter five revealed that in *Pristionchus* nematodes such genes can emerge from both intronic and intergenic loci. The older gene originated within an intronic region of a conserved gene, intronic origin of *de novo* genes has been reported in other studies [188–190]. The younger *de novo* gene originated from an intergenic region within the *P. pacificus* lineage. *De novo* origin of novel genes has been discovered, generally in small numbers, in many different lineages including yeast, insects, primates and plants [7,8,10,26,53,66–74]. To my knowledge, this is the first instance of *de novo* gene identification in nematodes. *De novo* genes are reported to be expressed at a low level and in a tissue-restricted manner [72]. I speculate that the generation of more tissue and stage-specific transcriptome data will facilitate the validation of many more *de novo* genes. Moreover, in this study, a large fraction of *P. pacificus* SSOGs remain classified as Untraceable genes. I assume most *de novo* genes, which originated within *P. pacificus* lineage, have lost all traces of sequence similarity with other species and will thus fall into the Untraceable gene class. My assumption is supported by a study carried out on insect genomes, which concluded that the DNA corresponding to novel domains of *de novo* origin was rarely found in sister species [25]. Thus, I propose that in order to investigate *de novo* origin of more *P. pacificus* SSOGs, genomes of several divergent *P. pacificus* strains should be assembled [111].

In summary, I have established that the majority of *P. pacificus* orphan genes are real protein-coding genes by using a robust bioinformatic pipeline that tests the level of selection acting at the protein level and by tracing the origin of most *P. pacificus* orphan genes on a ladder-like phylogeny, which was especially created for this study. This has facilitated the identification of a large number of TROGs, which are likely to be involved in lineage-specific adaptations [191]. The ladder-like phylogeny of *Pristionchus* nematodes has also allowed me to uncover the evolutionary dynamics that shape their novel gene families. The high phylogenetic resolution of my data allowed me to study the mechanism of origin for most *P. pacificus* SSOGs. Ultimately, I have shown that both *de novo* and divergence mechanisms play a role in the birth of new genes and in some cases both mechanisms can contribute collectively towards this phenomenon.

Chapter 7: Appendix

Contributions

In the following section, I list all the publications and manuscripts I have been part of during my doctoral research and describe in detail, as far as possible, my contribution and that of everyone else.

Publications

N Prabh, W Roeseler, H Witte, G Eberhardt, R J Sommer, C Rödelsperger: *Deep taxon sampling reveals the evolutionary dynamics of novel gene families in the Pristionchus genome*. Genome Research 2018

Contributions: N.P., R.J.S., and C.R. conceptualized the project. N.P. and C.R. developed the methodology. Formal analysis and visualization were the responsibility of N.P. Experiments were carried out by N.P., W.R., H.W., and G.E. W.R., H.W., G.E., and R.J.S. gathered resources. N.P. and C.R. wrote the original draft, and N.P., R.J.S., and C.R. helped with writing, review, and editing. C.R. supervised the project.

M S Werner, B Sieriebriennikov, **N Prabh**, T Loschko, C Lanz, R J Sommer: *Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation.* Genome Research 2018

Contributions: M.S.W. and R.J.S. conceived and designed all experiments. M.S.W. conducted ChIP- and ATAC-seq with assistance from T.L.; M.S.W, B.S., and C.L. performed Iso-Seq; N.P. conducted phylogenetic analysis and prepared evolutionary gene category datasets; M.S.W. performed all bioinformatic analysis. M.S.W. wrote the manuscript with assistance from R.J.S.

E Moreno, M Lenuzzi, C Rödelsperger, **N Prabh**, H Witte, W Roeseler, M Riebesell, R J Sommer: *DAF-19/RFX controls ciliogenesis and influences oxygen-induced social behaviours in Pristionchus pacificus.* Evolution & Development 2018

Contributions: N.P. did bioinformatics analysis with C.R.

C Rödelsperger, W Roeseler, **N Prabh**, K Yosida, C Weiler, M Hermann, R J Sommer: *Phylotranscriptomics of Pristionchus nematodes reveals parallel gene loss in six hermaphroditic lineages*. Current Biology 2018

Contributions: Conceptualization, C.R.; Investigation W.R.; Formal Analysis, C.R.; Resources,M.H., C.W., N.P., and K.Y.; Writing – Original Draft, C.R.; Writing – Review& Editing, C.R. and R.J.S; Funding Acquisition, R.J.S.

B Sieriebriennikov, **N Prabh***, M Dardiry*, H Witte, W Roeseler, M Kieninger, C Rödelsperger, R J Sommer: *A Developmental Switch Generating Phenotypic Plasticity Is Part of a Conserved Multi-gene Locus*. Cell Reports 2018 * Equal contribution

Contributions: Conceptualization, B.S., N.P., M.R.K., and R.J.S.; Methodology, B.S., N.P., M.D., C.R., M.R.K., and R.J.S.; Formal Analysis, B.S. and N.P.; Investigation, B.S., N.P., M.D., H.W., and W.R.; Resources, H.W. and W.R.; Writing – Original Draft, B.S. and R.J.S.; Writing – Review & Editing, B.S., N.P., M.D., and R.J.S.; Visualization, B.S. and N.P.; Supervision, C.R. and R.J.S.

C Rödelsperger, J M Meyer, **N Prabh**, C Lanz, F Bemm, R J Sommer: *Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode*

Model Organism Pristionchus pacificus. Cell Reports 2017

Contributions: C.R. and R.J.S. conceived and supervised the study. J.M.M., C.L., and N.P. performed the experiments. C.R, N.P., and F.B. analyzed the data. C.R., N.P., F.B., J.M.M., and R.J.S. wrote the manuscript.

N Prabh, C Rödelsperger: *Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs?*. BMC Bioinformatics 2016

Contributions: C.R. conceived and supervised this project. N.P. carried out the experiments. N.P. and C.R. analyzed the data. N.P. and C.R. wrote the manuscript. Both authors read and approved the final manuscript.

P Baskaran, C Rödelsperger, **N Prabh**, V Serobyan, G V Markov, A Hirsekorn, C Dieterich: *Ancient gene duplications have shaped developmental stage-specific expression in Pristionchus pacificus*. BMC Evolutionary Biology 2015

Contributions: C.D. conceived and supervised the project. A.H. carried out the staging and RNA-seq experiments. P.B., C.R., and C.D. analyzed the data. N.P. and V.S. performed the qRT-PCR experiments. G.V.M. contributed to the manual curation of orthologous gene datasets. P.B., C.R., and C.D. wrote the manuscript. All authors read and approved the final version of the manuscript.

Unpublished manuscripts

Submitted

N Prabh, C Rödelsperger: *The diversity of orphan gene origin - illustrated.*

Contributions: Conceptualisation, N.P. and C.R.; Methodology, N.P. ; Formal Analysis, N.P.; Writing – Original Draft, N.P.; Writing – Review & Editing, N.P. and C.R.; Visualisation, N.P.; Supervision, C.R.

In preparation

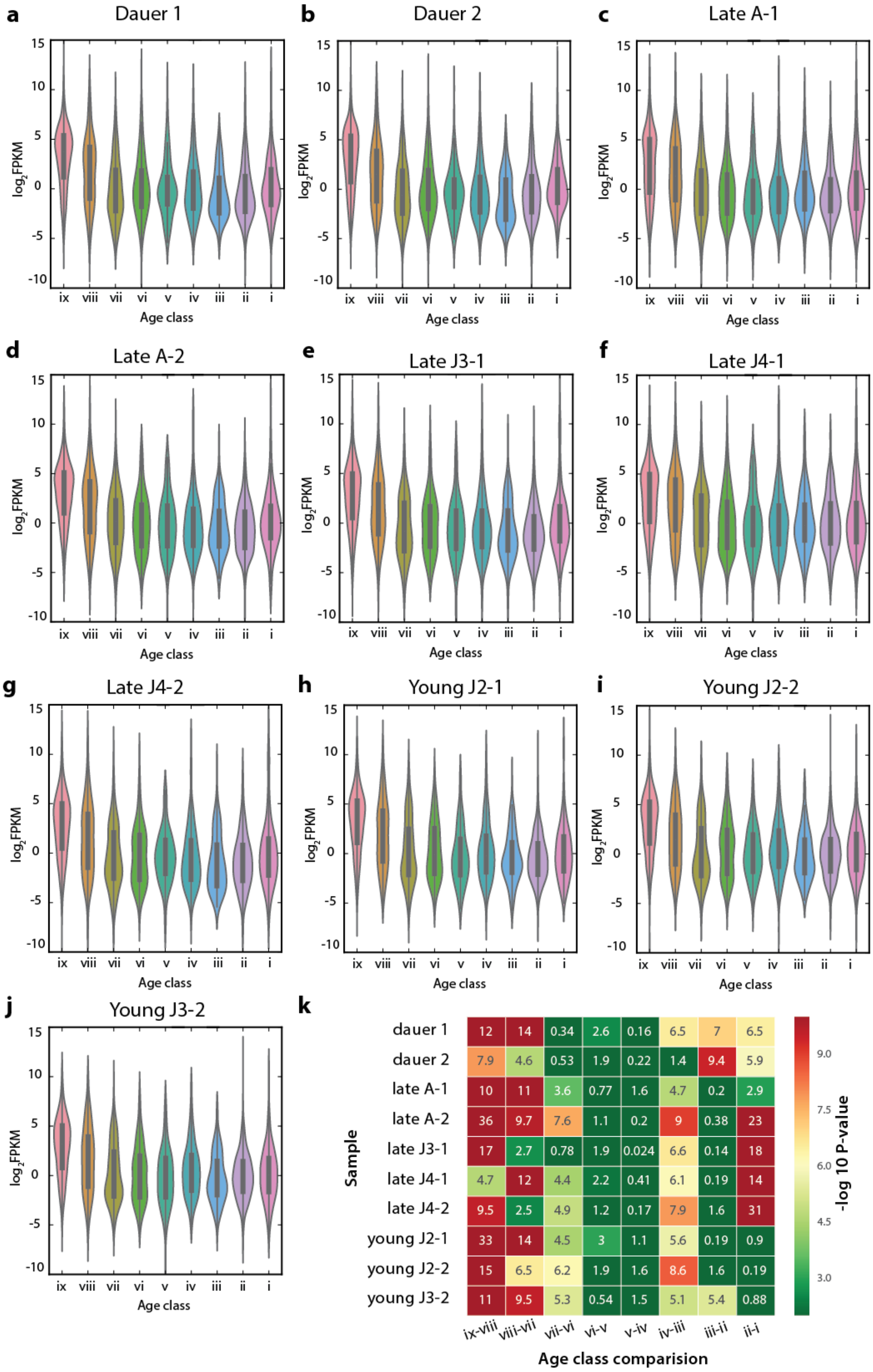
N Prabh, C Rödelsperger: *Phylogenomic analysis of P. pacificus nematodes using draft genomes of six new strains.*

Contributions: N.P. generated all data. C.R. is supervising the project.

N Prabh, H Witte, E Moreno, J Lightfoot, M Dardiry, C Rödelsperger, R J Sommer: *Tracing the origin of novel P. pacificus genes*.

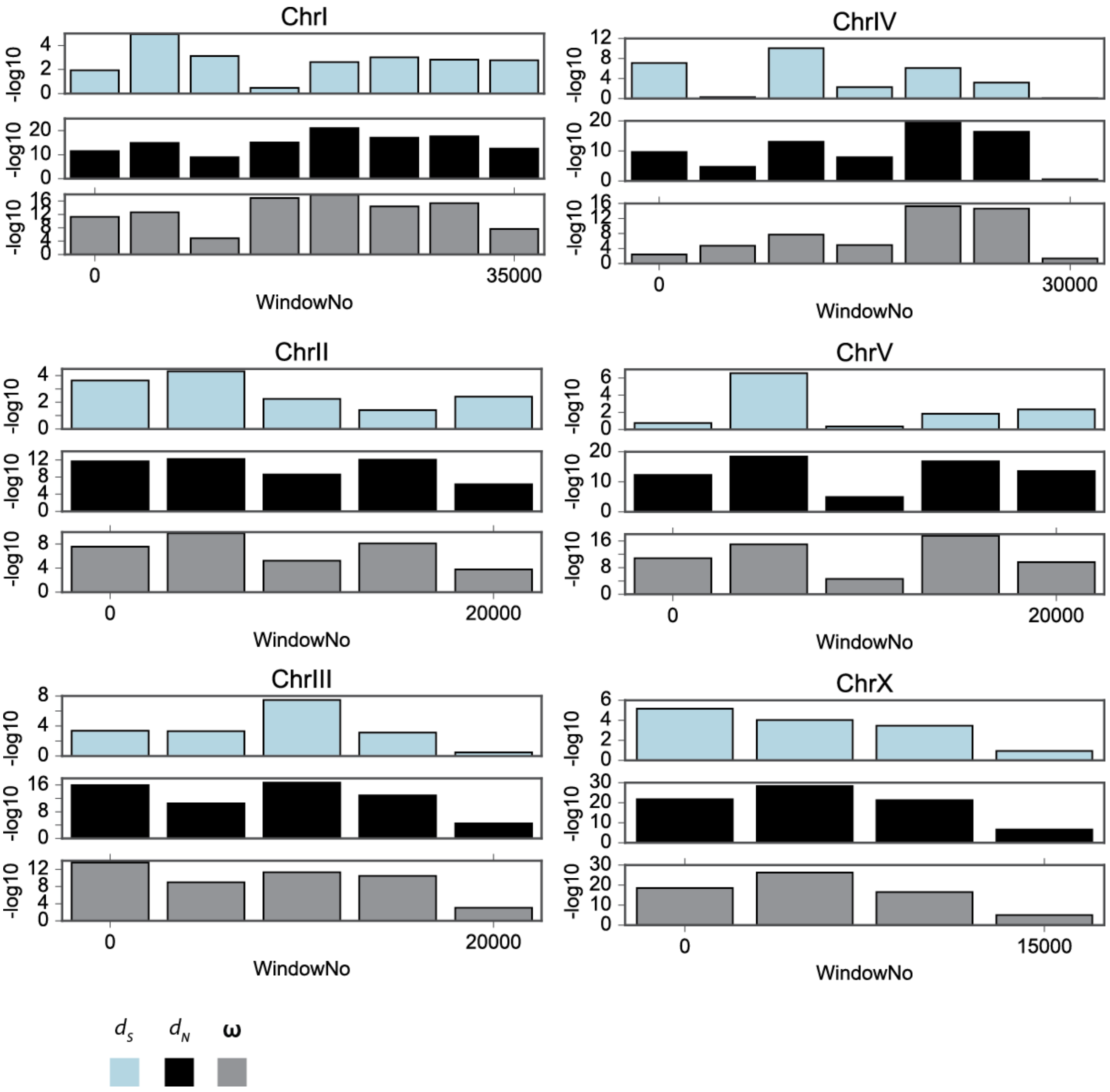
Contributions: N.P., C.R. and R.S. conceived the project. N.P., H.W., E.M. J.L., M.D. and R.J.S perform the experiments, N.P. has done and will do all data analysis, N.P. will write the manuscript with C.R. and R.J.S., R.J.S. is supervising the project.

Supplemental Figures



Sample	ix-viii	viii-vii	vii-vi	vi-v	v-iv	iv-iii	iii-ii	ii-i
dauer 1	12	14	0.34	2.6	0.16	6.5	7	6.5
dauer 2	7.9	4.6	0.53	1.9	0.22	1.4	9.4	5.9
late A-1	10	11	3.6	0.77	1.6	4.7	0.2	2.9
late A-2	36	9.7	7.6	1.1	0.2	9	0.38	23
late J3-1	17	2.7	0.78	1.9	0.024	6.6	0.14	18
late J4-1	4.7	12	4.4	2.2	0.41	6.1	0.19	14
late J4-2	9.5	2.5	4.9	1.2	0.17	7.9	1.6	31
young J2-1	33	14	4.5	3	1.1	5.6	0.19	0.9
young J2-2	15	6.5	6.2	1.9	1.6	8.6	1.6	0.19
young J3-2	11	9.5	5.3	0.54	1.5	5.1	5.4	0.88

Supplemental Figure S4.1: (a-j) Gene expression levels are correlated with gene age . Transcriptome expression values for *P. pacificus* genes from different Age classes in all 10 RNA seq samples. (k) Heat map of FDR corrected P-values of pairwise Wilcoxon ranksum test for expression values between subsequent age classes in all samples.



Supplemental Figure S4.2: On y-axis we have the P-values for Wilcoxon ranksum test for d_S, d_N and ω distributions between *P. pacificus* and *P. exspectatus* orthologs of Age class ix and Age class i-viii mapped on the chromosomes of *P. pacificus*, in a non-overlapping window of 5 Mb. Here, we see that both d_N and ω show significant difference between young and old genes in multiple windows and d_S is least different.

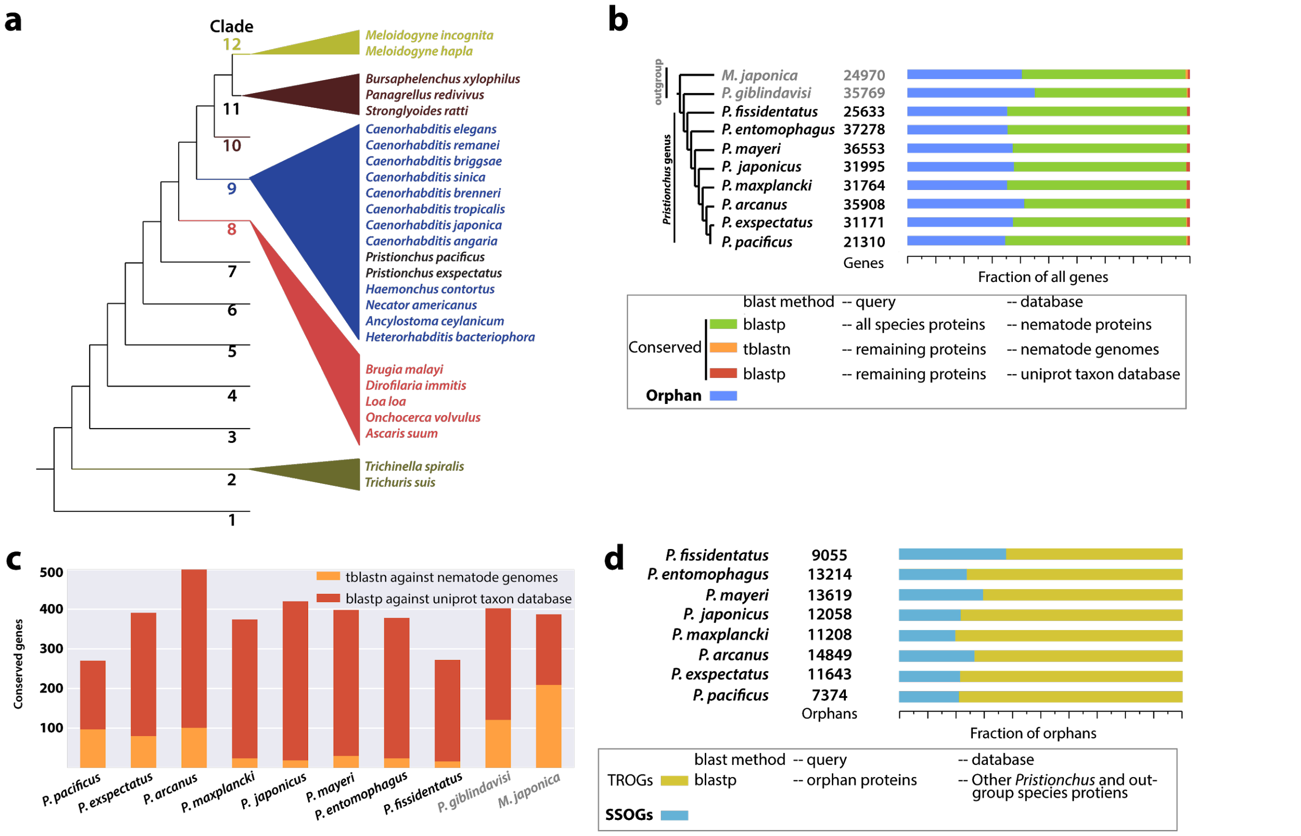


Figure S5.1: *Pristionchus* orphan gene identification. (a) Cartoon representing the distribution of nematode species with assembled genomes on Wormbase till 2017. The two *Pristionchus* species are labeled in black. **(b)** The total number of protein-coding genes for the eight *Pristionchus* species and the two non-*pristionchus* diplogastrid species is shown, followed by the fraction of orphan and conserved genes as horizontally stacked bars. The box shows the different blast methods and databases used to identify the conserved genes in panel a and b. Nematode proteins do not include proteins from the diplogastrid family nematodes **(c)** Number of conserved genes identified using the additional filtering steps. **(d)** TROGs and SSOGs as a fraction of orphan genes in each *Pristionchus* species. The box shows the different blast methods and databases used.

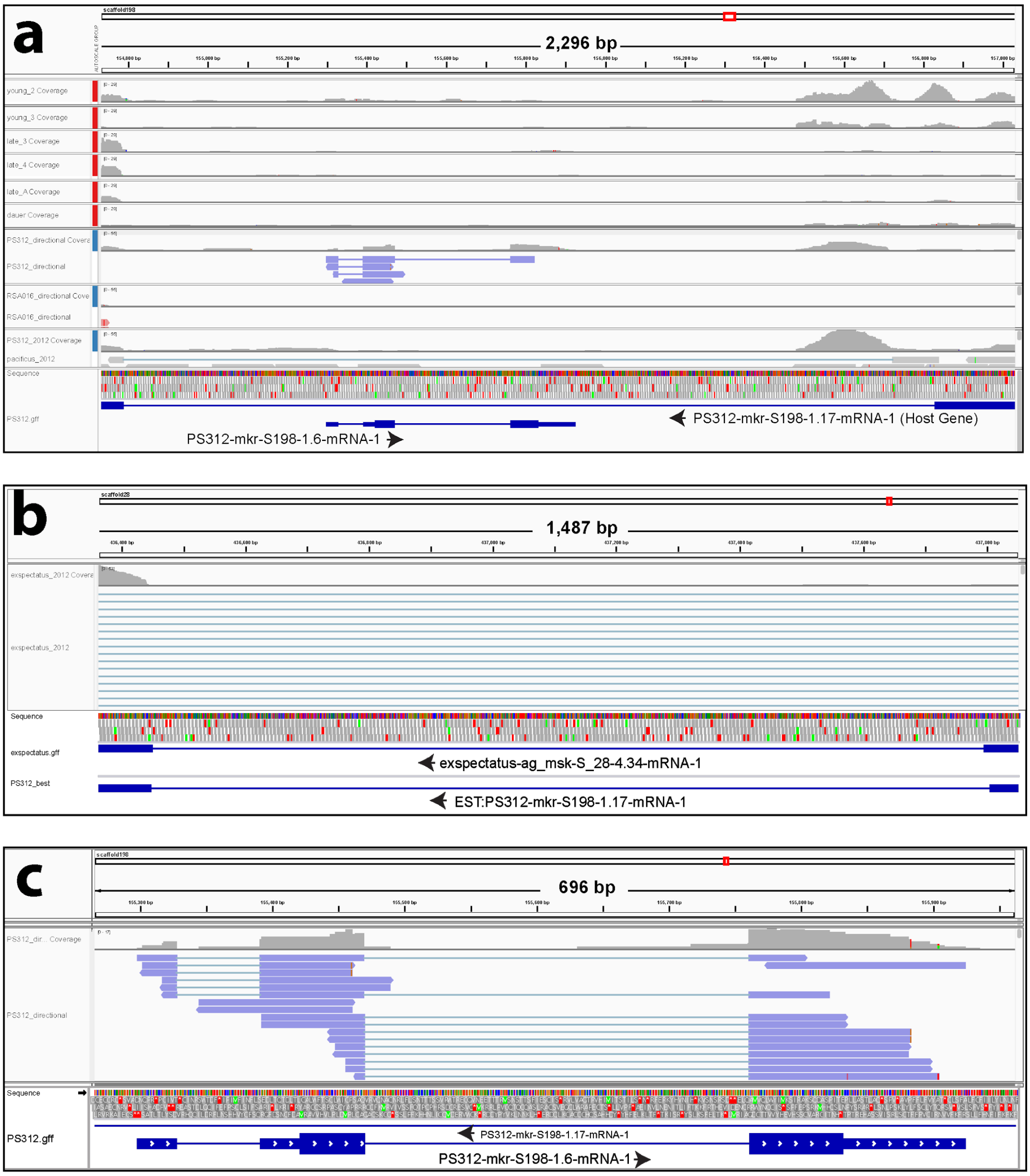


Figure S5.2: Novel gene formation by duplication and insertion of exonic sequences into an intron. (a) This IGV screenshot shows a 2.3 kb region on scaffold198 of the *P. pacificus genome*. Different tracks denote gene annotations, coverage profiles and alignments of various RNA-seq samples. The *P. pacificus* candidate SSOG PS312-mkr-S198-1.6-mRNA-1 is located within the intron of another gene (PS312-mkr-S198-1.17-mRNA-1, host gene). The same intron also contains a second transcriptionally active region (around position 156,600 bp) which presumably represents a short isoform of the host gene. **(b)** This screenshot shows the orthologous intron in *P. exspectatus* that was identified by exonerate alignment of the host gene. The genomic span is roughly 800 bp less compared with the *P. pacificus* region suggesting one or multiple insertions of a novel sequences in the *P. pacificus* lineage which gave rise to the candidate SSOG. **(c)** The genomic span carrying our candidate SSOG is roughly equal to the difference in the intron size between *P. pacificus* and *P. exspectatus*. Alignment of strand-specific raw reads shows that many spliced reads cover the two coding exons in the correct orientation.

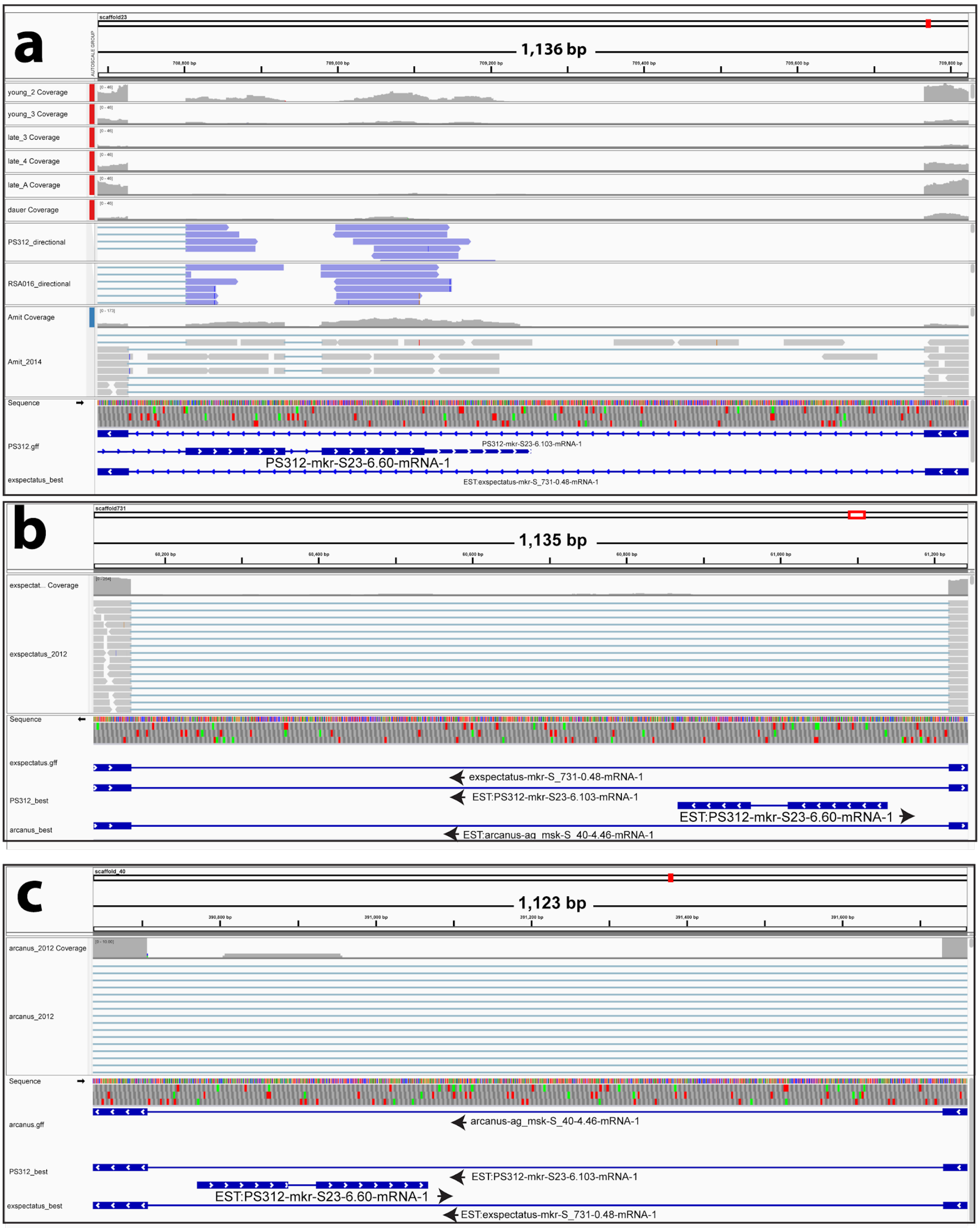


Figure S5.3: Intronic *de novo* gene. (a) This IGV screenshot shows a 1.1 kb region on scaffold23 of the *P. pacificus genome* harboring the candidate *de novo* SSOG, PS312-mkr-S23-6.60-mRNA-1. The candidate SSOG is within the intron of another gene (PS312-mkr-S23-6.103-mRNA-1, host gene). Strand-specific RNA-seq reads confirm that the gene is predicted in the correct orientation. Raw reads spanning the two coding exons are not found. The ends of spliced reads exceeding the left boundary of the displayed region align to the next intron and form the 5'UTR of the candidate SSOG. **(b, c)** The length of corresponding introns from *P. exspectatus* **(b)** and *P. arcanus* **(c)** genomes are comparable with the *P. pacificus* intron. The spliced alignment of our candidate genes onto the genome of sister species allows extraction of corresponding ORFs from these species. Except for a single unspliced read in *P. arcanus*, no transcriptional evidence is found in the two sister species.

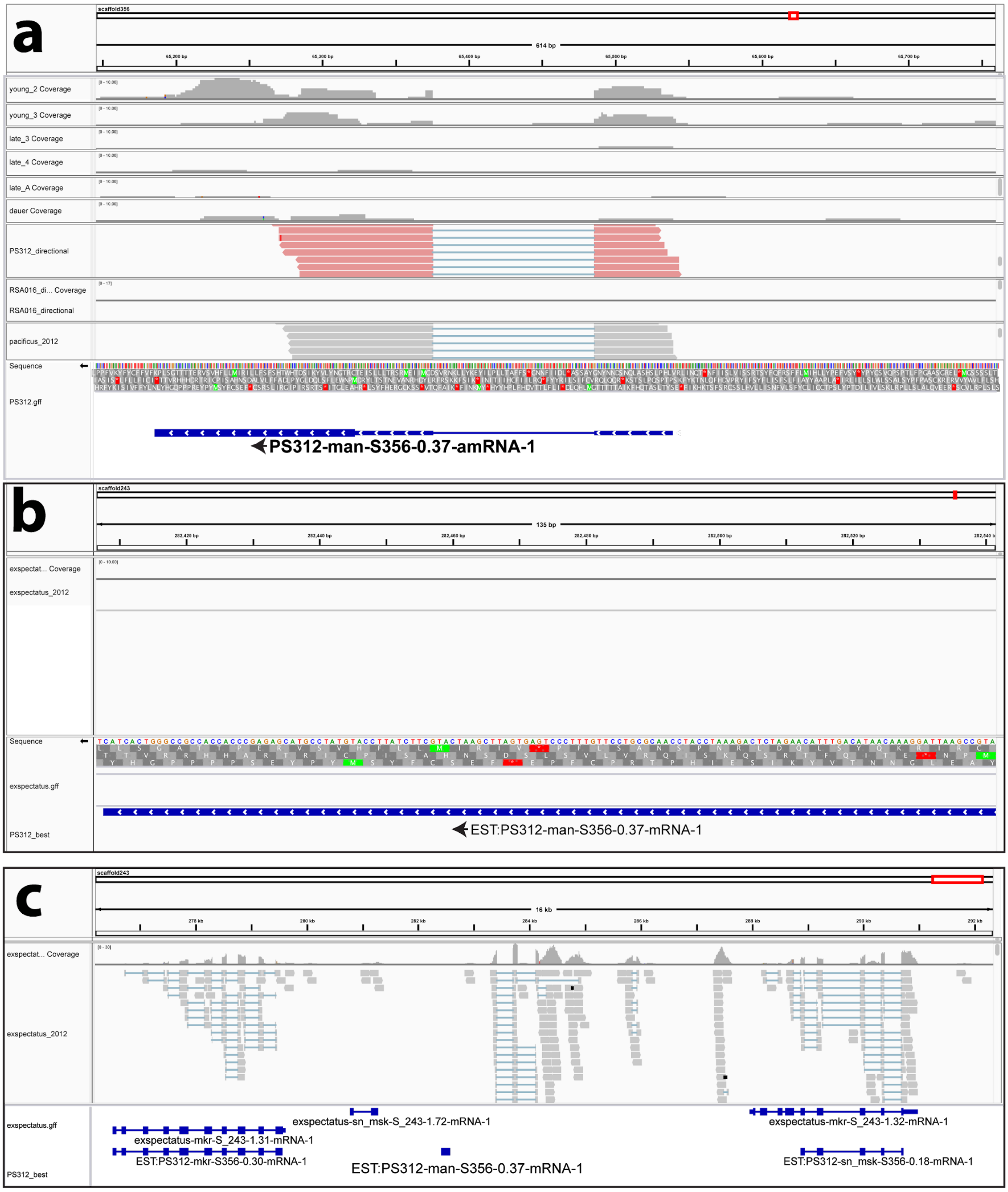


Figure S5.4: Intergenic *de novo* gene. (a) This IGV screenshot shows a 614bp region on scaffold356 of the *P. pacificus* genome harboring the candidate SSOG PS312-man-S356-0.37-mRNA-1. **(b)** Spliced alignment of our SSOG on the *P. exspectatus* genome shows no ORF exists in the sister species and raw RNA-seq reads do not align at this locus. **(c)** The neighboring *P. exspectatus* genes are syntenic with other *P. pacificus* genes mapped to the *P. exspectatus* genome and our candidate has emerged within this syntenic block.

Supplemental Tables

Table S3.1 – Primer pairs for the candidate genes

Gene Identifier	Primer Identifier	Primer sequence
Contig1-snapTAU.32	C1.32F1	TCTGTCCAGAGGAACGAATGGGATC
	C1.32R1	TGCACACTAACAAGTCTTCCCTCAG
	C1.32F2	CAGGAAAGATCGTCAAACAGGACCA
	C1.32R2	TGATTTCTCTTCAGGAGACACTCAG
Contig115-snapTAU.38	C115.38F1	GTCAGAGTGGAAATCAGTGCAACTG
	C115.38R1	TCACTTCCGTGTGTACGATTGACTT
	C115.38F2	ATGCCGAGCACAGAACAAATGCTGC
	C115.38R2	ACCGAGATTGCGGAAAACAGCGCAA
Contig159-snapTAU.23	C159.23F1	TTCATCGCTGACGATCACAGGCACA
	C159.23R1	AGATCATCATGCAGCCCTCCTTTGC
	C159.23F2	ATGCTCAAACTCCTCGTCTTCACCA
	C159.23R2	ACGATTTGACTGCGGGCTCTGCCTT
Contig162-snapTAU.8	C162.8F1	ATCAATGGCAATAAATCCGCTTACG
	C162.8R1	ATAAAGCCGTGAAGGTAATTCTCAT
	C162.8F2	AATAAATCCGCTTACGAACCAATCG
	C162.8R2	GGTAATTCTCATATTTGATGATTCC
Contig163-snapTAU.25	C163.25F1	GCAATCCCTCTACTGGCAGAATCTC
	C163.25R1	ATTGCATGGAGAGTACGTATCCGAC
	C163.25F2	AACTATGAAGGCGGTGATTCATTGG
	C163.25R2	GTTCGTTGAAAATCCACACTTTTCG
Contig27-snapTAU.5	C27.5F1	ACAAGAAGGCATACATGATGTACCC
	C27.5R1	AGTAGTCGAGGTGATGCTGTCAGGA
	C27.5F2	AACTGCATCTCAGACGCATCGGACA
	C27.5R2	TTTGACCTTGAACGCTTTCCTCCCG
Contig51-snapTAU.126	C51.126F1	ATGCTTGCGTGCATTGGGATCATCG
	C51.126R1	TAGCTCATTGAGATCAATGTCTTCG
	C51.126F2	TGACCTTCCTCGGCGGATGTTCCA
	C51.126R2	AGTTCACTTAGGCTCTCAAATGAGG
Contig57-snapTAU.76	C57.76F1	AGGAGATGATCGATAAACACAAAGCC
	C57.76R1	TCTTCTTCTGCAGCTGATTTGCCAC
	C57.76F2	TCGACAAGTGCTTCAAAGCCGAGCT
	C57.76R2	AAGATCCTCAAACTTCTCGCTGTG
Contig62-snapTAU.17	C62.76F1	TGCAAGTTGCACATCTCAACCACCT
	C62.76R1	ACACTTGGTTTCTTGAATGAGCTAAC
	C62.76F2	TGGGGATATCAAGTGCAAAGGCACTG
	C62.76R2	TTGGCTGGTTGGCTCTCGAATACTG
Contig67-snapTAU.30	C67.30F1	ATTCGACGTCTACTCTCACGCAACA
	C67.30R1	ATACGAAGTACAACATCACCTTGAG
	C67.30F2	TTCCGGCACACTTCTCATCATTCTC
	C67.30R2	AAATGAACGAGTACAACAGTAAACC
Contig68-snapTAU.138	C68.138F1	ACTGATTGCTGCTCATACAGATCGA
	C68.138R1	ACTGAGGAGCATCGTAAGCTGACTC
	C68.138F2	TCTTATTGGCTATACTGATTGCTGC
	C68.138R2	ATCCACTTTCCTGTCGAATTGACGC

Supplemental Table S4.1: The table gives an overview about the genome annotations in all 10 species. The first two columns indicate the number of RNA sequencing experiments and the size of the resulting transcriptome assembly (including isoforms). The last three columns show the number of final gene annotations along with their total exon and intron length.

Species	RNA sequencing experiments	Assembled Transcriptome Size (Mb)	Gene Count	Exon Length (Mb)	Intron Length (Mb)
P. pacificus	6	44	21,311	30	84
P. exspectatus	6	64	31,172	39	90
P. arcanus	6	71	35,909	45	112
P. maxplancki	1	51	31,765	43	109
P. japonicus	1	49	31,996	40	93
P. mayeri	1	62	36,554	37	114
P. entomophagus	1	49	37,279	38	107
P. fissidentatus	1	45	25,634	30	110
P. giblindavisi	1	36	35,770	30	91
M. japonica	1	60	24,971	26	99

References

1. Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. The complete DNA sequence of yeast chromosome III. Nature. 1992;357: 38–46.

2. Dujon B. The yeast genome project: what did we learn? Trends Genet. 1996;12: 263–270.

3. Casari G, De Daruvar A, Sander C, Schneider R. Bioinformatics and the discovery of gene function. Trends Genet. 1996;12: 244–245.

4. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics. 1999;15: 759–762.

5. Schmid KJ, Aquadro CF. The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics. 2001;159: 589–598.

6. Schmid KJ, Tautz D. A screen for fast evolving genes from Drosophila. Proc Natl Acad Sci U S A. 1997;94: 9746–9750.

7. Heinen TJAJ, Tobias J A, Staubach F, Häming D, Tautz D. Emergence of a New Gene from an Intergenic Region. Curr Biol. 2009;19: 1527–1531.

8. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009;19: 1752–1759.

9. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci. 2015;370: 20140332.

10. Schmitz JF, Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. F1000Res. 2017;6: 57.

11. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25: 404–413.

12. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12: 692–702.

13. Dunn CW, Ryan JF. The evolution of animal genomes. Curr Opin Genet Dev. 2015;35: 25–32.

14. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014;10: e1003998.

15. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12: 491.

16. Ferrada E, Wagner A. Evolutionary Innovations and the Organization of Protein Functions in Genotype Space. PLoS One. 2010;5: e14172.

17. Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proc Natl Acad Sci U S A. 2018;115: E409–E417.

18. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39: 309–338.

19. Brookfield JF. Genetic redundancy: screening for selection in yeast. Curr Biol. 1997;7: R366–8.

20. Borchert N, Dieterich C, Krug K, Schütz W, Jung S, Nordheim A, et al. Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. Genome Res. 2010;20: 837–846.

21. Sommer RJ, Carta LK, Kim S-Y, Sternberg PW. Morphological, genetic and molecular description of Pristionchus pacificus. Fundam Appl Nematol. 1996;19: 511–521.

22. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, et al. The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet. 2008;40: 1193–1198.

23. Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in Pristionchus nematodes. Genome Res. 2018;In Press.

24. Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Res. 2018; doi:10.1101/gr.234872.118

25. Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. Origins and Structural Properties of Novel and De Novo Protein Domains During Insect Evolution. FEBS J. 2018; doi:10.1111/febs.14504

26. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012;487: 370–374.

27. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, et al. The ecoresponsive genome of Daphnia pulex. Science. 2011;331: 555–561.

28. Babonis LS, Martindale MQ, Ryan JF. Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis. BMC Evol Biol. 2016;16: 114.

29. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 2013;9: e1003860.

30. Johnson BR. Taxonomically Restricted Genes Are Fundamental to Biology and Evolution. Front Genet. 2018;9: 407.

31. Luis Villanueva-Cañas J, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Albà MM. New Genes and Functional Innovation in Mammals. Genome Biol Evol. 2017;9: 1886–1900.

32. Kawasaki K, -G. Lafont A, -Y. Sire J. The Evolution of Milk Casein Genes from Tooth Genes before the Origin of Mammals. Mol Biol Evol. 2011;28: 2053–2061.

33. Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TCG. Characterization of taxonomically restricted genes in a phylum-restricted cell type. Genome Biol. 2009;10: R8.

34. Aguilera F, McDougall C, Degnan BM. Co-option and de novo gene evolution underlie molluscan shell diversity. Mol Biol Evol. 2017; msw294.

35. Kuwada Y. Maiosis in the Pollen Mother Cells of Zea Mays L. (With Plate V.). Shokubutsugaku Zasshi. 1911;25: 163–181.

36. Tischler G. Chromosomenzahl, Form und Individualität im Pflanzenreiche. Progr Rei Bot. 1915;20: 5–164.

37. Bridges CB. SALIVARY CHROMOSOME MAPS. J Hered. 1935;26: 60–64.

38. Haldane JBS. The Causes of Evolution. Cornell Univ. Press. 235 pp.; 1932.

39. Tautz D. The discovery of de novo gene evolution. Perspect Biol Med. 2014;57: 149–161.

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

41. Goldschmidt R. The Material Basis of Evolution. Yale University Press; 1940.

42. Gulick A. The Chemical Formulation of Gene Structure and Gene Action. Advances in Enzymology - and Related Areas of Molecular Biology. 1944. pp. 1–39.

43. Metz CW. Duplication of Chromosome Parts as a Factor in Evolution. Am Nat. 1947;81: 81–103.

44. Stephens SG. Possible Significance of Duplication in Evolution. Advances in Genetics. 1951. pp. 247–265.

45. Ohno S. Evolution by Gene Duplication. Springer, New York; 1970.

46. Muller HJ. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetica. 1935;17: 237–252.

47. Cortez D, Forterre P, Gribaldo S. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol. 2009;10: R65.

48. Zhaxybayeva O, Doolittle WF. Lateral gene transfer. Curr Biol. 2011;21: R242–6.

49. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 2008;9: 605–618.

50. Rödelsperger C, Sommer RJ. Computational archaeology of the Pristionchus pacificus genome reveals evidence of horizontal gene transfers from insects. BMC Evol Biol. 2011;11: 239.

51. Grassé P-P. Evolution of Living Organisms: Evidence for a New Theory of Transformation. Academic Press; 1977.

52. Jacob F. Evolution and tinkering. Science. 1977;196: 1161–1166.

53. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 2008;179: 487–496.

54. Ohno S. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci U S A. 1984;81: 2421–2425.

55. Kesse PK, Gibbs A. Origins of Genes: “Big Bang” or Continuous Creation? Proc Natl Acad Sci U S A. National Academy of Sciences; 1992;89: 9489–9493.

56. Chen L, DeVries AL, Cheng CH. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc Natl Acad Sci U S A. 1997;94: 3811–3816.

57. Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD. Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. PLoS Genet. 2005;1: e18.

58. Makalowska I, Lin C-F, Makalowski W. Overlapping genes in vertebrate genomes. Comput Biol Chem. 2005;29: 1–12.

59. Chung W-Y, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput Biol. 2007;3: e91.

60. Gontijo AM, Miguela V, Whiting MF, Woodruff RC, Dominguez M. Intron retention in the Drosophila melanogaster Rieske Iron Sulphur Protein gene generated a new protein. Nat Commun. 2011;2: 323.

61. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012;22: 2219–2229.

62. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012;29: 3767–3780.

63. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics. 2013;14: 117.

64. Guan Y, Liu L, Wang Q, Zhao J, Li P, Hu J, et al. Gene refashioning through innovative shifting of reading frames in mosses. Nat Commun. 2018;9: 1555.

65. Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol. 2011;3: 1245–1252.

66. Begun DJ, Lindfors HA, Thompson ME, Holloway AK. Recently Evolved Genes Identified From Drosophila yakuba and D. erecta Accessory Gland Expressed Sequence Tags. Genetics. 2005;172: 1675–1681.

67. Xiao W, Liu H, Li Y, Li X, Xu C, Long M, et al. A rice gene of de novo origin negatively regulates pathogen-induced defense response. PLoS One. 2009;4: e4603.

68. Neme R, Amador C, Yildirim B, McConnell E, Tautz D. Random sequences are an abundant source of bioactive RNAs or peptides. Nature Ecology & Evolution. 2017;1: 0127.

69. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 2014;343: 769–772.

70. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature Ecology & Evolution. 2017;1: 0146.

71. Light S, Basile W, Elofsson A. Orphans and new gene origination, a structural and evolutionary perspective. Curr Opin Struct Biol. 2014;26: 73–83.

72. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, et al. Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet. 2015;11: e1005721.

73. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014;3: e03523.

74. Wu B, Knudson A. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. MBio. 2018;9. doi:10.1128/mBio.01024-18

75. Hugot J-P, Baujard P, Morand S. Biodiversity in helminths and nematodes as a field of study: an overview. Nematology. Brill; 2001;3: 199–208.

76. Blaxter M. Nematodes: the worm and its relatives. PLoS Biol. 2011;9: e1001050.

77. Kiontke K, Fitch DHA. Nematodes. Curr Biol. 2013;23: R862–4.

78. Sommer RJ. Pristionchus pacificus: A Nematode Model for Comparative and Evolutionary Biology. BRILL; 2015.

79. Hong RL, Sommer RJ. Pristionchus pacificus: a well-rounded nematode. Bioessays. 2006;28: 651–659.

80. Sieriebriennikov B, Prabh N, Dardiry M, Witte H, Röseler W, Kieninger MR, et al. A Developmental Switch Generating Phenotypic Plasticity Is Part of a Conserved Multi-gene Locus. Cell Rep. 2018;23: 2835–2843.e4.

81. Witte H, Moreno E, Rödelsperger C, Kim J, Kim J-S, Streit A, et al. Gene inactivation using the CRISPR/Cas9 system in the nematode Pristionchus pacificus. Dev Genes Evol. 2015;225: 55–62.

82. Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model Organism Pristionchus pacificus. Cell Rep. 2017;21: 834–844.

83. Herrmann M, Mayer WE, Sommer RJ. Nematodes of the genus Pristionchus are closely associated with scarab beetles and the Colorado potato beetle in Western Europe. Zoology . 2006;109: 96–108.

84. Susoy V, Herrmann M, Kanzaki N, Kruger M, Nguyen CN, Rödelsperger C, et al. Large-scale diversification without genetic isolation in nematode symbionts of figs. Sci Adv. 2016;2: e1501031.

85. Harvey PH, Pagel MD. The comparative method in evolutionary biology. Oxford University Press, USA; 1998.

86. Eisen JA, Fraser CM. Phylogenomics: intersection of evolution and genomics. Science. 2003;300: 1706–1707.

87. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of the National Academy of Sciences. 1999;96: 4285–4288.

88. DeSalle R, Rosenfeld JA. Phylogenomics. Garland Science; 2012.

89. Verster AJ, Styles EB, Mateo A, Derry WB, Andrews BJ, Fraser AG. Taxonomically Restricted Genes with Essential Functions Frequently Play Roles in Chromosome Segregation in Caenorhabditis elegans and Saccharomyces cerevisiae. G3 . 2017;7: 3337–3347.

90. Johnson BR, Tsutsui ND. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics. 2011;12: 164.

91. Santos ME, Le Bouquin A, Crumière AJJ, Khila A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. Science. 2017;358: 386–390.

92. Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, et al. The genomic basis of parasitism in the Strongyloides clade of nematodes. Nat Genet. 2016;48: 299–307.

93. Kafatos M, Michalitsanos AG, Vardya MS. Mass loss, long-period variables, and the formation of circumnebular shells. Astrophys J. 1977;216: 526.

94. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267: 275–276.

95. Jukes TH, King JL. Evolutionary nucleotide replacements in DNA. Nature. 1979;281: 605–606.

96. Miyata T, Yasunaga T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol. 1980;16: 23–36.

97. Yang Z. Molecular Evolution: A Statistical Approach. Oxford University Press; 2014.

98. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences. 1986;17: 57–86.

99. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24: 1586–1591.

100. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 1985;2: 150–174.

101. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24: 637–644.

102. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5: 59.

103. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13: 329–342.

104. Hoff KJ, Stanke M. Current methods for automated annotation of protein-coding genes. Current Opinion in Insect Science. 2015;7: 8–14.

105. Baskaran P, Rödelsperger C, Prabh N, Serobyan V, Markov GV, Hirsekorn A, et al. Ancient gene duplications have shaped developmental stage-specific expression in Pristionchus pacificus. BMC Evol Biol. 2015;15: 185.

106. Ragsdale EJ, Müller MR, Rödelsperger C, Sommer RJ. A developmental switch coupled to the evolution of plasticity acts through a sulfatase. Cell. 2013;155: 922–933.

107. Schuster LN, Sommer RJ. Expressional and functional variation of horizontally acquired cellulases in the nematode Pristionchus pacificus. Gene. 2012;506: 274–282.

108. Sinha A, Rae R, Iatsenko I, Sommer RJ. System wide analysis of the evolution of innate immunity in the nematode model species Caenorhabditis elegans and Pristionchus pacificus. PLoS One. 2012;7: e44255.

109. Sinha A, Sommer RJ, Dieterich C. Divergent gene expression in the conserved dauer stage of the nematodes Pristionchus pacificus and Caenorhabditis elegans. BMC Genomics. 2012;13: 254.

110. Borchert N, Krug K, Gnad F, Sinha A, Sommer RJ, Macek B. Phosphoproteome of Pristionchus pacificus provides insights into architecture of signaling networks in nematode models. Mol Cell Proteomics. 2012;11: 1631–1639.

111. Rödelsperger C, Neher RA, Weller AM, Eberhardt G, Witte H, Mayer WE, et al. Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data. Genetics. 2014;196: 1153–1165.

112. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13: 2178–2189.

113. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

114. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34: W609–12.

115. Sommer R, Carta LK, Kim S-Y, Sternberg PW. Morphological, genetic and molecular description of Pristionchus pacificus sp. n.(Nematoda: Neodiplogasteridae). Fundam Appl Nematol. GAUTHIER-VILLARS/ORSTOM; 1996;19: 511–522.

116. Susoy V, Kanzaki N, Herrmann M. Description of the bark beetle associated nematodes Micoletzkya masseyi n. sp. and M. japonica n. sp. (Nematoda: Diplogastridae). Nematology. 2013;15: 213–231.

117. Fierst JL, Willis JH, Thomas CG, Wang W, Reynolds RM, Ahearne TE, et al. Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. PLoS Genet. 2015;11: e1005323.

118. Yin D, Schwarz EM, Thomas CG, Felde RL, Korf IF, Cutter AD, et al. Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. Science. 2018;359: 55–61.

119. Slos D, Sudhaus W, Stevens L, Bert W, Blaxter M. Caenorhabditis monodelphis sp. n.: defining the stem morphology and genomics of the genus Caenorhabditis. BMC Zoology. 2017;2. doi:10.1186/s40850-017-0013-2

120. Wang J, Chen P-J, Wang GJ, Keller L. Chromosome size differences may affect meiosis and genome size. Science. 2010;329: 293.

121. Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, et al. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol. 2009;10: R103.

122. Ekseth OK, Kuiper M, Mironov V. orthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics. 2014;30: 734–736.

123. Rödelsperger C. Comparative Genomics of Gene Loss and Gain in Caenorhabditis and Other Nematodes. Methods in Molecular Biology. 2018. pp. 419–432.

124. Gilabert A, Curran DM, Harvey SC, Wasmuth JD. Expanding the view on the evolution of the nematode dauer signalling pathways: refinement through gene gain and pathway co-option. BMC Genomics. 2016;17. doi:10.1186/s12864-016-2770-7

125. Melters DP, Paliulis LV, Korf IF, Chan SWL. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosome Res. 2012;20: 579–593.

126. Thomas JH. Analysis of homologous gene clusters in Caenorhabditis elegans reveals striking regional cluster domains. Genetics. 2006;172: 127–143.

127. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a

platform for investigating biology. Science. 1998;282: 2012–2018.

128. Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, et al. A transcriptomic analysis of the phylum Nematoda. Nat Genet. 2004;36: 1259–1267.

129. Johnston RJ. A novel C. elegans zinc finger transcription factor, lsy-2, required for the cell type-specific expression of the lsy-6 microRNA. Development. 2005;132: 5451–5460.

130. Thellmann M, Hatzold J, Conradt B. The Snail-like CES-1 protein of C. elegans can block the expression of the BH3-only cell-death activator gene egl-1 by antagonizing the function of bHLH proteins. Development. 2003;130: 4057–4071.

131. Cutter AD. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol Biol Evol. 2008;25: 778–786.

132. Yang L, Gaut BS. Factors that Contribute to Variation in Evolutionary Rate among Arabidopsis Genes. Mol Biol Evol. 2011;28: 2359–2369.

133. Weller AM, Rödelsperger C, Eberhardt G, Molnar RI, Sommer RJ. Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode Pristionchus pacificus. Genetics. 2014;196: 1145–1152.

134. Denver DR, Wilhelm LJ, Howe DK, Gafner K, Dolan PC, Baer CF. Variation in base-substitution mutation in experimental and natural lineages of Caenorhabditis nematodes. Genome Biol Evol. 2012;4: 513–522.

135. Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016;17: 226.

136. Juan D, Rico D, Marques-Bonet T, Fernández-Capetillo O, Valencia A. Late-replicating CNVs as a source of new genes. Biol Open. 2014;3. doi:10.1242/bio.20147815

137. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet. 2018;50: 285–296.

138. Baskaran P, Rödelsperger C. Microevolution of Duplications and Deletions and Their Impact on Gene Expression in the Nematode Pristionchus pacificus. PLoS One. 2015;10: e0131136.

139. Rogers RL, Shao L, Thornton KR. Tandem duplications lead to novel expression patterns through exon shuffling in Drosophila yakuba. PLoS Genet. 2017;13: e1006795.

140. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3. doi:10.7554/elife.01311

141. Chen S, Zhang YE, Long M. New genes in Drosophila quickly become essential. Science. 2010;330: 1682–1685.

142. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27: 578–579.

143. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

144. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21: 936–939.

145. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

146. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31: 3210–3212.

147. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29: 644–652.

148. Serobyan V, Xiao H, Namdeo S, Rödelsperger C, Sieriebriennikov B, Witte H, et al. Chromatin remodelling and antisense-mediated up-regulation of the developmental switch gene eud-1 control predatory feeding plasticity. Nat Commun. 2016;7: 12337.

149. Rödelsperger C, Menden K, Serobyan V, Witte H, Baskaran P. First insights into the nature and evolution of antisense transcription in nematodes. BMC Evol Biol. 2016;16: 165.

150. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017;45: D190–D199.

151. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6: 31.

152. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30: 1575–1584.

153. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14: R36.

154. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31: 46–53.

155. Kanzaki N, Ragsdale EJ, Herrmann M, Mayer WE, Sommer RJ. Description of three Pristionchus species (Nematoda: Diplogastridae) from Japan that form a cryptic species complex with the model organism P. pacificus. Zoolog Sci. 2012;29: 403–417.

156. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018;2: 1626–1632.

157. Domazet-Loso T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 2007;23: 533–539.

158. Jan CH, Friedman RC, Graham Ruby J, Bartel DP. Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. Nature. 2010;469: 97–101.

159. Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9: e1003569.

160. Rödelsperger C, Röseler W, Prabh N, Yoshida K, Weiler C, Herrmann M, et al. Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. Curr Biol. 2018; doi:10.1016/j.cub.2018.07.041

161. Rödelsperger C, Dieterich C. CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. PLoS One. 2010;5: e8861.

162. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14: 178–192.

163. Sinha A, Langnick C, Sommer RJ, Dieterich C. Genome-wide analysis of trans-splicing in the nematode Pristionchus pacificus unravels conserved gene functions for germline and dauer development in divergent operons. RNA. 2014;20: 1386–1397.

164. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

165. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol. 2009;27: 221–224.

166. Zhang G, Wang H, Shi J, Wang X, Zheng H, Wong GK-S, et al. Identification and characterization of insect-specific proteins by genome data analysis. BMC Genomics. 2007;8: 93.

167. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, et al. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 2009;26: 603–612.

168. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2010;2: 393–409.

169. Guillén Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, et al. Genomics of ecological adaptation in cactophilic Drosophila. Genome Biol Evol. 2014;7: 349–366.

170. Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, et al. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 2014;10: e1004830.

171. Pegueroles C, Laurie S, Albà MM. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol Biol Evol. 2013;30: 1830–1842.

172. O'Toole ÁN, Hurst LD, McLysaght A. Faster Evolving Primate Genes Are More Likely to Duplicate. Mol Biol Evol. 2018;35: 107–118.

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

174. Katju V, Lynch M. The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics. 2003;165: 1793–1803.

175. Long M, VanKuren NW, Chen S, Vibranovski MD. New gene evolution: little did we know. Annu Rev Genet. 2013;47: 307–333.

176. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most "Dark Matter" Transcripts Are Associated With Known Genes. PLoS Biol. Public Library of Science; 2010;8: e1000371.

177. Wu D-D, Irwin DM, Zhang Y-P. De novo origin of human protein-coding genes. PLoS Genet. 2011;7: e1002379.

178. Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, et al. On the Origin of De Novo Genes in Arabidopsis thaliana Populations. Genome Biol Evol. 2016;8: 2190–2202.

179. Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G. Multiple links between transcription and splicing. RNA. 2004;10: 1489–1498.

180. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5: 435–445.

181. Toll-Riera M, Rado-Trilla N, Martys F, Alba MM. Role of Low-Complexity Sequences in the Formation of Novel Protein-coding Sequences. Mol Biol Evol. 2011;29: 883–886.

182. Haerty W, Brian Golding G. Rapid evolution of low complexity sequences and single amino acid repeats across eukaryotes. Rapidly Evolving Genes and Genetic Systems. 2012. pp. 55–63.

183. Rogers RL, Bedford T, Lyons AM, Hartl DL. Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2010;107: 10943–10948.

184. Rogers RL, Hartl DL. Chimeric genes as a source of rapid evolution in Drosophila melanogaster. Mol Biol Evol. 2012;29: 517–529.

185. Rogers RL, Hartl DL. Rapid evolution via chimeric genes. Rapidly Evolving Genes and Genetic Systems. 2012. pp. 94–100.

186. Bornberg-Bauer E, Schmitz J, Heberlein M. Emergence of de novo proteins from “dark genomic matter” by “grow slow and moult.” Biochem Soc Trans. 2015;43: 867–873.

187. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nat Rev Genet. 2013;14: 645–660.

188. Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450: 203–218.

189. Assis R, Kondrashov AS, Koonin EV, Kondrashov FA. Nested genes and increasing organizational complexity of metazoan genomes. Trends Genet. 2008;24: 475–478.

190. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J Virol. 2009;83: 10719–10736.

191. Mayer MG, Rödelsperger C, Witte H, Riebesell M, Sommer RJ. The Orphan Gene dauerless Regulates Dauer Development and Intraspecific Competition in Nematodes by Copy Number Variation. PLoS Genet. 2015;11: e1005146.