

A genomic perspective
on variations in the molecular toolkit for development
and
on the evolution of parthenogenesis in Nematoda

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Philipp H. Schiffer

aus Köln

Hundt Druck GmbH, Köln

2015

Berichterstatter: Prof. Dr. Einhard Schierenberg
(Gutachter) Prof. Dr. Einhard Schierenberg

Prof. Dr. Guenter Plickert

Dissertationsgesuch: 17.11.2014

Disputation: 16.01.2015

Widmung

To Corinna,
Luca & Lilian

Danksagung

Als allererstes danke ich meiner Familie für ihre Geduld, Unterstützung und Liebe. Ohne euch wäre all das hier ein Muster ohne Wert.

Dank gilt besonders meinen Eltern sowie Monika für all ihre Unterstützung. Ebenso Theresia, Barbara und Harald.
Special thanks go to Marike for reading this TWICE and everything else.

Dom sei gedankt für Mensagänge, viele Diskussionen und Freundschaft. Gleiches gilt für Peter: RHEINLAND.

Gedankt sei Einhard. Vor allem dafür, mich machen zu lassen, gegebenenfalls zur Zurückhaltung zu mahnen, und die volle Unterstützung in allem.

Danke Michael, für deine Ideen und Gedanken und Guenter für die Hilfe, den Rat und die Unterstützung.

Many, many thanks to Mark for “adopting” and supporting me. Would have been far from what it is without you.

Dank gilt Thomas dafür, dass er mich aufnahm und mich pushte.

Many thanks to Paul and Irma for hosting me and good times in California.

In particular thanks go to Sujai for, fun, scripting, and more fun, and friendship, and Georgios (the Greek) for fun, many scientific arguments, and friendship.

Ich danke meinen Kollegen für die Unterstützung und gute Arbeitsatmosphäre. Im Besonderen Chris und Maarten, aber natürlich auch Julia und Ndifon.

Mein Dank gilt Hans-Georg Herbig, der mir schon in der Diplomarbeit große Freiheiten ließ, mich auch während der Doktorarbeit unterstützte und immer für ein gutes Gespräch zur Verfügung stand.

Last but not least gilt mein besonderer Dank der VolkswagenStiftung, ohne die all dies nie möglich gewesen wäre.

A genomic perspective
on variations in the molecular toolkit for development
and
on the evolution of parthenogenesis in Nematoda

Philipp H. Schiffer

March 18, 2015

Contents

Abstract	iv
1 Introduction	1
1.1 Model systems to understand the genetic and molecular underpinnings of life	1
1.2 <i>C. elegans</i> development has been regarded as archetypical for the phylum Nematoda	2
1.3 Nematoda is a tremendously diverse phylum with a conserved Bauplan model	5
1.4 Evolution: Developmental System Drift and Gene Regulatory Networks and Davidson’s theory	6
1.5 Sex and no sex: molecular plasticity and the evolution of parthenogenesis . .	8
1.6 Questions	11
1.7 Methods	12
1.7.1 Introduction to 2 nd Generation Sequencing and large scale biological data analysis	12
1.7.2 Orthology	15
2 Manuscripts	20
2.1 Published manuscripts	20
2.2 Extended manuscripts	49
3 Discussion	144
3.1 Discussion of P2 : A clade I genome exemplifies the huge differences in the developmental genetic toolkit of Nematoda	145
Discussion of M3 : Evolutionary dynamics of small RNA pathways demonstrate the genomic plasticity of the nematode systems	146
3.2 Discussion of P3 : Closely related taxa show variations in developmental gene expression	148
3.3 Discussion of M1 : DSD has changed GRNs in clade IV nematodes and might facilitate the evolution of parthenogenesis	149
3.3.1 Early developmental processes and the machinery of sex determination and fertilisation is divergent in clade IV nematodes	149
3.3.2 Genes acting in anhydrobiosis are found in <i>Panagrolaimus</i>	151
3.3.3 Horizontal Gene Transfer might have favoured the evolution of anhydrobiosis in <i>Panagrolaimus</i>	152
3.3.4 Parthenogenetic <i>Panagrolaimus</i> species appear to be polyploid hybrids	153
3.3.5 The use of two orthology detection pipelines enhances specificity and sensitivity of the approach	155

3.4	Discussion of M2 : Genes from the Cambrian Explosion unite Bilateria in a developmental phase	162
3.5	Discussion of P1 : We need to sequence more genomes in collaborative efforts	165
3.6	Conclusions and Outlook	166
3.6.1	Concluding remarks	166
3.6.2	Outlook	169
	Understanding the molecular backbone of development in Enoplea	170
	Understanding the molecular underpinnings of parthenogenesis	172
4	Bibliography	174
	Zusammenfassung in deutscher Sprache	191
	Beteiligung an den angeführten Publikationen	193
	Erklärung und Lebenslauf	195

List of Excursus

1	Genome assembly pipeline	18
2	Mutation Rates in parthenogenetic nematodes	156

List of Figures

1	Some Nematodes	3
2	The Nematode Tree	4
3	The Nematoda Bauplan	6
4	Parthenogenesis outcompetes sexual reproduction	9
5	de Bruijn Graph Assembly	14
6	Different classes of homology	15
7	Genome Assembly Pipeline	18
8	The MA line experiment	157
9	Fitness assay	159
10	Nematode Kinome Evolution	170

Abstract

The phylum Nematoda is characterised by a huge diversity of species that exploit almost all habitats on earth. Despite their prevalence in a wide range of different ecosystems, nematodes adhere to a strikingly strict Bauplan with only minor variations, even between two large groups that split more than 400 million years ago. The conservation of the final adult body form is quite special and not common in other animal taxa; this exceptional conservatism in the Bauplan, and the very similar patterns of early development observed in the model organisms *C. elegans*, *P. pacificus* and *Ascaris* together, have led scientist to suggest that these mechanisms of early development are archetypical for the phylum. However, analysis pioneered in the Schierenberg laboratory throughout the last 25 years challenged this view by describing considerable variations of early development in several species from different branches of the phylum. These observations together with data from divergent species in Panarthropoda gave rise to the question whether the molecular toolkit for nematode development could be subject to change as well. In the thesis presented here, this question is addressed from a genomic perspective, assembling and analysing large-scale data from species on different taxonomical levels. Based on these data comparisons have been made ranging from species at phylogenetically antipodal positions in the phylum separated by hundreds of millions of years of evolution to genera in one specific clade of the nematode tree, and finally the comparison of two closely related genera. In all these taxa Gene Regulatory Networks (GRNs) of early development are analysed and set into perspective with the common model of the nematode developmental toolkit drawn from *C. elegans*. I used these assays to test whether recently widely discussed theories on the role of GRNs for development deliver valid predictions for the evolution of early development in Nematoda. In fact, I find that the emerging picture supports such hypotheses of GRN evolution: in many pathways intermediate genetic switches appear to be exchanged by processes collectively called “Developmental System Drift (DSD), while upstream and downstream acting genes are more likely to be conserved. Despite this disparity across Nematoda, an analysis of genes retained across all Bilateria shows that this hugely diverse taxon, comprising Nematoda, could be characterized by a process of minimal divergence namely the phase in development when the adult body form is constructed. In Nematoda, parthenogenesis evolved in several genera, with a hotspot in clade IV of the phylum. The data sampled to assess the evolution of development in this thesis are used to elucidate the origin and molecular mechanisms underlying parthenogenesis in the genus *Panagrolaimus*. While the establishment of a re-shuffling mechanisms of GRNs through DSD does not yet allow us to unravel the distinct molecular mechanisms underpinning the establishment and maintenance of parthenogenesis, we have good evidence that parthenogenetic species in the genus *Panagrolaimus* are polyploid hybrids. This finding supports the hypothesis that hybridisation is a common route to parthenogenesis in Nematoda, as found in many other taxa as well. Parthenogenesis has also been linked to survival in novel and extreme environments, this would be facilitated in the *Panagrolaimus* species as they are capable of undergoing cryptobiosis (complete desiccation) in contrast to *C.elegans* and most other nematodes tested. Exploring the trait from a genomic perspective, we found genes known to be acting in this process in *Panagrolaimus*, but more importantly an intriguing link to Horizontal Gene Transfer (HGT) was found. Genes acquired through HGT appear to lend *Panagrolaimus* an adaptive advantage in extreme environments by acting in DNA repair mechanisms, which are important during rehydration.

This illustrates the previously underestimated importance of HGT in Metazoa. The genomic and transcriptomic data sampled and assembled for this thesis can serve as a basis for future projects analysing the evolution of developmental systems with regards to GRNs and DSD, as well as detailed analyses of anhydrobiosis and the molecular background of parthenogenesis.

Chapter 1

Introduction

1.1 Model systems to understand the genetic and molecular underpinnings of life

In their quest to order life into units of kinship taxonomists have traditionally relied on a hierarchical top-down system based on Linnaeus' "Systema Naturae". This hierarchy starts at the highest level with life itself. Taxa are then ordered by size, moving through domains (formerly kingdoms; e.g. Metazoa, Woese et al., 1990), from phyla (Nematoda), via superfamilies, families, orders, down to the smallest units, genera and species. In this hierarchy, morphological markers and developmental traits (ontology, including larval stages) have traditionally been used for the classification of animal taxa. This basis of classification has only recently been replaced by molecular phylogenies (Edgecombe et al., 2011). This shift from a morphology- and Bauplan-based taxonomy to phylogenetic trees based on conserved gene and protein sequences held some surprises for taxonomists, for example the erection of the taxon Ecdysozoa. Annelida and Arthropoda have classically been united as one taxon, based on mainly their segmented organs and partly coelomatic Bauplan model (Budd, 2001, chapter 5 ff. in Telford and Littlewood, 2009). We know today that molecular data from the sequence of the small ribosomal subunit gene (18S or SSU in short) clearly separate both taxa. Contemporary methods now group Panarthropoda with Nematoda in the taxon Ecdysozoa, the moulting animals (Aguinaldo et al., 1997), which also contains Kinorhyncha and Priapulida (Dunn et al., 2008). Please note that, Panarthropoda are still debated as a taxon (chapters 8, 11 in Telford and Littlewood, 2009), but used for convenience in this text. Annelida are now grouped with molluscs and some other taxa in the phylum Lophotrochozoa, that is animals possessing a lophophore for feeding (e.g. Brachiopods) or a trochophore larva (in the first-named taxa; Halanych et al., 1995). To gain insight into the genetic and molecular background of life, researchers established several model organisms across the tree of life. Among these classical models are the bacterium *Echerichia coli*, the plant *Arabidopsis thaliana*, the vertebrate *Mus musculus*, the arthropod *Drosophila*

melanogaster, and finally the nematode *Caenorhabditis elegans*. These animals were not primarily selected for their taxonomic position, or for being good representatives for a larger taxonomic unit, but because of ease of culturing, rapid development, and other traits that made them easily accessible to the experimenter. The nematode *Caenorhabditis elegans*, for example, was originally introduced as a model system by Sydney Brenner because it possesses several invaluable traits (Brenner, 1974): it is a self-fertilising hermaphrodite, allowing certain shortcuts in the genetic mapping of mutations, but can be outcrossed to generate lines with a distinct genetic makeup, due to the rare occurrence of males. It has a short generation time of about 3.5 days and grows in massive numbers on agar plates with *E. coli* bacteria as a food source. This is why the “The Worm” has become a favourite laboratory pet for biologists working on evolution, development, evolution of development (EvoDevo), neurobiology, and many other diverse research areas (Blaxter, 2011).

1.2 *C. elegans* development has been regarded as archetypical for the phylum Nematoda

The most important milestones in *C. elegans* research may be the description of the complete cell lineage during its development from single cell to larva to adulthood (Deppe et al., 1978; Kimble and Hirsh, 1979; Sulston and Horvitz, 1977; Sulston et al., 1983), the introduction of green fluorescent protein as a marker for gene expression (in transgene animals) (Chalfie et al., 1994), and the application of RNA interference (**RNAi**, Fire et al., 1998; Timmons and Fire, 1998). Within the phylum Nematoda, *C. elegans* is positioned in a crown-clade, distant from the root of the phylum, and can be regarded as highly derived in its biology (Blaxter, 2011). Despite this, its ubiquitous use in research has led to the impression that *C. elegans* development is archetypical for Nematodes. This view was re-enforced when only minor deviations from its development were found in both, *Ascaris* development, whose early cell lineage has been described already in the 19th century (Boveri, 1899; Müller, 1903), and also development of the other model nematode, *Pristionchus pacificus* (Sommer, 2006; Vangestel et al., 2008). A similar situation has been prevalent in arthropods where research dealing with the development and the evolution of development has focussed on *Drosophila melanogaster* (Sommer, 2009). These views are now under revision and while work on both, *C. elegans* and *D. melanogaster*, has allowed scientists to unravel major genetic pathways orchestrating development, scientists in the field now appreciate that the developmental background across the Panarthropoda differs considerably on the cellular-morphological level. This has led to a renewed interest in establishing new models from across the phylum (Sommer, 2009). One such new model is the red flour beetle *Tribolium castaneum*, which has been used as a laboratory organism for a long time, see for example Bywaters et al. (1959). But also other, previously not analysed species (e.g. the wasp *Naso-*

nia vitripennis, see Werren and Loehlin, 2009), are now used to shed light on developmental variations present in the phylum. This research into variations in development in various species has recently brought the role of Gene Regulatory Networks into focus.

Divergence in Gene Regulatory Networks (**GRNs**), which orchestrate developmental programmes (see section 1.4 on page 6, and Davidson and Erwin, 2006) between these arthropods has been reported for example for the establishment of the dorso-ventral axis (Lynch and Roth, 2011). Another example for a change in GRNs between Arthropoda is the formation of anterior-posterior body axes and body regions: In contrast to *D. melanogaster* the establishment of “posterior” in Arthropods seems to depend on the action of delta-notch and Wnt genes, which are following the putatively ancestral mode of short germ development (McGregor et al., 2009). The Wnt-signalling pathway is of further interest as it illustrates the derived status of the model organisms “worm” and “fly”. Both are depleted in Wnt genes, in comparison to for example in spiders at the base of Panarthropoda. Most strikingly, a much larger set of Wnts is present even outside Bilateria in the sea anemone *Nematostella vectensis*, (Janssen et al., 2010), where the pathway plays a role in axis formation (Marlow, 2013). This shows that GRNs are crucial in explaining changes between developmental programs, and analysing them sheds light on the underlying evolutionary principles. The here presented thesis will be concerned with such changes in the GRNs and the interactions of proteins in development, and I will come back to this shortly.



Figure 1: A: male *Panagrolaimus* sp., B: female *Panagrolaimus* sp., C: *Romanomermis culicivorax*, courtesy of E. Schierenberg. Not the different scale bar, *R. culicivorax* is much larger than the *Panagrolaimus* species, but the vermiform morphology is very similar.

Over the last two decades, the pioneering work on different nematodes in the Schierenberg lab (see for example Schulze and Schierenberg, 2009; Skiba and Schierenberg, 1992; Wiegner and Schierenberg, 1998, 1999) has shown that *C. elegans* is only one worm among many when it comes to early development in Nematoda. The included publications will pay special attention to these developmental shifts in the Panagrolaimids and the genome of the enoplean species *Romanomermis culicivorax*, showing that several microscopically observable developmental traits differ between nematodes. Comparing *C. elegans* and *R. culicivorax* is interesting as these worms are positioned at opposing positions in the the phylogenetic tree of the phylum Nematoda. Multiple differences are evident, such as that the first cell division, which is equal in *R. culicivorax* while it is asymmetric in *C. elegans*, and the dorso-ventral axis polarity in the former is inverted in comparison to the latter (Schulze and Schierenberg, 2008, 2009). Further, the formation of hypodermis in *R. culicivorax* is most peculiar, as it involves the generation of repetitive cell-rings (Schulze and Schierenberg, 2009), that

are reminiscent of arthropod segmentation stripes. Taking these findings into account, as well as cross-phylum comparisons of development (Schulze and Schierenberg, 2011), and the research on *Plectus sambesii* (Schulze et al., 2012), it becomes clear that variations on the cellular level are the norm, not an aberration, among Nematoda.

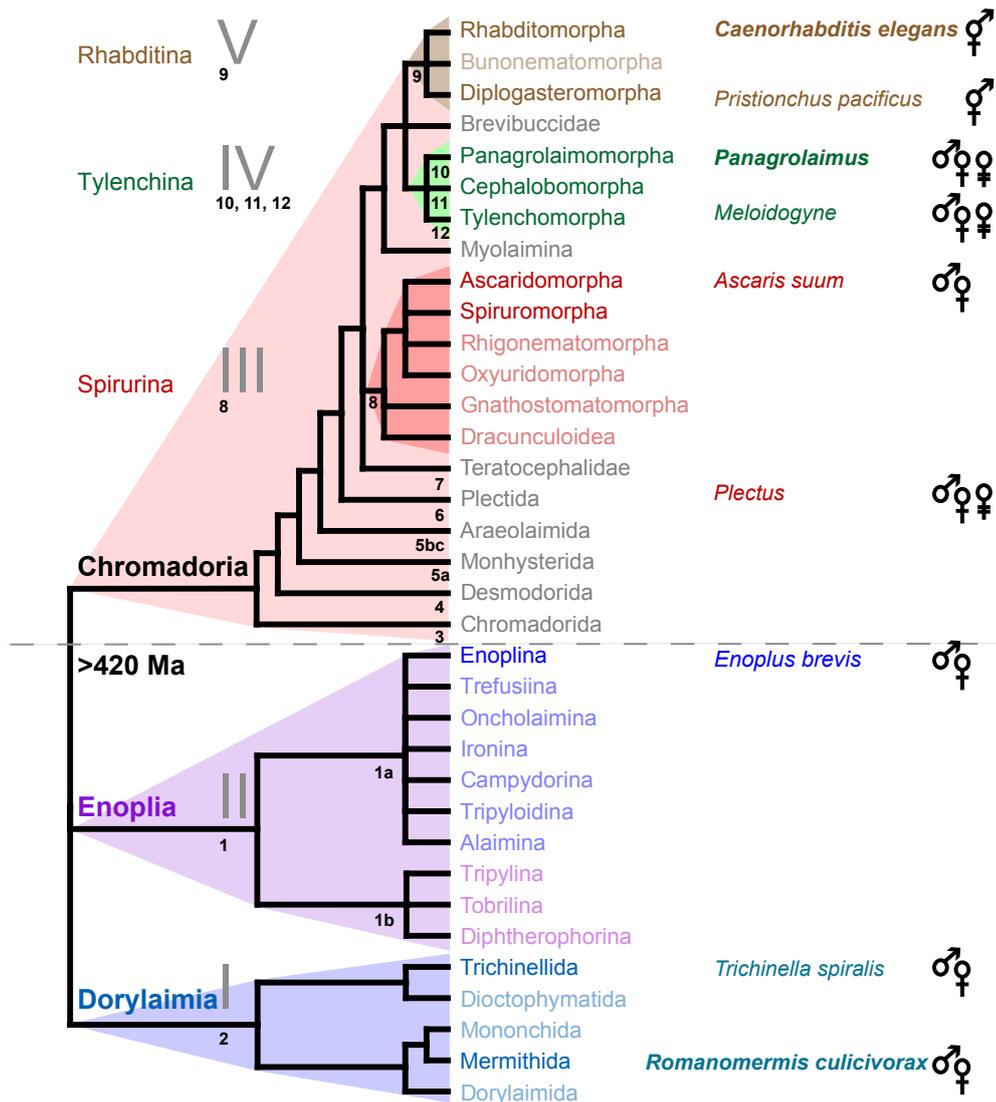


Figure 2: The phylum Nematoda with all major subtaxa. Enoplea (containing Enoplia and Dorylaimia) and Chromadorea (in the tripartite system displayed here named Chromadoria) split about ~420 Ma ago, when the latter presumably conquered land. Main species mentioned in this thesis are indicated on the right, with genera given where several species are dealt with. Modes of reproduction for named species or within genera are displayed to the right. Reproduced with modifications from Blaxter (2011).

1.3 Nematoda is a tremendously diverse phylum with a conserved Bauplan model

Before returning to the main questions focussing on the evolution of developmental systems and changes in GRNs, the main objects of study in the presented research should be introduced. The widely ramified phylum Nematoda could encompass an estimated 1 million species (Lambshhead and Boucher, 2003), with some data even suggesting that the true number could be 10 times higher (Mark Blaxter, personal communication). Nematoda split from the phylum Panarthropoda, with which it is joined in the Ecdysozoa (Aguinaldo et al., 1997) more than 600 million years ago in the Ediacaran age (Rota-Stabelli et al., 2013). Based on some morphological characters the phylum Nematoda itself can be subdivided into the two large classes Enoplea and Chromadorea. Examples for their differences are that Enoplea have a cylindrical or bottle-shaped oesophagus, while this organ usually is divided into bulbs in Chromadorea. Further, important chemoreceptors called amphids have a different structure in both classes. The excretory system in Enoplea is very simple (a single cell in most cases), whereas Chromadorea have gland cells or a system with canals (e.g. in *C. elegans*). As a last example, in Enoplea males and females have a bi-armed germline, while this is either single- or bi-armed in female Chromadorea. See <http://plpnemweb.ucdavis.edu/nemaplex/Taxadata/Classes.htm> for a short list of differences and various chapters in Lee (2002) for details. Both, Enoplea and Chromadorea, have again been subdivided into larger sub-taxa by various authors using morphological characters and molecular phylogeny. Despite the discrepancies in numbers and placing of single clades, these sub-divisions nevertheless retain a similar general tree topology (Blaxter et al., 1998; Holterman et al., 2006; van Megen et al., 2009). For the sake of simplicity, I will use the phylogenetically robust five clade system introduced by Blaxter et al. (1998) for the remainder of this thesis and the included publications (figure 2 on the preceding page). It has been hypothesised that Chromadorean nematodes could have first conquered land as parasites of arthropods when these left the oceans in the Silurian age (>420 Ma ago; **Ma = megaannus**), but other scenarios are possible, including an early sea-land transition following the first land plants (Pisani et al., 2004; Poinar et al., 2008; Rota-Stabelli et al., 2013). The huge molecular divergence in Nematoda has been appreciated for some time. This divergence is indicated by very long branches leading to many taxa in phylogenies based on conserved genes (van Megen et al., 2009) and is possibly owing to fast evolution within sub-taxa across the phylum. But just how much genetic variability *can* arise in a single species has only recently been discovered. One study showed that populations in the pan-tropical species *Caenorhabditis brenneri* are indeed hyperdiverse, i.e., this species shows more molecular diversity than certain bacteria (Dey et al., 2013), while retaining full mating compatibility (thus constituting one biological species

sensu Mayr, see Coyne and Orr, 2004). The mutational divergence inherent to single clades and between clades in the phylum has also been indicated by studies on gene family turnover in Nematoda. These showed high numbers in birth and death rates of these families in the lineages leading to sequenced species (Godel et al., 2012; Mitreva et al., 2011).

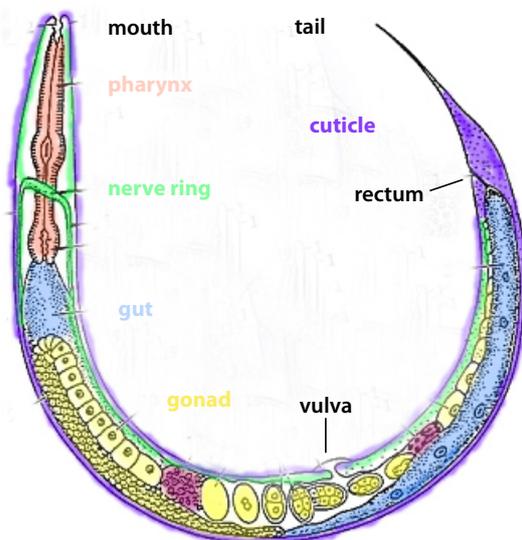


Figure 3: The general Bauplan of all species in the phylum Nematoda is highly invariant. Here only major Bauplan features shared across Enopleia and Chromadorea are named (see main text for details). Some species have two armed gonads, like *C. elegans*, while others have one, like *Panagrolaimus* Skiba and Schierenberg (1992), but structural changes as for example observed in the phyla Panarthropoda or Lophotrochozoa is absent (see main text for detailed examples). Illustration based on *C. elegans* courtesy of E. Schierenberg.

trachea. In comparison, the huge molecular divergence in Nematoda appears not to be reflected in morphological Bauplan-divergence to the extent found in other phyla.

1.4 Evolution: Developmental System Drift and Gene Regulatory Networks and Davidson’s theory

Development is orchestrated by a plethora of genes organised in interacting networks. Constituents of these developmental toolkits have been analysed in the classical model systems and their homologs have then been identified in other organisms. Through these analyses it has become clear that a set of cell-signalling pathways are conserved and acting in at least all bilaterians, with some being conserved across Metazoa or in all eukaryotes (Pires-daSilva and Sommer, 2003). Wnt-signalling, one such pathway, has already been shortly mentioned above. While the cnidarian *Nematostella vectensis*, as well as *Homo sapiens*, and the crustacean *Daphnia pulex* possess a set of at least 12 wnt ligands each, this number is decreased

Most strikingly however, all nematodes share a common vermiform Bauplan, which has been conserved for hundreds of millions of years (Poinar et al., 2008). While variations on a morphological (or phenotypic) scale such as in the form and ornamentation of the cuticle, appendages, the mouth and tail region, as well as in body size exist (de Ley, 2006), the whole phylum is united by a very strict and invariant vermiform Bauplan model (figure 3). This “assemblage of homologous architectural features” Valentine (1986) is largely shared by the closest outgroup, the nematomorpha. This long-term Bauplan conservatism is especially puzzling when compared to the sister phylum Arthropoda, which unites eight-legged spiders with decapods in Crustacea and the Hexapoda, and where respiratory systems range from book lungs to gills to

in the evolutionary lineages leading to both crown groups *D. melanogaster*, 7 genes, and *C. elegans*, 5 genes, (Janssen et al., 2010). Intriguingly, in *C. elegans* the opposite end of this signalling cascade, i.e. the DNA-binding factor β -catenin, has multiplied from the canonical one into four genes that are used in different processes (Eisenmann, 2005). Indeed, *C. elegans* lacks several important genes from the developmental toolkit present in arthropods and vertebrates: most prominent is the reduction of the HOX genes, with only four of the canonical bilaterian set retained in *C. elegans*, which has however acquired a fifth HOX gene through a nematode specific duplication event (Aboobaker and Blaxter, 2003). Other examples include the loss of the BMP antagonist chordin in the lineage leading to *C. elegans*. This is striking, as the gene is highly conserved between arthropods and vertebrates. It is acting in the dorso-ventral axis formation (Piccolo et al., 1996) and it is possible to replicate its function in early development in the frog *Xenopus laevis* by injection of the *D. melanogaster* mRNA (the fly's chordin orthologue is called short gastrulation, Sog; Schmidt et al., 1995). Still, there are differences between the Deuterostome (frog) and the Protostome (fly) systems, like the well-known inversion of the dorso-ventral axis between both taxa, and the possible restriction of expression of conserved genes to different germ layers in the organism (Ferguson, 1996). Thus, even these highly conserved orthologues show some plasticity in their expression between taxa and the processual implications thereof for the respective species. Such plasticity in developmental programmes, i.e. the change in utilisation, neo-functionalisation, or complete exchange of genes, between species in a process, which is not necessarily adaptive, has been termed "Developmental System Drift" (DSD, True and Haag, 2001). DSD has particularly been described in sex determination, a fundamental developmental process signified by considerable change in the molecular underpinnings of its regulatory mechanisms between bilaterian taxa such as *C. elegans*, *D. melanogaster*, and *H. sapiens* (Beukeboom and Perrin, 2014). Nevertheless, some genes in the involved signalling cascades are functionally conserved across vast evolutionary distances. For example, parts of the Hedgehog signalling system are involved in sex determination in the fly, worm, and human (Franco and Yao, 2012). While a bona-fide hedgehog gene is missing in *C. elegans* (Burglin, 2006), the gene *tra-1* (transformer 1) has a direct orthologue in the *D. melanogaster* gene *ci* (*cubitus interruptus*) and Gli-1 in *H. sapiens*: all are active in sex determination in one way or another (ibidem). DSD has also been intensively studied with respect to the formation of the vulva, the nematode egg-laying organ. Here, DSD was found to have acted in several branches of the rhabditid group, containing *C. elegans*, (Kiontke et al., 2007). The formation of the vulva is also under divergent genetic control between *C. elegans* and *P. pacificus* (Sommer et al., 1998; Tian et al., 2008). These examples suggest a pattern for genes acting in the developmental toolkit of animals: some key players are conserved, while other effectors are under considerable evolutionary turnover. Davidson (2006) and others see this as a fundamental process in animal evolution (Davidson and Erwin, 2006). They conjecture that GRNs orchestrating development can be broken down into smaller

elements where especially the terminal players cannot easily be changed, but intermediate switches underly considerable evolutionary modification. Indeed, in their theory a set of conserved sub-circuits termed kernels is deployed into different pathways to fulfil specific conserved developmental processes (Davidson, 2006). One possibly pan-bilaterian example Davidson gives in this account, is the system that specifies a basic tripartite structure of the brain; here a set of genes including *Orthodenticle* is found acting in homology in mouse and *Drosophila*.

According to this hypothesis the genetic architecture surrounding kernels is subject to evolutionary turnover, i.e. genes and their proteins that transmit information between kernel modules are constantly altered during speciations (Davidson, 2006; Davidson and Erwin, 2006). The degree of plasticity in such a systemic toolkit, or the amount of DSD affecting GRNs, across the phylum Nematoda has, however, not been explored so far. This is a striking void in our understanding especially because the molecular divergence and the birth/death rates of genes and gene families reported for Nematoda (see section 1.3 on page 5) suggest a high degree of DSD through the change and turnover in genetic toolkits. This is further reinforced by the observation of considerable change in the cellular patterns of development as described above. Explained in more detail in the introduction to 2nd generation sequencing below (page 12 ff.) one problem in comparing the genetic toolkits from different branches of the nematode phylum has been the scarcity of genomic data and the bias towards certain clades in the available data. One aim of this thesis is therefore to expand the genomic scope in this taxon by studying more representatives, including those from under-represented branches. To this end, the genome of *Romanormis culicivora*, a member of Dorylaimia in clade I, was sequenced and analysed (discussed in section 3.1 on page 145)

1.5 Sex and no sex: molecular plasticity and the evolution of parthenogenesis

Sex is by far the most abundant form of reproduction in Metazoa. The origin of outcrossing and meiosis is closely associated with the evolution of eukaryotes (Maynard Smith, 1978), where sexual reproduction is also predominant. However, under similar ecological and genetic conditions, an individual undergoing parthenogenetic reproduction without outcrossing will generate more offspring (each of which is by itself capable of generating offspring) than a sexual sibling (figure 4 on the next page). This cost of producing males, finding mates, courtship, or intercourse itself should give parthenogenetic taxa a huge evolutionary advantage (Otto and Lenormand, 2002). The seeming paradox, between the dominance of sexual reproduction and the apparent advantages of parthenogenesis, has always puzzled evolutionary biologists, leading Graham Bell to call it the "Queen of Evolutionary Questions"

(Bell, 1982). Consequently, much work has been devoted to the question of the predominance of sex, see for example Schön et al. (2009). The main theoretical explanation for the evolutionary benefit of sex assumes that sexual reproduction allows species to disseminate and combine (novel, beneficial) genotypes through meiosis followed by outcrossing (Maynard Smith, 1978). So far, most of the work attempting to prove the superiority of sex has been either theoretical or descriptive (Schön et al., 2009), but recently some studies produced evidence showing under which environmental conditions sex is favourable. For example, Becks and Agrawal (2010) found that in heterogeneous environments sexual rotifers outcompete parthenogenetic ones. This appears to be true as well for adaptation to changing environments (Becks and Agrawal, 2012). In summary, it is the sum of all factors which appears to favour sex evolutionary as the mode of reproduction (Beukeboom and Perrin, 2014). A further interesting theory states that it can not be abolished easily in vertebrates, as their molecular and cellular system might be too complex for such a drastic reorganisation (Avisé, 2008). This might serve as a further explanation for why parthenogenetic taxa are rare among animals (Otto and Lenormand, 2002), but common in plants.

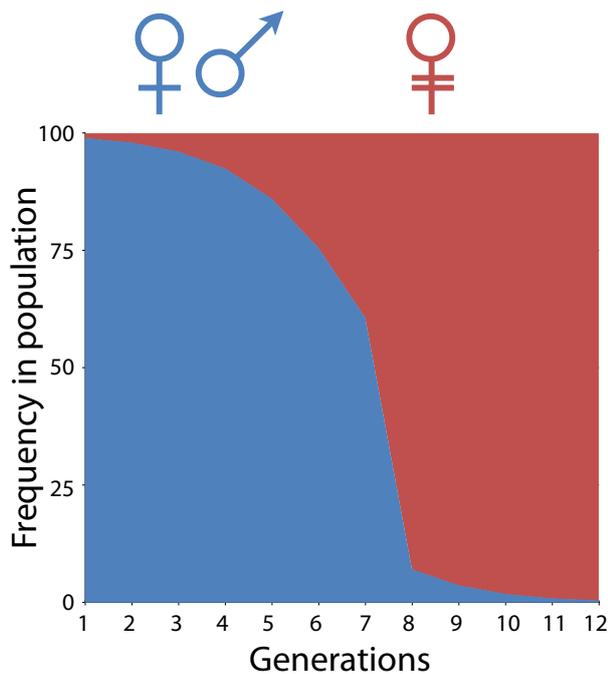


Figure 4: Parthenogenetic species outcompete sexual congeners in few generations. Starting from one parthenogenetic specimen and 99 amphimictic ones the novel parthenogenetic species would almost completely outnumber the amphimictic one after only 12 generations. Even if the initial proportion would be far worse for the parthenogenetic species (1 to several ten thousands) a reversal of the proportions would take place within ~ 40 generations. Calculations are based on Maynard Smith (1978).

(male/female) species *C. remanei* (Baldi et al., 2009). However, in parthenogenetic species, a major re-programming of the reproductive system is necessary for it to function without genetic contribution of males, or the activation of the oocyte induced by sperm entry

Thus, it is an interesting evolutionary trait within the phylum Nematoda that parthenogenesis arose in several branches with a particular hotspot in clade IV (see Denver et al. 2011, for a phylogenetic classification of parthenogenetic taxa in Nematoda). To implement the switch from a male/female (amphimictic) to a “female only” system, a newly evolved parthenogen has to overcome major cellular and molecular challenges to successfully reproduce. These are more severe than the switch to hermaphroditism, which does not need the abolishment of the male phenotype - otherwise a necessary evolutionary step to avoid the cost of sex in the absence of any benefits. Indeed, it has been shown that altering the expression of only two genes is sufficient to generate spermatid producing pseudo-hermaphrodites in the sexual

(Goldstein et al., 1998). Meiotic arrest must be released, diploidy restored, mitotic spindles formed, and polarity established (Engelstadter, 2008) to allow for successful reproduction. In general, Gene Regulatory Networks orchestrating sex determination and fertilisation must be rapidly re-wired to allow for a successful establishment of parthenogenesis. Data how this evolutionary switch between reproductive modes is achieved are scarce. Heger et al. (2010) found an apparently functional gene for major sperm protein in the parthenogenetic nematode *A. nanus*, but did not detect the corresponding protein. They also showed that MAP-kinase activation in oocyte maturation is conserved in this parthenogenetic species. Other authors have tried to induce parthenogenetic development (which inevitably fails) in vertebrate models, like mice (Siracusa et al., 1978). Schwander et al. (2010) modelled how a newly arisen ability to undergo parthenogenesis could lead to the establishment of all-female lines, and correlated their theoretical findings to data from *Timema* stick insects, but did not explore the molecular genetic background of these. Hence, it is not clear in which way the system of GRNs is changed and how “plastic” (i.e. liable to fast evolutionary change) the system might be (McGhee, 2011). Indeed, it is a long-standing question whether closely related taxa evolve similar phenotypes in a truly parallel way, i.e. by homologous change(s) to the same gene(s) or even nucleotide(s), or by a convergent path from a divergent basis, i.e. the evolutionary outcome is the same, but different, non-homologous, molecular routes are taken (Elmer and Meyer, 2011). Parthenogenesis (as a trait or phenotype) in closely related species could in theory evolve by mutations in the same genes; similarly as indicated by pseudo-hermaphrodite *C. remanei*, a few changes at crucial genome positions might for example stop the production of males or enable the activation of eggs. This would be parallel evolution. But it is also possible that in each instance where parthenogenesis evolves, the genetic programme is changed at different positions. There are for example quite a few genes affecting sex determination in *C. elegans* Stothard and Pilgrim (2003), which could become relevant when abolishing males. This would be convergent evolution. Convergence in character evolution has been reported for development in rhabditid nematodes, where for example the signal for the induction of vulva formation and the cell-lineage patterns which build this important sex-specific organ are divergent among sister taxa, but reoccur on a wider phylogenetic scale (Kiontke et al., 2007). This is also an example for DSD between closely related genera. In the genus *Caenorhabditis* hermaphroditism arose at least three times independently from amphimictic ancestors (Kiontke et al., 2004, 2011), but despite extensive sampling (45 described species in the genus; Karin Kiontke personal communication) so far no closely related parthenogenetic species have been described. To develop a system where the evolution of parthenogenesis could be analysed from a genomic perspective, we needed a different model taxon. Consequently, as part of the work conducted for this thesis I constructed the genomic backbone of species in the clade IV genus *Panagrolaimus*, where amphimictic and parthenogenetic species are found in close phylogenetic association. Additionally, as indicated above, in clade IV of the nematode tree partheno-

genetic species have evolved in several genera, which will in the future allow us to address questions about convergence and parallelism and correlate these to general patterns of DSD and GRN evolution.

1.6 Questions

The presented papers address similar questions comparing different taxonomic levels. These questions are addressed by comparisons between very closely related genera in the manuscript entitled “Developmental variations among Panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit”. A comparison (**P3**), across clades within the phylum in “How to survive the extreme: a multi genome analysis reveals evolutionary traits of cryptobiosis and routes to parthenogenesis” (**M1**), spanning the phylogenetic breadth of the phylum is conducted in “The genome of *Romanormis culicivorax*: Revealing fundamental changes in the core developmental genetic toolkit in Nematoda” (**P2**), and partly in “Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages” (**M3**). Finally, a comparison based on the evolution across Bilateria as a huge taxon encompassing a large variety of diverse phyla is presented in “Proteins from the Cambrian Explosion” (**M2**). The main focus of the described analyses is variation (and conservation) in the molecular toolkit of (early) development. This is then set into context of the evolution of parthenogenesis, for which some processes in early development have to be adapted. These results are further set into context of the evolution of parthenogenesis, a process that requires the adaptation of some aspects of early development. This adaptation is not only necessary because of the above-mentioned variations in nematode embryogenesis on the cellular level, but also because early development is arguably a key phase in animal life, where the most dramatic changes are taking place. Further on, changes in developmental mechanisms and the underlying genetic repertoire are expected to act as early boundaries during speciation and shortly after new species are established (Coyne and Orr, 2004). These changes might thus be a major driving force of evolution, which has led some researcher to call for an investigation of phylum (or taxon) evolution in the context of this phase of life (Davidson, 2006; Nei, 2013). Consequently, the main questions posed in the thesis at hand are as follows:

- Is the highly conservative Bauplan model of nematodes reflected by an equally conservative genetic toolkit for early development, or is the apparent plasticity on the cellular level found in some developmental processes reflected by rapid genomic evolution? In other words, how much plasticity in the genetic toolkits, or DSD, is inherent to specific taxonomic levels within the phylum?
- If genes known for their crucial function in *C. elegans* early development (and maybe in outgroup species, e.g. *Drosophila*) are conserved, are the respective interaction

partners of these important genes retained as well (on different taxonomic levels)? This question explores the divergence of GRNs.

- How is potential plasticity involved in the evolution of parthenogenesis, and are convergent or parallel genomic pathways taken in the process?
- Assuming drastic changes in the GRNs, as proposed by Davidson (see above): is there nevertheless a set of universally conserved genes functioning in comparable life-cycle processes across the diversity of bilaterian species?

A further question concerning the evolution of the Bauplan of Nematodea (a taxon uniting Nematodes and Nematomorphs; chapter 8 in Telford and Littlewood, 2009) will be addressed in a forthcoming manuscript, which will be shortly previewed in the **Outlook** section. The assay that will be described in this manuscript is the first to analyse to what extent genes that are missing in *C. elegans* and its close allies and were presumed to be absent in all nematodes, are actually retained in early branching roundworms. An answer to this question will allow a better estimate of the genetic properties and body form of the last common Ecdysozoan ancestor and the evolutionary route that led to Nematodea.

1.7 Methods

Only by making use of the novel kind of large scale data from 2nd generation sequencing assays, rather than going gene by gene from PCRs and cDNA fishing in the laboratory, was it possible to address key questions about the evolution of and within the phylum Nematoda presented in the included manuscripts. Second generation sequencing is still a very novel and rapidly changing approach. Here, I will depict the applied methods in more detail than in the concise Methods sections of the included publications. A major part of my PhD project was to acquire, employ and adapt these techniques in the context of the addressed questions.

1.7.1 Introduction to 2nd Generation Sequencing and large scale biological data analysis

In 1999, *C. elegans* became the first metazoan whose genome was fully sequenced (*C. elegans* Sequencing Consortium, 1998). Since then, several other model organisms, for example the fruit fly *Drosophila melanogaster*, have been genome sequenced. However, only the advent of 2nd generation sequencing methods has allowed researchers to explore the genomes of a variety of organisms across all branches of life. The originally exorbitant costs for sequencing of a whole genome have decreased dramatically (and still do so), making it attractive to analyse non-model species (Kumar et al., 2011). However, there is still no standard protocol

to create a reliable and complete genome sequence from 2nd generation data. In short, 2nd generation sequencing methods are based on the massively parallel sequencing of short stretches of DNA, which are generated by random fragmentation of genomic DNA. The currently most widely used technique developed by Solexa/Illumina can generate about 40 **Gigabases (Gb)** of raw data in about 2 weeks of time from DNA extraction to sequencing read. To set this into scale, 40Gb are approximately equals 12X the human genome, which when first analysed (about a decade ago; Venter, 2003; Venter et al., 2001) cost many millions of Euros and took several years for large teams to be completed by the so-called Sanger shot-gun-sequencing method. The main challenge for biologists remains to assemble genomes from 2nd generation sequencing data. Traditional Sanger-sequencing reads, by virtue of the low error rate and comparatively long fragments (800 - 1100bp), could be combined to contigs and scaffolds by looking for and then aligning overlapping regions at the ends of reads. Such an overlap-based assembly from the huge amount of short read data generated for by 2nd generation sequencers is unfortunately computationally not possible. This means that the problem cannot be solved computationally in a efficient amount of time. Thus, especially as the numbers of 2nd generation reads included into an assembly (termed the **read coverage**) have to be high to counteract the technique's intrinsic error rate. New algorithms had to be developed to overcome the assembly problem, see e.g. (Zerbino, 2009). Currently almost all assembler programs are based on the de Bruijn Graph methods, methodologically introduced by Pevzner (2001); Pevzner et al. (2001), see figure 5 on the following page. A de Bruijn graphs is constructed by splitting sequence reads into even shorter fragments, so called **kmers**, which are represented as edges in the graph. Continuous stretches of sequence (called **contigs**) are then build by finding those kmers, which differ by only one additional base at the end, see figure 5 on the next page. In this way, contigs are extended base by base during the assembly process. Unfortunately, with this procedure we cannot (yet) routinely generate genome assemblies consisting of few very large sequence stretches (hundreds of kilobases or even megabases). Sequencing errors, as well as repetitive regions, other complex genomic areas, and not last the diversity in the genome itself, lead to so called bubbles in the graph (see figure 5), which cannot be resolved in all cases, see for example Miller et al. (2010) or Schatz et al. (2012). Contigs build from 2nd generation sequencing are thus in general shorter than those originating from traditional Sanger sequencing. Some methods, as for example mate pair sequencing, see Schatz et al. (2010), or the generation of very long (but even more error prone) reads (Ferrarini et al., 2013; Schatz et al., 2012), exist to extend contigs to longer scaffolds. But the process is still far from being standardised and invokes a drastic increase of the monetary cost for each genome. Upcoming technology like Oxford Nanopore single-molecule sequencing promise to remediate these problems by generating much greater read lengths (up to several kilobases). But it is already now possible to construct good genomes from short reads. These assemblies contain enough information to generate a reliable representation of the gene content of a

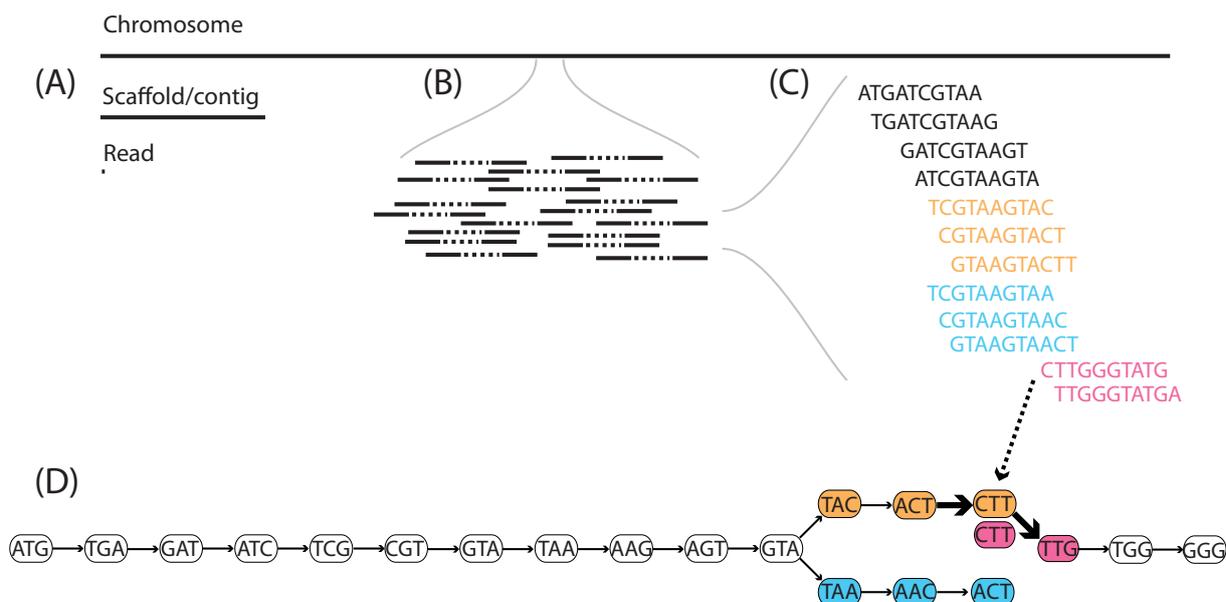


Figure 5: Short read assembly methods rely on the construction of de Bruijn graphs. A: The principle challenge is to construct chromosome size structures in the range of tens of Megabases from reads which were first 36bp and are now usually 100 - 150 bp long. B: To bridge at least short repetitive regions, DNA is sequenced in from both ends of fragments between 180 and 700bp long and paired information of reads belonging to one fragment is retained throughout the assembly process. This is called paired end sequencing. C: Substrings, called kmers, (here size 10, but usually up to 2/3 read length) are computed from reads and combined in a way that only the last base of each kmer is different. D: Sequencing errors or polymorphisms lead the graph to fork and form bubbles. If not resolved contigs will break at this point. Following higher kmer coverage (here $k=3$) the graph can, however, be resolved and the contig continued. Partly redrawn from Schatz et al. (2010)

given species. To predict genes in de-novo sequenced genomes, specialised software was developed (e.g. Augustus Stanke and Waack, 2003 or the MAKER pipeline Cantarel et al., 2008). These programs can make use of external evidence like RNASeq data or protein-to-DNA alignments from closely related species. This method considerably improves gene detection. Establishing an assembly process from reads to annotated genomes was key for two of the manuscripts included in this thesis and thus a major achievement of my work. Therefore, a more detailed description is provided in figure 7 on page 18 and its description in the **Excursus 1** on page 18.

The current sequencing assays, however, hinge on the availability of enough DNA, which especially in tiny organisms like nematodes with a small number of cells and thus a small per individual DNA amount, can be a considerable challenge. Species that can be cultured in the laboratory can usually be grown to massive population sizes to yield enough DNA. But obtaining large enough numbers from wild isolates that cannot be readily cultured is problematic. In addition, even when it is possible to collect many specimens from their natural habitat, the genetic diversity (heterozygosity) inherent to natural populations is a huge problem. The best assemblies are generated from near isogenic (inbred) lines. As in Nematoda, morphological differences between closely related species are often minute (see introduction of Nematoda on page 5 and **P3** and **M1** and the corresponding discussions

on pp. 148, 149), complicating their correct identification, while their genomes are much more diverse. This poses a further obstacle for sequencing when specimens have to be collected from the wild. For example for the manuscript, which will be briefly discussed in the **Outlook** section, we tried to sequence and assemble the genome of a wild isolate of the marine species *Enoplus brevis*. Assembling the genome, however, yielded poor results, presumably because two closely related species had been inadvertently included into the sequencing assay. For such problematic cases, sequencing the transcriptome with Illumina RNASeq technology is a valuable and reliable method to get a good representation of at least those genes that are actively expressed, see (Mortazavi et al., 2008). By sequencing mRNA, coding for exons, which are more conserved, the diversity in populations or even closely related species can be buffered against.

1.7.2 Orthology

In all manuscripts included in this thesis, homologous proteins are searched for in divergent species. However, the division of homologous proteins into sub-classes based on their evolutionary descent and relationship is neither trivial nor are these categories universally defined. For the purpose of this work the following definitions based on Koonin (2005) will be used:

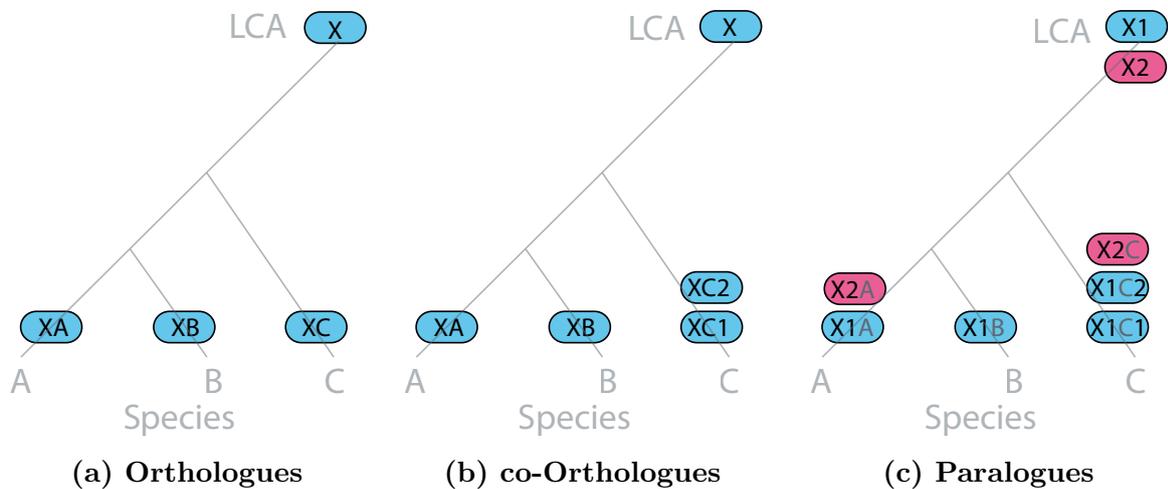


Figure 6: Different classes of homology modified from Koonin (2005). LCA stands for last common ancestor. See the following text for a detailed description.

- Two proteins which are related through a speciation event, i.e. the ancestral species had one protein and each of the derived species retained one copy of this protein, are called **orthologues**. This kind of relationship can be seen in figure 6a. For the ancestral protein X, present in the last common ancestor (LCA) of the derived species A, B, and C, one copy is retained in each of these species. Thus the proteins XA, XB, and XC are orthologues.

- **Co-orthologues** are proteins which diverged after a speciation event by duplication. This implicates that, as above, the ancestral species had one copy, while each of the descendent species might have one or several copies. Multiple copies within one species are then called in-paralogs. Figure 6b shows this. For the protein X in the LCA one copy is retained in species A and B, but species C has a mutation, a duplication of the protein. The proteins XC1 and XC2 are called co-orthologues to XA and XB, while at the same time they are **in-paralogues** of each other.
- Finally **out-paralogues** are proteins which were already present in multiple copies before the speciation event. Thus the ancestral species might have had two (divergent) copies of a given protein (in-paralogues), while the derived species retained either both proteins or independently lost one or the other. As depicted in figure 6c on the preceding page the ancient protein X mutated into duplication already in the LCA of species A, B, and C, creating X1 and X2. While X1A, X1B, and X1C are orthologues in the respective species (species C having a co-orthologue/in-paralogue to X1), the fate of the out-paralogue to X1, namely X2 is different. The gene was lost in species B.

While it is of course genes that duplicate and segregate into offspring, what is usually compared are the proteins¹. Among the given classes of homology, orthologous proteins are of interest for research on the evolution of development. Although the definition of orthology just given is a strictly phylogenetic classification, orthologues are often assumed to retain a similar function even across large evolutionary distances (Koonin, 2005). However, following Ohno's ground-breaking theory, it is expected that after a duplication leading to in-paralogues one copy can freely evolve and possibly gain new functions (Ohno, 1970). As he put it in this book: "Only when a redundant gene locus is created by duplication is it allowed to accumulate formerly forbidden mutations and emerge as a new gene locus with hitherto unknown function." We therefore typically compare orthologues that have been found acting in a given model organism. Conversely, the discovery of paralogues might lead to the discovery of novelty.

Several programs have been developed to conduct the non-trivial task of finding orthologues from genomic data. From the available programs based on BLAST searches OrthoMCL has been shown to be most reliable in finding orthologues that might share biological function (Chen et al., 2007). However, OrthoMCL is also very strict in clustering families of orthologues and might thus miss some connections between more distantly related species. The Orthoinspector is able to uncover such spurious relationships and it has the additional benefit of finding more in-paralogous relationships (Linard et al., 2011), which are important for biological questions concerning evolution via gene duplication and family

¹Due to the redundancy of the genetic code proteins are more similar and thus more easily compared than genes. In addition, by comparing genes one would have to take into account the possibility of divergent splice forms of a single gene.

expansions. However, Orthoinspector also is more lenient (including more distantly related proteins from divergent species), which means that false positives might be included. Such false positives can then only be identified by constructing alignments and building phylogenetic trees. In the manuscripts included into this thesis OrthoMCL has been used as the main tool to find orthologues possibly retaining function, while in one paper (**M1**) Orthoinspector was used to identify additional, divergent proteins.

Excursus 1: Genome assembly pipeline

Figure 7 depicts the general process implemented to assemble the *Panagrolaimus* and *Propanagrolaimus* genomes described in **M1**. The pipeline depicted is more elaborate than the one used for **P2**, which is detailed in the publication included in this thesis. More details are given in the **M1** manuscript. In short, the process is as follows: In a first step, obtained raw read libraries are checked for their quality and cleaned from residual adapter sequences. Low quality reads are discarded. Next, preliminary data to estimate parameters for the assembly (e.g. obtained read insert size, optimal kmer sizes) are evaluated using the preqc (Simpson, 2014) pipeline^a. The reads are then used to construct a preliminary assembly without regarding their paired-end nature (see figure 5 on page 14 for paired-end sequencing).

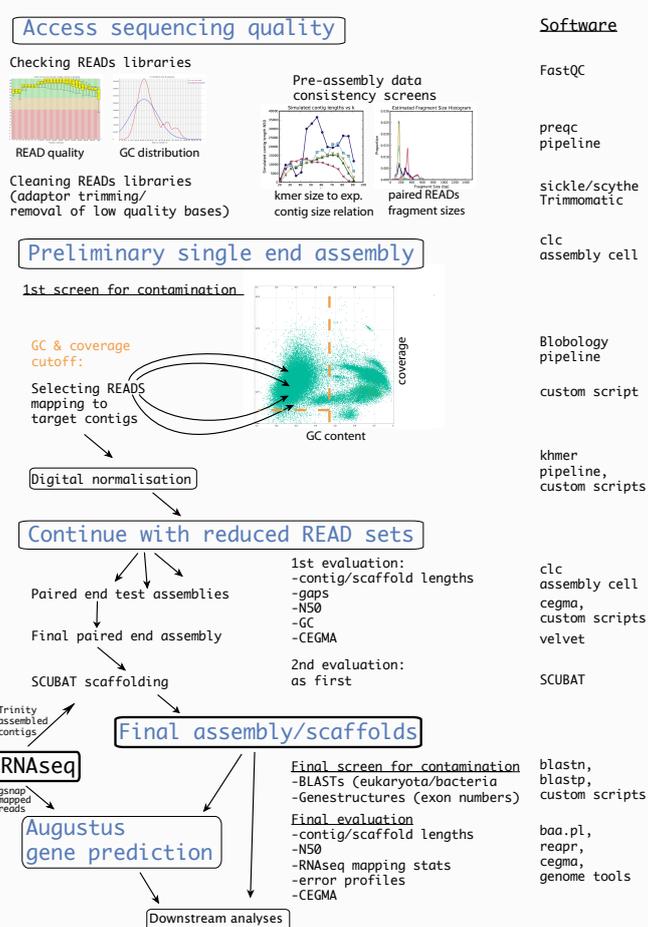


Figure 7: An illustration of the novel genome assembly pipeline. A larger version of the figure is included in the manuscript describing the M1 genome project.

This is to first assemble a gap-free genome to test for contaminations from microorganisms using the Blobology pipeline (Kumar et al., 2013). We found that despite extensive washing steps (including antibiotics treatment and starving of the worms for several days) bacterial DNA will almost always be co-sequenced. Using Blobology, a subset of contigs is blasted against the NCBI database to query. All contigs are then visualised plotting GC vs coverage, as these measures have been found to very often discriminate between target genome and contamination, and contigs with a BLAST hit are coloured (each dot in the plot in the figure is one contig, the coloured ones are so rare that they are hardly visible). Based on this information, a threshold can be defined for which contigs are most likely bacterial. These are discarded from the assembly. Next all reads are mapped against the remaining contigs that are likely to be from the target species genome. All mapping reads are extracted from the original sets and subsequently used to build further draft assemblies. Unfortunately, in 2nd generation sequencing, not all regions of each genome are evenly captured in the library preparation and sequencing process. Thus, in each obtained

Excursus 1 (Cont.)

sequencing library some genomic regions are over-represented, while others have much lower coverage (as can be seen from the blobology plot). High coverage in some and low coverage in other regions poses problems to most assembler programs. Therefore, I used the khmer pipeline in this genome assembly assay which normalises read coverage (it screens for sequences that are overrepresented in the sequencing reads and a fraction of these can then be discarded to gain a level coverage distribution over all regions, e.g. if a sequence is found to have 180X coverage this could be reduced to 30X, to be comparable to the average of all regions). That such a strategy is beneficial for the *Panagrolaimus* genomes could only be empirically tested by conducting repeated re-assemblies with different assemblers (clc assembly cell and velvet). Essentially, each assembly is a scientific experiment. Finally, one assembly is picked and frozen as the working draft. For the *Panagrolaimus genome project*, **M1**, RNASeq data was available which could be used to scaffold the assemblies by combining contigs that contained different regions of one coding sequence reconstructed from the sequenced mRNA using the SCUBAT pipeline (<https://github.com/elswob/SCUBAT>). Augustus (Stanke and Waack, 2003) was used to predict genes from these draft genomes. In this approach I also incorporated the RNASeq data, which was mapped at the read level using gsnap (Wu and Nacu, 2010) to improve the splice site prediction capability of Augustus. Despite the blobology based contamination screening some bacterial contigs were still found in the draft genomes, presumably due to either being close in GC to the respective nematode, or owing to bacterial reads being similar enough to map to the worm contigs. To remove these contigs and the proteins predicted by them, I developed a pipeline based on blast searches of both the contigs and the proteins and checked for exon numbers of the genes on the contigs (assuming that bacterial genes will be largely intron-less), as well as the number of potentially bacterial genes per contigs (i.e. not discarding contigs which had several eukaryotic and only one bacterial gene on them, as these were potential HGT candidates, see Discussion section 3.3 on page 149). In the case of *Panagrolaimus* sp. ES5 very large bacterial contigs with a GC content close to that of the nematode were still in the final genome assembly. These were detected by full genome alignments to several bacterial strains and subsequently removed.

Finally, the cleaned genomes and gene sets could be submitted to downstream analyses, like orthology screening and gene annotation, which are for example detailed in the included manuscript **M1**, as well as in the Discussion of this starting on 149.

^aReferences for used software not given in the text or figure 7 can be found in **M1**.

Chapter 2

Manuscripts

Designators for the manuscripts are given in bold face. These connect the manuscripts to their respective section in the Discussion and are used in the text to refer to the manuscript as well as to the respective discussion section.

2.1 Published manuscripts

1. 959 Nematode Genomes: a semantic wiki for coordinating sequencing projects (**P1** discussed on page 165)
2. The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda (**P2** discussed on page 145)
3. Developmental variations among Panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit (**P3** discussed on page 148)

Here publications are included, followed by the extended manuscripts on page 49. **The Discussion will start on page 144.**

959 Nematode Genomes: a semantic wiki for coordinating sequencing projects

Sujai Kumar^{1,*}, Philipp H. Schiffer² and Mark Blaxter^{1,*}

¹Institute of Evolutionary Biology, The University of Edinburgh, Edinburgh EH9 3JT, UK and ²Zoological Institute, Biocenter Cologne, Zuelpicher Strasse 47b, University of Cologne, 50674 Cologne, Germany

Received August 23, 2011; Accepted September 16, 2011

ABSTRACT

Genome sequencing has been democratized by second-generation technologies, and even small labs can sequence metazoan genomes now. In this article, we describe ‘959 Nematode Genomes’—a community-curated semantic wiki to coordinate the sequencing efforts of individual labs to collectively sequence 959 genomes spanning the phylum *Nematoda*. The main goal of the wiki is to track sequencing projects that have been proposed, are in progress, or have been completed. Wiki pages for species and strains are linked to pages for people and organizations, using machine- and human-readable metadata that users can query to see the status of their favourite worm. The site is based on the same platform that runs Wikipedia, with semantic extensions that allow the underlying taxonomy and data storage models to be maintained and updated with ease compared with a conventional database-driven web site. The wiki also provides a way to track and share preliminary data if those data are not polished enough to be submitted to the official sequence repositories. In just over a year, this wiki has already fostered new international collaborations and attracted newcomers to the enthusiastic community of nematode genomicists. www.nematodegenomes.org.

INTRODUCTION

The nematode *Caenorhabditis elegans* was the first animal to have its genome completely sequenced in 1998 (1). Since then, second-generation sequencing technologies have revolutionized and democratized the field of genome sequencing. Even small labs can now sequence their favourite nematodes in a few weeks for a few thousand dollars.

By 2012, we anticipate that more than 100 nematode genomes will be sequenced, a happy state of affairs for those of us who study this most abundant and diverse Metazoan phylum.

The only problem with rapid and inexpensive sequencing is that it is becoming harder to keep track of which genomes are being sequenced, who is sequencing them, what stage the genome projects are at, and where one can get early access to the data. The nucleotide sequence archives (GenBank/EMBL/DDBJ) (2) are the *de facto* storehouses for complete and published genomes. However, as the bottleneck of a genome project has shifted from sequencing to analysis, which can take months, it has become imperative to have a place to share information about the project before it is published. Inspired by ArthropodBase (www.arthropodgenomes.org), the 959 Nematode Genomes (959NG) wiki was created in early 2010 to meet this need and can be accessed at www.nematodegenomes.org.

959NG is unlike existing genome and transcriptome database web sites such as WormBase (3) and NemBase (4) because, instead of storing the relationships between genes, proteins and DNA sequences, it stores the relationships between people, institutions and sequencing projects at various stages of completion. The goal is to connect users, and make it easy for them to form collaborations and share data. The platform choice reflects this goal as we describe in the ‘Software’ section.

Why (Only) 959NG?

Unlike the 1000 Human Genomes (www.1000genomes.org) or Genome 10 K (genome10k.soe.ucsc.edu) sequencing projects, the effort to sequence as many nematodes as possible is a distributed, bottom-up enterprise. We picked 959 as an initial target because all adult female hermaphrodite *C. elegans* have exactly 959 somatic cells. The definition of the embryonic lineage of *C. elegans* from

*To whom correspondence should be addressed. Tel: +44 131 650 6761; Fax: +44 131 650 5455; Email: mark.blaxter@ed.ac.uk
Correspondence may also be addressed to Sujai Kumar. Tel: +44 131 650 7403; Fax: +44 131 650 5455; Email: sujai.kumar@ed.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

fertilized zygote to fertile adult was a milestone in *C. elegans* developmental biology. Just as the tree of the *C. elegans* embryonic lineage was a key underpinning of later work on this model nematode, we hope that a nematode phylogeny with 959 genome-sequenced taxa will underpin the investigation of nematode biology in general. Obviously, we do not limit the vision to these few genomes: with 23 000 species described, and an estimated 1–2 million species undescribed, the scope for genomic exploration of *Nematoda* is vast.

FEATURES

959NG is a wiki and thus very easy for end-users to edit and interact with. As it is based on the Semantic MediaWiki (SMW) platform, it also allows pages to store properties and relationships to other pages. These properties and relationships can be queried by anyone.

Editable Taxonomy

We offer a view of the taxonomy of the phylum *Nematoda*, pre-loaded with all species that have data present in EMBL/GenBank/DDBJ. Clicking on any node in the taxonomic tree of nematodes shows the sequencing status of all species below that taxon. Each node also provides links to the NCBI page for that taxon and the Expressed Sequence Tags (ESTs) available for any species within that taxon (Figure 1). The initial tree was populated using the NCBI taxonomy (www.ncbi.nlm.nih.gov/taxonomy) but the more widely used Blaxter clades (5) and Helder clades (6) were easy to incorporate into the tree because of the SMW architecture. Users can add new species. See the ‘Software’ section for more details.

Species and Strain Information

For each species, several pieces of information are stored and displayed, such as a short description, its NCBI taxonomic identifier, a picture, as well as some facts about genome size and nucleotide frequency, if known. Species pages also store names of people interested in that species. Each species can have one or more strains with a genome and transcriptome sequencing status that includes links to the funding bodies and the sequencing centres contributing to the sequencing projects (Figure 2).

All page properties are stored internally as Resource Description Framework (RDF) triples which are expressions with three parts: subject, predicate and object. An example of an RDF triple is ‘*Brugia malayi* TRS: Strain genome status: Published’. Although some properties are integer or text values, other properties define relationships to pages, such as ‘*Trichinella spiralis*: Has interested party: Makedonka Mitreva’ which links to a person page.

Persons and Organizations

Because the main goal of 959NG is to connect users, people and organization pages are as important as species pages. These pages store personal and institutional URLs, contact information as well as relationships to the species

such as ‘is genome contact for’ and ‘is interested in species’.

Queries

SMW sites allow users to add new properties that the original web site creators may not have thought of. These properties and relationships can be queried to generate useful dynamic tables. Using the species, strain, people and organization properties, any user can create queries to collate and display information. The following queries are already implemented and linked to from the home page as potentially useful starting points:

- species with published genomes;
- species with genomes being sequenced; and
- species for which sequencing has been proposed.

In addition, clicking on a node in the taxonomic tree displays the result of the query ‘Species under this taxon that have their sequencing status set to anything other than “None”’ (Figure 3).

New queries and information mash-ups can be added by users on any page if they know the SMW query syntax. For example, the following queries are trivial to run from the ‘Semantic Search’ page:

- List of strains sequenced by the funding body NIH: `[[Strain_genome_funder::NIH]]`
- Species in Blaxter clade III with Adenine-Thymine content greater than 70%: `[[Category:Species]]`
`[[Species_genome_at::>70]]`
`[[Species_bclade::Bclade_III]]`

All the pages and the relationships in 959NG can also be exported in XML and RDF format, respectively, using the Special:Export and Special:ExportRDF sections of the web site.

Blast Server For Genomes in Progress

One of the most used features of 959NG is the BLAST (7) server for intermediate genome assemblies. Although generating sequence data is no longer the bottleneck in a sequencing project, quality checks, assembly, annotation and analysis of the data can take several months. The 959NG BLAST server provides a place to park intermediate data so that interested researchers can start looking for their genes or features of interest and speed up the process of research, especially in time-critical areas such as drug-target and vaccine-candidate discovery. Completed genomes will be submitted to centralized repositories (GenBank/EMBL/DDBJ) and to specialized databases such as WormBase, at which point the intermediate assemblies can be removed from the 959NG BLAST server.

SOFTWARE

SMW (semantic-mediawiki.org) is an extension to the popular MediaWiki (mediawiki.org) platform that powers Wikipedia. We chose it for the 959NG web site

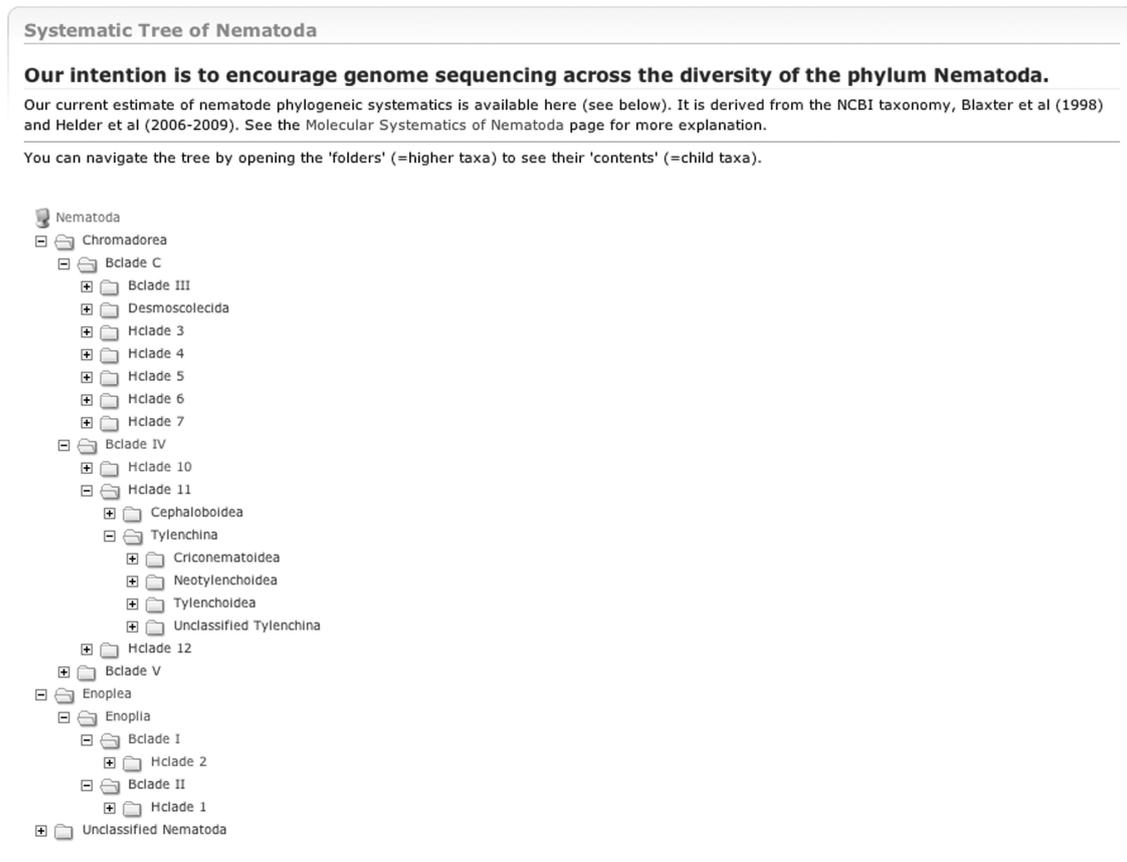


Figure 1. Systematic tree of *Nematoda*, with a few taxonomic nodes expanded to show how the Blaxter and Helder classifications were incorporated into the tree.

because (i) users are familiar with wikis and comfortable with creating and editing pages and (ii) we were not sure at the outset about the information we wanted to capture for each species and its genome sequencing status. As we show in this section, SMWs are better than traditional databases when the data model may change.

SMW Concepts

The initial setup requires an understanding of the following SMW concepts:

- **Categories:** all pages on the site are in one of the following Categories: (i) Genome Sequencing Centre, (ii) Person, (iii), Species, (iv) Strain and (v) Taxon. A category would typically correspond to a table in a relational database.
- **Forms:** each category normally has a specialized form to enter information for that type of page. For example, a Taxon form will have fields for 'NCBI taxon id' and 'Taxon parent', which are specific to Taxon pages.
- **Pages and Properties:** a page is analogous to an object in a database or a row in a database table. Page

properties in SMW are conceptually equivalent to object values or to columns in a database table.

- **Templates:** templates display information about a page or a property. Each category will usually have a template that determines how the information for those types of pages should be displayed. Templates also transform values into displays. For example, the 'PubmedID Linkout' template takes a PubMed ID such as 20980554 and displays a URL to that article on PubMed.

Advantages of SMW

Traditional database-driven web sites have fixed data models that are defined by the developers, and end-users typically only add data within the existing framework to such web sites. One of the main advantages of our SMW site is that, as sequencing technologies and needs change, even end-users can change the types of data stored for each entity (species, person, organization, etc.). For example, when we started the web site in early 2010, we did not have strain-specific pages because only one strain was sequenced per species. However, with sequencing becoming more accessible, different strains are now

4 *Nucleic Acids Research*, 2011

Caenorhabditis elegans

Strain Genome Sequencing:

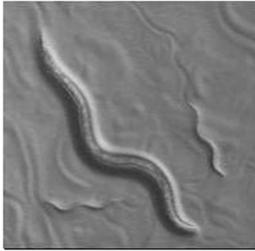
Strain	Status	Contact	Funder	Institution	Sequencing Centre	Plan/Status	URL	Reference Strain
Caenorhabditis elegans N2	published		Wellcome Trust NIH		WTSI Pathogen Genomics The Genome Centre at Washington University	The genome was sequenced from a combination of mapped cosmids and fosmids, mapped YACs and a few long-range PCR products, using Sanger dideoxy technology. The genome is essentially complete (telomere to telomere).	http://www.wormbase.org/	X

Strain Transcriptome Sequencing:

Strain	Status	Contact	Funder	Institution	Sequencing Centre	Plan/Status	URL	Reference Strain
Caenorhabditis elegans N2	published						http://www.wormbase.org/	X

To add/edit information about sequencing strains of the species **Caenorhabditis elegans**, use the following form with the full strain name, e.g., **Caenorhabditis elegans ABC123**:

Species: Caenorhabditis elegans



Parent taxon: Caenorhabditis

NCBI Taxonomy Page: [NCBI txid:6239](#)

Description: free-living bacteriovore

BClade: Bclade V

Interested people: John Sulston, Sydney Brenner

Genome size: 100.2 Mb

Genome size source: Sulston and Brenner PMID: 4858229

Genome AT%: 64 %

Genome AT% source: genome sequence

Genome Publications: PMID:1538779 [PMID:9851916](#)

Transcriptome Publications: PMID:1302005 [PMID:1302004](#) [PMID:8650370](#)

Figure 2. Species page for *C. elegans* displaying information for the species as well as the status of strains that have been sequenced.

being sequenced for the same species, so we used the web interface to add a new ‘Strain’ category, created a new template and a new form for strains, and thus changed the fundamental data model without once touching a database table.

The taxonomy tree is another example of how an end-user can change the data hierarchy without knowing anything about how the back-end is implemented. On our site, each taxon is a wiki page with a ‘Taxon parent’

property pointing to another taxon page, and the tree is generated dynamically based on this single property. Therefore, all we had to do to include additional sub-classifications such as Blaxter clades and Helder clades was to edit a few high-level taxon pages so that their ‘Taxon parent’ properties pointed to a new Blaxter or Helder clade page.

Another advantage of SMW is that it sits on the MediaWiki platform, which is a mature and scalable

Enoplia

ESTs for NCBI Taxid 33218 [↗](#)
 NCBI Taxonomy Browser: NCBI txid:33218 [↗](#)

Species in Enoplia: 600

Strains that are descendants of this taxon, and have their transcriptome or genome status set to any value other than **None**.

<input type="checkbox"/>	<input type="checkbox"/> Strain genome status	<input type="checkbox"/> Strain transcriptome status
Enoplus brevis Sylt/wild	ongoing	none
Romanomermis culicivora Ed Platzer	in annotation	ongoing
Romanomermis iyengari Not specified	proposed	none
Tobrillus sp Not specified	proposed	none
Trichinella spiralis Not specified	published	none
Trichuris muris E isolate	ongoing	none

Phylogenetic context:

- Nematoda
 - Chromadorea
 - Enoplea
 - Enoplia**
 - Unclassified Nematoda

Category: Taxon

Figure 3. Page for the taxonomic node *Enoplia*, showing NCBI Taxonomy and NCBI EST link-outs, as well as the results of the query ‘Species and strains under this taxon with their sequencing status set to anything other than “None”’.

engine for serving high-capacity web sites and has a large developer base. Setting up the initial web site took only three person-days, thanks to the examples of templates and forms on another similar site (arthropodgenomes.org). The BioDBCORE description of the wiki is provided in the [Supplementary Data](#) section.

FUTURE DIRECTIONS

As more genomes are sequenced and the 959NG site grows, we hope that the evolving data model for nematode genome sequencing projects will also inform other genome sequencing efforts. Most genomes these days are not finished, but are published as high-quality draft sequences, so we will need to not only store the Minimum Information about a Genome Sequence (gensc.org) and Minimum Information about a high-throughput Sequencing Experiment (www.mged.org/minseqe), but also additional values such as CEGMA scores (8) to measure how complete the genome is. We will also develop descriptors for genome-scale genetic mapping data, derived from technologies such as restriction-site-associated DNA sequencing (RADSeq) (9), genotyping by sequencing (GBS) (10) and other methods (11), across many strains or isolates of a species.

Currently, site visitors can interrogate intermediate draft assemblies of genomes in progress only through the BLAST server. In addition, we would like to provide a basic, automatic annotation service for these

incomplete genomes using RNASeq alignments and gene predictors.

CONCLUSIONS

The 959 Nematode Genomes wiki has already inspired international collaborations to sequence, annotate and interpret the genomes of key species. We know of two cases where groups who did not know of each other’s efforts are now merging expertise and effort in a unified project. As additional genomes are proposed, new collaborations can be forged and cross-species analyses coordinated. We also hope that the existence of the wiki, and the enthusiastic community behind it, will serve to attract new researchers into this field. As nematode genomics moves into population genomics, this register of strains and sources will become ever more useful. SMW technology builds a system that is easy to navigate, easy to edit and, importantly, easy to develop as needs, knowledge and possibilities change.

Genomics research on nematodes (particularly *C. elegans*) has already delivered important information on core biological processes. Adding additional nematode genomes will allow the specific instance of *C. elegans* to be contextualized, and will, we hope, feed research on comparative genomics of nematodes, the evolutionary biology of genome change, and the biology of (many) parasitic nematodes, among other fields. We hope 959NG will become a one-stop site in which to forge collaborations, learn about best practice in assembly and annotation,

6 *Nucleic Acids Research*, 2011

record insights and advances and explore the genomic diversity of *Nematoda*.

ACKNOWLEDGEMENTS

We would like to thank Dan Lawson at EBI for inspiring us with his SMW site ArthropodBase and allowing us to use his templates and forms as a starting point. Dan Bolser at the University of Dundee, Yaron Koren of WikiWorks and the rest of the SMW community were very helpful and patiently answered questions on online forums. The University of Edinburgh provides hosting space for nematodegenomes.org.

FUNDING

This work was supported by the School of Biological Sciences at the University of Edinburgh. Funding for open access charge: Natural Environment Research Council (NERC).

Conflict of interest statement. None declared.

REFERENCES

1. C elegans Genome Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, **282**, 2012–2018.
2. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
3. Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J.W., De La Cruz, N., Davis, P., Duesbury, M., Fang, R. *et al.* (2010) WormBase: A comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
4. Elsworth, B., Wasmuth, J. and Blaxter, M. (2011) NEMBASE4: The nematode transcriptome resource. *Int. J. Parasitol.*, **41**, 881–894.
5. Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
6. Holterman, M., van der Wurff, A., van den Elsen, S., van Megen, H., Bongers, T., Holovachov, O., Bakker, J. and Helder, J. (2006) Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. *Mol. Biol. Evol.*, **23**, 1792–1800.
7. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
8. Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.
9. Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
10. Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T. and Stern, D.L. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.*, **21**, 610–617.
11. Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, **12**, 499–510.

Developmental variations among Panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit

Philipp H. Schiffer · Ndifon A. Nsah ·
Henny Grotehusmann · Michael Kroiher ·
Curtis Loer · Einhard Schierenberg

Received: 14 February 2014 / Accepted: 29 April 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract Comparative studies of nematode embryogenesis among different clades revealed considerable variations. However, to what extent developmental differences exist between closely related species has mostly remained nebulous. Here, we explore the correlation between phylogenetic neighborhood and developmental variation in a restricted and morphologically particularly uniform taxonomic group (Panagrolaimidae) to determine to what extent (1) morphological and developmental characters go along with molecular data and thus can serve as diagnostic tools for the definition of kinship and (2) developmental system drift (DSD; modifications of developmental patterns without corresponding morphological changes) can be found within a small taxonomic unit. Our molecular approaches firmly support subdivision of Panagrolaimid nematodes into two monophyletic groups. These can be discrimi-

nated by distinct peculiarities in early embryonic cell lineages and a mirror-image expression pattern of the gene *skn-1*. This suggests major changes in the logic of cell specification and the action of DSD in the studied representatives of the two neighboring nematode taxa.

Keywords Nematoda · Molecular phylogeny · Cell lineage · In situ hybridization · *skn-1* · Developmental system drift

Introduction

Nematodes exhibit a highly conserved body plan. In contrast, embryogenesis varies considerably among distinct branches of the widely ramified phylum (see, e.g., Schulze and Schierenberg 2011). This phenomenon, called developmental system drift (True and Haag 2001), indicates changes in the employment of underlying gene regulatory networks (GRNs). Nematodes appear particularly suitable models to study the linkage between evolution and development (Sommer and Bumbarger 2012). They can be subdivided into 12 different clades (Holterman et al. 2006). The model system *Caenorhabditis elegans* belongs to clade 9, while Panagrolaimidae belong to clade 10. The latter possess few diagnostic anatomical characters useful for detailed taxonomy, and phylogenies using small subunit (SSU) rDNA have left this taxon more dubious than its sister groups. The aim of the present report is to explore as a case study the possibility that developmental system drift (DSD) acts even within such an evolutionary confined taxonomic unit. For this, structural and developmental characters are related to phylogenetic position among cognate Panagrolaimids.

Communicated by: Volker G. Hartenstein

ORCID: 0000-0001-6776-0934

Electronic supplementary material The online version of this article (doi:10.1007/s00427-014-0471-2) contains supplementary material, which is available to authorized users.

P. H. Schiffer (✉) · N. A. Nsah · H. Grotehusmann · M. Kroiher ·
E. Schierenberg
Zoological Institute, Cologne Biocenter, University of Cologne,
Cologne, Germany
e-mail: philipp.schiffer@gmail.com

P. H. Schiffer
Institute for Genetics, Cologne Biocenter, University of Cologne,
Cologne, Germany

C. Loer
Department of Biology, University of San Diego, San Diego, CA,
USA

Materials and methods

Strains

Strain	Origin	Reproduction	Source
WTM1	Underwater cave; Romania	H	WB
LC91	Littoral river mud; San Diego, CA	H	CL
JU765	Edge of rice paddy; Guangxi, China	H	MAF
PS1159	Soil sample; North Carolina	P	AB
DL137	Soil sample; Corvallis, OR	P	DL
PS1579	Soil sample; Pasadena, CA	P	DL
ES5	Dead blackberry; Bonn, Germany	G	ES
<i>P. detritophagus</i> (BSS8)	Iceland	H	AB
<i>A. nanus</i> (ES501)	Soil sample; Peru	P	ES

G, gonochoristic; *H* hermaphroditic; *P* parthenogenetic
AB Ann Burnell, National University of Ireland, Maynooth, Ireland

CL Curtis Loer, University of San Diego, USA

DL Dee Denver, University of Oregon, USA

ES Einhard Schierenberg, University of Cologne, Germany

MAF Marie-Anne Felix, University Paris-Diderot, France

WB Walter Traunspurger, University of Bielefeld, Germany

Measurements

Measurements were performed under a Zeiss Imager microscope using “ImageJ” software. For each strain, at least 10 adult individuals were examined.

Molecular phylogeny

To robustly infer the phylogenetic relationships, we used seven proteins (Supplementary Table 1) predicted from RNAseq data with the CEGMA pipeline (Parra et al. 2007). Additionally, large fragments of the *Propanagrolaimus* SSU and large subunit (LSU) genes were amplified with high-fidelity polymerase. PCR-amplified DNA sequences have been deposited in the NCBI database under the Accession numbers KJ434174–KJ434177. Protein sequences predicted from RNAseq data using CEGMA are publicly available on figshare (<http://dx.doi.org/10.6084/m9.figshare.980719>; complete RNAseq data to be published elsewhere, registered under ENA PRJEB5767).

Data for outgroups were downloaded from GenBank and WormBase (www.wormbase.org) and kindly provided by P.

Sternberg, CalTech, Pasadena and I. Yanai, Technion, Haifa. Sequences were aligned with Clustal Omega v.1.2 (Sievers et al. 2011) and visually controlled. Evolutionary models for the seven protein alignment were explored with ProtTest (v.3.2) (Darriba et al. 2011). Phylogenies were inferred with MrBayes (v.3.2.1) (Ronquist and Huelsenbeck 2003) and RAxML (v.7.2.6) (Stamatakis 2006) on the local CHEOPS computer cluster. For the seven protein alignment, MrBayes was run for 5,000,000 generations (discarding 10 % burnin) in four runs on four chains allowing the program to apply a mixed model of evolution optimizing for the gamma parameter. The ribosomal alignments were run under the GTR model for 1,000,000 generation in otherwise similar settings. The GTR model was specified for RAxML runs, and an automated bootstrapping algorithm implemented in the program was used. Trees were visualized with FigTree v.1.4.0 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Cell lineage analysis

Cell lineage studies were performed using the SIMI Biocell software (Unterschleißheim, Germany) after 4-D recording essentially as described in Schulze and Schierenberg (2011). From each strain/species, at least three recordings were made.

In situ hybridization

We isolated the *Panagrolaimus skn-1* sequences first by RACE PCR from cDNA and later reconfirmed these by screening our assembled RNAseq data with BLAST+. Corresponding coding and protein sequences predicted from RNAseq data are available on figshare (<http://dx.doi.org/10.6084/m9.figshare.980718>). Reciprocal best-hit BLAST+ searches on the RNAseq data were also employed to screen for genes acting downstream of *skn-1* in *C. elegans*.

Digoxigenin-labeled sense and antisense RNA probes were generated from linearized pBluescript vectors (Stratagene, La Jolla, CA, USA) containing a fragment of the *Panagrolaimus* or *Cephalobid skn-1* homolog via run off in vitro transcription with T7 or T3 RNA-polymerase (Roche, Mannheim, Germany) according to the manufacturer's protocol. Homologs of *skn-1* were identified from total genome sequencing or RNAseq data (our unpublished data), amplified by PCR, cloned into pBs vectors, and finally verified by BigDye terminator cycle sequencing (PE Biosystems). For visualization of gene expression in embryos, we followed essentially the protocol for *C. elegans* as given in Broitman-Maduro and Maduro (2011). However, to overcome problems with egg-shell penetration in our strains, embryos were pre-treated with alkaline bleach solution (4.25 % sodium hypochlorite, 0.75 M KOH for 2 min) followed by freeze-cracking.

Results and discussion

Phylogenetic evaluation

Lewis et al. (2009) subdivided the genus *Panagrolaimus* into two groups (PI and PII). Their analysis, however, had limited resolution as only few short-gene fragments were sequenced. To better resolve the phylogenetic relationships of species whose development was studied here, we sampled additional molecular data. Our extended analysis, based on seven genes (nearly 9,500 amino acids) predicted from RNAseq data, robustly confirms the split between the Panagrolaimid groups. We find that the PII group occupies an intermediate position between representatives of the *Panagrolaimus* PI group, and the recently sequenced *Panagrellus redivivus* and with respect to *Strongyloides ratti* and *Acrobelloides nanus* used as outgroups (see “Materials and methods”). All applied methods placed the PII species JU765 as an outgroup to *P. redivivus* and the PI species (Fig. 1).

Based on morphological criteria, Andr ssy (2005) defined a sister genus to *Panagrolaimus* named *Propanagrolaimus*, characterized by a particularly slender body shape, barely separated lips, non-protruded vulva, an extended tapering tail, and in addition, a limnic habitat. Based on these criteria, we classified two novel isolates LC91 and WTM1 as *Propanagrolaimus* sensu Andr ssy, both showing a ratio of body length:width (rlw) >40. JU765 shares morphological similarities (tail, vulva, lips) with WTM1 and LC91 and likewise shows differences to ES5 and PS1159 (rlw ca. 18.5). See Supplementary Fig. 1 for morphology of studied species in comparison to *C. elegans*.

However, not all morphological and behavioral variations follow the suggested subdivision into *Panagrolaimus* and *Propanagrolaimus*. With respect to slenderness, JU765 occupies an intermediate position (rlw ca. 25). The position of the vulva varies considerably between JU765 (55 % body length) and the two slender isolates (ca. 72 %) but not

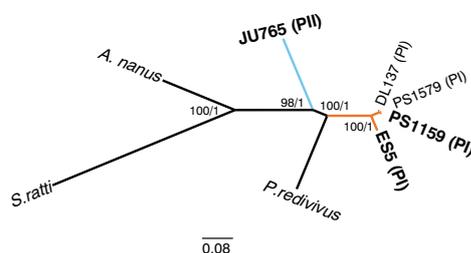


Fig. 1 Phylogenetic revision of *Panagrolaimus* groups PI and PII sensu Lewis et al. (2009). Based on the analysis of seven CEGMA predicted proteins comprising nearly 9,500 residues based on maximum likelihood (RAxML) and Bayesian inference (MrBayes), *P. redivivus* is placed between the two *Panagrolaimus* groups. This phylogeny is well supported by bootstrapping and posterior probabilities (values indicated at nodes in this order). Species analyzed for early development are shown in **bold**

significantly within individual cultures. LC91 and WTM1 are ovoviviparous, while all other strains lay early-stage eggs. The particularly slender body shape and ovovivipary of *Propanagrolaimus* LC91 and WTM1 absent in *Propanagrolaimus* JU765 may be explained with their adaptation to a limnic lifestyle (Andr ssy 2005).

In the absence of second-generation sequencing data for the two bona fide *Propanagrolaimus* isolates, we sequenced SSU and LSU ribosomal genes to elucidate the phylogenetic position of PII relative to *Propanagrolaimus*. These data strongly support a monophyletic group that includes the PII members JU765 and *Panagrolaimus detritophagus* as well as WTM1 and LC91 (Fig. 2).

Thus, our morphological and molecular data are compatible with a subdivision of the studied species into two separate sister taxa, *Panagrolaimus* and *Propanagrolaimus*, as suggested by Andr ssy. The largely uniform morphology as well as short branch lengths in the phylogenetic analysis suggest that these closely related representatives diverged rather recently making it attractive to look there for DSD and changes in GRNs.

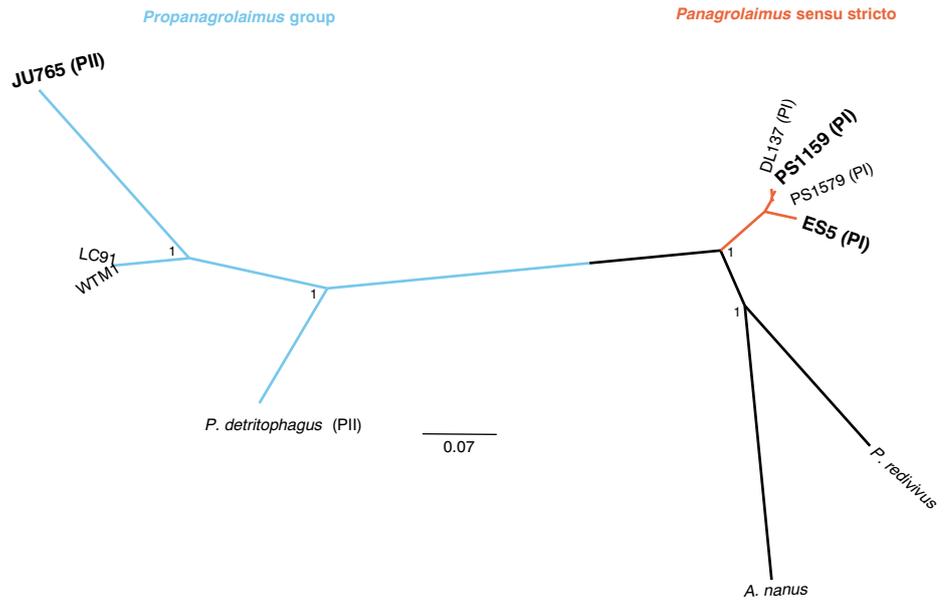
Nematode embryogenesis and phylogenetic position: cell lineage and gene expression pattern

We showed previously that the order of embryonic cleavages follows taxon-specific patterns, particularly concerning the timing of germ cell divisions (Schulze and Schierenberg 2011). As our sequence data place the *Propanagrolaimus* between *Panagrolaimus* (clade 10) and *Acrobelloides* (clade 11; Fig. 2), we analyzed early cell lineages of five Panagrolaimid strains plus *Acrobelloides nanus* and compared these to the reference *C. elegans*. The variations in the division order were found to be essentially restricted to the germline cells P₃ and P₄; nevertheless, they allowed a clear distinction between the *Panagrolaimus* (PI group) and *Propanagrolaimus* (PII group; Table 1). These differences may reflect variations in the availability of maternal gene products as suggested for other nematodes (Laugsch and Schierenberg 2004).

In *C. elegans*, the gene *skn-1* encodes a transcription factor that is the key switch for proper specification of the endomesoderm founder cell, the EMS blastomere (Bowerman et al. 1992). We searched and cloned this gene in Panagrolaimid and Cephalobid nematodes. Reciprocal best BLAST+ hits of the *C. elegans* protein to the isolated and predicted sequences and of these sequences to *Caenorhabditis skn-1* in the NCBI database confirmed orthology precluding the inadvertent use of a paralogue (e.g., *C. elegans sknr-1*) in our in situ analyses (see Supplementary Fig. 2).

PS1159 (PI) and JU765 (PII) SKN-1 orthologs are similar and share ~34 % overall identity with the *C. elegans* protein. The functionally crucial DNA-binding domain is highly

Fig. 2 Bayesian phylogeny of the 18S ribosomal gene (~1,700 bp). The inferred phylogeny places the PII species JU765 and *P. detritophagus* in a monophyletic group with LC91 and WTM1. A similar tree topology is found in the LSU tree (not shown). Posterior probabilities are shown at important nodes; species analyzed for developmental patterns are shown in *bold* type



conserved (>60 % identity), while the DIDLID region important for transcriptional activation shows a lower similarity to *C. elegans* (see Supplementary Fig. 2 for alignment). Localization of *skn-1* messenger RNA (mRNA) in *Panagrolaimus* sp. PS1159 (Fig. 3a–c)

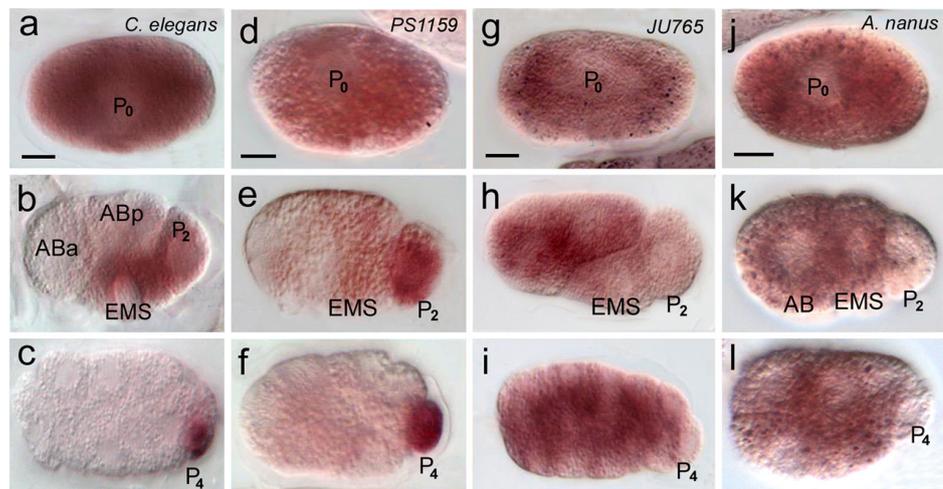
resembles that in *C. elegans*. In both species, *skn-1* mRNA segregates into the germline. In striking contrast, JU765 *skn-1* mRNA accumulates in somatic cells (Fig. 3d–f) like in *A. nanus* (Fig. 3g–i) and several other *Acrobeloides* species (N.A.N., unpublished data).

Table 1 Sequence of early embryonic cleavages

<i>C. elegans</i>	<i>Pana.</i> ES5	<i>Pana.</i> PS1159	<i>Propana.</i> JU765	<i>Propana.</i> WTM1	<i>Propana.</i> LC91	<i>A. nanus</i>
P₀	P₀	P₀	P₀	P₀	P₀	P₀
AB	P₁	P₁	P₁	P₁	P₁	P₁
P₁	AB	AB	AB	AB	AB	P₂
² AB ⁴	P₂	P₂	P₂	P₂	P₂	AB
EMS	² AB ⁴	² AB ⁴	² AB ⁴	² AB ⁴	² AB ⁴	P₃
P₂	EMS	EMS	EMS	EMS	EMS	² AB ⁴
⁴ AB ⁸	P₃	P₃	⁴ AB ⁸	⁴ AB ⁸	⁴ AB ⁸	EMS
MS	⁴ AB ⁸	⁴ AB ⁸	P₃	MS	P₃	⁴ AB ⁸
E	MS	MS	MS	P₃	MS	MS
C	E	E	E	E	E	E
P₃	⁸ AB ¹⁶	C	C	C	C	C
⁸ AB ¹⁶	C	⁸ AB ¹⁶	⁸ AB ¹⁶	⁸ AB ¹⁶	⁸ AB ¹⁶	⁸ AB ¹⁶
² MS ⁴	² MS ⁴	² MS ⁴	² MS ⁴	² MS ⁴	² MS ⁴	² MS ⁴
² C ⁴	P₄	P₄	¹⁶ AB ³²	¹⁶ AB ³²	¹⁶ AB ³²	¹⁶ AB ³²
¹⁶ AB ³²			² C ⁴	² C ⁴	² C ⁴	² C ⁴
² E ⁴			⁴ MS ⁸	⁴ MS ⁸	⁴ MS ⁸	² E ⁴
D			² E ⁴	² E ⁴	² E ⁴	⁴ MS ⁸
⁴ MS ⁸			D	D	P₄	<i>no emb. division of P₄</i>
⁴ C ⁸			P₄	P₄		
³² AB ⁶⁴						
P₄						

Germline cells are in bold. Superscripts indicate cell numbers in this lineage before and after division

Fig. 3 Embryonic localization of *skn-1* mRNA in four nematode strains. **a–c** *C. elegans*, segregation into germline cells; **d–f** *Panagrolaimus* PS 1159, segregation into germline cells; **g–i** *Propanagrolaimus* JU765, segregation into somatic blastomeres; **j–l** *Acrobeloides nanus*, segregation into somatic blastomeres. For each of the depicted staining patterns, at least 10 cases were found in our preparations of >200 embryos/strain of all developmental stages. Sense probes showed no staining at all. Scale bar 10 μ m; orientation anterior, left



This unexpected result could indicate that the central role of the SKN-1 transcription factor in endomesoderm specification (Maduro 2006), in addition to its other multifaceted functions, was established in a phylogenetic branch comprising *C. elegans* and *Panagrolaimus* but not *Propanagrolaimus* and *Acrobeloides* (Figs. 1 and 2). Alternatively, the observed variants could have been established independently in two phylogenetic branches. As even between *C. elegans* and its close relative *C. briggsae*, the gene regulatory network of endomesoderm specification is somewhat differently employed (Lin et al. 2009); it appears possible that this process is particularly prone to rapid and independent evolutionary change. To decide which of the two observed expression patterns is apomorphic, an outgroup comparison with more basal nematodes like *Romanomermis culicivorax* (Schulze and Schierenberg 2011) could be helpful.

The case of *skn-1* in *C. elegans* demonstrates that localization of mRNA does not necessarily indicate the expression domain of the respective protein (see Bowerman et al. (1992, 1993)) Hence, from our current data, we cannot infer the role of SKN-1 protein in the studied Panagrolaimid and Cephalobid species, but the variable RNA distribution nevertheless indicates an evolutionary change.

We screened our data for downstream targets of SKN-1 known from *C. elegans* and found credible homologues of PHA-4 and candidates for the GATA factor ELT-2 in both *Panagrolaimus* and *Propanagrolaimus*. Nevertheless, further studies are needed to determine to what extent the gene regulatory network around *skn-1* is conserved in these species.

The observed variations within a small taxonomic unit (two closely related genera) highlight the evolutionary plasticity of the nematode developmental program. DSD appears to act and re-shuffle cellular interactions rather quickly on an evolutionary timescale, while adult morphology is essentially retained. Similarly, it appears that underlying GRNs in development are evolving fast in Nematoda (Schiffer et al. 2013).

We hypothesize that such changes constitute a driving force for speciation and thus can be one explanation for the species' richness in this phylum. Extending studies with additional representatives, particularly closely related species, should not only further unravel the plasticity of underlying networks but also help to elucidate the labyrinthine pathway that led to the many idiosyncrasies of *C. elegans* development.

Acknowledgments P.H.S. was funded by a personal grant of the Volkswagen Foundation in the framework of the Initiative for Evolutionary Biology and by the German Research Foundation (DFG) through the grant SFB680 to T. Wiehe, Institute for Genetics, University of Cologne.

C. L. was funded through a Fletcher Jones endowment and a USD International Opportunity Grant. The authors are grateful to Walter Traunspurger for sharing the *Propanagrolaimus* strain WTM1. C.L. thanks Terry Bird, Keith MacDonald, and the Biology 342 Microbiology lab students (USD) that first found LC91 in Winogradsky columns made from the San Diego River mud.

Competing interests The authors declare that they have no competing interests.

References

- Andrássy I (2005) Free-living nematodes of Hungary (*Nematoda errantia*), 1st edn. Pedozoologica Hungarica, Hungarian Natural History Museum, Budapest
- Bowerman B, Eaton BA, Priess JR (1992) *skn-1*, a maternally expressed gene required to specify the fate of ventral blastomeres in the early *C. elegans* embryo. *Cell* 68:1061–1075
- Bowerman B, Draper BW, Mello CC, Priess JR (1993) The maternal gene *skn-1* encodes a protein that is distributed unequally in early *C. elegans* embryos. *Cell* 74:443–452
- Broitman-Maduro G, Maduro MF (2011) In situ hybridization of embryos with antisense RNA probes. *Meth Cell Biol* 106:253–270
- Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27: 1164–1165
- Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J (2006) Phylum-wide analysis

- of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. *Mol Biol Evol* 23:1792–1800
- Lausch M, Schierenberg E (2004) Differences in maternal supply and early development of closely related nematode species. *Int J Dev Biol* 48:655–662
- Lewis S, Dyal L, Hilburn C, Weitz S, Liao W, LaMunyon C, Denver D (2009) Molecular evolution in *Panagrolaimus* nematodes: origins of parthenogenesis, hermaphroditism and the Antarctic species *P. davidi*. *BMC Evol Biol* 9:15
- Lin KT-H, Broitman-Maduro G, Hung WWK, Cervantes S, Maduro MF (2009) Knockdown of SKN-1 and the Wnt effector TCF/POP-1 reveals differences in endomesoderm specification in *C. briggsae* as compared with *C. elegans*. *Dev Biol* 325:296–306
- Maduro MF (2006) Endomesoderm specification in *Caenorhabditis elegans* and other nematodes. *BioEssays* 28:1010–1022
- Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067
- Ronquist F, Huelsenbeck J (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574
- Schiffer PH, Kroiher M, Kraus C, Koutsovoulos GD, Kumar S, Camps JIR, Nsah NA, Stappert D, Morris K, Heger P, Altmüller J, Frommolt P, Nürnberg P, Thomas WK, Blaxter ML, Schierenberg E (2013) The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda. *BMC Genomics* 14:923
- Schulze J, Schierenberg E (2011) Evolution of embryonic development in nematodes. *EvoDevo* 2:18
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Ding JSO, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:1–6
- Sommer RJ, Bumbarger DJ (2012) Nematode model systems in evolution and development. *Wiley Interdiscip Rev Dev Biol* 1:389–400
- Stamatakis AV (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690
- True JR, Haag ES (2001) Developmental system drift and flexibility in evolutionary trajectories. *Evol Dev* 3:109–119

RESEARCH ARTICLE

Open Access

The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda

Philipp H Schiffer^{1*}, Michael Kroiher¹, Christopher Kraus¹, Georgios D Koutsovoulos², Sujai Kumar², Julia I R Camps¹, Ndifon A Nsah¹, Dominik Stappert³, Krystalynne Morris⁴, Peter Heger¹, Janine Altmüller⁵, Peter Frommolt⁵, Peter Nürnberg⁵, W Kelley Thomas⁴, Mark L Blaxter² and Einhard Schierenberg¹

Abstract

Background: The genetics of development in the nematode *Caenorhabditis elegans* has been described in exquisite detail. The phylum Nematoda has two classes: Chromadorea (which includes *C. elegans*) and the Enoplea. While the development of many chromadorean species resembles closely that of *C. elegans*, enoplean nematodes show markedly different patterns of early cell division and cell fate assignment. Embryogenesis of the enoplean *Romanomermis culicivorax* has been studied in detail, but the genetic circuitry underpinning development in this species has not been explored.

Results: We generated a draft genome for *R. culicivorax* and compared its gene content with that of *C. elegans*, a second enoplean, the vertebrate parasite *Trichinella spiralis*, and a representative arthropod, *Tribolium castaneum*. This comparison revealed that *R. culicivorax* has retained components of the conserved ecdysozoan developmental gene toolkit lost in *C. elegans*. *T. spiralis* has independently lost even more of this toolkit than has *C. elegans*. However, the *C. elegans* toolkit is not simply depauperate, as many novel genes essential for embryogenesis in *C. elegans* are not found in, or have only extremely divergent homologues in *R. culicivorax* and *T. spiralis*. Our data imply fundamental differences in the genetic programmes not only for early cell specification but also others such as vulva formation and sex determination.

Conclusions: Despite the apparent morphological conservatism, major differences in the molecular logic of development have evolved within the phylum Nematoda. *R. culicivorax* serves as a tractable system to contrast *C. elegans* and understand how divergent genomic and thus regulatory backgrounds nevertheless generate a conserved phenotype. The *R. culicivorax* draft genome will promote use of this species as a research model.

Keywords: Nematode, Genome, Evolution, Development, *Caenorhabditis*, Mermithida, *Romanomermis*

Background

Nematodes have a generally conserved body plan. Their typical form is dictated by the presence of a single-chamber hydroskeleton, where longitudinal muscles act against an inextensible extracellular cuticle. The conservation of organ systems between nematode species is even

more striking, with, for example, the nervous system, the somatic gonad and the vulva having very similar general organisations and cellular morphologies across the phylum. It might be thought that these similarities arise from highly stereotypical developmental programmes, but comparative studies challenge this “all nematodes are equal” view.

Embryonic development of the nematode *Caenorhabditis elegans* has become a paradigmatic model for studying developmental processes in animals, including early soma-germline separation, fate specification including

*Correspondence: philipp.schiffer@googlemail.com

¹Zoologisches Institut, Universität zu Köln, Cologne, NRW, Germany, ORCID:0000-0001-6776-0934

Full list of author information is available at the end of the article

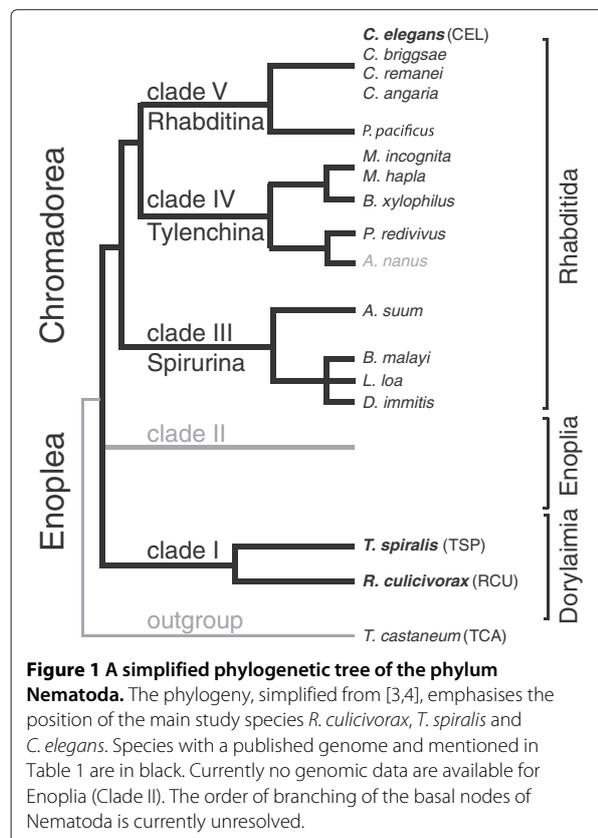
inductive interactions, and tissue-specific differentiation. The particular mode of development of *C. elegans* is distinct within the major metazoan model organisms, but much of the regulatory logic of its development is comparable to that observed in other phyla. One key aspect in which *C. elegans* differs from vertebrate and arthropod models is that *C. elegans* shows a strictly determined development [1], with a largely invariant cell-lineage giving rise to predictable sets of differentiated cells [2]. Inductive cell-cell interactions are, nevertheless, essential for its correct development [1]. *C. elegans* is a rhabditid nematode, one of approximately 23,000 described and 1 million estimated nematode species. Molecular and morphological systematics of the phylum Nematoda identify two classes: Chromadorea (including Rhabditida, and thus *C. elegans*), and Enoplea (subdivided into Dorylaimia and Enoplia) [3,4] (Figure 1). *C. elegans* is a chromadorean, and most investigation of developmental biology of nematodes has been carried out on Chromadorean species. The first description of the early embryonic cell-lineage of a nematode, that of *Ascaris* (Spirurina within Chromadorea) in the 1880's [5,6], conforms to the *C. elegans* model. Early development across all suborders of the Rhabditida is very similar [7,8]. In general, only minor variations of the division pattern observed in *C. elegans*

have been described in these nematodes [9,10], including heterochrony in the timing of cell divisions, and restrictions in cell-cell interaction due to different placement of blastomeres in the developing embryo. From these observations it might be assumed that all nematodes follow a *C. elegans*-like pattern of development. However, deviations from the *C. elegans* pattern observed in other rhabditid nematodes show that the strictly determined mode of development is subject to evolutionary change, making it particularly attractive for the study of underpinning regulatory logic of developmental mechanisms. Indeed, a greater role for regulative interactions in early development has been demonstrated in another rhabditid, *Acrobeloides nanus* (Tylenchina) [11,12].

Regulative development is common among Metazoa, and is also observed in other Ecdysozoa, including Arthropoda. Indeed, in several enoplean species, early embryos have been found to not display polarised early divisions, arguing against a strongly determined mode of development in this group [13,14]. The determined mode found in *C. elegans* is thus likely to be derived even within Nematoda [15], implying that the core developmental system in Nematoda has changed, while maintaining a very similar organismal output. This phenomenon, termed "developmental system drift" [16], reveals independent selection on the developmental mechanism and the final form produced.

To explore the genetics of development of enoplean and other non-rhabditid nematodes requires tractable experimental systems with a suitable set of methodological tools and extensive genomic data. While *C. elegans* and its embryos are relatively easily manipulated and observed, and the *C. elegans* genome has been fully sequenced [17], embryos from the Enoplia and Dorylaimia are much harder to culture and manipulate. Few viable laboratory cultures exist and obtaining large numbers of embryos from wild material is difficult. Functional molecular analysis of most nematodes, in particular Enoplea, is further hindered by the lack of genetic tools such as mutant analysis or gene-knockdown via RNAi. Performing detailed comparative experimental embryology on a phylogenetically representative set of species across the phylum Nematoda thus remains a distant goal.

The genetic toolkit utilised by a species is represented in its genome, and direct assessment of the genetic capabilities of an organism can thus be assessed through analysis of genome data. Using the background knowledge of pathways and modules used in other taxa, the underpinning logic of a species' developmental system can be inferred from its genome, and the developmental toolkits of different species can be compared. These comparisons can reveal changes in developmental logic between taxa by identifying gene losses during evolution that must result in changed pathway functioning, and similarly identify



genes recruited to developmental regulatory roles in particular lineages.

Efficient generation of genomic resources for non-model species, and the inference of developmental regulatory pathways from the encoded gene sets, is now possible. The majority of the fifteen nematode genomes published to date have been from Rhabditida (Figure 1) [18-26]. The single enoplean genome sequences is from the mammalian parasite *Trichinella spiralis* (Dorylaimia; order Trichocephalida) [27]. *T. spiralis* is ovoviviparous, proper development requires intrauterine environment, and early blastomeres are extremely transparent [28] such that individual nuclei are hard to identify (E.S., unpublished observations). Hence, this species is of very limited value for light microscopical image analysis and experimental investigation correlating cell dynamics with the molecular circuitry regulating early development.

Although the genomes of many additional nematode species are being sequenced [29,30], even in this wider sampling of the phylum, Enoplea remains neglected. The enoplean *Romanomermis culicivorax* (order Mermithida within Dorylaimia) has been established in culture for decades. It infects and kills the larvae of many different mosquito species [31], and is being investigated for its potential as a biocontrol agent of malaria and other disease vectors [31,32]. *R. culicivorax* and *T. spiralis* differ fundamentally in many life-cycle and phenotypic characters. *R. culicivorax* reproduces sexually. A single female can produce more than a thousand eggs, and embryos are easily studied under laboratory conditions. They display a developmental pattern that differs markedly from *C. elegans*. As in other Enoplea [14,33] the first division is equal, and not asymmetric as in *C. elegans*. *R. culicivorax* also shows an inversion of dorso-ventral axis polarity compared to *C. elegans*, while a predominantly monoclonal fate distribution indicates fewer modifying inductions between blastomeres [33,34]. Generation of the hypodermis involves repetitive cell elements extending from posterior to anterior over the remainder of the embryo, a process distinct from that observed in *C. elegans* [34].

We here catalogue the *R. culicivorax* developmental toolkit derived from annotation of a draft genome sequence. We contrast genes and proteins identified in *R. culicivorax* and *T. spiralis* with those of *C. elegans*, and other Ecdysozoa, represented by the arthropod *Tribolium castaneum*. We conclude that major changes in the regulatory logic of development have taken place during nematode evolution, possibly as a consequence of developmental system drift, and that the model species *C. elegans* is considerably derived compared to an ecdysozoan (and possibly metazoan) ground system. However, we are still able to define conserved gene sets that may act in "phylogenic" developmental stages.

Results and discussion

Romanomermis culicivorax has a large and repetitive genome

A draft genome assembly for *R. culicivorax* was generated from 26.9 gigabases (Gb) of raw data (filtered from a total of 41 Gb sequenced; Additional file 1: Table S1). The assembly has a contig span of 267 million base pairs (Mb) and a scaffold span of 323 Mb. The 52 Mb of spanned gaps are likely inflated estimates derived from use of the SSPACE scaffolder. We do not currently have a validated independent estimate of genome size for *R. culicivorax*, but preliminary measurements with Feulgen densitometry suggest a size greater than 320 Mb (Elizabeth Martínez Salazar pers. comm.). The *R. culicivorax* genome is thus three times bigger than that of *C. elegans*, and five times that of *T. spiralis* (Table 1). The assembly is currently in 62,537 scaffolds and contigs larger than 500 bp, with an N50 of 17.6 kilobases (kb). The N50 for scaffolds larger than 10 kb is 29.9 kb, and the largest scaffold is over 200 kb. The GC content is 36%, comparable to 38% of *C. elegans* and 34% in *T. spiralis*. We identified 47% of the *R. culicivorax* genome as repetitive. To validate this estimate we applied our repeat-finding approach to previously published genomes and achieved good accordance with these data (Table 1). The non-repetitive content of the *R. culicivorax* genome is thus approximately twice that of *C. elegans* and three times that of *T. spiralis*. *T. spiralis* thus stands out as having the least complex nematode genome sequenced so far, and the contrast with *R. culicivorax* indicates that small genomes are not characteristic of Dorylaimia.

We generated 454 Sequencing transcriptome data from mixed adults, and assembled 29,095 isotigs in 22,418 isogroups, spanning 23 Mb. These are likely to be a reasonable estimate of the *R. culicivorax* transcriptome. Using BLAT [35], 21,204 of the isotigs were found to be present (with matches covering >80% of the isotig) in single contigs or scaffolds of the genome assembly, suggesting reasonable biological completeness and contiguity of the genome. We also used the CEGMA [36] approach to assess quality of the genome assembly, and found a high representation (90% partial, 75% complete) and a low proportion of duplicates (1.1 fold) (Table 2). Automated gene prediction with iterative rounds of the MAKER pipeline [37], using the transcriptome data as evidence both directly and through GenomeThreader-derived mapping, yielded a total of nearly 50,000 gene models. These were reduced to 48,171 by merging those with identities >99% using Cd-hit [38]. Within the 48,171 models, 12,026 were derived from the AUGUSTUS modeller [39] and 36,145 from SNAP. Because AUGUSTUS predictions conservatively require some external evidence (transcript mapping and/or sequence similarity to other known proteins), we regarded these as the most reliable

Table 1 Genome statistics

Species	Approximate [#] genome size	Estimated repeat content	Median [†] exon length	Median [†] intron length	GC content	Source
<i>C.elegans</i>	100Mb	17% (16.5%)	145bp	69bp	38%	[17,18]
<i>P.pacificus</i>	165Mb	15.3% (17%)	85bp	141bp	42%	[20,25]
<i>A.suum</i>	334Mb	4.4%	144bp	907bp	37.9%	[21,40]
<i>B.malayi</i>	95Mb	16.5% (15%)	140bp	219bp	30%	[22]
<i>B.xylophilus</i>	69Mb	22.5%	183bp	69bp	40%	[25]
<i>M.incognita</i>	~200Mb	36.7%	136bp	82bp	31%	[24]
<i>T.spiralis</i>	63Mb	19.8% (18%)	128bp	283bp	34%	[27]
<i>R.culicivorax</i>	>270Mb	48.2%	161bp	405bp	36%	this work

Repeat content of different nematode genomes appears not to be directly correlated with genome size. Re-calculation in selected genomes shows little deviance from published data (in parentheses)* and thus indicates the validity of our inference for *R. culicivorax*.

*For *B. xylophilus* and *M. incognita* only reference data is given as the same programs were used for initial inference (see references); *A. suum* not re-calculated.

[#]*M. incognita* genome size given as 86Mbp in [24] has been re-estimated to about 150Mbp (E. Danchin pers. comm.).

[†]Median lengths for *A. suum* and *T. spiralis* were calculated in this work as these data are not given in the cited publications.

and biologically complete. In comparison *C. elegans* has ~22,000 genes, and *T. spiralis* has ~16,000. The satellite model nematode *Pristionchus pacificus* has ~27,000 genes [20]. Exons of the AUGUSTUS-predicted genes in *R. culicivorax* had a median length of 161 bp, slightly larger than those in *C. elegans* (137bp) and *T. spiralis* (128bp). Introns of the *R. culicivorax* AUGUSTUS models, with a median of 405 bp, were much larger than those of *C. elegans*

(69 bp) or *T. spiralis* (283bp). The small introns observed in *C. elegans* and other rhabditid nematodes (Table 2) are thus likely to be a derived feature.

We annotated 1,443 tRNAs in the *R. culicivorax* genome using INFERNAL [41] and tRNAscan-SE [42], of which 382 were pseudogenes (see Additional file 1: Table S2 for details). In comparison, *T. spiralis* has 134 tRNAs of which 7 are pseudogenes, while *C. elegans* has 606 tRNAs with 36 pseudogenes [43]. Threonine (Thr) tRNAs were particularly overrepresented (676 copies), a finding echoed in the genomes of *Meloidogyne incognita* and *Meloidogyne floridensis* [24,43] and in *P. pacificus* [20]. The latter has also an overrepresentation of Arginine tRNAs [43].

We have made available the annotated *R. culicivorax* genome, with functional categorisations of predicted genes and proteins and annotation features, in a dedicated genome browser at <http://romanomermis.nematod.es>.

The *R. culicivorax* gene set is more representative of *Dorylaimia* than *T. spiralis*

The phylogenetic placement of *R. culicivorax* makes its genome attractive for exploring the likely genetic complexity of an ancestral nematode. With *T. spiralis*, it can be used to reveal the idiosyncrasies of the several genomes available for Rhabditida. To polarise this comparison, we used the arthropod *Tribolium castaneum*, for which a high quality genome sequence is available [44]. *T. castaneum* development is considered less derived than that of the major arthropod model *Drosophila melanogaster* [45]. The OrthoMCL pipeline accurately clusters orthologous proteins, facilitating the complex task of grouping proteins that are likely to share biological function in divergent organisms [46], and performs better than approaches that simply use domain presence information or aggregative approaches such as psiBLAST [47]. We used the

Table 2 Assembly and annotation statistics

Metric	Result
Contigs > 100bp span	267,342,457bp
Scaffolds > 500bp span	322,765,761bp
Num. contigs/scaffolds	62,537
N50 contigs/scaffolds > 500bp	17,632 bp
N50 scaffolds > 500bp	29,995bp
Max contig length	28,847bp
Max scaffold length	201,054bp
Mean transcript length	593bp
Mean protein length	190aa
MAKER AUGUSTUS predictions	12,026 proteins
MAKER SNAP predictions	36,145 proteins
Num. ESTs (isogroups)	22,418 ESTs
Mean EST length	330bp
80% BLAT sequence coverage	21,204 ESTs
CEGMA compl. completeness	75.40%
CEGMA Group 1 part. compl.	81.82%
CEGMA Group 2 part. compl.	91.07%
CEGMA Group 3 part. compl.	91.80%
CEGMA Group 4 part. compl.	95.38%

OrthoMCL pipeline to generate a set of protein clusters for the four species (*R. culicivora*, *T. spiralis*, *C. elegans* and *T. castaneum*). While the large divergence between these species may obscure relationships between protein sequences, making inference of orthology problematic [48-50], the parameters used were most inclusive [50-52]. Additionally, as the *R. culicivora* genome assembly may not be complete we based inference of absence on shared loss in both *R. culicivora* and *T. spiralis*. Additionally, we validated inferences of absence from the OrthoMCL analyses by performing detailed sequence comparisons using BLAST+ [53] (Additional file 2).

We identified 3,274 clusters that contained protein representatives from all three nematodes, and 2,833 of these also contained at least one *T. castaneum* representative (Figure 2). These 2,833 clusters represent a conserved ecdysozoan (and possibly metazoan) core proteome. Many clusters had *T. castaneum* members, and members from some but not all of the three nematodes, representing candidate examples of loss in one or more nematode lineages of ancient proteins. For example, we identified clusters containing proteins from only one of the nematode species. *T. spiralis* had the lowest number of these (975), while *C. elegans* and *R. culicivora* each had over two thousand. Interestingly, of the 2,747 clusters with only *R. culicivora* proteins from Nematoda, 324 included *T. castaneum* orthologues, whereas *C. elegans* only shared 283 clusters uniquely with the beetle. *T. spiralis* has lost more of these phylogenetically ancient genes than has either *R. culicivora* or *C. elegans*. *T. spiralis* and *C. elegans* shared only 412 clusters exclusive of *R. culicivora* members, while *R. culicivora* and *C. elegans* shared about 1300 clusters exclusive of *T. spiralis*. Despite their phylogenetic affinity, *R. culicivora*

and *T. spiralis* only shared 600 clusters exclusive of *C. elegans* (Figure 2). We suggest that *T. spiralis* genome is not typical of dorylaeids. In comparison to other nematodes it is smaller, has fewer genes overall, and has fewer phylogenetically ancient genes. This is congruent with the previously reported loss of proteins with metabolic function in *T. spiralis* [27]. This reduction in genetic complexity could be due to evolutionary pressures following acquisition of a lifestyle that lacks a free-living stage. Many parasitic and endosymbiotic prokaryotes and eukaryotes have reduced genome sizes, though this is not an absolute rule [54].

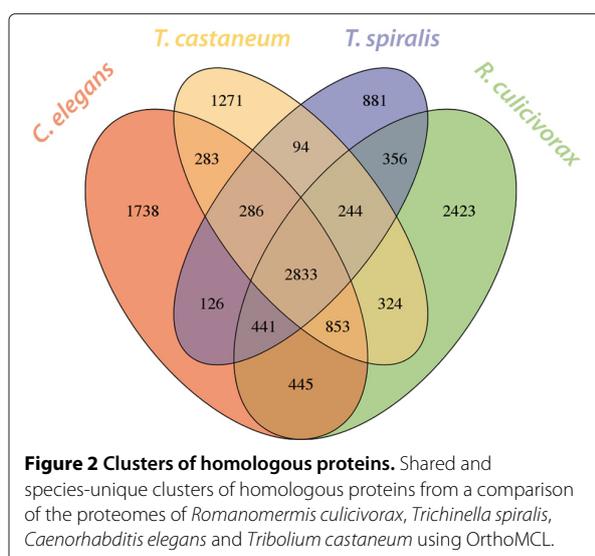
Clusters containing only *R. culicivora* and *T. spiralis* proteins might identify functions distinct to these dorylaeid nematodes. In the 461 *T. spiralis* and 806 *R. culicivora* proteins in these clusters, a total of 65 GO terms were found to be overrepresented (single test $p < 0.05$ by Fisher's exact test). While *C. elegans* has a reduced ability to methylate DNA [55], we found four methylation-associated GO terms among the 64 overrepresented. We also detected significant enrichment (single test $p < 0.05$) for GO terms describing chromatin and DNA methylation functions in the set of *R. culicivora* proteins that lacked homologues in *C. elegans* (see Additional file 3). Important roles for methylation and changes in methylation patterns during *T. spiralis* development have been inferred from transcriptional profiling [56]. Methylation is important for the silencing of transposable elements [57,58] and could play a crucial role in the highly repetitive *R. culicivora* genome.

The clusters that contained *R. culicivora*, *T. spiralis* and *T. castaneum* proteins but no *C. elegans* orthologues might contain proteins involved in core ecdysozoan processes lost in *C. elegans*. In these clusters we identified 40 GO terms overrepresented (single test $p < 0.05$) compared to the *C. elegans* proteome (see Additional file 3). Some of these GO terms were linked to chromatin remodelling and methylation (e.g. Ino80 complex, histone arginine methylation). Other overrepresented GO terms were related to cell signalling (the Wnt receptor pathway; the *C. elegans* Wnt signalling system is distinct from other metazoa [59]), and ecdysone receptor holocomplex (potentially a basic ecdysozoan function [60]).

The genetic background of development in *R. culicivora* and *T. spiralis* differs markedly from that of *C. elegans*

In a recent multi-species developmental time course expression analysis within several *Caenorhabditis* species, conserved sets of genes were found to have conserved patterns of differential expression in discrete phases in the timeline from zygote to the hatching larva [61].

Nearly half (845) of these 1725 conserved, differentially expressed *C. elegans* proteins were not clustered with *R. culicivora* or *T. spiralis* proteins using OrthoMCL. We

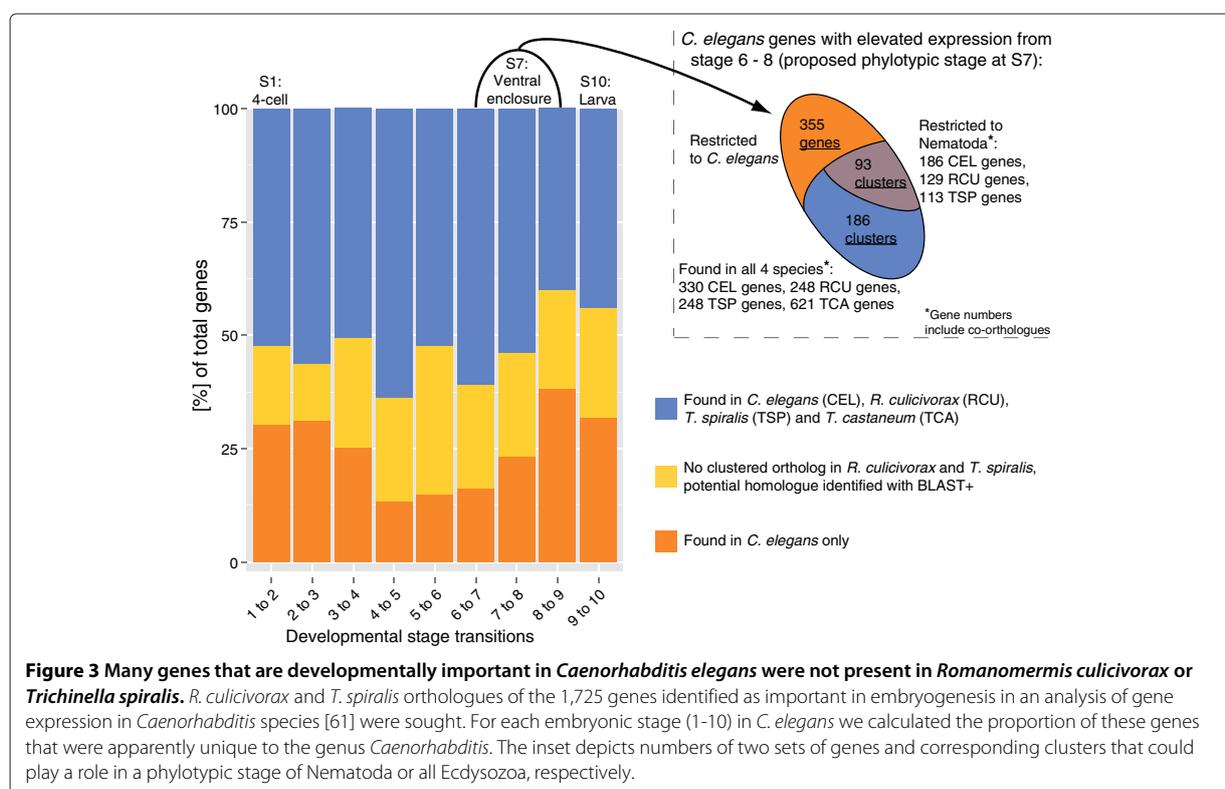


were unable to identify any sequence similarity for 450 of these *C. elegans* proteins, while 395 had only marginal similarities insufficient for OrthoMCL clustering. Eighteen of these 395 are members of *C. elegans* nuclear hormone receptor subfamilies, 5 are innexin type gap-junction proteins, 6 are TWiK potassium channel proteins and 5 are acetylcholine receptor proteins. These protein families are particularly diverse and expanded in *C. elegans* [62-65] and we suggest that they represent rapidly evolved, divergent duplications within the lineage leading to *C. elegans*. The proportion of *Caenorhabditis*-restricted genes across the developmental time course examined by Levin et al. [61] varied from 36% to 60% (Figure 3 and Additional file 4). Thus a surprisingly high proportion of *Caenorhabditis* genes with conserved expression during embryogenesis appear to be unique to the genus or are so divergent that we could not detect possible orthologues in the dorylaims. The pattern of higher retention of conserved genes in *R. culicivora*x compared to *T. spiralis* was also evident in these conserved-expression developmental genes, as 238 had *R. culicivora*x orthologues but lacked a *T. spiralis* orthologues. Given the conservatism of body plan evolution in nematodes, these dramatic genetic differences suggest extensive, largely phenotypically "silent" changes in the genetic programmes orchestrating nematode development.

Core developmental pathways differ between nematodes

There are important differences in cell behaviour during early embryogenesis between *R. culicivora*x and *C. elegans* [33,34]. We used the genomic data to follow up on some of the striking contrasts between the dorylaim and the rhabditid patterns of development: establishment of primary axis polarity, segregation of maternal message within the early embryo, hypodermis formation, the vulval specification pathways, epigenetic pathways (especially DNA methylation), sex determination and light sensing (see Additional file 1).

The mechanisms of sex determination differ considerably among animals and it has been claimed to be one of the developmental programs most influenced by developmental system drift [16]. Divergence in sex determination pathways is thus not unexpected. While sex is determined by X to autosome ratio in *C. elegans* [66], sex ratios in *R. culicivora*x are likely to be environmentally determined through in-host nematode density [67]. Environmental sex determination is found in many nematode taxa, including Strongyloididae and Meloidogyninae (both Tylenchina), taxa more closely related to *C. elegans*. In *C. elegans*, the X to autosome ratio is read by the master switch XOL-1 [68], which acts through the three *sdc* genes [69-71] to regulate the secretion of HER-1, a ligand for the TRA-2 receptor [72-74]. TRA-2 in turn



negatively regulates a complex of *fem* genes, which regulates nuclear translocation of TRA-1, the final shared step in the pathway that switches between male and hermaphrodite systems. No credible homologues of XOL-1, SDC-1, SDC-2, SDC-3, HER-1 or TRA-2 in either *T. spiralis* or *R. culicivora*x were detected through OrthoMCL and re-confirmation with BLAST+ (Table 3; Additional file 2), and thus these species are unlikely to use the HER-1 – TRA-2 ligand-receptor system to coordinate sexual differentiation.

Other developmental processes are however more conserved between metazoan taxa. In *C. elegans* and many other animals *par* genes are essential for cell polarisation [75]. Polarised distribution of PAR proteins results in the restriction of mitotic spindle rotation to the germline cell in the *C. elegans* two-cell stage [76-78]. This rotation is not observed in *R. culicivora*x [33]. The division pattern of *C. elegans* mutants lacking *par-2* and *par-3* genes resembles that of the early *R. culicivora*x embryo [33,79]. The *par-2* gene was absent from both *R. culicivora*x and *T. spiralis* (Figure 4; Table 3). Additionally, no orthologues for the *par-2*-interacting genes *let-99*, *gpr-1* or *gpr-2*, required for proper embryonic spindle orientation in *C. elegans* [80], were identified in the dorylaeims using OrthoMCL clustering or sensitive BLAST+ searches. Although we identified a protein with weak similarity to *par-3* in *R. culicivora*x, this was so divergent from *C. elegans*, *T. castaneum* and *T. spiralis* *par-3* that it was not clustered in our analysis. In *D. melanogaster* a *par-3* orthologue, *bazooka*, functions in anterior-posterior axis formation [81], but *par-2* is absent from the fly. Thus, we hypothesise that the PAR-3/PAR-2 system for regulating spindle positioning evolved within the lineage leading to the genus *Caenorhabditis*. If the divergent *par-3*-like gene in *R. culicivora*x is involved in axis formation, it probably interacts with different partner proteins.

Once polarity has been established in the early *C. elegans* embryo, many maternal messages are differentially segregated into anterior or posterior blastomeres [78,82]. MEX-3 is an RNA-binding protein translated from maternally-provisioned mRNAs found predominantly in early anterior blastomeres [83,84]. We identified a highly divergent MEX-3 orthologue in *R. culicivora*x, but no orthologue in *T. spiralis*. We explored embryonic expression of *mex-3* in *R. culicivora*x embryos using *in situ* hybridisation (Figure 5). In the fertilized egg the *mex-3* mRNA is initially equally distributed. Prior to first cleavage it is segregated to the anterior pole and thus becomes essentially restricted to the somatic S1 blastomere (for nomenclature, see [14]). With the division of S1 it is localized to both daughter cells. After the 4-cell stage the signal disappears gradually. This expression pattern is similar to that of *C. elegans* *mex-3*, affirming that the *R. culicivora*x gene is likely to be an orthologue retaining

Table 3 Presence and absence of selected* *C. elegans* proteins in Dorylaeims

Protein	<i>T. spiralis</i>	<i>R. culicivora</i> x
Early asymmetry		
CDC-42	+	+
PKC-3	+	+
GPR-1	-	-
GPR-2	-	-
PAR-2	-	-
PAR-6	+	+
MES-6	+	+
MES-3	-	-
MES-4	-	-
GFL-1	+	+
LET-70	+	+
Axis formation		
NUM-1	+	+
ZIM-1	-	-
MES-2	-	-
POS-1	-	-
SMA-6	+	+
SET-2	-	-
UBC-18	+	+
LET-99	-	-
OOC-3	-	-
OOC-5	+	+
GPA-16	+	+
PAR-5	-	-
ATX-2	-	-
MEX-5	-	-
MEX-6	-	-
UNC-120	-	-
NOS-2	-	-
OMA-1	-	-
RME-2	+	+
SPN-4	-	-
Sex determination		
XOL-1	-	-
HER-1	-	-
SEX-1	+	+
FOX-1	+	+
SDC-1	-	-
SDC-2	-	-
SDC-3	-	-
TRA-2	-	-
FEM-1	+	+
FEM-2	+	+

Table 3 Presence and absence of selected* *C. elegans* proteins in Dorylaimia (Continued)

Protein	<i>T. spiralis</i>	<i>R. culicivora</i> x
Hypodermis and vulva formation		
AFF-1	-	-
BAR-1	-	-
CEH-2	-	-
CEH-27	-	-
GRL-15	-	-
INX-5	-	-
LIN-1	-	-
PEB-1	-	-
ELT-3	-	-
ELT-1	+	+
SMA-3	-	-
SMA-5	-	-

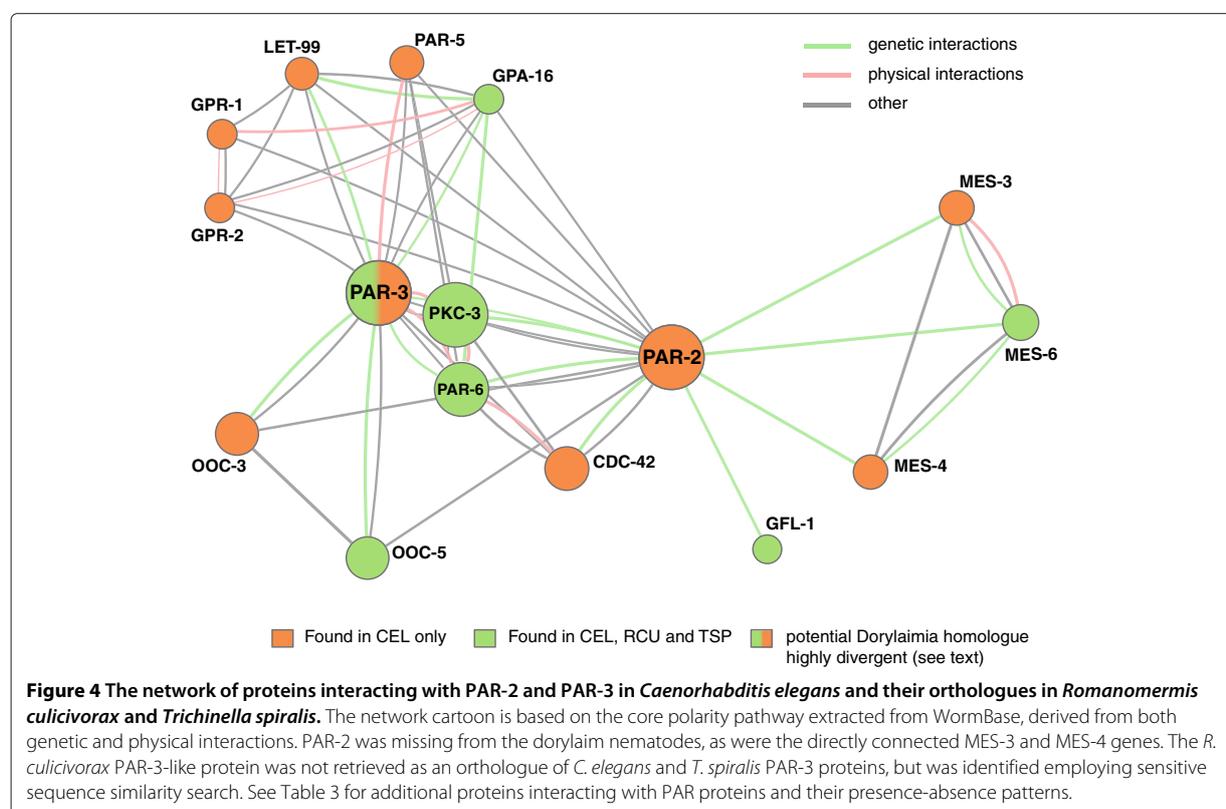
*For additional proteins see Additional files 2 and 4.

similar functions. However, despite the presence, and apparent conservation of the *mex-3* expression pattern, we were unable to identify other interacting partners of the *C. elegans* MEX-3 protein, such as MEX-5, MEX-6 and SPN-4 in either dorylaim species. While MEX-5

and MEX-6 are important for controlled MEX-3 expression in *C. elegans* [85], the apparent absence of SPN-4 in *R. culicivora*x and *T. spiralis* is particularly intriguing. SPN-4 links embryonic polarity conferred by the *par* genes and partners to cell fate specification through maternally deposited mRNAs and proteins [86,87]. Our findings suggest that the core regulatory logic of the early control of axis formation and cell fate specification must differ significantly between the dorylaim species and *C. elegans*.

The hypodermis in *C. elegans* is derived from specific descendants of the anterior and posterior founder cells [88]. In contrast, in *R. culicivora*x hypodermis is derived from descendants of a single cell [34]. Several *C. elegans* genes expressed in the hypodermis or associated with hypodermal development were absent from *R. culicivora*x and *T. spiralis* (see Table 3 and Additional file 3). For example the GATA-like transcription factors ELT-1 and ELT-3 act redundantly in *C. elegans* [89]. ELT-3 was absent from the dorylaim species, but ELT-1 was conserved in *R. culicivora*x, *T. spiralis* and *T. castaneum*. Thus, ELT-3 appears to be an innovation in the rhabditid lineage, suggesting changes of interaction complexity during nematode evolution.

In *C. elegans*, vulva formation is highly dependent on initial inductive signals from the anchor cell that activate



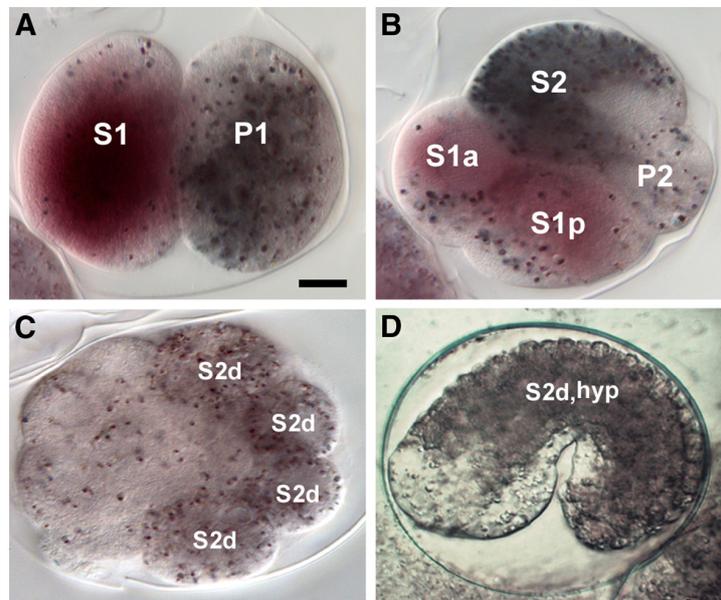


Figure 5 *In situ* hybridisation mapping of *mex-3* mRNA distribution in early embryos of *Romanomermis culicivorax*. We used the *R. culicivorax mex-3* gene to prove application of the *in situ* technique in this species and investigated the segregation patterns of segregation of this maternal RNA in early development. The *R. culicivorax mex-3* expression pattern is similar to that of *C. elegans* [83]. *R. culicivorax* embryos contain dark pigment granules that are asymmetrically segregated in development. **(A)** At the 2-cell stage, maternal *mex-3* mRNA is detected in the S1 blastomere. The cytoplasmic pigment granules are predominantly in the P1 blastomere. **(B)** At the 4-cell stage, *mex-3* mRNA is detected in daughters of the anterior S1 cell. Cytoplasmic pigment granules are predominantly in the S2 blastomere. **(C)** At a later stage (>20 cells), *mex-3* mRNA is absent. The pigment granules are found in descendants of S2 (S2d). **(D)** During early morphogenesis, the pigment granules are found in S2 descendants forming hypodermis, (S2d, hyp). **(A-C)** fixed embryos; **(D)** live embryo. Scale bar 10 μ m. Orientation: anterior left.

a complex gene regulatory network, which drives tissue specific cell division and differentiation. The evolutionary plasticity of this system has been explored in rhabditid nematodes, revealing the changing relative importance of cell-cell interactions, inductions, and lineage-autonomous specifications [90,91]. The signal transduction pathways include a RTK/RAS/MAPK cascade, activated by EGF- and wnt-signalling [92]. Among the downstream targets in *C. elegans* are for example LIN-1 and the β -catenin BAR-1, which in turn regulates the HOX-5 orthologue LIN-39 [93-95]. These important regulators of vulva development are completely absent from the genomes of *R. culicivorax* and *T. spiralis* (Table 3 and Additional file 2). We identified a *R. culicivorax* protein with low similarity to *C. elegans* BAR-1 (24% sequence identity). However, this protein is not clustered with other dorylaim proteins, and appears to be either a duplication of the β -catenin ortholog HMP-2 or another armadillo repeat-containing protein rather than an orthologue of BAR-1 (see Additional file 5). These shared patterns of absence again suggest that similar morphological structures can be generated with very different genetic underpinnings. Vulva formation in the dorylaims may be regulated

without the BAR-1 – LIN-39 interaction, as observed in *P. pacificus* [96]. In *C. elegans* Hox gene expression is cell-lineage dependent [97,98], organised so that the cells that express specific Hox genes are clustered along the anterior-posterior axis (see e.g. [99]). It will be informative to test whether in *R. culicivorax* and other non-rhabditid nematodes Hox genes act in an axis position-dependent, but cell lineage-independent manner, as observed in many other animals, notably arthropods [100,101]. Epigenetic regulation is key to developmental processes in many animals, but its roles in *C. elegans* are more muted (see above). Notably *C. elegans* is depleted for chromatin remodelling genes of the Polycomb and Trithorax groups [102]. It is intriguing that we found orthologues of *T. castaneum pleiohomeotic* in *R. culicivorax* and *T. spiralis*, and orthologs of *T. castaneum trithorax* and *Sex comb on midleg (Scm)* in *R. culicivorax*. This suggests that dorylaim chromatin restructuring mechanisms may be more arthropod-like than in *C. elegans*. The presence of an intact methylation machinery and conserved chromatin re-modelling factors opens the prospects for a role for epigenetic modification in developmental regulation of dorylaim nematodes.

Defining a set of potential phylotypic stage genes

While the examples above demonstrate considerable developmental system drift in Nematoda, we also identified many sets of orthologous proteins conserved between *Dorylaimia* and *C. elegans*. We asked if these could be correlated with functions in distinct developmental phases with a conserved phenotype. Shortly before the start of morphogenesis, at the point of ventral enclosure, nematode embryos from Chromadorea and Enoplea share a similar morphology [14]. Levin et al. [61] found that in five *Caenorhabditis* species a distinct set of genes had elevated expression around the ventral enclosure stage (their stage 7) (Figure 3) and proposed that this constitutes a "phylotypic stage" for nematodes. We used *T. spiralis* and *R. culicivora*x gene sets to refine and restrict this set of phylotypic stage genes. Of the 834 *C. elegans* genes with elevated expression between stages 6 to 8 [61], 355 had no orthologue in *R. culicivora*x, *T. spiralis* or *T. castaneum*. The remaining 479 phylotypic stage candidates from *C. elegans* were present in 279 of our OrthoMCL clusters. Of these clusters 93 were nematode-restricted containing 186 *C. elegans* proteins grouped with 129 *R. culicivora*x and 113 *T. spiralis* homologues. The remaining 186 clusters were part of the conserved ecdysozoan core proteome (see above) and contained 330 *C. elegans* proteins together with 248 *R. culicivora*x, 248 *T. spiralis* and 621 *T. castaneum* proteins (Figure 3; The total number of *C. elegans* candidates is larger than 479 due to the inclusion of co-orthologues in this species). In the set of phylotypic stage genes identified by Levin et al. [61] are proteins functioning in processes such as muscle and neuron formation, signalling between cells, and morphogenesis. This pattern was retained in the conserved clusters (see Additional file 5). Although time-resolved expression data will be needed to confirm the activity of these genes in developmental stages of *R. culicivora*x, their retention in the *Dorylaimia* supports their general importance. We can now sub-classify the set of conserved proteins expressed at the potential nematode phylotypic stage. A first, nematode-restricted set includes many proteins that are important for cuticle formation (e.g. collagen proteins) and some hedgehog-like proteins, expressed in the *C. elegans* hypodermis [103]. As cuticle formation follows ventral enclosure in nematodes, these proteins may be involved in this nematode-specific function. The second set, comprising clusters conserved between the nematodes and *T. castaneum*, contains many important developmental transcription factors, such as the Hox gene *mab-5*, other homeobox genes, and helix-loop-helix and C2H2-type zinc finger transcription factors. This second set may represent a genetic backbone driving formation of phylotypic stage in diverse animal taxa, in accordance with the recent extension of the concept to Metazoa [104-106].

Conclusions

To be useful as a contrasting system to the canonical *C. elegans* model, any nematode species must be accessible to both descriptive and manipulative investigation. The reference genome for *R. culicivora*x lays bare the core machinery available for developmental regulation, and we have demonstrated that *in situ* hybridisation approaches are feasible for this species. Along with the long established, robust laboratory cultures, this makes *R. culicivora*x an attractive and tractable alternative model for understanding the evolutionary dynamics of nematode development. By combining the *R. culicivora*x genome with that of *T. spiralis*, we have been able to explore the molecular diversity of *Dorylaimia*, and provide robust contrasts with the intensively studied Rhabditida. Particularly surprising are the differences between *R. culicivora*x and *T. spiralis*. The *R. culicivora*x genome is much larger than that of *T. spiralis*, and contained a high proportion of repetitive sequence, including many transposable elements. Despite the phylogenetic and lifestyle affinities between the two dorylaims compared to *C. elegans*, the *R. culicivora*x genome retained many more genes in common with *C. elegans* than did *T. spiralis*. We suggest that *T. spiralis* may be an atypical representative of dorylaim nematodes, perhaps due to its highly derived life cycle.

Our analyses identified many genes apparently absent from the dorylaim genomes, despite relaxed analysis parameters. In particular, for genes identified as critical to *C. elegans* development but apparently absent from the dorylaims, we were unable to identify credible orthologues using sensitive search strategies. In this phylum-spanning comparison, inferences of gene orthology can be obscured by levels of divergence. In addition, the gene family birth rate in the chromadorean lineage leading to *C. elegans* is high [25,27], and therefore *C. elegans* was expected to have many genes absent from the dorylaim species. Thus, we might not have found a *R. culicivora*x orthologue for a specific gene for three reasons: it may have arisen in the branch leading to *C. elegans*; its sequence divergence may be too great to permit clustering with potential homologs; or it was not assembled in the draft dorylaim genomes. The case of *C. elegans* PAR-3 and *D. melanogaster bazooka* illustrate some of these difficulties: the possible *R. culicivora*x orthologue was highly divergent. Whether or not we have been able to identify all the orthologues of the key *C. elegans* genes present in the *R. culicivora*x and *T. spiralis* genomes, the absence of an identified orthologue maximally implies loss from the genome, and minimally implies significant sequence divergence. In the latter case this would most likely cause changes in the networks and pathways in which genes interact to deliver biological function.

Between the model nematode *C. elegans* and arthropod models such as *T. castaneum* many key mechanisms

governing early cell patterning are divergent [76]. Our data strongly support the view that major variation also exists within Nematoda. *T. spiralis* and *R. culicivora* both lack orthologues of genes involved in core developmental processes in *C. elegans*, and many of these *C. elegans* genes appear to be restricted to the Rhabditida. It is thus doubtful that these processes are regulated by same molecular interactions across the phylum. We suggest that developmental system drift has played a major role in nematode evolution. The phenotypic conservatism associated with the vermiform morphology of nematodes [107] has fostered unjustified expectations concerning the conservation of genetic programmes that determine these morphologies. Despite this divergence in developmental systems, we were able to define two sets of conserved genes possibly active in a taxon-specific phase of ventral enclosure and cuticle formation in Nematoda, and in a potential phylotypic stage of Ecdysozoa. The advent of robust, affordable and rapid genome sequencing also opens the vista of large-scale comparative genomics of development across the phylum Nematoda [29] to better understand the diversity of the phylum and also place the remarkable *C. elegans* model in context of its peers. It will next be necessary to extend these studies to a broader sampling of developmental pathway genes from a wider and representative sampling of nematode genomes across the full diversity of the phylum. We have highlighted a few of the possible avenues a research programme could follow: early axis formation and polarisation, the specification of hypodermis, sex determination, vulva formation, the roles of epigenetic processes in developmental regulation and the confirmation of potential “phylotypic stage genes” with expression analysis in *R. culicivora*.

Methods

Sequencing and genome assembly

Genomic DNA was extracted from several hundred, mixed-sex, adult *R. culicivora* specimens from a culture first established in Ed Platzer's laboratory in Riverside, California. Illumina paired end and mate pair sequencing with libraries of varying insert sizes, and Roche 454 single end sequencing, was performed at the Cologne Center for Genomics (CCG: <http://www.ccg.uni-koeln.de>). A Roche 454 dataset of transcriptome reads from cDNA synthesised from mixed developmental stages and sexes was also generated (see Additional file 1: Table S1 for details of data generated).

The quality of the raw data was assessed with FastQC (v.0.9; <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Adapter sequences and low quality data were trimmed from the Illumina paired end data with custom scripts (see <http://github.com/sujaikumar/assembly>) and from the mate pair libraries with Cutadapt (v.1.0) [108]. We constructed a preliminary

genome assembly, with relaxed insert size parameters, from the paired end Illumina libraries with the de-novo-assemble option of the *clAssembleCell* (v.4.03b) [109]. We validated the actual insert sizes of our libraries by mapping back the reads to this preliminary assembly using *clAssembleCell*. The preliminary assembly was also used to screen out bacterial and other contaminant data [110]. The transcriptome data were assembled with Roche GSAssembler (Newbler; version 2.5). For the production assembly, we explored assembly parameters using different mixes of our data, evaluating each for total span, maximal contig lengths, N50, number of contigs, representation of the transcriptome, and conserved eukaryotic gene content (using the CEGMA pipeline v.2.1 [36]). The most promising assembly was scaffolded with the filtered Illumina mate pair read sets using SSPACE (v.1.2) [111]. As our genomic DNA derived from a population of nematodes of unknown genetic diversity, we removed short contigs that mapped entirely within larger ones using Cd-hit (v.4.5.7) [38] at a 95% cutoff. A final round of superscaffolding was performed, linking scaffolds that had logically consistent matches to the transcriptome data based on BLAT [35] hits and processed with SCUBAT (B. Elsworth, pers. comm.; <http://github.com/elswob/SCUBAT>). The final genome assembly was again assessed for completeness by assessing the mapping of the transcriptome contigs and with the CEGMA pipeline [36].

Genome annotation

RepeatMasker (v.3.3.0) [112,113], RepeatFinder [114] and RepeatModeler (v.1.0.5; <http://www.repeatmasker.org/RepeatModeler.html>; combining RECON (v.1.07) [115] and RepeatScout (v.1.05) [116]), were used to identify known and novel repetitive elements in the *R. culicivora* genome. We employed the MAKER pipeline [37] to find genes in the *R. culicivora* genome assembly. In a first pass, the SNAP gene predictor included in MAKER was trained with a CEGMA [36] derived output of predicted highly conserved genes. As additional evidence we included the transcriptome assembly and a set of approximately 15,000 conserved nematode proteins derived from the NEMBASE4 database [117] (recalculated by J. Parkinson; pers. comm.). In the second, definitive, pass we used the gene set derived from this first MAKER iteration to train AUGUSTUS [39] inside the MAKER pipeline for a second run, also including evidence from transcriptome to genome mapping obtained with GenomeThreader [118]. Codon usage in *R. culicivora*, *T. spiralis*, and *C. elegans* was calculated using INCA (v.2.1) [119]. Results were then compared to data from [120] (see Additional files 1 and 6).

We used Blast2GO (Blast2GO4Pipe, v.2.5, January 2012 database issue) [121] to annotate the gene set with Gene

Ontology terms [122], based on BLAST matches with expect values less than $1e^{-5}$ to the UniProt/SwissProt database (March 2012 snapshot), and domain annotations derived from the InterPro database [123]. Comparison of annotations between three nematode species (*R. culicivora*, *C. elegans*, and *T. spiralis*) and, as a reference outgroup, the holometabolous coleopteran arthropod *Tribolium castaneum*, was based on GO Slim data retrieved with Blast2GO. RNA genes were predicted using INFERNAL (v.1.0.2)[41] and the Rfam database [124], and tRNAscan-SE (v.1.3.1) [42].

Orthology screen

We inferred clusters of orthologous proteins between *R. culicivora*, *T. spiralis*, and *C. elegans*, and the beetle *T. castaneum* using OrthoMCL (v.2.0.3) [125]. *T. spiralis*, *C. elegans* and *T. castaneum* protein sets were downloaded from NCBI and WormBase (see Additional file 1: Table S3) and redundancy screened with Cd-hit at the 99% threshold. We selected an inflation parameter of 1.5 for MCL clustering (based on [126,127]) within OrthoMCL to generate an inclusive clusterings in our analysis likely to contain even highly diverged representatives from the four species. In analyses of selected developmental genes, clusters were manually validated using NCBI-BLAST+ [53]. We affirmed the uniqueness of *C. elegans* proteins identified as lacking homologues in the enoplean nematodes by comparing them to the *R. culicivora* proteome using BLAST. Those with no significant matches at all (all matches with E-values $> 1e^{-5}$) were classified as confirmed absent. Those having matches with E-values $< 1e^{-5}$ were investigated further by surveying the cluster memberships of the *R. culicivora* matches. If the *R. culicivora* protein was found to cluster with a different *C. elegans* protein, the uniqueness to *C. elegans* was again confirmed. If the *R. culicivora* protein did not cluster with an alternative *C. elegans* protein, we reviewed the BLAST statistics (E-value, identity and sequence coverage) of the match and searched the GenBank non-redundant protein database for additional evidence of possible orthology. Only if these tests yielded no indication of direct orthology was the *C. elegans* protein designated absent from the enoplean set. Further details of the process are given in Additional file 5.

We identified the protein sequences of 1,725 genes differentially expressed in *C. elegans* developmental stages [61] and selected, using our OrthoMCL clustering, those apparently lacking orthologues in *R. culicivora* and *T. spiralis* (verified as above). Using Wormbase (<http://www.wormbase.org>, release WS233) we surveyed the *C. elegans*-restricted genes for their experimentally-defined roles in development.

Custom Perl scripts were used to group orthoMCL clusters on the basis of species membership patterns. The

sets of clusters that contained (i) both *T. spiralis* and *R. culicivora* members but no *C. elegans* members and (ii) *T. spiralis* and *R. culicivora* and *T. castaneum* members but no *C. elegans* members were surveyed for GO annotations enriched in comparison to the whole *C. elegans* proteome (sets i and ii) and the *T. castaneum* proteome (set i), conducting Fisher's exact test as implemented in Blast2GO. Due to the small size of both sets compared to the large reference set, p-values could not be corrected for multiple testing. To improve annotation reliability, these proteins were re-compared (using BLAST) to the UniProt/SwissProt database and run through the Blast2GO pipeline as described above.

Whole-mount in situ hybridization

For in situ hybridisation we modified the freeze-crack procedure described previously for *C. elegans* [128] and revised by Maduro et al. (2007; <http://www.faculty.ucr.edu/~mmaduro/resources.htm>). In particular, to achieve reliable penetration of the durable *R. culicivora* egg envelopes we initially partly removed the protective layer by incubation in alkaline bleach solution (see [33]). Digoxigenine-labeled sense and antisense RNA probes were generated from linearized pBs vectors (Stratagene, La Jolla, USA) containing a 400 bp fragment of *R. culicivora mex-3* via run off *in vitro* transcription with T7 or T3 RNA-polymerase according to the manufacturer's protocol (Roche, Mannheim, Germany). The concentration of the labelled probes was about $300 \text{ ng} \times \text{ml}^{-1}$.

Additional files

Additional file 1: Supplementary data figures and tables.

Additional file 2: Analysis of OrthoMCL output by BLAST+. BLAST+ results for specific *C. elegans* proteins not found in a cluster with Dorylaimia proteins.

Additional file 3: Fisher's exact test data. GO terms enriched in a set of protein clusters shared between Dorylaimia in comparison to (i) *C. elegans* and (ii) *T. castaneum* proteomes.

Additional file 4: Levin data. Genes identified as being differentially expressed in *Caenorhabditis* development by Levin et al. [61].

Additional file 5: Analysis of Phylotypic stage genes. *C. elegans* orthologues of genes possibly acting in (i) a potential nematode specific phylotypic stage and (ii) a metazoan phylotypic stage.

Additional file 6: Codon usage in *R. culicivora*. Codon usage data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PHS conceived study, assembled and annotated the genome, conducted analyses and wrote paper; MK conceived study, conducted analyses and wrote part of the paper; CK conceived part of the study, conducted analyses on developmental expression set and wrote part of the paper; GDK helped with genome assembly and annotation; SK helped with genome assembly and wrote/provided Perl scripts; JIRC analysed MEX-3 dataset; NAN analysed PAR dataset; DS analysed SEX determination dataset; KM conducted RNA sequencing and initial EST assembly; PH performed preparative laboratory

experiments and conceived sequencing strategy; JA conceived sequencing strategy and conducted genome sequencing; PF helped with initial genome pre-assembly; PN initiated study and conceived sequencing strategy; WKT conceived parts of study; MLB conceived study and wrote paper; ES initiated and conceived study and wrote paper. All authors read and approved the final manuscript.

Acknowledgements

We are indebted to E. Platzer, Riverside, for the continuous supply with *R. culicivora* nematodes. We thank J. Schulze, Cologne, for advice on nematode cultivation and C. Becker and K. Konrad for expert technical assistance in the genome sequencing experiments. We are also grateful to H. Oezden, Cologne for assistance with In-situ hybridisations. We thank J. Parkinson, Toronto, for providing a conserved NEMBASE4 protein set, Elizabeth Martínez Salazar, Zacatecas, Mexico, for Feulgen C-value data and A. H. Jay Burr, Vancouver, Canada for sharing preliminary results on phototaxis in *R. culicivora*. Assemblies and other computations were conducted on the HPC cluster "CHEOPS" at the University of Cologne (<http://rrzk.uni-koeln.de/cheops.html>).

Funding

This work was partly funded through the SFB 680: "Molecular Basis of Evolutionary Innovations". Philipp H. Schiffer is funded by the VolkswagenStiftung in the "Förderinitiative Evolutionsbiologie". Gerogios D. Koutsovoulos is funded by a UK BBSRC Research Studentship and an Overseas Research Studentship from the University of Edinburgh. Additional Funding came through the BMBF-Projekt "NGSgoesHPC".

Accession numbers

Raw genome and transcriptome sequence data reported in this manuscript have been deposited in the ERA under accession ERP002111 (<http://www.ebi.ac.uk/ena/data/view/ERP002111>), and assembled genomic contigs deposited in the ENA INSDC database under accession numbers CAQS01000001-CAQS01062537. Annotation information and additional data are available through (<http://romanomermis.nematod.es>).

Author details

¹Zoologisches Institut, Universität zu Köln, Cologne, NRW, Germany, ORCID:0000-0001-6776-0934. ²Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh, Scotland, UK. ³Institute für Entwicklungsbiologie, Universität zu Köln, Cologne, NRW, Germany. ⁴Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH, USA. ⁵Cologne Center for Genomics, Universität zu Köln, Cologne, NRW, Germany.

Received: 7 May 2013 Accepted: 17 December 2013
Published: 27 December 2013

References

- Maduro MF: **Cell fate specification in the *C. elegans* embryo.** *Dev Dyn* 2010, **239**:1315–1329.
- Sulston JE, Schierenberg E, White JG, Thomson JN: **The embryonic cell lineage of the nematode *Caenorhabditis elegans*.** *Dev Biol* 1983, **100**:64–119.
- Blaxter M, de Ley P, Garey J, Liu L, Scheldeman P, Vierstraete A, Vanfleteren J, Mackey L, Dorris M, Frisse L: **A molecular evolutionary framework for the phylum Nematoda.** *Nature* 1998, **392**(6671):71–75.
- Meldal B, Debenham N, de Ley P, de Ley I, Vanfleteren J, Vierstraete A, Bert W, Borgonie G, Moens T, Tyler P, Austen M, Blaxter M, Rodgers A, Lambshead P: **An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa.** *Mol Phylogenet Evol* 2007, **42**(3):622–636.
- Boveri T: *Die Entwicklung von Ascaris megaloccephala mit besonderer Rücksicht auf die Kernverhältnisse.* Fischer: Festschrift fuer Carl von Kupffer; Jena; 1899.
- Müller H: **Beitrag zur Embryonalentwicklung von Ascaris megaloccephala.** *Zoologica* 1903, **41**:60.
- Vangestel S, Houthoofd W, Bert W, Borgonie G: **The early embryonic development of the satellite organism *Pristionchus pacificus*: differences and similarities with *Caenorhabditis elegans*.** *Nematology* 2008, **10**:301–312.
- Skiba F, Schierenberg E: **Cell lineages, developmental timing, and spatial pattern formation in embryos of free-living soil nematodes.** *Dev Biol* 1992, **151**(2):597–610.
- Lahl V, Schulze J, Schierenberg E: **Differences in embryonic pattern formation between *Caenorhabditis elegans* and its close parthenogenetic relative *Diploscapter coronatus*.** *Int J Dev Biol* 2009, **53**(4):507–515.
- Brauchle M, Kiontke K, Macmenamin P, Fitch DHA, Piano F: **Evolution of early embryogenesis in rhabditid nematodes.** *Dev Biol* 2009, **335**:253–262.
- Wiegner O, Schierenberg E: **Specification of gut cell fate differs significantly between the Nematodes *Acroboloides nanus* and *Caenorhabditis elegans*.** *Dev Biol* 1998, **204**:3–14.
- Wiegner O, Schierenberg E: **Regulative development in a nematode embryo: a hierarchy of cell fate transformations.** *Dev Biol* 1999, **215**:1–12.
- Voronov DA, Panchin YV: **Cell lineage in marine nematode *Enoplus brevis*.** *Dev* 1998, **125**:143–150.
- Schulze J, Schierenberg E: **Evolution of embryonic development in nematodes.** *EvoDevo* 2011, **2**:18.
- Schulze J, Houthoofd W, Uenk J, Vangestel S, Schierenberg E: **Plectus - a stepping stone in embryonic cell lineage evolution of nematodes.** *EvoDevo* 2012, **3**:13.
- True JR, Haag ES: **Developmental system drift and flexibility in evolutionary trajectories.** *Evol Dev* 2001, **3**(2):109–119.
- C elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282**(5396):2012–2018.
- Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'eustachio P, Fitch DHA, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, et al.: **The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics.** *PLoS Biol* 2003, **1**(2):E45.
- Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW: **Scaffolding a *Caenorhabditis* nematode genome with RNA-seq.** *Genome Res* 2010, **20**(12):1740–1747.
- Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P, Mitreva M, Roeseler W, Tian H, Witte H, Yang SP, Wilson RK, Sommer RJ: **The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism.** *Nature Genet* 2008, **40**(10):1193–1198.
- Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z, Chen F, Wu X, Zhang G, Fang X, Kang Y, Anderson GA, Harris TW, Campbell BE, Vlaminck J, Wang T, Cantacessi C, Schwarz EM, Ranganathan S, Geldhof P, Nejsum P, Sternberg PW, Yang H, Wang J, Wang J, Gasser RB: ***Ascaris suum* draft genome.** *Nature* 2011, **479**(7374):529–533.
- Ghedini E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, Angiuoli SV, Creasy T, Amedeo P, Haas B, El-Sayed NM, Wortman JR, Feldblyum T, Tallon L, Schatz M, Shumway M, Koo H, Salzberg SL, Schobel S, Perteau M, Pop M, White O, Barton GJ, Carlow CKS, Crawford MJ, Daub J, et al.: **Draft genome of the filarial nematode parasite *Brugia malayi*.** *Science* 2007, **317**(5845):1756–1760.
- Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S, Bringaud F, Wolstenholme A, Bandi C, Epe C, Kaminsky R, Blaxter M, Mäser P: **The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets.** *FASEB J* 2012, **26**(11):4650–4661.
- Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EGJ, Deleury E, Perfus-Barbeoch L, Anthonard V, Artiguenave F, Blok VC, Caillaud MC, Coutinho PM, Dasilva C, De Luca F, Deau F, Esquibet M, Flutre T, Goldstone JV, Hamamouch N, Hewezi T, Jaillon O, Jubin C, Leonetti P, Magliano M, Maier TR, Markov GV, Mcveigh P, Pesole G, Poulain J, Robinson-Rechavi M, et al.: **Genome sequence of the metazoan plant-parasitic nematode, *Meloidogyne incognita*.** *Nature Biotechnol* 2008, **26**(8):909–915.
- Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, Mcveigh P, Takashi T, Tsai IJ, Assefa SA, Cock PJA, Otto TD, Hunt M, Reid AJ, Sanchez-Flores A, Tsuchihara K, Yokoi T, Larsson MC, Miwa J, Maule AG, Sahashi N, Jones JT, Berriman M: **Genomic insights into the origin of**

- parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. *PLoS Pathogens* 2011, **7**(9):e1002219.
26. Srinivasan J, Dillman AR, Macchietto MG, Heikkinen L, Lakso M, Fracchia KM, Antoshechkin I, Mortazavi A, Wong G, Sternberg PW: **The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle.** *Genetics* 2013, **193**(4):1279–1295.
 27. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton LA, Minx P, Yang SP, Warren WC, Fulton RS, Bhonagiri V, Zhang X, Hallsworth-Pepin K, Clifton SW, Mccarter JP, Appleton J, Mardis ER, Wilson RK: **The draft genome of the parasitic nematode *Trichinella spiralis*.** *Nature Genet* 2011, **43**(3):228–235.
 28. Hope IA: **Embryology, Developmental Biology and the Genome.** In *The Biology of Nematodes*. 1st edition. Edited by Lee DL. New York: Taylor & Francis; 2002:121–145.
 29. Kumar S, Schiffer PH, Blaxter M: **959 Nematode Genomes: a semantic wiki for coordinating sequencing projects.** *Nucl Acids Res* 2012, **40**(D1):D1295–D1300.
 30. Kumar S, Koutsouvelos G, Kaur G, Blaxter M: **Toward 959 nematode genomes.** *Worm* 2012, **1**:0–8.
 31. Petersen JJ: **Nematodes as biological control agents: Part I. Mermithidae.** In *Advances in Parasitology*. Edited by Baker JR, Muller R. London: Academic Press; 1985:307–346.
 32. Petersen JJ, Chapman HC, Willis OR, Fukuda T: **Release of *Romanomermis culicivorax* for the control of *Anopheles albimanus* in El Salvador II. Application of the nematode.** *Ame J Trop Med Hyg* 1978, **27**(6):1268–1273.
 33. Schulze J, Schierenberg E: **Cellular pattern formation, establishment of polarity and segregation of colored cytoplasm in embryos of the nematode *Romanomermis culicivorax*.** *Dev Biol* 2008, **315**(2):426–436.
 34. Schulze J, Schierenberg E: **Embryogenesis of *Romanomermis culicivorax*: an alternative way to construct a nematode.** *Dev Biol* 2009, **334**:10–21.
 35. Kent W: **BLAT—the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656.
 36. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**(9):1061–1067.
 37. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18**:188–196.
 38. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**(13):1658–1659.
 39. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**(2):ii215–ii225.
 40. Wang J, Mitreva M, Berriman M, Thorne A, Magrini V, Koutsouvelos G, Kumar S, Blaxter ML, Davis RE: **Silencing of germline-expressed genes by DNA elimination in somatic cells.** *Dev Cell* 2012, **23**(5):1072–1080.
 41. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–1337.
 42. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucl Acids Res* 1997, **25**(5):0955–0964.
 43. Kumar S: **Data for PhD thesis on next generation nematode genomes.** 2012. <http://dx.doi.org/10.6084/m9.figshare.96089>.
 44. Richards S, The *T. castaneum* Genome Consortium: **The genome of the model beetle and pest *Tribolium castaneum*.** *Nature* 2008, **452**(7190):949–955.
 45. Schröder R, Beermann A, Wittkopp N, Lutz R: **From development to biodiversity—*Tribolium castaneum*, an insect model organism for short germband development.** *Dev Genes Evol* 2008, **218**(3–4):119–126.
 46. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing performance of orthology detection strategies applied to eukaryotic genomes.** *PLoS ONE* 2007, **2**(4):e383.
 47. Altschul S, Madden T, Schäffer A: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids ...* 1997. [<http://nar.oxfordjournals.org/content/25/17/3389.short>]
 48. Jensen RA: **Orthologs and paralogs - we need to get it right.** *Genome Biol* 2001, **2**(8):interactions1002.1–1002.3.
 49. Koonin E: **Orthologs, paralogs, and evolutionary genomics.** *Ann Rev Genet* 2005, **39**:309–338.
 50. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits.** *Bioinformatics* 2008, **24**(3):319–324.
 51. Shaye DS, Greenwald I: **OrthoList: A compendium of *C. elegans* Genes with human Orthologs.** *PLoS ONE* 2011, **6**(5):e20085.
 52. Tautz D, Domazet-Lošo T: **The evolutionary origin of orphan genes.** *Nat Rev Genet* 2011, **12**(10):692–702.
 53. Altschul SF, Gish W, Miller W, Myers EW: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
 54. Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, Akiyoshi DE, Tzipori S: **The reduced genome of the parasitic microsporidian enterocytozoan *bieneusi* lacks genes for core carbon metabolism.** *Genome Biol Evol* 2010, **2**(0):304–309.
 55. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes & Dev* 2002, **16**:6–21.
 56. Gao F, Liu X, Wu XP, Wang XL, Gong D, Lu H, Xia Y, Song Y, Wang J, Du J, Liu S, Han X, Tang Y, Yang H, Jin Q, Zhang X, Liu M: **Differential DNA methylation in discrete developmental stages of the parasitic nematode *Trichinella spiralis*.** *Genome Biol* 2012, **13**(10):R100.
 57. Tran RK, Zilberman D, de Bustos C, Ditt RF, Henikoff JG, Lindroth AM, Delrow J, Boyle T, Kwong S, Bryson TD, Jacobsen SE, Henikoff S: **Chromatin and siRNA pathways cooperate to maintain DNA methylation of small transposable elements in *Arabidopsis*.** *Genome Biol* 2005, **6**(11):R90.
 58. Martienssen RA, Colot V: **DNA methylation and epigenetic inheritance in plants and filamentous fungi.** *Science* 2001, **293**(5532):1070–1074.
 59. Eisenmann DM: **Wnt signaling.** *WormBook* 2005. [http://www.wormbook.org/chapters/www_wntsignaling/wntsignaling.html]
 60. Graham LD, Kotze AC, Fernley RT, Hill RJ: **Molecular & biochemical parasitology.** *Mol Biochem Parasitol* 2010, **171**(2):104–107. [<http://dx.doi.org/10.1016/j.molbiopara.2010.03.003>]
 61. Levin M, Hashimshony T, Wagner F, Yanai I: **Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo.** *Dev Cell* 2012, **22**(5):1101–1108.
 62. Phelan P: **Innexins: members of an evolutionarily conserved family of gap-junction proteins.** *Biochimica et Biophysica Acta (BBA) - Biomembranes* 2005, **1711**(2):225–245.
 63. Jones AK, Sattelle DB: **Functional genomics of the nicotinic acetylcholine receptor gene family of the nematode, *Caenorhabditis elegans*.** *BioEssays* 2003, **26**:39–49.
 64. Antebi A: **Nuclear hormone receptors in *C. elegans*.** In *Wormbook*. Edited by The *C. elegans* Research Community. WormBook; 2006. <http://www.wormbook.org>.
 65. Altun ZF, Chen B, Wang ZW, Hall DH: **High resolution map of *Caenorhabditis elegans* gap junction proteins.** *Dev Dyn* 2009, **238**(8):1936–1950.
 66. Haag E: **The evolution of nematode sex determination: *C. elegans* as a reference point for comparative biology.** In *WormBook*. Edited by The *C. elegans* Research Community. WormBook; 2005. <http://www.wormbook.org>.
 67. Tingley GA, Anderson RM: **Environmental sex determination and density-dependent population regulation in the entomogenous nematode *Romanomermis culicivorax*.** *Parasitol* 1986, **92**:431–449.
 68. Powell JR, Jow MM, Meyer BJ: **The T-box transcription factor SEA-1 is an autosomal element of the X:A signal that determines *C. elegans* sex.** *Dev Cell* 2005, **9**(3):339–349.
 69. Chu DS, Dawes HE, Lieb JD, Chan RC, Kuo AF, Meyer BJ: **A molecular link between gene-specific and chromosome-wide transcriptional repression.** *Genes & Dev* 2002, **16**(7):796–805.
 70. Meyer B: **X-Chromosome dosage compensation.** In *WormBook*. Edited by The *C. elegans* Research Community. WormBook; 2005. <http://www.wormbook.org>.
 71. Zarkower D: **Somatic sex determination.** In *WormBook*. Edited by The *C. elegans* Research Community. WormBook; 2006. <http://www.wormbook.org>.
 72. Kuwabara PE, Okkema PG, Kimble J: **tra-2 encodes a membrane protein and may mediate cell communication in the**

- Caenorhabditis elegans sex determination pathway. *Mol Biol Cell* 1992, **3**(4):461–473.**
73. Goodwin EB, Ellis RE: **Turning clustering loops: sex determination in *Caenorhabditis elegans*.** *Curr Biol* 2002, **12**(3):R111–R120.
74. Baldi C, Cho S, Ellis RE: **Mutations in two independent pathways are sufficient to create hermaphroditic nematodes.** *Science* 2009, **326**(5955):1002–1005.
75. Goldstein B, Macara IG: **The PAR proteins: fundamental players in animal cell polarization.** *Dev Cell* 2007, **13**(5):609–622.
76. Bowerman B: **Embryonic polarity: Protein stability in asymmetric cell division.** *Curr Biol* 2000, **10**(17):R637–R641.
77. Severson AF, Bowerman B. *J Cell Biol* 2003, **161**:21–26.
78. Gönczy P, Rose LS: **Asymmetric cell division and axis formation in the embryo.** In *WormBook*. Edited by The C. elegans Research Community. WormBook:2005. <http://www.wormbook.org>.
79. Cheng NN, Kirby CM, Kempthues KJ: **Control of cleavage spindle orientation in *Caenorhabditis elegans*: the role of the genes par-2 and par-3.** *Genetics* 1995, **139**(2):549–559.
80. Wu JC, Rose LS: **PAR-3 and PAR-1 inhibit LET-99 localization to generate a cortical band important for spindle positioning in *Caenorhabditis elegans* embryos.** *Mol Biol Cell* 2007, **18**(11):4470–4482.
81. Doerflinger H, Vogt N, Torres IL, Mirouse V, Koch I, Nusslein-Volhard C, St Johnston D: **Bazooka is required for polarisation of the *Drosophila* anterior-posterior axis.** *Development* 2010, **137**(10):1765–1773.
82. Goldstein B, Frisse L, Thomas W: **Embryonic axis specification in nematodes: evolution of the first step in development.** *Curr Biol* 1998, **8**(3):157–160.
83. Draper BW, Mello CC, Bowerman B, Hardin J, Priess JR: **MEX-3 is a KH domain protein that regulates blastomere identity in early *C. elegans* embryos.** *Cell* 1996, **87**(2):205–216.
84. Huang N, Mootz D, Walhout A, Vidal M, Hunter CP: **MEX-3 interacting proteins link cell polarity to asymmetric gene expression in *Caenorhabditis elegans*.** *Development* 2002, **129**(3):747–759.
85. Evans TC, Hunter CP: **Translational control of maternal RNAs.** In *WormBook*. Edited by The C. elegans Research Community. WormBook:2005. <http://www.wormbook.org>.
86. Gomes JE, Encalada SE, Swan KA, Shelton CA, Carter JC, Bowerman B: **The maternal gene *spn-4* encodes a predicted RRM protein required for mitotic spindle orientation and cell fate patterning in early *C. elegans* embryos.** *Dev* 2001, **128**(21):4301–4314.
87. Labbé JC, Goldstein B: **Embryonic development: A New SPN on cell fate specification.** *Curr Biol* 2002, **12**(11):R396–R398.
88. Simske JS, Hardin J: **Getting into shape: epidermal morphogenesis in *Caenorhabditis elegans* embryos.** *BioEssays* 2001, **23**:12–23.
89. Gilleard JS, McGhee JD: **Activation of hypodermal differentiation in the *Caenorhabditis elegans* embryo by GATA transcription factors ELT-1 and ELT-3.** *Mol Cell Biol* 2001, **21**(7):2533–2544.
90. Sommer R: **As good as they get: cells in nematode vulva development and evolution.** *Curr Opin Cell Biol* 2001, **13**(6):715–720.
91. Kiontke K, Barriere A, Kolotuev I, Podbilewicz B, Sommer R, Fitch DHA, Félix MA: **Trends, stasis, and drift in the evolution of nematode vulva development.** *Curr Biol: CB* 2007, **17**(22):1925–1937.
92. Sternberg PW: **Vulval development.** In *WormBook*. Edited by The C. elegans Research Community. WormBook; 2005. <http://www.wormbook.org>.
93. Salsler SJ, Loer CM, Kenyon C: **Multiple HOM-C gene interactions specify cell fates in the nematode central nervous system.** *Genes & Dev* 1993, **7**(9):1714–1724.
94. Eisenmann DM, Kim SK: **Protruding vulva mutants identify novel loci and Wnt signaling factors that function during *Caenorhabditis elegans* vulva development.** *Genetics* 2000, **156**(3):1097–1116.
95. Shemer G, Podbilewicz B: **LIN-39/Hox triggers cell division and represses EFF-1/fusogen-dependent vulval cell fusion.** *Genes & Dev* 2002, **16**(24):3136–3141.
96. Tian H, Schlager B, Xiao H, Sommer RJ: **Wnt signaling induces vulva development in the nematode *Pristionchus pacificus*.** *Curr Biol* 2008, **18**(2):142–146.
97. Streit A, Kohler R, Marty T, Belfiore M, Takacs-Vellai K, Vignano MA, Schnabel R, Affolter M, Müller F: **Conserved Regulation of the *Caenorhabditis elegans* labial/Hox1 Gene *ceh-13*.** *Dev Biol* 2002, **242**(2):96–108.
98. Aboobaker AA, Blaxter ML: **Hox Gene Loss during dynamic evolution of the nematode cluster.** *Curr Biol* 2003, **13**:37–40.
99. Chisholm A: **Control of cell fate in the tail region of *C. elegans* by the gene *egl-5*.** *Deve* 1991, **111**(4):921–932.
100. Aboobaker A, Blaxter M: **Hox gene evolution in nematodes: novelty conserved.** *Current Opinion in Genet & Dev* 2003, **13**:593–598.
101. Lemons D, McGinnis W: **Genomic evolution of Hox gene clusters.** *Science* 2006, **313**(5795):1918–1922.
102. Chamberlin HM, Thomas JH: **The bromodomain protein LIN-49 and trithorax-related protein LIN-59 affect development and gene expression in *Caenorhabditis elegans*.** *Development* 2000, **127**(4):713–723.
103. Aspöck G, Kagoshima H, Niklaus G, Bürglin TR: ***Caenorhabditis elegans* Has scores of hedgehog related genes: sequence and expression analysis.** *Genome Res* 1999, **9**:909–923. genome.cshlp.org.
104. Domazet-Lošo T, Tautz D: **A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns.** *Nature* 2010, **468**(7325):815–818.
105. Richardson MK: **A phylotypic stage for all animals?** *Dev Cell* 2012, **22**(5):903–904.
106. Kalinka AT, Tomancak P: **The evolution of early animal embryos: conservation or divergence?** *Trends In Ecology & Evolution* 2012, **27**(7):385–393.
107. De Ley P: **A quick tour of nematode diversity and the backbone of nematode phylogeny.** In *WormBook*. Edited by The C. elegans Research Community. WormBook; 2006. <http://www.wormbook.org>.
108. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet.Journal* 2011, **17**(1).
109. CLCbio: **White Paper on de Novo Assembly in CLC Assembly Cell.** CLC Bio: Whitepaper; 2010.
110. Kumar S, Blaxter ML: **Simultaneous genome sequencing of symbionts and their hosts.** *Symbiosis* 2012, **55**(3):119–126.
111. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27**(4):578–579.
112. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0. 1996-2010.** <http://www.repeatmasker.org>.
113. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repeat Update, a database of eukaryotic repetitive elements.** *Cytogenetic and Genome Res* 2005, **110**(1-4):462–467.
114. Volfovsky N, Haas BJ, Salzberg SL: **A clustering method for repeat analysis in DNA sequences.** *Genome Biol* 2001, **2**(8):research0027.1–research0027.11.
115. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.** *Genome Res* 2002, **12**(8):1269–1276.
116. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21**(Suppl 1):i351–i358.
117. Parkinson J, Mitreva M, Whitton C, Thomson M: **A transcriptomic analysis of the phylum Nematoda.** *Nature Genet* 2004, **36**:1259–1267.
118. Gremme G, Brendel V, Sparks ME, Kurtz S: **Engineering a software tool for gene structure prediction in higher organisms.** *Information and Software Technol* 2005, **47**(15):965–978.
119. Supek F, Vlahovicek K: **INCA: synonymous codon usage analysis and clustering by means of self-organizing map.** *Bioinformatics* 2004, **20**(14):2329–2330.
120. Cutter AD, Wasmuth JD, Blaxter ML: **The evolution of biased codon and amino acid usage in nematode genomes.** *Mol Biol Evol* 2006, **23**(12):2303–2315.
121. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(18):3674–3676.
122. The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25–29.
123. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucl Acids Res* 2005, **33**(Web Server):W116–W120.

124. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucl Acids Res* 2004, **33**(Database issue):D121–D124.
125. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Research* 2003, **13**(9):2178–2189.
126. van Dongen S: **A Cluster algorithm for graphs.** *Report - Information Syst* 2000, **10**:1–40.
127. van Dongen S: **Graph Clustering by Flow Simulation.** *PhD thesis*, University of Utrecht 2000. <http://www.wormbook.org>.
128. Seydoux G, Fire A: **Whole-mount in situ hybridization for the detection of RNA in *Caenorhabditis elegans* embryos.** *Methods in cell biology* 1995, **48**:323–337.

doi:10.1186/1471-2164-14-923

Cite this article as: Schiffer *et al.*: The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda. *BMC Genomics* 2013 **14**:923.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



2.2 Extended manuscripts

1. How to survive the extreme: a multi genome analysis reveals evolutionary traits of cryptobiosis and routes to parthenogenesis (**M1** discussed on page 149)
2. Genes from the Cambrian Explosion (**M2** discussed on page 162)
3. Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages (**M3** discussed on page 146) **The manuscript has since been published and the final version is now included.**

Note that both, **M1** and **M2** are included in an “as is” state, i.e. the **M2** manuscript is almost finished and will be submitted in the coming days, while the **M1** manuscript lacks some parts where data from collaborators is needed. However, both manuscripts show all results, data, and figures, which are discussed in the Discussion section of the presented thesis. Especially the **M1** manuscript is of importance, since work on the *Panagrolaimus* genomes was a core part of my projects. Supplementary material to these manuscripts can be found on the CD enclosed with this thesis.

**How to survive the extreme:
a multi genome analysis reveals evolutionary traits of cryptobiosis
and evolutionary routes to parthenogenesis**

Philipp H. Schiffer, Etienne G. Danchin, Simon Wong, Chris Creevey, Anne-Marike Schiffer, Georgina O'Mahony, Michael Thorne, Corinne Rancurel, Gary Stier, Michael Kroihner, Einhard Schierenberg, Ann Burnell, Thomas Wiehe, Mark Blaxter

Introduction

Nematodes are a species-rich phylum with representatives in almost all the ecological niches available on earth. They are found in soils, lacustrine, as well as marine environments and have evolved successful strategies to exploit extreme environments like deep, hot underground mines (Borgonie et al., 2011), arid soils (Demeure, Freckman, & Van Gundy, 1979), polar and sub-polar soils (Convey & Worland, 2000). The phylum also contains economically important parasites of plants and animals (Jasmer, Goverse, & Smant, 2003). In clade IV of the nematode tree (phylogeny after Blaxter, see (Blaxter, 2011)) many of these eco-modes occur. In this clade vertebrate (e.g. genus *Strongyloides*) and insect (family Steinernematidae) parasites are grouped with highly potent plant parasites (*Meloidogyne* spp.), as well as free-living bacterial feeding species (Cephalobidae and Panagrolaimidae, and fungus feeding species (Aphelenchida). The family Panagrolaimidae comprises *Pangrolaimus*, *Propanagrolaimus*, *Panagrellus* (Andrássy, 2005), and potentially *Halicephalobus* (Lewis et al., 2009). In the genus *Panagrolaimus* the potential for cryptobiosis (anhydrobiosis and cryobiosis) (Ricci & Pagani, 1997; A. Shannon, Browne, Boyd, Fitzpatrick, & Burnell, 2005; Wharton & Ferns, 1995) has evolved. Anhydrobiotic nematodes can survive extreme desiccation by entering into a reversible ametabolic state and during this process they may lose up to 95-99% of their body water

{Crowe:1992vm}. Anhydrobiotic phenotypes are typical of *Panagrolaimus*, (Alpert, 2006; Raffi V Aroian & Sternberg, 1993; Wharton & Barclay, 1993). Anhydrobiotes synthesise a variety of molecules to protect their macromolecules and cellular structures during desiccation (A. M. Burnell & Tunnacliffe, 2011; Farrant & Moore, 2011). These include water replacement molecules such as trehalose, molecular chaperones, particularly small heat shock proteins (sHSPs), and LEA (= “late embryogenesis abundant) proteins, which prevent their aggregation; antioxidants, DNA repair proteins and fatty acid desaturases which adjust membrane fluidity. Most anhydrobiotic organisms need to be dehydrated slowly to allow time for these metabolic adjustments to occur, but some “fast-dehydration strategist” mosses (Proctor et al., 2007) and nematodes, including *P. superbus* (Shannon et al., 2005) can survive rapid dehydration. In addition to its anhydrobiotic capacity, *P. davidi* from maritime Antarctica can also survive freezing when fully hydrated (Wharton & Ferns, 1995) and this cryobiotic phenotype is associated with the synthesis of an antifreeze ice-active proteins which inhibits the growth of ice crystals (Wharton, Barrett, Goodall, Marshall, & Ramløv, 2005). A similar cryotolerant capacity also exists in other hydrated *Panagrolaimus* species isolated in temperate and continental regions (McGill, al., submitted).

Recently it was found that frequent desiccation promoted the intake of horizontally transferred genetic material in the rotifer *Adineta vaga* (Flot et al., 2013). In general Horizontal Gene Transfer (HGT) appears to play a much larger role in metazoans than previously anticipated (Boto, 2014). It has been described in plant parasitic clade iv nematodes, Tylenchina, where for example horizontally acquired Cellulases constitute an adaptive advantage for the acceptor species (Danchin et al., 2010). It has been hypothesised that the integration of a divergent group of Cellulases into the *P. pacificus* genome (Dieterich et al., 2008) allows this nematode to enhance its food range (Haegeman, Jones, & Danchin, 2011). Although found to be largely non-functional, horizontal acquisition of exogenous DNA from *Wolbachia* parasites in several parasitic nematodes (Koutsovoulos, Makepeace, Tanya, & Blaxter, 2014) illustrate the abundance of the general mechanism. The large number of genes acquired by these and in particular the Tylenchina species raise the question if other roundworms might equally profit from HGT in adapting to their environment. Especially in the adaptation to harsh environments (Srinivasan et al., 2013), where the desiccating *Panagrolaimus* species abound HGT might be evolutionary beneficial. In the Bdelloid rotifers rampant HGT has also been brought into connection with evolution under parthenogenesis (Gladyshev, Meselson, & Arkhipova, 2008). Survival under unfavourable environmental conditions has also been connected with the evolution of parthenogenesis (termed ‘geographical parthenogenesis’) (Kearney, 2005). It is thus intriguing to note that parthenogenesis indeed evolved in the genus *Panagrolaimus*. Many parthenogenetic species sampled from around the globe can be found in the genus and phylogenetic data, although based on few gene fragments, indicate a single origin of this mode of reproduction in *Panagrolaimus* (Lewis et al., 2009). A large body of work on the evolution of sex and its general prevalence has been compiled during the past century (Schön, Martens, Dijk, & van Dijk, 2009). While many previous studies were theoretical, emerging experimental data

(see e.g. (Becks & Agrawal, 2010; 2012)) appears to now demonstrate why and that sexual reproduction is superior to parthenogenesis. However, it is still not clear if most newly evolved parthenogens are on a fast track to extinction or if they have the ability to adapt and exploit new (possibly unfavourable) environments. It appears that many parthenogenetic species are hybrids of some kind (Simon, Delmotte, Rispe, & Crease, 2003) and a hybrid background has recently been described for species in the genus *Meloidogyne* (Lunt, Kumar, Koutsovoulos, & Blaxter, 2014). Regardless if through a hybrid or spontaneous origin a newly arisen parthenogenetic species has to overcome certain cellular and molecular challenges. In particular egg activation, which is usually coupled to sperm entry (Runft, Jaffe, & Mehlmann, 2002), the restoration of full chromosome sets (Avise, 2008), and importantly the abolishment of the male phenotype to gain the full advantage of parthenogenesis. An intricate connection between early development and mechanistic alterations possibly needed for the successful establishment of parthenogenesis is also illustrated by the fixation of the anterior posterior axis by the sperm entry point in *C. elegans* (Goldstein & Hird, 1996). It is not clear which genomic prerequisites have to be met for a successful transition from sexual to strictly parthenogenetic propagation overcoming these hurdles. To address these questions clearly a laboratory model is in need for, which can be accessed with molecular tools, and where the genomic background is available.

While a diversity of mating and propagation systems is realised in the phylum Nematoda (Denver, Clark, & Raboin, 2011), the model organism *C. elegans*, however, is an androdioecious hermaphrodite and its congeners either follow this mode of reproduction or are amphimictic. Thus to unravel the genomic background a model system with male female and parthenogenetic species is needed. The connection of parthenogenesis and early development is intriguing in regard to Panagrolaimid nematodes, as we recently found indications for a re-shuffling of Gene Regulatory Networks (GRNs) possibly due to Developmental System Drift (DSD) between species in the genus *Panagrolaimus* and *Propanagrolaimus* (Schiffer et al., 2014). The observed shift in cellular pattern formation and gene expression in early development between these species is even more interesting as the product of development, the final adult morphology, is very uniform in *Panagrolaimus* sensu lato, leading previous authors to assume these to belong to one genus (Lewis et al., 2009).

However, given the functional divergence of proteins across nematodes (Schiffer et al. 2013), and in Panagrolaimids in particular (Schiffer et al. 2014) it cannot be regarded obvious that these genes are acting in similar pathways and gene regulatory networks in the clade IV species. This will also be true for the molecular basis of sex determination, which might be affected in evolutionary processes leading to the establishment of parthenogenesis. The molecular system is again very well resolved in *C. elegans*, but many of the genetic components vary across the nematodes {Schiffer et al. 2013} (while some genes appear conserved across large phylogenetic distances (P. Hunt, 2011)).

In comparison to *Meloidogyne*, which can only be cultured in close association with plants, or vertebrate parasites like *Strongyloides*, which might have highly derived genomes as often found in parasites, free living *Panagrolaimus* species appear ideal candidates for

comparative exploration of developmental idiosyncrasies in clade IV. Even more so, as cultures are easily reared under standard laboratory conditions, can be stored for a long time under anhydrobiosis, and worms can be studied on the cellular level (e.g. lineage through 4D microscopy), as well subjected to molecular analysis like in-situ hybridisations and genetic tools like RNAi (A. J. Shannon, Tyson, Dix, Boyd, & Burnell, 2008). Thus, establishing the genomic backbone of species in the taxon is highly interesting for a wide array of evolutionary questions.

In this study we present 4 novel genome assemblies from one parthenogenetic, two amphimictic *Pangrolaimus* and a hermaphrodite *Pronagrolaimus* species. This is supplemented by an improved re-assembly of the parthenogenetic *P. davidi* genome based on data from (Thorne, Kagoshima, Clark, Marshall, & Wharton, 2014). Generating RNAseq data for our species we were able to compute robust gene prediction for a comparative assay. From this we are able to deduce a possible hybrid origin incorporating triploidisation in parthenogenetic *Panagrolaimus* species. We find an array of important *C. elegans* genes acting in development of the model organism that are not present in many clade IV species and thus illustrating fascinating new biology to be discovered there. We were able to analyse the adaptive potential of *Panagrolaimus* and especially the parthenogenetic species on the genomic level by comparing with close outgroups like the semi-aquatic *Propanagrolaimus*. We highlight potential gene candidates allowing adaptation to unfavourable environments employing anhydrobiosis. Intriguingly several of these genes appear to be acquired horizontally, illustrating the evolutionary power of HGT. The presented data will serve as a genomic basis for future in detail functional studies into the evolutionary origin of parthenogenesis in Nematoda.

Material and Methods

Genome size and chromosome numbers

We used Feulgen stainings performed by Elizabeth Martínez Salazar (Zacatecas, Mexico) to get an estimate of genome sizes in various *Panagrolaimus* strains, and conducted flow cytometry experiments as well. We complemented these measurements with a RT-D-PCR based assay developed by (Wilhelm, Pingoud, & Hahn, 2003). For this we used the *C. elegans* locus ZK682.5 (WBGene00022789), which had been identified as nematode wide single copy gene (Mitreva et al., 2011). Chromosome numbers were counted in DAPI stained whole body mounts of adult specimens and eggs, from which pictures of gonads and single cell stages were acquired and analysed (see Supplementary Methods).

Genome Assembly

We constructed genome assemblies from Illumina short read libraries of different insert sizes (Supplementary Table 1). DNA was extracted with standard protocols (including liq. N2

freeze-cracking, phenol-chloroform extraction or the Qiagen Genomic Tip Kit in a modified two step protocol, see Supplementary Methods) from large numbers of adults, larvae and eggs, which were washed from plates, cleaned in several washing steps (partly including an antibiotics, Antimycotic treatment). Detailed protocols are available from the first author. The parthenogenetic PS1159 and the hermaphrodite JU765 had been bottlenecked for several generations in independent lines, while the other 3 species (ES5, *P. superbus*, *P. davidi*) were standard lab cultures.

We first cleaned Illumina reads from residual adapters and removed low quality bases either with or Trimmomatic (versions ≤ 0.32) (Bolger, Lohse, & Usadel, 2014), or a combination of sickle (version ≤ 1.33) {Joshi:ui} and scythe (<https://github.com/vsbuffalo/scythe>). The genomic 454 library available for *P. superbus* was cleaned with QTrim (v.1.1) (Shrestha et al., 2014). We then evaluated our raw data with the sga preqc pipeline (Simpson, 2014).

After running first pass assemblies with the clc assembly cell v.4.2 we applied the khmer pipeline (Brown, Howe, Zhang, & Pyrkosz, 2012) to digitally normalise read coverage on ES5, *P. superbus*, JU765 and PS1159 data. It is almost impossible to sequence nematode genomes without the bacteria they feed on and which reside on their cuticle, particularly as the low total number of cells per nematode leads to a skewed rate of nematode cells to bacterial cells even after repeated washing steps prior to DNA extraction. We thus used the initial assembly to purge bacterial sequences from our data following a protocol by Kumar et al. (Kumar, Jones, Koutsovoulos, Clarke, & Blaxter, 2013). This left us with a reduced set of reads near exclusively originating from the target genome. From these reads we constructed a second genome using the velvet assembler (Zerbino & Birney, 2008) testing kmer sizes of 51 and 61. In the ES5 assembly HaploMerger (Huang et al., 2012) was used as an intermediate step. We then employed RNASeq derived mRNA predictions (see below) to scaffold the genomes with the SCUBAT pipeline (<https://github.com/elswob/SCUBAT>). For *P. superbus* the genome input to SCUBAT originated from a clc assembly, as this proved better than the tested velvet assemblies. Using our established pipeline we also re-assembled the *P. davidi* genome, where the original assembly {Thorne 2014} was of low completeness (CEGMA: 39% complete KOGs). The final assemblies were assessed for completeness employing CEGMA (Parra, Bradnam, & Korf, 2007) and the baa.pl script (Ryan, 2013), which uses mapped RNASeq derived contigs to check for intrinsic gene content. We also employed the REAPR (M. Hunt et al., 2013) pipeline to evaluate our assemblies.

Hybridisation and Repeats

Hybridisation between closely related species appears to be a common origin of parthenogenesis (Simon et al., 2003). To test the possible presence of a third, divergent copy of the genome in the parthenogenetic strains we first mapped reads against the assembled SSU sequences in each species and then visually analysed the distribution of variants (see Supplementary Figure SSU). To further test the phenomenon we took advantage of having different genomic ground states in our assay. The hermaphrodite JU765, expected to be diploid,

and the potentially triploid parthenogenetic PS1159, were both run through bottlenecking in single offspring lines and thus expected to be partly homogenised genomes. The amphimictic ES5 and *P. superbus*, as well as the *P. davidi* were expected to be heterozygous as large populations from multiple plates containing several generations were used for DNA extraction. We extracted all Augustus predicted genes (exons and introns, no promotor regions/UTRs) for each species and mapped read sets used in the assemblies against these using the *clc* mapper (v.4.2.0), while requiring a sequence identity of 90% and a read length threshold of 80%. We then used *bamtools* (v.2.3.0) (Barnett, Garrison, Quinlan, Strömberg, & Marth, 2011) to sort the mappings, *samtools* (v.1.0-13) (“PNAS-1994-Shang-8373-7,” 2011) to create *mpileups* and *VarScan* (v.2.3.6) (Koboldt et al., 2012) to call variants. We then counted the occurrence of all distinct variant frequencies, e.g. how often a distribution 20% variant to 80% reference, or 60% variant to 40% was observed in all mapping instances.

We used a two-pronged approach to screen for repetitive elements in the final draft genomes. First we used *RepeatModeler* (v. open-1.0) (<http://www.repeatmasker.org/RepeatModeler.html>) to identify and *RepeatMasker* (v. open-4.0.5) {RepeatMaskerOpen:ov95ByFr} to mask genomes using the BLAST based search engine. We also evaluated *LTRharvest* (Ellinghaus, Kurtz, & Willhoeft, 2008) implemented in the *GenomeTools* (Gremme, Steinbiss, & Kurtz, 2013) to screen for transposons possibly missed by the *RepeatModeller*, but could not detect any additional.

As this approach is not necessarily sufficient to identify all repetitive regions in 2nd generation genomes and similar repeats are often assembled into one contig we also employed a second, a kmer based assay. In this we mainly followed a pipeline used previously by established Dan Bolser (personal comment, see <https://github.com/dbolser/PGSC/blob/master/kmer-filter/README>), which is based on tallymer within the *GenomeTools* package (Gremme et al., 2013). We first counted the frequency of all unique and repeated kmers at sizes 10 - 50 in the genomes and plotted these (Supplementary figure Gkmers). For each genome we then picked a kmer size close to the value where frequency curves reached a plateau and counted the occurrence of the kmer at coverage levels 10 - 50. We then identified the corresponding sequences in the genome and, extending from the described tallymer pipeline, analysed the per base coverage for these genomic regions based on *clc* mappings with the aid of *bedtools* (Quinlan & Hall, 2010). We compared the obtained coverage distributions with genome wide coverage and based on this selected a set of potential repeats to be intersected with the previous sequence data from *RepeatModeler*. This unified set was then fed back into *RepeatMasker* for a second round of masking.

Transcriptome assembly

To aid gene annotation and confirm expression of important genes we sequenced transcriptomes of mixed life cycle stages in PS1159, ES5, and JU765 using Illumina RNAseq technology. Obtained reads (Sup Table M1) were adapter and quality screened with *Trimmomatic* (versions ≤ 0.32). We then used *Trinity* (Grabherr et al., 2011) to assemble reads

into contigs and predicted ORFs with the corresponding module of Trinity. To complement our data from the parthenogenetic group within *Panagrolaimus* we sequenced the transcriptomes of two further strains, DL137 and PS1579 (see also (Schiffer et al., 2014)). For *P. superbis* a transcriptome enriched for nematodes in the anhydrobiotic stage was sequenced using the 454 platform. Using CEGMA we assessed the completeness of our assemblies.

Gene Prediction and final contamination cleaning

We relied on Augustus (Stanke & Waack, 2003) for gene predictions. We used an iterative approach to generate most credible predictions from our genomic data. Augustus was first trained with CEGMA predicted genes. Where available (ES5, PS1159, JU765), we then to map RNASeq data using gsnap (Wu & Nacu, 2010), which was shown to perform well in finding correct splice sites (Engström et al., 2013). These mappings then served as evidence for Augustus in a second round of prediction. For *P. superbis* in a first round of Augustus training we employed exonerate (Slater & Birney, 2005) to align the predicted *P. sp.* PS1159 proteins, as no RNASeq data was available. We then mapped findorf (Krasileva et al., 2013) predicted ORFs from the 454 sequencing derived transcriptomic data (see above) using blat (Kent, 2002) and created Augustus evidence from this. For *P. davidi* unfortunately no RNASeq or other transcriptomic data was available at the time of our analyses. Thus, we choose to use the *P. sp.* PS1159 species model in Augustus along with the *P. davidi* CEGMA genes as hints. According to Lewis et al. (Lewis et al., 2009) analysis (Figure tree) and our extended revision in this work both species are phylogenetically close.

Open reading frames were predicted from de novo transcriptome assemblies (*P. superbis*, DL137, PS1579) using the findorf program (<https://github.com/vsbuffalo/findorf> and (Krasileva et al., 2013)). The findorf approach involved comparison of transcripts with proteomes from other nematode species (*B. malayi*, *C. briggsae*, *C. elegans*, ES5, JU765, *M. hapla*, *P. redivius*, PS1159, and preliminary data from *Plectus sambesii*) using BLASTX to identify frameshifts, premature stop codons, etc. It also involved the detection of Hidden Markov Model (HMM) Pfam domains among the transcripts that may be used to infer ORF start positions. In total, findorf inferred 21,381 *P. superbis* ORFs, 34,868 DL137 ORFs and 47,754 PS1579 ORFs.

To purge remaining bacterial contigs and predicted proteins on them that had evaded our GC and coverage based cleaning of the reads we used a BLAST+ based strategy. We developed custom Python scripts to compare blast results of proteins against all bacterial and all metazoan data in the protein section of Genbank, respectively. We also blasted each contig/scaffolds in the genomes against the bacterial nucleotides stored in Genbank. When finally compiling lists of which contigs and proteins to purge we also used custom IPython (Pérez & Granger, 2007) developed scripts to evaluate exon numbers and numbers of potentially bacterial proteins per contig to facilitate our decisions. In *P. sp.* ES5 we found many long bacterial contigs and scaffolds close in GC to the nematode retained in the genome. Here we employed mugsy (Angiuoli & Salzberg, 2011) to perform genome alignments of contamination-candidate

nematode contigs/scaffolds against several full bacterial genomes we had identified as potential contamination in the BLAST+ screen. In this way it was possible to remove several contigs, which showed homology to Bacteria over extended stretches or the full sequence. When in doubt we opted to retain contigs and proteins to re-evaluate after our HGT screen (see below).

Codon usage on the predicted genes was analysed with GCUA (McInerney, 1998). We also used GCUA to compare codon usage between genes gained through HGT (see below) and the genomic background. We used tRNAscan-SE v1.3.1 (Lowe & Eddy, 1997) to screen for tRNAs selecting the implemented Cove as search engine.

Functional Proteome Annotation

Putative domains in 13 nematode proteomes were inferred using InterProScan version 5.7-48.0 (<https://code.google.com/p/interproscan/> and (P. Jones et al., 2014)). The tested proteomes included all our *Panagrolaimus* species (both transcriptomic and genomic data for *P. superbus*), our *Propanagrolaimus* outgroup, and the two remote outgroup species *C. elegans* and *A. suum*. The software scans input proteomes using a variety of search algorithms and models to infer the presence of functional domains and sites. A summary of Pfam (Punta et al., 2012) and PANTHER (Mi et al., 2010) domains identified in our 13 proteomes can be found in (Supplementary Table Interpro).

We parsed the inferred annotation with custom Python scripts, mainly relying on information from Pfam and PANTHER.db. We also used the tools available through PANTHER.db to calculate enrichments of particular functions comparing the *Panagrolaimid* species to data from the model *C. elegans* from that database. We used a custom Python script to calculate statistical overrepresentation performing a two-sided Fisher's exact test (NHST) on the retrieved domains for groups of species. We then corrected for multiple testing using the Benjamin-Hochberg algorithm ("FDR") implemented in the R statistical language. We choose to include the transcriptomic data from *Panagrolaimus* sp. DL137 and PS1159, as well as the duplication containing *P. davidi* gene predictions in our test to retrieve proteins potentially missed in the genome wide annotations of PS1159 alone.

To validate our NHST measures, especially with regard to recent discussions of the weakness of this method only rejecting the null-hypothesis (Cohen, 2001), we implemented a machine learning approach in Matlab. A support vector machine (SVM), trained on the number of expressions of 30 annotations for respective species (e.g. *Panagrolaimus* vs. *Propanagrolaimus* and *Panagrellus*) was able to classify species correctly. As additional control for the NHST measure we excluded the transcriptomic data as well as the *P. davidi* data in the SVM assay. Sampling annotations that were most often part of successful classification iterations, we obtained lists of annotations, which were diagnostic for each group (see Supplementary Methods). These lists were then intersected with results from the NHST analysis.

Detection of Horizontal Gene Transfers

To detect candidate horizontal gene transfers (HGT), we used the Alieness (Rancurel, Da Rocha, & Danchin, 2014) software to calculate an Alien Index (AI) as described in (Flot et al., 2013; Gladyshev et al., 2008). Briefly, all the Panagrolaimidae predicted proteins were compared against the NCBI's nr library using BLASTp (Altschul, Madden, & Schäffer, 1997) with an e-value threshold of $1e^{-3}$ and no SEG filtering. BLAST hits were parsed to retrieve associated taxonomic information, using the NCBI's taxonomy as a reference. For every Panagrolaimidae protein returning at least one hit in either a metazoan or non-metazoan species, we calculated an AI, according to this formula:

$$AI = \log(\text{best metazoan } e_value + e^{-200}) - \log(\text{best non_metazoan } e_value + e^{-200})$$

When either no metazoan or non-metazoan significant BLAST hit was found, an e-value of 1 was automatically assigned as the best metazoan and non-metazoan e-values, respectively. To allow detection of HGT events that took place in an ancestor of Panagrolaimidae and its close relatives, BLAST hits to Panagrolaimoidea (TaxID: 55746) and Aphelenchina (TaxID: 1182516) were skipped for the calculation of AI. No AI value could be calculated for proteins returning no significant hit at all in NR. An AI > 0 indicates a better hit to a non-metazoan species than to a metazoan species and thus a possible acquisition via HGT. An AI > 30 corresponds to a difference of magnitude e^{10} between the best non-metazoan and best metazoan e-values and is estimated to be a strong indication of a HGT event (Flot et al., 2013).

All *Panagrolaimidae* proteins that returned an AI > 0 and that aligned with ≥ 70 % identity to a non-metazoan protein were considered as possible contaminants and were discarded from the analysis.

Reconstruction of the timing of acquisition

We used the Mesquite software v.3.01 suite to reconstruct timing of gene acquisition via HGT (W. P. Maddison & Maddison, n.d.). Based on a matrix of presence/absence of the genes, mapped on a reference phylogeny of the Panagrolaimidae species included here + 2 outgroups (*A. nanus* and *B. xylophylus*), Mesquite traced back ancestral states at each node. We used both a parsimony model and a maximum likelihood model. For the parsimony model, when presence/absence of a gene family at a given node was equally parsimonious, we arbitrarily considered the family as present, because it intuitively appears more likely to lose a gene that had been acquired by HGT in an ancestor than gaining it multiple times independently via HGT. We used the same assumption for the likelihood approach and whenever equal probabilities were given for presence/absence of a gene at a given node, we favored its presence.

Functional analysis of candidate HGT

Pfam annotations were retrieved for all the proteins putatively acquired via HGT from our Interproscan data. We used the online Venn diagram building tool at <http://bioinformatics.psb.ugent.be/webtools/Venn/> to identify Pfam domains that were conserved

among the set of candidate HGT proteins of several *Pana* species. We also predicted putative functions based on the presence of Pfam domains based on the pfam2go file that associates Pfam domains to Gene Ontology terms (available at: geneontology.org/external2go/pfam2go) and using in house perl scripts. To make Gene Ontology annotations comparable, we mapped raw GO terms to the generic GOSlim ontology that contains only parent terms, so that all the terms are compared at a same depth level in the ontology. We used the online tool "GOSlimViewer" developed as part of the AgBase (McCarthy et al., 2006).

Phylogenetics

To robustly confirm the phylogenetic setting of species in our analyses we extended a previous phylogeny {Schiffer, 2014} by analysing all 452 proteins predicted by the CEGMA pipeline (see Figure 1). We constructed individual alignments for each CEGMA KOG from our species and outgroups with clustalomega v.1.2 (Sievers et al., 2011) and then used trimal (v.1.4.rev15) (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009) to exclude overly ambiguous regions. A supermatrix was then constructed with the aid of phyutility (v.2.2.6) {Smith:2007vc}.

Phylogenies were inferred using the mpi version of Phylobayesf1.4f (Lartillot, Blanquart, & Lepage, 2011) with the CAT model implemented therein and RAxML (v.7.7.2) (Stamatakis, 2006) using the VT model under the GAMMA parameter with empirical base frequencies, as predicted by ProTest (v.3.2) (Darriba, Taboada, Doallo, & Posada, 2011). Phylogenetic inferences for single proteins, or groups of orthologs (e.g. sHSP), were conducted with clustalomega, trimal, protest3, and RAxML in the way as described for the CEGMA KOGs. We employed Jalview (Waterhouse, Procter, Martin, Clamp, & Barton, 2009) and seaview to visualise alignments (Galtier, Gouy, & Gautier, 1996), and used SplitsTree4 (Huson, 2008) to explore relationships in gene families. Phylogenetic inferences were run on the CHEOPS cluster at the University of Cologne.

To acquire an age estimate of splits between species in the genus *Panagrolaimus* we calculated the Kr value (Haubold, Pfaffelhuber, o, & Wiehe, 2009) with *genomdiff* implemented in the *GenomeTools*. We used this alignment free methods to compare the *Panagrolaimus* species among each other, as well as the several species from *Caenorhabditis*, for which divergence times had been published (Cutter, 2008). We were thus able to correlate the Kr values in *Caenorhabditis* to the divergence times in that genus and transfer this to data from *Panagrolaimus*.

Orthology

To infer orthologues between our *Panagrolaimus* sensu lato strains, other species in Clade IV, including preliminary data from the Cephalobus species *Acrobelloides nanus* (mainly

for the HGT analysis; courtesy of Itai Yanai, Haifa; to be published elsewhere), as well as the model *C. elegans* and the clade III species *Ascaris suum* we used OrthoMCL (v.2.0.8) (Li:2003en; Van Dongen, 2000). We complemented this with the Orthoinspector (v.2.11) (Linard, Thompson, Poch, & Lecompte, 2011), which we found to be especially useful to validate OrthoMCL's inferences of absence/divergence - in our extended tests with larger datasets Orthoinspector appears to find some connections, which are over-looked in OrthoMCL, but appear to be valid based on functional comparative studies. Before initiating the all vs. all blasts (BLAST+ v.2.28) we removed redundant proteins from the proteomes using cd-hit (v.4.6) (Li & Godzik, 2006). We tested several thresholds (99, 97, 95, 90%) and decided to choose the 99% to not to over-exclude recent in-paralogues. We used both programs in conjunction to screen for homologous potentially implicated in the Gene Regulatory Networks (GRNs) of early development, the reproductive system, anhydro- and cryobiosis (independently), and in particular based on our previous work (Schiffer et al., 2014) in gut development. In cases where OrthoMCL and Orthoinspector grossly disagreed or the presence absence pattern across several species appear inconsistent we applied the HMM profile based figmop pipeline (Curran, Gilleard, & Wasmuth, 2014) to directly screen genomes.

Adaptive evolution

To screen for recent selection on genes that (a) might affect the ability to undergo anhydrobiosis in *Panagrolaimus* and (b) be implicated in the evolution of parthenogenesis we used the CRANN software (Creevey & McInerney, 2003). We first created more than 20,000 protein alignments based on our Orthology clusters with clustalomega, and converted these into DNA alignments with trimal and the help of samtools ("PNAS-1994-Shang-8373-7," 2011). Focussing on those alignments, which had parthenogenetic and amphimictic *Panagrolaimus* species we removed short and outlier sequences we arrived at ~7,200 alignments, for which we calculated K_a/K_s estimates with CRANN.

Results

Genome and transcriptome assemblies

We found *Panagrolaimus* sp. PS1159 to possess 12 chromosomes in the diploid set up in deposited single cell eggs. The sexual species *Panagrolaimus* sp. ES5 had $2n=8$ (Figure 2) and another amphimictic species we tested (*P.* sp. "Bromber") had the same number of chromosomes. These numbers seems to be valid for parthenogenetic and amphimictic *Panagrolaimus* species in general (Ann Burnell, personal observation; to be published elsewhere). The hermaphrodite *Propanagrolaimus* sp. had $n=5$, which is the same as in

Panagrellus redivivus (Srinivasan et al., 2013). Our genome size measurements based on RT-D-PCR indicated a haploid c-value of 0,09 (+/-5%, n=3) for ES5, and c=0,16 (+/- 5%; n=4) in PS1159. We were not able to acquire Feulgen data for ES5, but could perform several measurements in a closely related species (see {Lewis2009}) *P. sp.* “Bromber”. For this species we calculated a mean diploid c-value of 0.14 (+/- 0.01; n =6). Unfortunately only one PS1159 sample could be accessed, which had a diploid c-value of **0.27**. We could however obtain two measurements from the very closely related PS1579, which we also use in our analysis, and which had a diploid c-value of 0.23. Other preliminary Feulgen staining data from parthenogenetic *Panagrolaimus* species indicate values in the range between PS1579 and PS1159 (data not shown). It thus appears that the difference in chromosome numbers between parthenogenetic and amphimictic *Panagrolaimus* species, 12 vs. 8, is roughly reflected by genome sizes, with the parthenogenetic lying between 115 and 160Mb, and the amphimictic between 70 and 90Mb (note, due to the uncertainty between measurements c-values are not exactly converted, but only approximations for the amount of base pairs given). For JU765 we found c=0.08 (+/- 0.003; n =4) in our Feulgen staining, making this the largest difference between staining and RT-D-PCR, were we measured c=0,101 (+/-4%).

We found that for our species the best assembly results were obtained after normalising read sets with the khmer pipeline and then using the velvet assembler to construct genomes (except for *P. superbus* where the clc assembly cell was best). RNASeq data used in the SCUBAT pipeline was helpful in scaffolding. All read sets were found to have bacterial contamination, with the highest amount in *P. sp.* ES5 and *P. superbus*, but after implementing the blobology pipeline, as well as post assembly cleaning and a final evaluation during AlienIndex calculation we are confident to have removed most of the contaminants. We achieved a varying grade of reduction to the haploid state for the newly assembled genomes. The *Propanagrolaimus sp.* JU765 assembled into a final in-silico genome size of ~64Mb, thus somewhere in-between the haploid measure from Feulgen staining (~80Mb) and the one calculated with RT-D-PCR (~50Mb). We credit the comparably low number of contigs and large maximal scaffold length (with 907Kbp the longest achieved for any of our species) to the bottlenecking of the worms for many generations before DNA was extracted. This should have homogenised genomes along independently bottlenecked lines and thus reduced some potential break points in the assembly process. Recombination will however have led to some structural variants in the bottlenecked lines. The *Panagrolaimus sp.* ES5, as well as *P. superbus* appear both to be in the general range of the measured c-values with, 90Mb, and 76Mb respectively. As we suspect the *P. superbus* genome to retain some bacterial contamination, for example evidenced through the higher GC content of 31.8%, this genome might be lacking some target species regions. Interestingly, the parthenogenetic species PS1159 assembled to a size of ~84Mb, which is roughly the value measured for the haploid genomes of the sexual species and considerably lower than what was physically measured. In contrast to this we obtained a 221Mb genome for *P. davidi*. It thus appears the bottlenecking in PS1159 lead to a high grade of

homogenisation, while the non-bottlenecked *P. davidi* specimens retained the heterozygous wild-type state that led to many break points in the assembly.

Hybridisation and polyploidy

Our chromosome counts in combination with the genome size measurements could be taken as indicative of the parthenogenetic species being triploid in comparison to the amphimictic *Panagrolaimus* species. There are 12 chromosomes in our parthenogenetic species compared to 8 in the diploid stage of the analysed sexual ones. The genome sizes of the parthenogenetic species appear to be about in line with this. Comparing the diploid c-value of the sexual species of ~0.14 (Feulgen) to 0.18 (RT-D-PCR) with values between 0.23 and 0.27 in the parthenogenetic species is roughly in line with this ($0.14 + \frac{1}{2} * 0.14 = 0.21$; $0.18 + \frac{1}{2} * 0.18 = 0.27$). While the presumably heterozygous *P. davidi* fell into 221Mb and had a high amount of predicted genes, which could not be reconciled with cd-hit, PS1159 assembled into a genome size similar to the haploid state of the amphimictic species. As parthenogenetic species are often formed in hybridisation and become polyploid in the process (Simon et al., 2003), we asked whether we could find a signature of this based on our genome sequencing data.

We hypothesised that variant frequencies in a triploid parthenogenetic species might show a different profile from a diploid one, in that many variant positions will be supported by only 1/3 of the reads mapping to the genomic regions; these are the 1/3 reads coming from the additional chromosome set.

Our analyses of the variant frequency in genes of the sequenced species appears to support a different variant profile in the parthenogenetic PS1159 in comparison to *Propanagrolaimus* JU765, as well as the amphimictic *Panagrolaimus* ES5 and *P. superbus*. In particular we appear to find a shift of variant frequencies towards the 33% value in PS1159, while in JU765 the peak of frequencies is closer to 50% (Figure 2). This becomes even more striking when reads are mapped at full, and not at khmer normalised, coverage. In both amphimictic species we see no clear pattern, with frequencies more evenly distributed over all values. We also tried cross mapping PS1159 reads on ES5 and vice versa. Variant frequencies in the PS1159 on ES5 mappings are evenly distributed over values up to 90% before they peak at higher values. Mapping ES5 on PS1159 appears to be very similar to mapping ES5 on itself. When mapping across genera from PS1159 to JU765 and vice versa we see the expected pattern that when a variant is found, it is present in virtually 100% of the reads. Coverage for cross species and genera mappings is low in general, as expected.

A third set of chromosomes originating from a divergent parental species in a hybridisation event could also cause a signature in the global coverage profile. We tested this by comparing genome wide coverage profiles in the PS1159, JU765, ES5, and *P. davidi*. While both, JU765 and ES5 appear to have smooth profiles with one overall coverage peak, we observe a second, albeit lower peak at in the parthenogenetic species PS1159 (Figure 2). *P. davidi* does

not show this second peak, which we credit to the high heterozygosity and the resulting fragmentation of the genome assembly for this species .

Repeats

Based on our RepeatModeler and RepeatMasker approach we found only modest numbers of repeats in the *Panagrolaimus* and *Propanagrolaimus* genomes. *P. sp.* ES5 had 7.96%, *P. superbus* 8.43 %, PS1159 6.84% and JU765 9.10% masked bases (we did not evaluate the fragmented *P. davidi* genome). These data are similar to what was reported for the *P. redivivus* genome (7.1%; (Srinivasan et al., 2013)). Following the kmer based approach we found genomic regions that had already been identified and masked by the RepeatModeler and RepeatMasker pipeline, but also some additional candidates elevating the total repeat counts in the genomes by about 1 – 2%.

In our screen for anhydrobiosis related genes (see below) we detected a Helitron transposon in one *P. redivivus* sHSP. Helitrons are eukaryotic transposons that are predicted to amplify by a rolling-circle mechanism (considered to be acquired by horizontal gene transfer) (Kapitonov & Jurka, 2001). Further searches revealed Helitron insertions to be quite abundant in *Panagrolaimus* proteins with first screens indicating 27 or more in PS1159 and 49 or more in *P. superbus*. Thus, Helitrons appear to be possibly more abundant in *Panagrolaimus* protein coding genes than in the *C. elegans* proteome.

Gene predictions, annotation

Our RNASeq enhanced Augustus gene predictions gave a varying number of models for the Panagrolaimid species. *Propanagrolaimus* JU765, which has the smallest genome, also has the lowest numbers of gene models after purging contamination, with 24,878 predicted genes. In total, findorf inferred 21,381 *P. superbus* ORFs, 34,868 DL137 ORFs and 47,754 PS1579 ORFs. Using a standalone version of the Interpro pipeline and database we were able to retrieve annotations for 77% - 88% of the predicted genes in our species (see Table 1). We found mean transcript lengths of around 1460bp in PS1159 and 1400bp in ES5, while *P. superbus* and *P. davidi* had 1100 and 1140 respectively. We attribute this difference to the more broken state of the latter two assemblies, which will inevitably have lead to less complete genes. Interestingly this is not reflected by the CEGMA values for *P. superbus*. The *Propangrolaimus* JU765 appears to have somewhat longer transcripts with a mean of ~1560bp. This seems to be due to longer exon and intron lengths in this species compared to the others, which are similar to the ones found in *Panagrellus redivivus* (Srinivasan et al., 2013). It also seems that the amphimictic *Panagrolaimus* species have slightly longer exons, than the parthenogenetic species (see Table 1). Notably all our species, as well as *P. redivivus* have shorter introns, but longer exons than *C. elegans* (Srinivasan et al., 2013). We found 328 rRNAs in ES5, 55 of which were predicted as pseudogenes and 30 had introns, in PS1159 there were 331 (39, 52) tRNAs, and *P. superbus* had 354 (64, 15), while the *P. davidi* genome was too fragmented for a sensible

measurement. In the *Propanagrolaimus* JU765 we found 314 tRNAs, 6 pseudogenes and 31 with introns.

Phylogenetics and divergence

We used all 452 KOGs predicted with the CEGMA pipeline for our phylogenetic inferences. We retrieved the majority of CEGMA KOGs in at least 8 of the species included in our phylogeny, with the largest number of proteins present in all 10 species (see Figure 1). Of the Panagrolaimid species our *P. davidi* re-assembly and the outgroup *Panagrellus redivivus* had the smallest number of KOGs present with 385 and 389 proteins respectively. The *P. davidi* re-assembly was however improved in comparison to the previous version (Thorne et al., 2014). Our *Propanagrolaimus* JU765 had 427, while the ES5 genome had 395 and the *P. superbus* transcriptome had 431. The parthenogenetic PS1159 had 433, and the transcriptomes from the parthenogenetic species PS1579 and DL137 had 444 and 445 of the proteins.

Prottest 3 inferred the VT+G+F model as fitting under the AIC criterion for our super matrix alignment and we used this in RAxML. We let RAxML automatically detect the amount of bootstrap replicates needed and it found 56 trees after 861 bootstraps. Our Bayesian approach calculated with the mpi version of Phylobayes could not converge on a single tree even after running for more than 8000 generations in four parallel chains. We found this to be due to *P. davidi* swapping between two positions within the group of parthenogenetic *Panagrolaimus* species. However, tree topologies of both methods are otherwise congruent. As previous inferences with much less total alignment lengths (Lewis et al., 2009; Schiffer et al., 2014) our data support that parthenogenesis has a common origin within *Panagrolaimus*. We robustly confirm that *Propanagrolaimus* is not the closest outgroup to *Panagrolaimus*, but *Panagrellus* is.

We were interested to acquire a measure of divergence time between the parthenogenetic and amphimictic *Panagrolaimus* species and possibly to the outgroups. As calibration of molecular clocks without fossils and mutation rates is not reliable we used a more indirect approach employing the alignment free K_r measure of genome wide divergence {Haubold} in comparison to measures from *Caenorhabditis* species (where mutation rates are available) (Cutter, 2008). We found K_r values between the amphimictic and parthenogenetic species to be ~ 0.18 , thus lower than between the two closest *Caenorhabditis* species, *C. briggsae* and *C. species 5*, with 0.22 (See Supplementary Excel file KR). Given a divergence time estimate of ~ 14 Ma between these two species the parthenogenetic *Panagrolaimus* species appear to have originated more recently. The two amphimictic species were even closer with $K_r \sim 0.15$. Values for the outgroup species were varying and in part much higher, and thus not useful to be translated into divergence time measures. This however is expected due to the non-linear behaviour of the K_r measure for higher pairwise distances.

	amphimict <i>P. superbus</i> DF5050	amphimict <i>P. sp.</i> ES5	parthenogen <i>P. sp.</i> PS1159	parthenogen <i>P. davidi</i> CB1	hermaphrodite <i>Prop. sp.</i> JU765	parthenogen <i>P. sp.</i> PS1579	parthenogen <i>P. sp.</i> DL137	amphimict <i>Panagrellus</i> <i>redivivus</i> [#]
Measured haploid genome size (Mb)	-	~80	130 - 160	-	50 - 80	~230†	-	~64
Number of chromosomes	-	4	12	-	5	-	-	5
GC%	31.8	29	28.2	28	36	-	-	~44
Transcript GC%	36	33.8	33.2	33.1	38.3	33.5	33.7	49.4*
Number of scaffolds	53,192	25,284	17,628	47,551	13,268	-	-	940*
Longest scaffold (bp)	29,093	96,280	142,873	91,919	907,536	-	-	2,280,433*
Scaffold N50 (for contigs >500bp)	1,894	6,312	9,924	8,895	10,861	-	-	262,414
Assembled genome span (Mbp)	76	90	85	222	64	-	-	65
Span of gaps in scaffolds (Mbp)	1,1	1,6	1,7	11,3	0,8	-	-	3
Repeats	8.43 %	7.96%	6.84%	-	9.10%	-	-	7.1%
Number of transcript models from genome	23,319	24,756	27,350	48,215	24,878	-	-	26,372
proteine models cd-hit (99%)	22,860	24,363	26,083	46,385	23,195	-	-	25,268
mRNA models from transcriptomes	21,381	-	-	-	-	47,754	34,869	-
mRNA models cd-hit (99%)	19,506	-	-	-	-	37,818	30,215	-
Mean/median transcript length (bp)	1107/828	1401/1057	1459/1099	1140/857	1566/1162	-	-	-
mean/median exon length (bp)	240/179	246/187	224/170	213/167	275/195	-	-	288/-
mean/median intron length (bp)	88/53	121/54	114/52	129/54	131/47	-	-	163/-
Exons per gene	-	-	-	-	-	-	-	-
Interpro annotations	16,235	18,932	21,592	35,613	20,548	32,746	25,046	19,768*

Table 1: [#]*P. redivivus* data after (Srinivasan et al., 2013). * *P. redivivus* wwn calculations. † Feulgen based measurements not on haploid cells, see main text.

Functional classification

We first used null hypothesis significance testing (NHST) to obtain a measure of which protein domains and the connected biological functions might be over-represented in the analysed groups of species. However, as NHST bears the problem of not being able to make a statement about the data, it only rejects the null-hypothesis (Cohen, 2001), we validated our results using a different statistical approach. We constructed a support vector machine (SVM) implemented in Matlab, trained it on a subset of annotations, and then tested groups of species for such annotations which are successful in classifying between groups.

In the comparison between the genus *Panagrolaimus* on one and the closest outgroup containing *Propanagrolaimus* JU765 and *Panagrellus redivivus* on the other side 43 Pfam domains were detected as overrepresented in *Panagrolaimus* by NHST and 64 found to have an above-average

frequency among the most successful domains in the SVM classification. Of these 13 were shared in both methods. In this intersection of both methods we found several annotations overrepresented in / successfully classifying *Panagrolaimus* that can be brought into connection with anhydrobiosis. There were for example Hsp70 protein domains, serine protease inhibitors, and Ubiquitin family domains, which are important to ensure protein stability and degradation. We also found Subtilase family and Matrixin domains, which act as peptidases and in peptide cleavage as overrepresented by the NHST method alone. Additional both methods found Lipase domains and the NHST method Enoyl-reductase domains, which are connected to fat metabolisms. From both methods we also retrieved Piwi domain and RNA dependent RNA polymerase as overrepresented/important in *Panagrolaimus*. *Piwi* proteins and RNA dependent RNA polymerases have recently been found to mediate transposon silencing in nematodes outside of clade V (Sarkies et al. in review). A set of annotations found overrepresented through NHST alone (PIF1-like helicase, Deoxyribonuclease II, linker histone H1 and H5 family) can also be brought in connection with maintaining genome integrity, particularly important during and after rehydration. See table 2.

Pfam	Description	Potential role
PF00012	Hsp70 protein	Anhydrobiosis: protein protection
PF00079	Serpin (serine protease inhibitor)	Anhydrobiosis: protein protection
PF02987	Late embryogenesis abundant protein	Anhydrobiosis: protein protection
PF00082	Subtilase family	Anhydrobiosis: peptidase
PF00413	Matrixin	Anhydrobiosis: peptide cleavage
PF00240	Ubiquitin family	Anhydrobiosis: protein degradation
PF02798	Glutathione S-transferase	Life-cycle: countering Xenobiotics
PF01764	Lipase	Life-cycle: fat metabolism
PF13561	Enoyl-(Acyl carrier protein) reductase	Life-cycle: fat metabolism
PF05970	PIF1-like helicase	Genome integrity: DNA replication
PF03265	Deoxyribonuclease II	Genome integrity: hydrolyses DNA
PF00538	linker histone H1 and H5 family	Genome integrity: higher order chromatin
PF02171	Piwi domain	Genome integrity: transposon silencing
PF05183	RNA dependent RNA polymerase	Genome integrity: transposon silencing

Table 2: White background confirmed by NHST and SVM, grey only by NHST.

Developmental Systems (and GRNs)

We had previously shown that Developmental System Drift at least modified the Gene Regulatory Network (GRN) of endo-/mesoderm formation between *Panagrolaimus* and *Propanagrolaimus* by observing a shift in the expression pattern of the *skn-1* ortholog (Schiffer et al., 2014), a key gene in determining *C. elegans* cell-fates. We found that combining the stringent clustering of OrthoMCL, thought to capture mostly orthologues which retain function, with Orthoinspector, which is able to detect more remotely connected paralogues (Linard et al., 2014), enhanced our analysis. For example in the endo-/mesoderm induction pathway an *elt-2* orthologue is not detected in any species outside *C. elegans* by OrthoMCL, but is found by Orthoinspector. In general however we found many genes of the pathway missing in all species of clade IV with both programs, see Figure 3. Hierarchically seen the intermediate switches of the pathway are missing. These are in particular the *med* and *end* genes, as well as *tbx-35*, which are all acting downstream of *skn-1* (Maduro & Rothman, 2002).

The sex determination pathway has special relevance for the evolution of parthenogenesis, as species have to abolish the male phenotype to not pay part of the cost of sex without even gaining the advantage of combining genotypes. Male defining genes would thus be ideal-candidates for downstream functional analyses of the evolution of parthenogenesis. We found many components of the *C. elegans* sex determination pathway not present in either the *Panagrolaimids*, or most of the other clade IV species analysed (Supplementary file Genelist). Again we found a difference between OrthoMCL and Orthoinspector, while both detected no orthologues of the F-Box protein *fog-2* are present in any of the clade iv species or *A. sum*, only Orthoinspector showed the huge number of F-Box in-paralogues known to be present in *C. elegans* (Kipreos & Pagano, 2000). The global picture resembles that of the endo-mesoderm GRN with upstream (*sex-1*) and downstream players (*fem-2*, *tra-1*) conserved, while intermediate switches are missing (e.g. *her-1*, *tra-2*), see table (Supplementary file Genelist). We do find orthologues to the *mab-3* (*Drosophila* Doublesex) and related proteins (*dmd-4*, *dmd-5*).

Interestingly the *sdc* genes, acting in *C. elegans* dosage compensation and found absent in basally branching Enoplea nematodes (Schiffer et al., 2013), are also absent from clade IV. Similarly *let-99* and *par-2*, genes acting in early axis specification and both missing from the Enoplea, are also not found in the clade IV species. As the *skn-1* is expressed in the germline specification in *Panagrolaimus*, but not *Propanagrolaimus* and the male germline can be abolished in parthenogens we were also interested to see how conserved the general *C. elegans* GRN for this system (including spermatogenesis) is among the *Panagrolaimid* species. This process, as well as dauer formation, which had already been analysed in the *Panagrellus* genome report (Srinivasan et al., 2013), as there are no dauer stages in the *Panagrolaimids*, appears to show the same general pattern of presence and absence as described above, see supplementary file Genelist.

Spe-41 is a crucial sperm specific gene mediating gamete fusion in *C. elegans* (Xu & Sternberg, 2003). We detected orthologues of the gene the clade IV species, and importantly

found transcripts in the RNASeq data of the parthenogenetic *Panagrolaimus* species. Similarly we find expression of major sperm proteins in the parthenogenetic species.

DNA repair

Efficient DNA repair is crucially important for species undergoing frequent desiccation leading to strand breaks (Hespeels et al., 2014). It could be equally important under parthenogenesis, to avoid mutational meltdown by Muller's ratchet (Felsenstein, 1974; Muller, 1964). Our assessment of the *C. elegans* DNA repair system in clade IV, and in particular between amphimictic and parthenogenetic *Panagrolaimus* species does however reveal the same pattern as described for developmental processes above. Genes are either completely absent, or present in all species. We found one interesting gene, *mlh-1*, coding for a *C. elegans* mismatch repair enzyme to be inferred lost in the parthenogenetic *Panagrolaimus* species, while it is found present in the amphimictic species as well as in the other clade IV species and *A. suum*. As this pattern of absence in only a few of the clade IV species was uncommon in our assay we used the figmap pipeline to build a gene model based on motifs in the *C. elegans* gene and then screen the PS1159 genome for these. We did find one candidate in this genome and the gene turned out to be inferred as a paralogue to genes found as orthologues of *mlh-1* in ES5, *P. superbus*, and the outgroups by OrthoInspector. Studying an alignment of all these proteins together we identified a *Panagrolaimus* specific region at the N-terminal end of the protein, as well as residues specific to the parthenogenetic species. A splits network constructed for the proteins also indicates *Panagrolaimus* specific divergence.

Adaptive divergence

Interestingly we find further evidence for divergence in the DNA repair or replication system in *Panagrolaimus*. Our genome wide screen for genes under selection using the KaKs measure revealed a one OrthoMCL cluster to have genes with a median value of 8.5 in the *Panagrolaimus* species. The *C. elegans* gene in this cluster was identified as a DNA polymerase (WBGene00021344). Further genes with connection to genome integrity, where the KaKs value was above one and thus indicating possible positive selection were the *C. elegans* gene *fcd-2*, which acts on cross-linked DNA and the un-named WBGene00010061, a helicase.

Detection of putative HGT in Panagrolaimidae genomes and transcriptomes

Using the Alieness (Rancurel et al., 2014) approach described in methods, we calculated Alien Indexes (AI) to detect candidate HGT in 5 Panagrolaimidae genomes (*P. sp.* ES5, *P. sp.* PS1159, *P. davidi*, *Propanagrolaimus sp.* JU765, and *Panagrellus redivivus*) and one transcriptome (of *P. superbus*). We found proteins with AI >30 and less than 70% identity to non-metazoan proteins for all species. The number of putative HGT of non-metazoan origin ranged from 22 in *P. redivivus* to 342 *P. sp.* ES5. All proteins that returned an AI >30 and more than 70% identical to a non-metazoan protein were considered as possible contaminants and

discarded. The number of putative contaminants ranged from 12 in *P. redivivus* to 141 in *P. sp.* ES5.

Abb. name	Full name	# prots	AI >30	AI >0	Conta
PSU	<i>Panagrolaimus superbus</i>	21381	88	528	20
ES5	<i>Panagrolaimus sp.</i> ES5	24756	232	975	141
PS1159	<i>Panagrolaimus sp.</i> PS1159	27350	182	899	19
PDA	<i>Panagrolaimus davidi</i>	48215	342	1915	30
JU765	<i>Propanagrolaimus sp.</i> JU765	24878	44	530	39
PRED	<i>Panagrellus redivivus</i>	26372	22	516	12

Table 3: Alieness scores.

In case putative donors are of bacterial origin, another source of evidence to discard contamination from the set of predicted HGT is to look for the presence of spliceosomal introns. This data is available for the 5 *Panagrolaimus* genomes. In PS1159, 90 putative HGT are of bacterial origin, 59 of the corresponding gene models have at least one spliceosomal intron.

Comparison of sets of predicted HGT in Panagrolaimidae and timing of acquisition

Based on an OrthoMCL analysis that has been performed on the 6 *Panagrolaimus* species and other nematodes, we crossed information of groups of orthology with alien indexes. We searched OrthoMCL groups in which all *Panagrolaimus* proteins processed by Alieness had an AI >30 and less than 70% id with non-metazoan proteins. We found that a total of 142 MCL groups contained *Panagrolaimus* proteins that had all AI strongly indicating acquisition via HGT. Hence, it can be assumed that at least 142 gene families, grouping 430 genes of the 6 *Panagrolaimus* species have been putatively acquired via HGT of non metazoan origin. As could be expected for candidate HGT, the majority of these MCL groups (89/142 = 63%) were Panagrolaimidae-specific (See Supplementary Excel file crossing-MCL-AI.xls). Furthermore, 46 additional MCL groups were represented in one single other species and in 45 of the 46 cases, the only non-*Panagrolaimus* species was *Acrobeloides nanus*, a species poorly represented in the NCBI's NR database (only 4 protein sequences as of November 2014) and one of the closest relatives of *Panagrolaimus* species in our set of compared nematode genomes. The 46th case was one sequence present in 6 copies in *B. xylophilus* and present in 7 different *Panagrolaimus* species. In contrast, 7 MCL groups had representatives in 2 or more non-*Panagrolaimus* species and were discarded from the analysis. (See Supplementary Excel file crossing-2.xls). Thus, in total, 135 MCL groups contain *Panagrolaimus* proteins that all have AI >30 when data available and are either Pana-specific or shared with only one close outgroup.

A total of 26 of these MCL groups contained only one *Panagrolaimus* species and thus either represented species-specific HGT events or more ancient HGT that were secondarily lost (but in the absence of strong evidence for conservation in other *Panagrolaimus*, they were considered as

species-specific). Finally, 109 MCL groups were present in at least 2 *Panagrolaimus* species and were more likely acquired in an ancestor of the corresponding species.

Using Mesquite, we reconstructed the putative timing of acquisitions via HGT for the 109 MCL families and traced back the history on a phylogenetic tree (Figure 4). Using a parsimony approach, we found that 102 HGT events took place in an ancestor of 2 or more *Panagrolaimus* species. This corresponded to 101 different MCL families (one family, *Pana_OMCL20286* has been acquired twice independently at nodes 9 and 14, according to the parsimony model). Only seven families were considered as acquired in an ancestor of all the 8 *Panagrolaimus* species included in this analysis, of which 5 were already present in an ancestor with their close outgroups (2 were novel acquisitions at node 4 = common ancestral node of *Panagrolaimus* species). The highest number of acquisitions took place at node number 6, with 49 HGT events occurring there. In contrast, 8 MCL families were considered as species-specific multiple independent acquisitions (represented by bars in Figure 4). All the results are summarized and available in table all-mesquite-analysis-and-groups.xlsx.

Putative functions of horizontally-acquired genes

For each *Panagrolaimus* species scanned with Alieness, we retrieved Pfam domains that had been predicted in proteins considered as originating from HGT. A total of 192 different Pfam domains were found among the 910 proteins with AI >30 in the 6 *Pana* species. They were distributed as follows in the table below:

Abb name	Full name	# prots	AI >30	# Pfam	uniq Pfam
PSU	<i>Panagrolaimus superbus</i>	21381	88	91	56
ES5	<i>Panagrolaimus sp.</i> ES5	24756	232	429	136
PS1159	<i>Panagrolaimus sp.</i> PS1159	27350	182	268	93
PDA	<i>Panagrolaimus davidi</i>	48215	342	465	103
JU76	<i>Propanagrolaimus sp.</i> JU765	24878	44	100	23
PRED	<i>Panagrellus redivivus</i>	26372	22	18	12

Table 4: Pfams found for HGT candidates.

Using a Venn diagram analysis, we identified Pfam domains that were conserved in the set of HGT proteins of several different *Pana* species (See Supplementary Excel file Comp-Pfam-HGT.xlsx). We found 3 Pfam domains present in the HGT sets of all the 6 *Panagrolaimus* species, (Alcohol dehydrogenase GroES-like domain, Pyridine nucleotide-disulphide oxidoreductase and Zinc-binding dehydrogenase), all corresponding to enzymatic functions. Furthermore, 4 other domains were present in 5 of the 6 *Panagrolaimus* species studied. Interestingly, a group of 30 domains was conserved in *P. superbus*, ES5, *P. davidi* and PS1159. This also mainly corresponded to enzymatic functions, including glycoside hydrolases such as

GH43, GH28 and GH32, already described as acquired via HGT in other nematodes (Haegeman et al., 2011).

One intriguing gene, which we found is a photolyase. Photolyases are ancient proteins capable of repairing DNA from UV induced damage (Sancar, 1990) and are thought to be lost in various branches of metazoa during (Lucas-Lledo & Lynch, 2009). In our data we find two different photolyases, one in the *Panagrolaimus* species and a different in the distant outgroups *A. sum* and *A. nanus*, while there is none in the closest outgroups *Propanagrolaimus* or *P. redivivus* (Figure 5). This phylogenetic pattern would argue for an independent gain, not loss of the gene. Analysing the gene structure in PS1159 we find the enzyme on a contig with 5 eukaryotic genes and to possess 2 introns, showing a full integration into the genome.

We also analyzed putative functions of genes acquired via HGT, using the gene ontology as described in methods. At the molecular function level, frequently found GO terms corresponded to enzymatic activities (e.g. oxidoreductase activity, hydrolase activity acting on glycosyl bonds, methyltransferase activity) and this is consistent with Pfam domains identified in HGT proteins. The terms "DNA binding" as well as "Ion binding" were also frequently present but this corresponded to enzymes that depend on ATP or ion binding for their activities. At the biological process point of view, the most conserved and highly frequent terms were ("carbohydrate metabolic process" and "cellular nitrogen compound metabolic process"), again illustrating enzymatic activities as the functions of the most frequently transferred genes.

Anhydrobiosis specific genes

Apart from our overrepresentation/classification analysis based on functional annotation we also directly screened for genes known to be relevant for cryptobiosis. Small Heat Shock Proteins (sHSPs) are known to conduct chaperone functions in anhydriobiotic animals (Gusev, Cornette, Kikawada, & Okuda, 2011). Our phylogenetic analyses of the sHSP indicates that the *Panagrolaimid* species possess an additional group of these genes, which is divergent from the ones which underwent a lineage specific expansion in *C. elegans* (Aevermann & Waters, 2008). However, we do not find a *Panagrolaimus* specific set of sHSP genes, all proteins detected by us appear to have orthologs in other clade IV species, like *B. xylophilus*.

Discussion

Organisms have to adapt to environmental challenges to survive in evolution and nematodes are successful in exploiting a large variety to extreme environments. The genome sequence of *Panagrellus redivivus*, a member of the Panagrolaimidae and used as an outgroup to *Panagrolaimus* by us, had already revealed some gene family expansions potentially giving and adaptive advantage for these free-living roundworms (Srinivasan et al., 2013). Our three new draft genomes of amphimictic and parthenogenetic *Panagrolaimus* species combined with an improved re-assembly of the parthenogenetic *P. davidi*, and novel RNASeq transcriptomes of two further parthenogenetic strains in comparison with one new draft genome from a

hermaphrodite *Propanagrolaimus* provided us with a platform to explore details of the evolution of several processes.

Genomes, Genes and Phylogenetics

Our draft genomes are, typical for many second-generation approaches, still fragmented in part. Especially the *P. davidi* genome has many short contigs and a high number of gene predictions, similarly in number with the RNASeq predicted ORFs in PS1579 and DL137, which we attribute to large heterozygosity in the sequenced *P. davidi* populations. Heterozygosity in the species was potentially increased by shuffling of the genomes after cryobiosis (Flot et al., 2013) in the long standing laboratory strain. For the three other *Panagrolaimus* and the *Propanagrolaimus* genome we were able to generate robust gene predictions with reasonable gene numbers (enhanced by our RNASeq data) of ~23,000 to ~26,000 genes. The *Panagrolaimus* genomes, as well as *Propanagrolaimus* and *Panagrellus redivivus* (Srinivasan et al., 2013) appear to be comparatively poor in repeat sequences with 6% - 9% of the genomes being repetitive sequences.

Our phylogenetic analysis based on ~400 genes robustly re-confirmed previous results on the position of individual species, which were based on much less information (Lewis et al., 2009; Schiffer et al., 2014). The obtained phylogeny thus shows that parthenogenesis arose once in *Panagrolaimus* and that *Panagrellus redivivus* is indeed closer to *Panagrolaimus*, than *Propanagrolaimus* is. The tree will be useful to map further strains and species to the taxon. In contrast to the phylogeny, the K_r measure shows more divergence between *P. redivivus* and the *Panagrolaimus* species, than the *Propanagrolaimus*. However, this might be explained by the full genome sequences being used for the calculation and the measure itself becoming unreliable at higher values. Interestingly the K_r values do indicate that both sexual species (*P. superbus* and ES5) are closer than the two parthenogenetic species tested (*P. davidi* and PS1159). The larger divergence between the two parthenogenetic species might however also be interpreted as an indication for the possible hybrid origin of these.

Origins of Parthenogenesis

We argue that our chromosome counts, genome size measurements, the obtained genomic variant and read coverage signatures speak in favour of a hybrid origin of parthenogenesis in *Panagrolaimus*, including polyploidisation of the parthenogens. It is well known that many parthenogenetic species are hybrids and polyploid, in particular nematodes in clade IV (Castagnone-Sereno & Danchin, 2014; Lunt et al., 2014). While many triploid parthenogenetic species are apomictic (Simon et al., 2003), some automictic (meiotic) cases are known (Avisé, 2008). In meiosis, which we assume to be present in *Panagrolaimus* due to the presence of one polar body, reducing a triploid set of chromosomes to one and then re-creating the triploid stage in all oocytes appears difficult. However, it is known that hybridisation leads to aberrant meiotic process that could favour parthenogenesis (Avisé, 2008) and there are possible routes to regain triploidy, for example by incorporating a round of chromosomal endoreplication (Lutes, Neaves, Baumann, Wiegraabe, & Baumann, 2010), which could lead to

oocytes containing three sets of chromatids, see (Awise, 2008) and chapter 4.2 in (Schön et al., 2009) for illustrations. Thus, while future laboratory studies, for example incorporation fluorescent in situ hybridisation (FISH), need to be conducted to resolve the exact cytological mechanism, we nevertheless suggest that our combined data speak in favour of a triploid genomic state in the parthenogenetic *Panagrolaimus* species.

DSD and GRNs

It appears that the re-shuffling of GRNs through DSD leads to considerable differences in the molecular toolkit of development in nematodes. While we had previously seen this when comparing species across large phylogenetic distances (the clade V model species *C. elegans* and the clade I species *R. culicivora*) (Schiffer et al., 2013) and had found some indications for divergence between *Panagrolaimid* species, in this assay we surprisingly found a large number of genes important in *C. elegans* missing from the *Panagrolaimids* and other clade IV species. As can best be seen from the data on the endo-/mesoderm pathway the missing genes are mainly acting at intermediate levels, thus indicating evolutionary processes as brought forward by Davidson (Davidson, 2006). Davidson proposed that exactly such intermediate switches are most liable to (fast) evolutionary change. For the gut development it is interesting to note that some data speak in favour of a *skn-1* and *med-1/2* independent pathway in *C. elegans* (Goszczyński, 2005), illustrated in Figure 3. Lacking also the *end-1/3* genes it is well possible that in clade IV species the whole cascade is different and other factors activate the focal *elt-2* gene, which acts on a large variety of other genes (McGhee et al., 2009). Notably, our previous finding that *skn-1* appears more weakly expressed outside the germline than in *C. elegans* could be in line with this (Schiffer et al., 2014). As wnt-signalling is acting in the *C. elegans* endo-mesoderm defining pathway upstream of the *end* genes and *tbx-35* (Lin, Broitman-Maduro, Hung, Cervantes, & Maduro, 2009) it is for example possible that this signalling cascade acts in species outside of *Caenorhabditis*.

In the light of these findings observing the expression of one sperm specific *C. elegans* gene (*spe-41*) in the parthenogenetic species cannot be taken as direct cue to launch analysis into egg activation in parthenogenetic species. However, the gene is a calcium channel and crucial for fertilisation in *C. elegans* (Xu & Sternberg, 2003). Calcium channels are generally important for fertilisation in animals (Stricker, 1999). Thus, finding the *spe-41*, as well as the major sperm proteins, which we also find expressed in the parthenogenetic species and which had been reported functional in other parthenogenetic nematodes (“Evolution of Gene Regulatory Networks Controlling Body Plan Development,” 2011) makes some functional connection at least likely. What is more, the expression of ‘male-genes’ in all female lines could argue for a cis-regulatory effects in the establishment of parthenogenesis, which would again be in line with theories on GRN evolution (Davidson, 2006). In particular in a hybrid system, even more so in a polyploid one, problems in dosage compensation and chromatin based silencing could lead to expression ‘leakage’ of such ‘male-genes’.

Anhydrobiosis and HGT

Extending from previous work on cryobiosis in *P. davidi* alone (Thorne et al., 2014) we explored the genetic background of anhydrobiosis in *Panagrolaimus*. Our study is especially powerful as it integrates data from the close outgroup *Propanagrolaimus*, which had been found much less amenable to desiccation, than the *Panagrolaimus* species (A. Shannon et al., 2005).

Our phylogenetic analysis did not reveal a lineage specific expansion in sHSPs genes in *Panagrolaimus*, which could have been assumed for a family of proteins important in anhydrobiotic animals (Gusev et al., 2011). To the contrary, the presence of a lineage specific expansion of these genes in *C. elegans* and *P. pacificus*, both not capable of anhydrobiosis, argues against a direct adaptive link of these genes to the process. Rather, it can be assumed that sHSPs are activated during desiccation as part of the normal cellular stress-response. Our overrepresentation/classification approach revealed several other gene families that might be inflated in *Panagrolaimus* (Table 2) and directly linked to protein stability and degradation and thus to anhydrobiosis. In the chironomid *Polypedilum vanderplanki*, genes annotated as Protein kinase, Proteasome/Ubiquitin, Protease inhibitor, Protease, LEA, DNA repair, Oxidative stress, and HSP have been found upregulated during desiccation and rehydration (Cornette et al., 2010). Finding similarly annotated protein domains to be more important in classifying and overrepresented in *Panagrolaimus*, in comparison to *Panagrellus* and *Propanagrolaimus* reveals an important link for functional studies into anhydrobiosis in these nematodes. One class of heat shock proteins, HSP-70, have been found active in *P. vanderplanki* larvae during desiccation (Gusev et al., 2011) and to protect (re-)hydrated Tardigrades against irradiation (Jönsson & Schill, 2007). We were intrigued by finding this class of genes to be important in classifying *Panagrolaimus* and by the link to DNA repair and genome integrity as other important Pfam domains we found also pointed towards the process. It has recently been appreciated how crucial DNA repair is after desiccation, (Flot et al., 2013; Hespels et al., 2014). However, we also appear to find that this process has led to the gain of important genes favouring anhydrobiosis in the *Panagrolaimus* species. Most prominently the photolyase detected by us will help species to repair genomes damaged by irradiation, which, as exemplified by the HSP-70s (Jönsson & Schill, 2007), appears important. Additionally, GO terms like “oxidoreductase activity”, “methyltransferase activity”, and “DNA binding”, which were associated with genes acquired through HGT do also point towards a positive feedback-loop between anhydrobiosis and the horizontal acquisition of genetic material.

In this work we have shown that parthenogenetic *Panagrolaimus* nematodes are likely polyploid hybrids, which adds further evidence to the theory that parthenogenesis might be an escape route for incompatible hybrids. While further systems will need to be analysed to estimate how common this process is, our genomic data already now allows to analyse the genetics and genomics of parthenogenesis, particularly as *Panagrolaimus* nematodes are easily cultured, assessable to molecular imaging techniques and appear amenable to gene knockouts via RNAi. However, our study on the GRNs of development and related processes display the great

disparity between the model organisms in Nematoda. Thus, our findings call for an analysis of early development in many genera across the phylum to evaluate which processes are really orchestrated by orthologues and which are under divergent control. Such analysis, can also be started in *Panagrolaimus* in clade IV and then extended to other genera. Finally we found an intriguing link between HGT and the evolution of anhydrobiosis in *Panagrolaimus*. These data illustrate how important the process really might be for metazoans in acquiring new adaptive potential and might lend further support to the hypothesis of a common gene pool for life (McInerney, Pisani, Baptiste, & O'Connell, 2011).

Figure Legends

Figure 1: Shown is a RAxML tree of *Panagrolaimus*, *Panagrellus*, *Propanagrolaimus*, and two outgroup species based on CEGMA KOGs (see main text). *Panagrellus* appears as closer outgroup to *Panagrolaimus*, than *Propanagrolaimus* and the single origin of parthenogenesis in *Panagrolaimus* is re-confirmed. We found most KOGs in at least 8 of the species and the newly sequenced species to have at least 394 of the KOGs, with 4 out of 5 having more than 430. This illustrates the completeness of our draft genome assemblies.

Figure 2: A: Depicted in is the occurrence of each variant frequency observed when mapping all sequencing data against the predicted gene sets of the respective species. The dashed line gives the 1/3 mark, at which a variant at a given position is observed in 1/3 of the mapped reads, while 2/3 of the reads support the genomics reference position. The parthenogenetic frequencies appear to be shifted towards the 1/3 value, most clearly seen in PS1159. The highly fragmented *P. davidi* genome peaks at even smaller values. For PS1159 and the amphimictic species ES5 karyotypes are displayed, with 12 and 8 Chromosomes respectively. **B:** A second peak in overall genome coverage was observed in the PS1159 data, while distributions for the other species did not show this (only JU765 shown, see supplement for others).

Figure 3: Crucial genes from the *C. elegans* endomesoderm pathway were not retrieved in the *Panagrolaimus* and other clade IV species. Especially genes acting as intermediate switches were missing, raising the possibility that an ancestral pathway without the involvement of these genes is at place, as was suggested for *Caenorhabditis* (Goszczynski, 2005). *C. elegans* has a *skn-1* co-orthologue, which was not found in the *Panagrolaimus* species.

Figure 4: Analysis of HGT gain/loss events in the lineage leading to *Panagrolaimus* based on a parsimony analysis with Mesquite.

Figure 5: A: An alignment of a photolyase gene in several *Panagrolaimus* species, as well as of a different photolyase found in *A. suum* and *A. nanus*. The gene was present in all *Panagrolaimus* species, but only a subset is shown for clarity. The gene is not present in the closest outgroup and the divergent photolyase found in the remote outgroups makes a gain by

LGT more likely than independent loss. **B:** The gene structure in PS1159 with three exons, and the position on a contig with 5 eukaryotic genes corroborates the LGT hypothesis. **C:** Interproscan derived PFam and ProSite data clearly identify the crucial photolyase domains.

Figure 6: Depicted is a neighbour net of sHPS genes from *C. elegans*, *P. pacificus* and other clade IV species. While lineage specific expansions present in *C. elegans* are not observed in *Panagrolaimus*, it becomes clear that *Panagrolaimus* and other clade iv species possess a set of sHPS genes not shared with either of the model organisms.

References

- Aevermann, B. D., & Waters, E. R. (2008). A comparative genomic analysis of the small heat shock proteins in *Caenorhabditis elegans* and *briggssae*. *Genetica*, *133*(3), 307–319. doi:10.1007/s10709-007-9215-9
- Alpert, P. (2006). Constraints of tolerance: why are desiccation-tolerant organisms so small or rare? *Journal Of Experimental Biology*, *209*(Pt 9), 1575–1584. doi:10.1242/jeb.02179
- Altschul, S., Madden, T., & Schäffer, A. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids ...*, *25*(17), 3389–3402.
- Andrássy, I. (2005). *Free-Living Nematodes of Hungary* (1st ed.). Budapest.
- Angiuoli, S. V., & Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, *27*(3), 334–342. doi:10.1093/bioinformatics/btq665
- Avise, J. (2008). *Clonality : The Genetics, Ecology, and Evolution of Sexual Abstinence in Vertebrate Animals*. Oxford University Press.
- Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, *27*(12), 1691–1692. doi:10.1093/bioinformatics/btr174
- Becks, L., & Agrawal, A. F. (2010). Higher rates of sex evolve in spatially heterogeneous environments. *Nature*, *468*(7320), 89–92. doi:10.1038/nature09449
- Becks, L., & Agrawal, A. F. (2012). The evolution of sex is favoured during adaptation to new environments. *PLoS Biology*, *10*(5), e1001317. doi:10.1371/journal.pbio.1001317
- Blaxter, M. (2011). Nematodes: The Worm and Its Relatives. *PLoS Biology*, *9*(4), e1001050.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170
- Borgonie, G., García-Moyano, A., Litthauer, D., Bert, W., Bester, A., van Heerden, E., et al. (2011). Nematoda from the terrestrial deep subsurface of South Africa. *Nature*, *474*(7349), 79–82. doi:10.1038/nature09974
- Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1777), 20132450. doi:10.1098/rspb.2013.2450
- Brown, C., Howe, A., Zhang, Q., & Pyrkosz, A. (2012). A single pass approach to reducing sampling variation, removing errors, and scaling `{\ em de novo}` assembly of shotgun sequences. *arXiv.org*.

- Burnell, A. M., & Tunnacliffe, A. (2011). Gene induction and desiccation stress in nematodes. *Molecular and physiological basis of ...*
- Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. doi:10.1093/bioinformatics/btp348
- Castagnone-Sereno, P., & Danchin, E. G. J. (2014). Parasitic success without sex - the nematode experience. *Journal of Evolutionary Biology*, *27*(7), 1323–1333. doi:10.1111/jeb.12337
- Cohen, J. (2001). The Earth Is Round ($p < .05$), 1–7.
- Convey, P., & Worland, M. R. (2000). *Google Scholar*. Cryo-Letters.
- Cornette, R., Kanamori, Y., Watanabe, M., Nakahara, Y., Gusev, O., Mitsumasu, K., et al. (2010). Identification of Anhydrobiosis-related Genes from an Expressed Sequence Tag Database in the Cryptobiotic Midge *Polypedilum vanderplanki* (Diptera; Chironomidae). *The Journal of biological chemistry*, *285*(46), 35889–35899. doi:10.1074/jbc.M110.150623
- Creevey, C. J., & McInerney, J. O. (2003). CRANN: detecting adaptive evolution in protein-coding DNA sequences. *Bioinformatics*, *19*(13), 1726–1726. doi:10.1093/bioinformatics/btg225
- Curran, D. M., Gilleard, J. S., & Wasmuth, J. D. (2014). Figmap: a profile HMM to identify genes and bypass troublesome gene models in draft genomes. *Bioinformatics*, *30*(22), 3266–3267. doi:10.1093/bioinformatics/btu544
- Cutter, A. D. (2008). Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Molecular Biology And Evolution*, *25*(4), 778–786. doi:10.1093/molbev/msn024
- Danchin, E. G. J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P. M., Henrissat, B., & Abad, P. (2010). Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *107*(41), 17651–17656. doi:10.1073/pnas.1008486107
- Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, *27*(8), 1164–1165. doi:10.1093/bioinformatics/btr088
- Demeure, Y., Freckman, D. W., & Van Gundy, S. D. (1979). Anhydrobiotic coiling of nematodes in soil. *Journal Of Nematology*, *11*(2), 189–195.
- Denver, D. R., Clark, K. A., & Raboin, M. J. (2011). Reproductive mode evolution in nematodes: insights from molecular phylogenies and recently discovered species. *Molecular Phylogenetics And Evolution*, *61*(2), 584–592. doi:10.1016/j.ympev.2011.07.007
- Dieterich, C., Clifton, S. W., Schuster, L. N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., et al. (2008). The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nature Genetics*, *40*(10), 1193–1198. doi:10.1038/ng.227
- Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008, January 14). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. BioMed Central Ltd. doi:10.1186/1471-2105-9-18
- Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Alioto, T., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, *10*(12), 1185–1191. doi:10.1038/nmeth.2722
- Evolution of Gene Regulatory Networks Controlling Body Plan Development. (2011). Evolution of Gene Regulatory Networks Controlling Body Plan Development, *144*(6), 970–985. doi:10.1016/j.cell.2011.02.017

- Farrant, J. M., & Moore, J. P. (2011). Programming desiccation-tolerance: from plants to seeds to resurrection plants. *Current opinion in plant biology*, *14*(3), 340–345. doi:10.1016/j.pbi.2011.03.018
- Felsenstein, J. (1974). The Evolutionary Advantage of Recombination. *Genetics*, *78*, 737–757.
- Flot, J.-F., Hespels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E. G. J., et al. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature*, *500*(7463), 453–457. doi:10.1038/nature12326
- Galtier, N., Gouy, M., & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*.
- Gladyshev, E. A., Meselson, M., & Arkhipova, I. R. (2008). Massive horizontal gene transfer in bdelloid rotifers. *Science*, *320*(5880), 1210–1213. doi:10.1126/science.1156407
- Goldstein, B., & Hird, S. N. (1996). Specification of the anteroposterior axis in *Caenorhabditis elegans*. *Development*, *122*(5), 1467–1474.
- Goszczyński, B. (2005). Reevaluation of the Role of the med-1 and med-2 Genes in Specifying the *Caenorhabditis elegans* Endoderm. *Genetics*, *171*(2), 545–555. doi:10.1534/genetics.105.044909
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. doi:10.1038/nbt.1883
- Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(3), 645–656. doi:10.1109/TCBB.2013.68
- Gusev, O., Cornette, R., Kikawada, T., & Okuda, T. (2011). Expression of heat shock protein-coding genes associated with anhydrobiosis in an African chironomid *Polypedilum vanderplanki*. *Cell Stress and Chaperones*, *16*(1), 81–90. doi:10.1007/s12192-010-0223-9
- Haegeman, A., Jones, J. T., & Danchin, E. G. J. (2011). Horizontal gene transfer in nematodes: a catalyst for plant parasitism? *Molecular plant-microbe interactions : MPMI*, *24*(8), 879–887. doi:10.1094/MPMI-03-11-0055
- Haubold, B., Pfaffelhuber, P., o, M. D.-L., & Wiehe, T. (2009). Estimating Mutation Distances from Unaligned Genomes. *dx.doi.org*, *16*(10), 1487–1500. doi:10.1089/cmb.2009.0106
- Hespels, B., Knapen, M., Hanot-Mambres, D., Heuskin, A. C., Pineux, F., Lucas, S., et al. (2014). Gateway to genetic exchange? DNA double-strand breaks in the bdelloid rotifer *Adineta vaga* submitted to desiccation. *Journal of Evolutionary Biology*, *27*(7), 1334–1345. doi:10.1111/jeb.12326
- Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., et al. (2012). HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Research*, *22*(8), 1581–1588. doi:10.1101/gr.133652.111
- Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013, May 27). REAPR: a universal tool for genome assembly evaluation. *Genome Biology*. BioMed Central Ltd. doi:10.1186/gb-2013-14-5-r47
- Hunt, P. (2011). OrthoList: a compendium of *C. elegans* genes with human orthologs. *PLoS ONE*, *6*(5), e20085. doi:10.1371/journal.pone.0020085
- Huson, D. (2008). Drawing explicit phylogenetic networks and their integration into SplitsTree. *BMC Evolutionary Biology*.
- Jasmer, D. P., Govere, A., & Smant, G. (2003). P ARASITICN EMATODEI NTERACTIONS

- WITHM AMMALS ANDP LANTS. *Annual Review of Phytopathology*, 41(1), 245–270.
doi:10.1146/annurev.phyto.41.052102.104023
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240.
doi:10.1093/bioinformatics/btu031
- Jönsson, K. I., & Schill, R. O. (2007). Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade *Richtersius coronifer*. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology*, 146(4), 456–460.
doi:10.1016/j.cbpb.2006.10.111
- Kapitonov, V. V., & Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 98(15), 8714–8719.
doi:10.1073/pnas.151269298
- Kearney, M. (2005). Hybridization, glaciation and geographical parthenogenesis. *Trends In Ecology & Evolution*, 20(9), 495–502. doi:10.1016/j.tree.2005.06.005
- Kent, W. (2002). BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4), 656.
- Kipreos, E. T., & Pagano, M. (2000). The F-box protein family. *Genome Biology*, 1(5), reviews3002.1–3002.7. doi:10.1186/gb-2000-1-5-reviews3002
- Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3), 568–576. doi:10.1101/gr.129684.111
- Koutsovoulos, G., Makepeace, B., Tanya, V. N., & Blaxter, M. (2014). Palaeosymbiosis revealed by genomic fossils of *Wolbachia* in a strongyloidean nematode. *PLoS Genetics*, 10(6), e1004397. doi:10.1371/journal.pgen.1004397
- Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., et al. (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biology*, 14(6), R66. doi:10.1186/gb-2013-14-6-r66
- Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., & Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in genetics*, 4, 237. doi:10.3389/fgene.2013.00237
- Lartillot, N., Blanquart, S., & Lepage, T. (2011). *PhyloBayes 3.3a Bayesian software for phylogenetic reconstruction and molecular dating using mixture models* (pp. 1–44).
- Lewis, S., Dyal, L., Hilburn, C., Weitz, S., Liau, W., LaMunyon, C., & Denver, D. (2009). Molecular evolution in *Panagrolaimus* nematodes: origins of parthenogenesis, hermaphroditism and the Antarctic species *P. davidi*. *BMC Evolutionary Biology*, 9(1), 15.
- Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658–1659.
doi:10.1093/bioinformatics/btl158
- Lin, K. T.-H., Broitman-Maduro, G., Hung, W. W. K., Cervantes, S., & Maduro, M. F. (2009). Knockdown of SKN-1 and the Wnt effector TCF/POP-1 reveals differences in endomesoderm specification in *C. briggsae* as compared with *C. elegans*. *Developmental Biology*, 325(1), 296–306. doi:10.1016/j.ydbio.2008.10.001
- Linard, B., Allot, A., Schneider, R., Morel, C., Ripp, R., Bigler, M., et al. (2014). OrthoInspector 2.0: software and database updates. *Bioinformatics*, btu642.
doi:10.1093/bioinformatics/btu642
- Linard, B., Thompson, J. D., Poch, O., & Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12(1), 11.

- doi:10.1186/1471-2105-12-11
- Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*(5), 0955–0964.
- Lucas-Lledo, J. I., & Lynch, M. (2009). Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Molecular Biology And Evolution*, *26*(5), 1143–1153. doi:10.1093/molbev/msp029
- Lunt, D. H., Kumar, S., Koutsovoulos, G., & Blaxter, M. L. (2014). The complex hybrid origins of the root knot nematodes revealed through comparative genomics. *PeerJ*, *2*(8), e356. doi:10.7717/peerj.356
- Lutes, A. A., Neaves, W. B., Baumann, D. P., Wiegraebe, W., & Baumann, P. (2010). Sister chromosome pairing maintains heterozygosity in parthenogenetic lizards. *Nature*, *464*(7286), 283–286. doi:10.1038/nature08818
- Maddison, W. P., & Maddison, D. R. (n.d.). Mesquite: a modular system for evolutionary analysis. Retrieved from <http://mesquiteproject.wikispaces.com/>
- Maduro, M. F., & Rothman, J. H. (2002). Making worm guts: the gene regulatory network of the *Caenorhabditis elegans* endoderm. *Developmental Biology*, *246*(1), 68–85.
- McCarthy, F. M., Wang, N., Magee, G. B., Nanduri, B., Lawrence, M. L., Camon, E. B., et al. (2006). AgBase: a functional genomics resource for agriculture. *BMC Genomics*, *7*(1), 229. doi:10.1186/1471-2164-7-229
- McGhee, J. D., Fukushige, T., Krause, M. W., Minnema, S. E., Goszczynski, B., Gaudet, J., et al. (2009). ELT-2 is the predominant transcription factor controlling differentiation and function of the *C. elegans* intestine, from embryo to adult. *Developmental Biology*, *327*(2), 551–565. doi:10.1016/j.ydbio.2008.11.034
- McInerney, J. O. (1998). GCUA: general codon usage analysis. *Bioinformatics*, *14*(4), 372–373. doi:10.1093/bioinformatics/14.4.372
- McInerney, J. O., Pisani, D., Baptiste, E., & O'Connell, M. J. (2011). The public goods hypothesis for the evolution of life on Earth. *Biology Direct*, *6*(1), 41. doi:10.1186/1745-6150-6-41
- Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., & Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, *38*(Database issue), D204–10. doi:10.1093/nar/gkp1019
- Mitreva, M., Jasmer, D. P., Zarlenga, D. S., Wang, Z., Abubucker, S., Martin, J., et al. (2011). The draft genome of the parasitic nematode *Trichinella spiralis*. *Nature Genetics*, *43*(3), 228–235. doi:10.1038/ng.769
- Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *1*(1), 2–9. doi:10.1016/0027-5107(64)90047-8
- Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. doi:10.1093/bioinformatics/btm071
- Pérez, F., & Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, *9*(3), 21–29. doi:10.1109/MCSE.2007.53
- PNAS-1994-Shang-8373-7. (2011). PNAS-1994-Shang-8373-7, 1–5.
- Proctor, M. C. F., Oliver, M. J., Wood, A. J., Alpert, P., Stark, L. R., Cleavitt, N. L., & Mishler,

- B. D. (2007). Desiccation-tolerance in bryophytes: a review. *The Bryologist*, 110(4), 595–621. doi:10.1639/0007-2745(2007)110[595:DIBAR]2.0.CO;2
- Punta, M., Coghill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Bournsnel, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40(Database issue), D290–301. doi:10.1093/nar/gkr1065
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. doi:10.1093/bioinformatics/btq033
- Raffi V Aroian, L. C. I. K., & Sternberg, P. W. (1993). A Free-living Panagrolaimus sp. from Armenia Can Survive in Anhydrobiosis for 8.7 Years. *Journal of Nematology*, 25(3), 500.
- Rancurel, C., Da Rocha, M., & Danchin, E. G. J. (2014, September). Alieness: rapid detection of horizontal gene transfers in metazoan genomes. *European Conference on Computational Biology*. Retrieved from <http://fl1000.com/posters/browse/summary/1096828>.
- Ricci, C., & Pagani, M. (1997). Desiccation of Panagrolaimus rigidus (Nematoda): survival, reproduction and the influence on the internal clock. *Hydrobiologia*, 347(1/3), 1–13. doi:10.1023/A:1002979522816
- Runft, L. L., Jaffe, L. A., & Mehlmann, L. M. (2002). Egg Activation at Fertilization: Where It All Begins. *Developmental Biology*, 245(2), 237–254. doi:10.1006/dbio.2002.0600
- Ryan, J. F. (2013, September 9). Baa.pl: A tool to evaluate de novo genome assemblies with RNA transcripts. *arXiv.org*.
- Sancar, G. B. (1990). DNA photolyases: physical properties, action mechanism, and roles in dark repair. *Mutation Research*, 236(2-3), 147–160.
- Schiffer, P. H., Kroiher, M., Kraus, C., Koutsovoulos, G. D., Kumar, S., Camps, J. I. R., et al. (2013). The genome of Romanomermis culicivorax: revealing fundamental changes in the core developmental genetic toolkit in Nematoda. *BMC Genomics*, 14(1), 923. doi:10.1186/1471-2164-14-923
- Schiffer, P. H., Nsah, N. A., Grotehusmann, H., Kroiher, M., Loer, C., & Schierenberg, E. (2014). Developmental variations among Panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit. *Development Genes and Evolution*. doi:10.1007/s00427-014-0471-2
- Schön, I., Martens, K., Dijk, P. J., & van Dijk, P. (2009). *Lost Sex*. Springer Science & Business Media. doi:10.1007/978-90-481-2770-2
- Shannon, A. J., Tyson, T., Dix, I., Boyd, J., & Burnell, A. M. (2008). Systemic RNAi mediated gene silencing in the anhydrobiotic nematode Panagrolaimus superbus. *BMC Molecular Biology*, 9, 58. doi:10.1186/1471-2199-9-58
- Shannon, A., Browne, J., Boyd, J., Fitzpatrick, D., & Burnell, A. (2005). The anhydrobiotic potential and molecular phylogenetics of species and strains of Panagrolaimus (Nematoda, Panagrolaimidae). *Journal Of Experimental Biology*, 208(12), 2433–2445. doi:10.1242/jeb.01629
- Shrestha, R. K., Lubinsky, B., Bansode, V. B., Moiz, M. B., McCormack, G. P., & Travers, S. A. (2014, January 30). QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics*. BioMed Central Ltd. doi:10.1186/1471-2105-15-33
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 1–6. doi:10.1038/msb.2011.75

- Simon, J., Delmotte, F., Rispe, C., & Crease, T. (2003). Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals. *Biological Journal Of The Linnean Society*, 79(1), 151–163.
- Simpson, J. T. (2014). Exploring Genome Characteristics and Sequence Quality Without a Reference. *Bioinformatics*. doi:10.1093/bioinformatics/btu023
- Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 31. doi:10.1186/1471-2105-6-31
- Srinivasan, J., Dillman, A. R., Macchietto, M. G., Heikkinen, L., Lakso, M., Fracchia, K. M., et al. (2013). The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle. *Genetics*, 193(4), 1279–1295. doi:10.1534/genetics.112.148809
- Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21), 2688–2690. doi:10.1093/bioinformatics/btl446
- Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2), ii215–ii225. doi:10.1093/bioinformatics/btg1080
- Stricker, S. A. (1999). Comparative Biology of Calcium Signaling during Fertilization and Egg Activation in Animals. *Developmental Biology*, 211(2), 157–176. doi:10.1006/dbio.1999.9340
- Thorne, M. A. S., Kagoshima, H., Clark, M. S., Marshall, C. J., & Wharton, D. A. (2014). Molecular analysis of the cold tolerant Antarctic nematode, *Panagrolaimus davidi*. *PLoS ONE*, 9(8), e104526. doi:10.1371/journal.pone.0104526
- Van Dongen, S. (2000). A Cluster algorithm for graphs. *Report - Information systems*, (10), 1–40.
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), 1189–1191. doi:10.1093/bioinformatics/btp033
- Wharton, D. A., & Barclay, S. (1993). Anhydrobiosis in the free-living antarctic nematode *Panagrolaimus davidi* (Nematoda): Rhabditida. *Fundamental and applied nematology*, 16(1), 17–22.
- Wharton, D. A., Barrett, J., Goodall, G., Marshall, C. J., & Ramløv, H. (2005). Ice-active proteins from the Antarctic nematode *Panagrolaimus davidi*. *Cryobiology*, 51(2), 198–207. doi:10.1016/j.cryobiol.2005.07.001
- Wharton, D., & Ferns, D. (1995). Survival of intracellular freezing by the Antarctic nematode *Panagrolaimus davidi*. *Journal Of Experimental Biology*, 198(Pt 6), 1381–1387.
- Wilhelm, J., Pingoud, A., & Hahn, M. (2003). Real-time PCR-based method for the estimation of genome sizes. *Nucleic Acids Research*, 31(10), e56.
- Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7), 873–881. doi:10.1093/bioinformatics/btq057
- Xu, X. Z. S., & Sternberg, P. W. (2003). A *C. elegans* Sperm TRP Protein Required for Sperm-Egg Interactions during Fertilization. *Cell*, 114(3), 285–297. doi:10.1016/S0092-8674(03)00565-8
- Zerbino, D., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821.

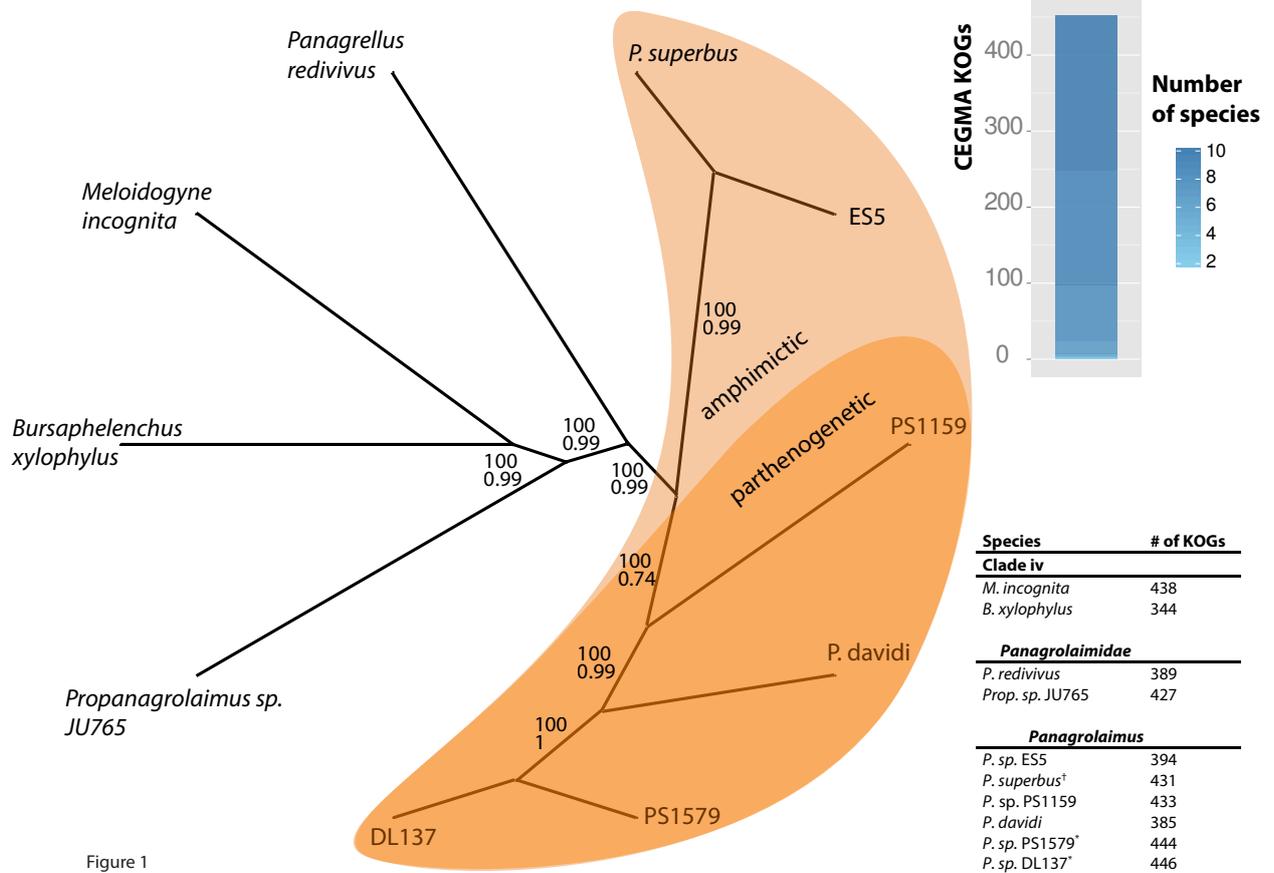


Figure 1

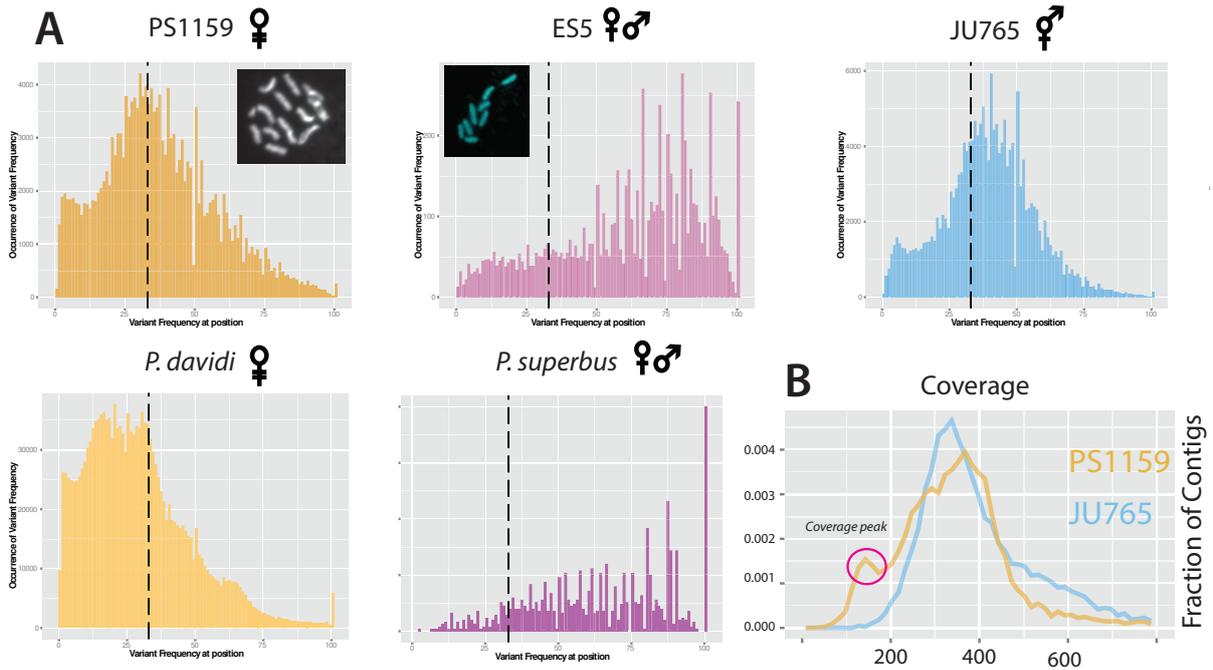


Figure 2

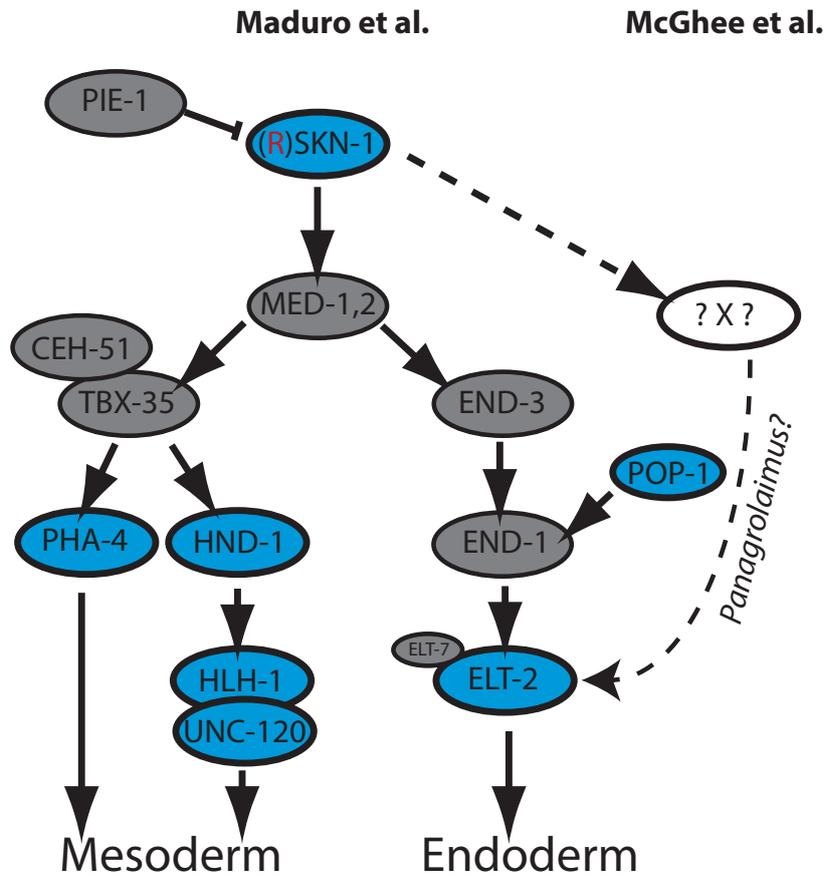


Figure 3

2.2 Extended manuscripts

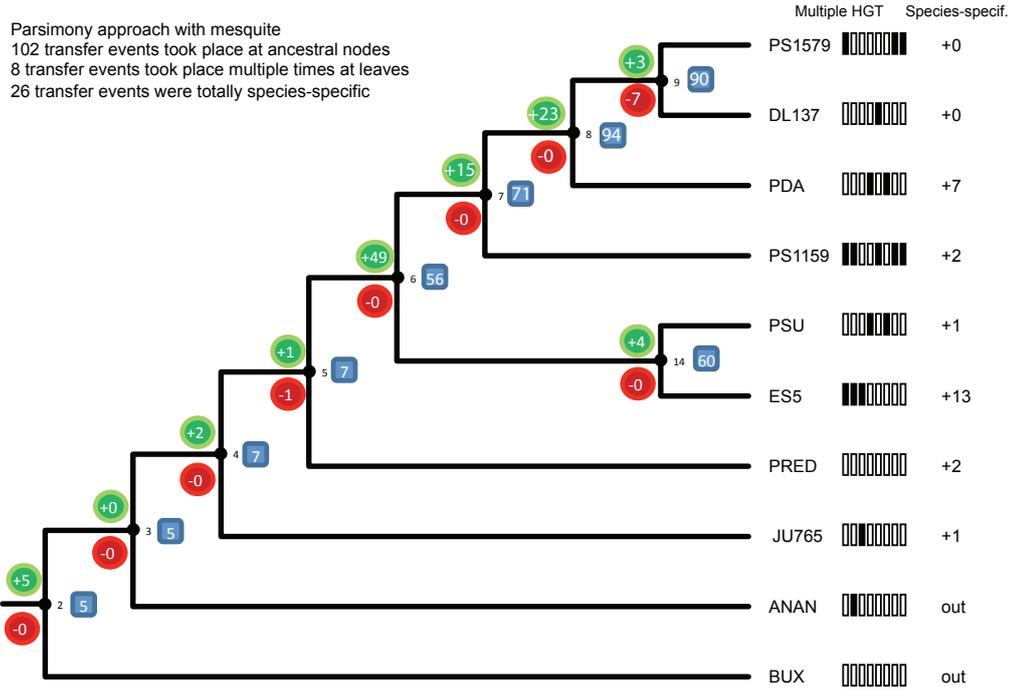


Figure 4

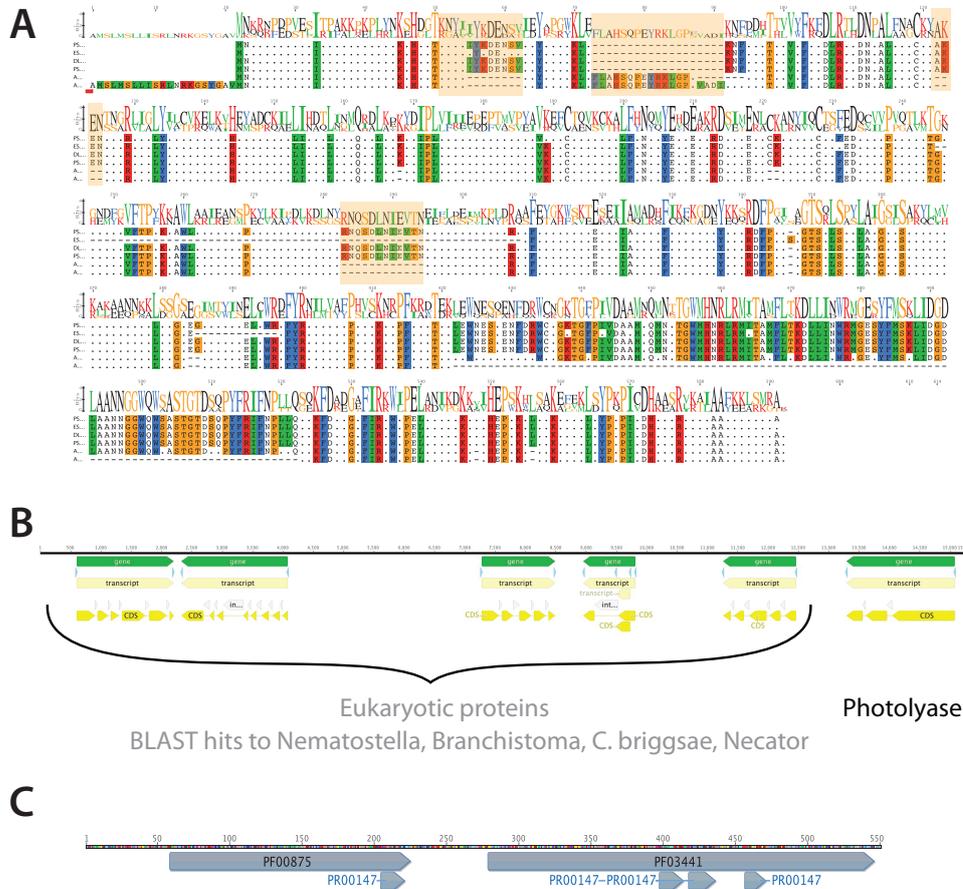


Figure 5

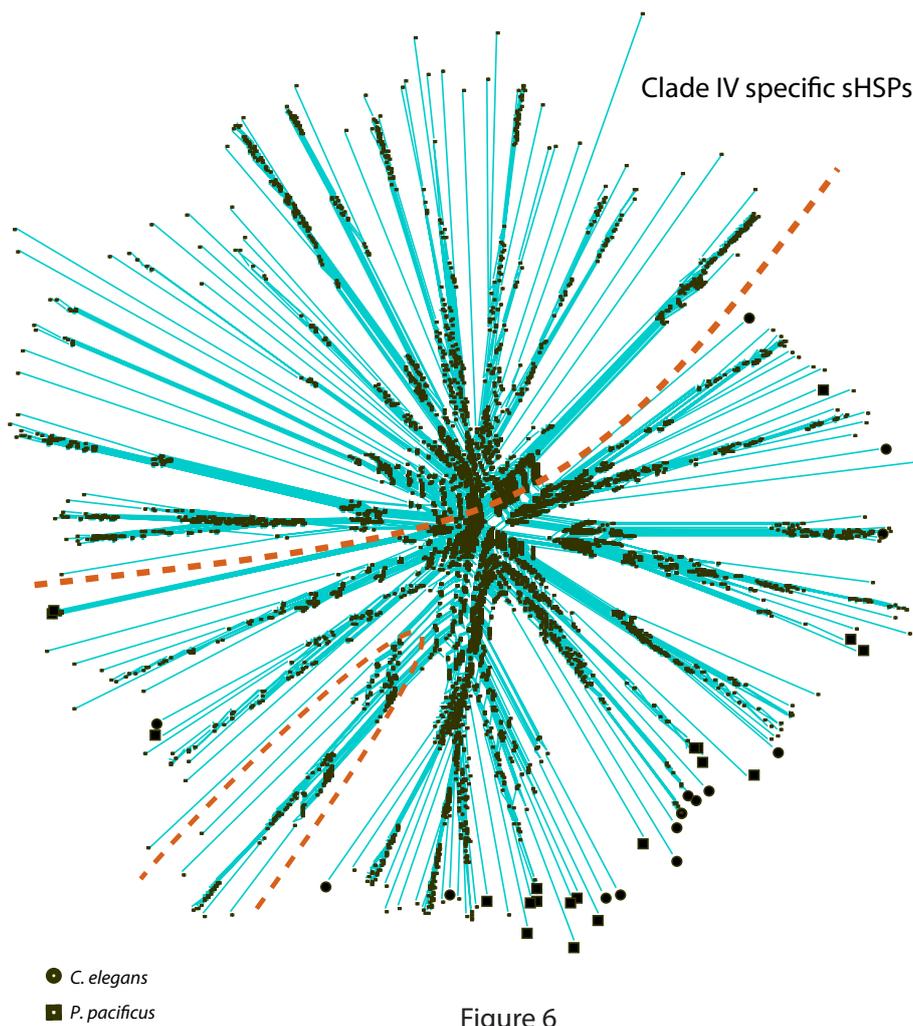


Figure 6

A set of orthologous proteins retained since the Cambrian Explosion drives Bilaterian development today

Andrea Krämer-Eis¹, Luca Ferretti², Philipp Schiffer¹, Peter Heger¹, Thomas Wiehe^{1,*}

1 Institut für Genetik, Universität zu Köln, 50674 Köln, Germany

2 Collège de France et Atelier de Bioinformatique, Université Pierre et Marie Curie, 75005 Paris, France

*** E-mail: Corresponding author@institute.edu**

Abstract

Since their initial radiation in the Cambrian Explosion Bilateria have shown remarkable success in evolution. The phylum has evolved a plethora of traits and body forms allowing these animals to reach ecological dominance in many habitats. Indeed the diversity in Bilateria today is intriguing, especially in contrast to few truly uniting shared morphological features retained across the phylum. It has been shown that the branches of the phylum are also characterized by considerable evolution and turnover in their genomic toolkit, e.g. high gene family birth/death rates. In the light of this divergence we were interested to see if a set of genes uniting Bilateria could be defined and more importantly, if such genes would be functionally related in only remotely related species. Using a conservative pyramidal approach of orthology inference we first employed reciprocal BLASTs to define a set of genes that were at least evolutionary distant from potential non-Bilaterians orthologs. We then narrowed this set to proteins present in all major branches of Bilateria using orthology predictors and functionally characterized the proteins in universally shared clusters. For this we made use of extensive expression data available for three model organisms. We find that some clusters of pan-bilaterian proteins appear to exist. Most exciting we could also find retained functions for the proteins in species development. Some of these molecules appear to act in defined and comparable developmental phases across huge evolutionary distances, when the adult body-plan is constructed. Our observations thus reinforce the idea that Bilateria are united by the last phase of development when the embryo-larva-adult transition gives rise to morphologically distinct adults.

Author Summary

Introduction

Bilateria present by far the largest monophyletic group in the animal kingdom, comprising about 99% of the eumetazoans today [1]. Morphologically the taxon Bilateria has been erected on possessing the name giving bilateral symmetry (at least realized in one life stage) and additional core features, such as triploblasty, an enhanced nervous system, a complex set of cell types forming different tissues [2], all in comparison to species outside the taxon.

However, what might be most striking in terms of defining this group of animals is the plethora of body forms and diverse morphologies, which have evolved since the initial massive radiation in the Cambrian explosion, some 545 million years ago [1]. Based on these morphologies today we count at least 32 bilaterian phyla [3], which conquered almost all habitats on earth. The ur-bilaterian animal has been suggested to have been a small unsegmented animal with a single opened gut, a radial, total cleavage pattern and a potentially regulatory development [4]. Since it seems unlikely that structures like photoreceptor organs, appendages, a heart, and metamerism (found in all larger bilaterian clades) have emerged several times [5,6], more recent phylogenies [7] might argue for an ur-bilaterian with more of the bauplan features seen in modern species. However, it is clear that the descendants of this animal have undergone several reductions and re-modellings of body forms in lineages leading to the modern crown clades. Nematoda for example lost their coelom, while Hexapoda lost some of the appendages found in crustacean [8]. At the same time they gained functional and ecological capacity by (possibly) modifying the gills of their crustacean ancestors into wings [9]. The tetrapod limbs might have gained their temporary climax in the evolution of the human hand, while the evolutionary history of cetaceans reversed this gain of morphological complexity in a body appendage [8].

A similar pattern of abundant gain and loss seems to be present in the bilaterian molecular toolkit. Expansions and deflations of gene families have been reported in all major branches of the phylum (see for example Mitreva et al. [10]). Even such important regulators like the HOX genes, which have been increasing in numbers in the lineage leading to the bilaterian ancestor [11] can be reduced in number as the nematode example shows [?]. It also appears that genes acting for example in development can be readily substituted in different lineages of the phyla [12].

It is notable that the emergence of the major signaling pathways, such as Wnt, TGF- β , Delta-Notch, hedgehog, Jak/STAT and Receptor Tyrosine Kinases, predates the evolution of the Bilateria [?]. This seems also be true for other genetic determinants important in shaping the bilaterian bodies [13]. Any present-day Metazoa holds a toolkit of, on average 20,000 protein coding genes [14]. The majority of those proteins were present already in the common ancestor of Cnidaria and Bilateria, where they have could have been involved in the formation of specialized cell types and patterning in specific regions of the body [14]. Recent works have have also documented a low level of conservation in transcriptional regulation across the different bilaterian clades [15]. It needs to be pointed out that new data from two comb jelly genomes [?,?], as well as the Cnidarian *Nematostella* illustrates the genomic complexity already present in these non-bilaterian taxa [?]. Previously, it had been already hypothesised that the cnidarian ancestor was triploblastic [?]. Indeed bilateral symmetry is present in Anthozoa ([?]) and complex body structures and cell types are also found in these animals [?].

Thus, while the monophylum Bilateria is robust in phylogenetic analyses [?], it appears hard to define morphological and functional molecular unifying bilaterian characters. Several partly competing theories for the apparently sudden emergence of this novel and diverse group of animals have been brought forward. Among these are geo-ecological ones, which argue for example that the end of large glaciation cycles (snow-ball earth) removed evolutionary constraints [1] and/or that a global increase in available atmospheric oxygen allowed animals to become larger (and thus visible in the fossil record) [16,17]. Partially connected with this, a theory stating that pre-Cambrian Bilateria were minute animals resembling the type 1 larvae seen in many modern marine species has been brought forward [?,?]. A stage of "maximal indirect development" [?], where adult forms are formed from these planktotrophic larva would then be key to the early evolution of bilaterian stem lineages Blackstone:2000fy. Such an evolutionary transition could have been mediated through changes within the arrangement and interplay of the genetic toolkit available [2]. Davidson et al. connected modifications in hierarchical gene regulatory network with the evolution of early Bilateria [18], especially in the control of developmental pattern formation [2]. These crucial gene networks, thought to play an indispensable role in preserving body form development, are assumed to be preserved for hundreds of millions of years [19].

While we previously found a possible important bilaterian restricted candidate gene acting in these networks [?], our current aim was to find out which parts of the initial bilaterian gene complement is retained across major branches in the phylum today. We endeavor to find genetic components uniting

this hugely diverse taxon - despite more than 540 Ma of disjoint evolution between the lineages leading to today's crown clades - and define their functional roles in modern animals to analyse whether they are related to developmental processes. Such an approach has only become feasible with the surge of data acquired from non-model organisms following the introduction of High-Throughput DNA sequencing methods. Only now it is possible to compare a reasonable number of well annotated genomes from within and outside of Bilateria, allowing to define a retained molecular toolkit for the taxon and conduct in-depth queries of described protein functions based on extensive data from model organisms. Here we present data from such an analysis, based first on reciprocal BLAST followed by stringent ortholog clustering. We are able to define a set of genes retained in the most distant species within Bilateria (crown clades). Comparing the expression patterns, functional annotations and the age classification of these proteins we find unifying characteristic of Bilateria. Particularly our data can be interpreted as reinforcing the theory that first Bilateria were minute organisms, similar to modern marine planktotrophic larval stages, and the acquisition of the bilaterian adult body-plan was a major asset in their evolution.

Results

Bilateria-specific ortholog clusters

We considered a total of 268,252 proteins from 10 bilaterian species and 115,334 proteins from 7 non-bilaterian species (Table 2 and Supplementary). Reciprocal BLAST filtering resulted in 13,582 Bilateria-specific proteins. Clustering them with OrthoMCL, we obtained 506 clusters of Bilateria-specific orthologs (see Table 1 and Supplementary). Based on different criteria for allowed losses along the branches, we created 4 groups from the ortholog clusters (see Figure 1):

- C** : each cluster contains at least one representative from all three major clades, Lophotrochozoa, Ecdysozoa and Deuterostomia.
- M** : each cluster contains at least all the model organisms, *D. rerio*, *D. melanogaster* and *C. elegans*, and one representative of Lophotrochozoa;
- L** : the distribution of species in each cluster can be explained by at most one loss event along the species tree, but all major clades are represented;
- A** : each cluster in this group contains representatives of all bilaterian species.

These sets are listed in increasing order according to conservation of orthologs across Bilateria, and consequently in decreasing order in the number of clusters (see table 1). The most retained set A contains only 34 ortholog clusters, about 1% of a typical bilaterian genome.

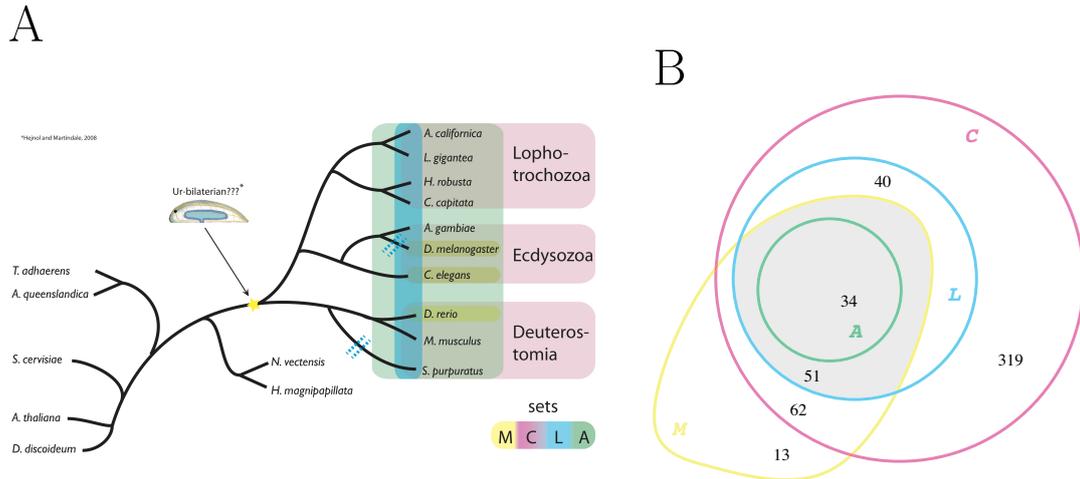


Figure 1. Panel **A**: Sketch of the genealogical relationship between the clades of Lophotrochozoa, Ecdysozoa and Deuterostomia and the species included in our analysis (see Table 2). Letter code: *M* – clusters of orthologous proteins (COPs) with proteins present in all **M**odel organisms from the three clades (textitD. rerio, D. melanogaster, C. elegans); *C* – COPs with proteins in all three **C**lades; *L* – proteins present in all nine or less species, allowing for at most one **L**oss event along the genealogical history; *A* – proteins present in **A**ll of 9 species (not necessarily in *A. californica*). Ur-Bilaterian after [?]. Panel **B**: Absolute numbers of COPs in the four sets. In detail analyzed (see text) are 85 clusters in set $L \cap M = L \cap M$.

The distribution of the Bilateria-specific ortholog clusters is listed in table 1. Disregarding Aplysia, which we found to have low quality in protein annotation, every species is represented on average in 66.2% of all orthologues clusters in set *C*. As for set *M*, which requires all model organisms to be represented in a cluster, this percentage of species jumps to 80.9%. For set *L*, which has more stringent criteria, the likelihood of species to be represented is 91.2%. *C. elegans* is the species with this smallest number of clusters in *L*, but the model species is known to be highly derived even in Nematoda [20].

For some of the downstream analyses, we focus on a small and well-retained set, namely the intersection $L \cap M$. In this set, the lowest scoring species (*S. purpuratus*) is represented in 63 of the 85 clusters, while 5 species are in 80 or more and all others in more than 75 clusters, resulting in an over all representation of species among the clusters of 93.1%. The requirement of including all model species (set *M*) opens a rich source of well-curated annotations and expression data to interpret our findings, while concentrating

Table 1. COP distribution in different data sets

species	set <i>C</i> (506)			set <i>M</i> (160)			set <i>L</i> (125)			<i>L</i> ∩ <i>M</i> (85)			set <i>A</i> (34)		
	#P	#C	ratio	#P	#C	ratio	#P	#C	ratio	#P	#C	ratio	#P	#C	ratio
Deuterostomia															
<i>D. rerio</i>	1169	415	2.82	690	160	3.89	555	125	4.44	436	85	5.13	205	34	6.03
<i>S. purpuratus</i>	636	254	2.50	233	89	2.62	288	103	2.80	175	63	2.78	90	34	2.65
<i>M. musculus</i>	855	349	2.45	398	134	2.97	428	121	3.54	301	81	3.72	163	34	4.79
Ecdysozoa															
<i>A. gambiae</i>	373	300	1.24	156	120	1.30	143	112	1.28	107	78	1.37	51	34	1.50
<i>D. melanogaster</i>	964	379	2.54	494	160	3.09	387	116	3.34	317	85	3.73	143	34	4.12
<i>C. elegans</i>	589	278	2.12	359	160	2.24	231	94	2.46	207	85	2.44	101	34	2.97
Lophotrochozoa															
<i>A. californica</i>	15	7	2.14	7	4	1.75	3	2	1.50	3	2	1.50	1	1	1.00
<i>Capitella spI</i>	642	376	1.71	247	116	2.13	198	119	1.66	153	79	1.94	65	34	1.91
<i>H. robusta</i>	420	325	1.29	165	109	1.51	160	115	1.39	110	75	1.47	48	34	1.41
<i>L. gigantea</i>	492	337	1.46	224	117	1.91	217	121	1.79	169	81	2.09	84	34	2.47

#P: number of proteins from given species in given set. #C: number of COPs with proteins from given species. ratio: #P/#C number of proteins per COP (=av. number of paralogs in COP)

on widely retained clusters (set *L*) which should be functionally relevant and less prone to harbour false positive (i.e. non-Bilateria specific) proteins.

The observed contingency ratios (see in the table 1 indicate that many of the proteins in the clusters are paralogs. Sets of more retained clusters tend to have higher ratios, i.e. more paralogs. Most paralogs are found in Deuterostomia, with the highest on average number for *D. rerio*. This is mostly likely a consequence of the genome duplication events in fish [21]. The lower rate of paralogs for Lophotrochozoa is consistent with the observation that gene divergence in the lophotrochozoan clade occurred at a lower rate compared to ecdysozoan [22].

Aiming for an approach to distinguish orthologs and paralogs, we extract from each cluster the most conserved ortholog (MCO) for each model organism. Assuming that conservation in sequence is correlated with conservation in function, the set of all MCOs should be enriched in true orthologs, in contrast with the set of all orthologs (AOs) which contain a large fraction of paralogs. Comparing the results of analyses on MCOs and AOs is therefore a way to assess the different function of orthologs and paralogs.

GO enrichment

We performed a Gene Ontology enrichment analysis across all three model species to explore the function of retained Bilateria genes. To this end, we developed a new method for Multi-Species Gene Enrichment Analysis (MSGEA) that is sensitive to ontology terms enriched in several species at the same time.

The results of our MSGEA analysis point to several different functions in which Bilateria-specific genes are involved. Complete tables of all significant GO terms and their enrichment p-values for all sets can be found in the Supplementary. The terms appearing in Bilateria-specific genes are illustrated in Figure 2 for the biological processes in the set $L \cap M$.

Some of the most prominent functions are related to developmental processes, and especially anatomical structure development, which are highly enriched especially among the MCOs. Related to that is the development of muscle tissue, organ and structure. Development of nervous system is significant as well as terms related to the nervous system like neurotransmitter receptor activity and neuron projection morphogenesis.

Many other GO terms are related to cell-cell communication and tissue complexity. Among them, the most prominent is transport (especially transmembrane and ion transport) and ion channel activity; similarly for lipid, peptide and protein binding terms, signalling and receptor activity and protein localization in the cell. Interestingly, MCOs tend to be less enriched in terms related to cell-cell communication. Finally, terms related to regulation of RNA-related processes like nucleotide metabolism, transcription, translation (elongation and fidelity) and ribosome constituents are also enriched in Bilateria.

We analysed also the trends across the sets A, L, M, C. Significant differences between the sets are reported in the Supplementary. Terms that show a consistent (increasing or decreasing) trend in relative abundance are shown in Figure [Supplementary Figure ADD].

Proteins and their functions

GO terms are helpful for an initial broad overview on protein function. However, they are often human (or mouse) derived and do not yield specific and sufficient information about function in a given process of other organisms. Thus, focussing on the orthologous clusters in set $L \cap M$, we analyzed functional descriptions for individually proteins by an extensive literature research. From this we created a system to group our proteins into 6 classes relevant to their described function in development (see supplementary

??). The six classes are “morphology related”, “muscle related”, “neuron related”, “signaling related”, “translation and mitochondrial related”, as well as “others”. Ultimately these are based on descriptions from laboratory assays like antibody studies, or RNAi knockdowns in different animals (see Supplementary File)

Class Morphology In eight out of the 85 clusters of set $L \cap M$ we find proteins related to morphology, molting and ecdysis. We decided to group these processes, as the formation of final, adult morphology incorporates molting and ecdysis in *D. melanogaster* and *C. elegans*.

One cluster contains the transcription factor Hr46, which activates various genes involved in developmental transition steps in *Drosophila* and was shown to be involved in the proper development of its ventral nerve cord [23]. Its ortholog in *C. elegans* is the nuclear hormone receptor (NHR) nhr-23, which is highly expressed in the first two *C. elegans* stages and again from the sevens stage onwards [24]. Described as a key positive regulator for genes implicated in molting, larval development and responsible for the direct or indirect activation of expression of many genes in the hypodermis, it is also required for the expression of the collagen gene and coordinates gene expression in epithelial cells [25] [26]. The nuclear receptor Vitamin D (VDR) represents the *D. rerio* orthologo and is found in the developing brain and plays a role in epithelial transport, bone and endocrine function [27].

We also classify the protein Disabled as ‘Morphology related’ as the *Drosophila* ortholog interacts with the abelson tyrosine kinase (Abl), an essential regulator of cell migration and morphogenesis [28], and is furthermore coexpressed within axons, mesoderm and body wall muscles, showing high expression at ‘4-6 hrs embryo’ [29]. *Dab* mutants display defects in epithelial morphogenesis and disruptions in the dorsal closure [28]. The corresponding *C. elegans* ortholog directly interacts with the Erythropoietin-producing hepatocellular protein-receptor-tyrosine kinase (VAB-1) [30] associated with both, the movement of neuroblasts during the closure of the ventral gastrulation cleft, as well as in the movement of epidermal cells during ventral enclosure of the embryo epidermis [31]. In *D. rerio* orthologo interacts with bone morphogenetic protein 2 (BMP2) [32], which is having an impact on neural crest progenitors, nonneural ectoderm [33] and muscle development during the segmentation [?].

Another cluster of the ‘Morphology related’ class contains SPARC. The protein is an important component of the basement membrane, a specialized extracellular matrices surrounding most tissues

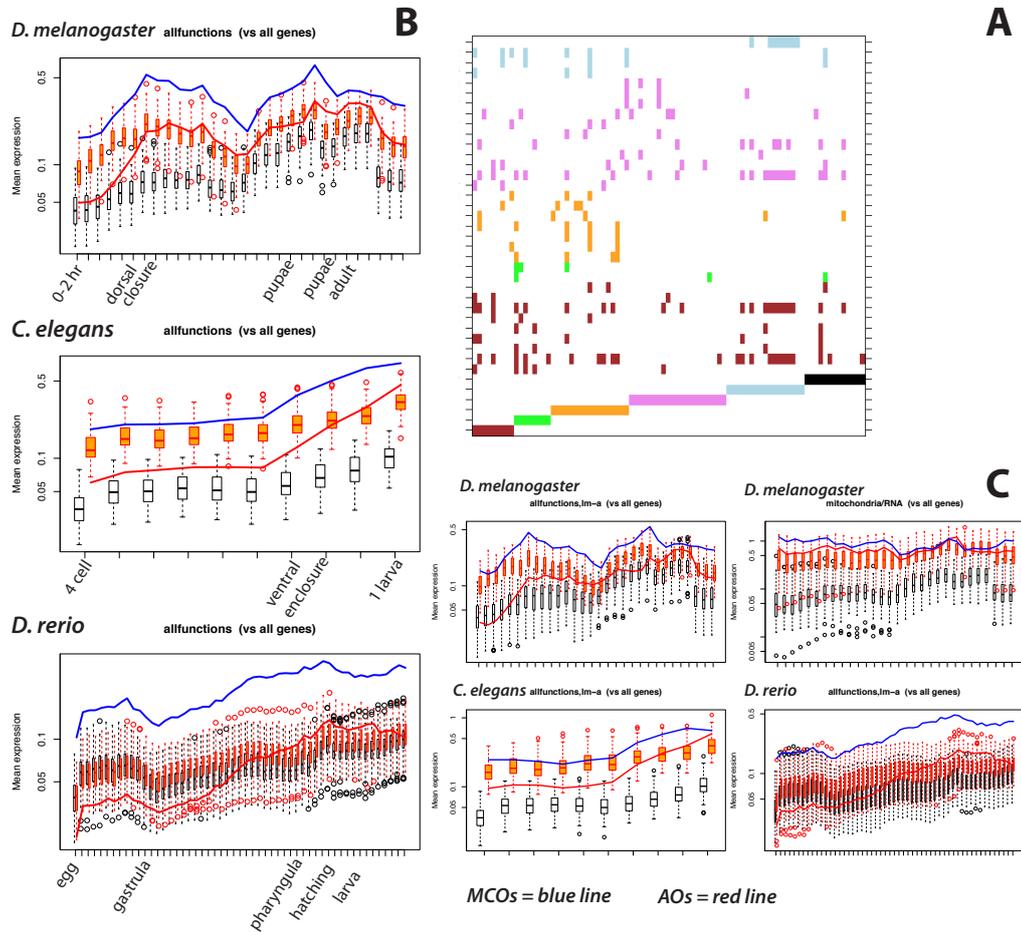


Figure 2. **A:** Assigning the 85 COPs (x -axis) of set $L \cap M$ to six functional groups (color coded bars in lower 6 lines of the plot) based on GO term abundance and on human inspection of primary literature (color coded rectangles). **B:** Expression profiles for all of the six functional groups. All Orthologs (AO) are displayed by the red, Most Conserved Orthologs (MCOs) by the blue line. Orange and white boxplots represent the distribution of the means of expression levels in background-samples (see text). Orange: samples of matched-function. White: random samples. **C:** Subtracting set A from set $L \cap M$ yields slightly modified curves. These genes are mostly of the "signalling class" and appear to be higher expressed throughout development and are especially switched on towards the beginning of *D. rerio* development. The *D. melanogaster* "mitochondrial/Ribosomal" group shows that these genes are unregulated throughout development.

in all metazoans and necessary for their organization, which is essential for normal development in *C. elegans* [34] [35] and *D. rerio* [36]. In *D. rerio* an increase of SPARC is observed when cells are exposed to solar UV radiation, causing developmental defects [37]. The *Drosophila* orthologs is described as an collagen-binding matrix glycoprotein with morphogenetic contribution to embryonic development, such as patterning and condensation of the ventral nerve cord [38], albeit SAPRC is shown to be a marker by which outcompeted cells are protected from direct Caspase activation during development [39]. Additionally to mention in this context is the *D. melanogaster* membrane protein Flower, another protein of our orthologues set classified as 'Signalling related', which plays in parallel to SPARC during the process of labeling cells as winners or loser, since the quantity of an specific isoform of Flower leads to the direct Caspase activation resulting in the death of the cell [40].

Class 'Signalling related' Within the class of 'Signalling related' one of the 18 clusters the STARD3 protein in *D. rerio*, which is involved in the regulation of steroidogenesis [41]. The *Drosophila* ortholog START1 is expressed in the embryonic progenitor cells and the prothoracic gland cells of larva showing wavelike expression during larval stages. In the prothoracic gland cells ecdysteroids are synthesized from cholesterol and START1 is involved in both, the steroid synthesis as well as the cholesterol traffic [42].

The Eaf factor, part of the 'Signalling related' set stimulates the RNA polymeras II (PolII) in *Drosophila*, which is especially important for developmentally regulated genes, as many show a paused PolII at their promoter which gets activated by the EAF and has been shown to be essential for viability of the embryo [43]. In *C. elegans* the orthologues proteins are expressed throughout the worm life cycle, controls development and contribute to the regulation of the extracellular matrix [44, 45]. In *D. rerio* eaf is described to mediate modulations in the mesoderm, extension and convergence movements, as well as neural patterning by interacting with the Wnt signalling in early embryonic processes [46].

One of the 'Signalling related' sets describes the connector enhancer of the kinase suppressor of RAS (CNK) which directly induces RAF catalytic function in fly and worm [47]. Unlike in *Drosophila* where CNK is required for viability and involved in the eye development, affects cell proliferation, as well as cell differentiation and migration [48] [49], the {*D. rerio* ortholog is important for development, but not essential for viability [50].

Class 'Muscle' Eleven of clusters were classified as 'Muscle'. While skeletal muscles are found in bilaterians, as well as in cnidarians and ctenophora [51], we find Troponin T (TnT) within one cluster

and moreover Troponin I (TnI) within another cluster, which are reported to exclusively be represented in bilateria [52]. The named Troponins build, together with Troponin C (TnC) an complex which acts as a regulatory switch for the striated muscle contraction [53].

While TnC is encoded by multiple genes, TnT and TnI are encoded by single genes in *Drosophila* [54] and are controlled mutual [55] [56] in the developing muscle cells [57]. Both Troponin are crucial for a correct muscle development during gastrulation in fly [58] [55], fish [59] [60] and worm [61] [62].

Another cluster of muscle related proteins holds the *C. elegans* protein HLH-1 and the *D. melanogaster* ortholog nautilus. Both proteins belong to the myogenic regulator factors (MRFs), which convert undifferentiated non-mesodermal cells into muscle cells [63]. It was shown that the induction of HLH-1 activity throughout the early *C. elegans* embryo is sufficient to convert most, if not all, somatic cells into a body wall muscle-like fate as [64].

Class 'Neuron' This class consists of 23 clusters, whereof one is describing the *D. melanogaster* protein Prospero, which alters the expression of homeobox genes and controls cell fate in the developing central nervous system [65], for example the fate of neuroblast progeny [66]. It is also expressed in the developing eye and in cells associated with the midgut [66]. The *C. elegans* ortholog Ceh-26 functions as a terminal selector in neurons [67] and is mainly found in the head of the developing animal [68]. Furthermore is described, that both transcription factors are additionally involved in the regulation of the extension of tubular networks during embryogenesis and are responsible for proper lumen development [69].

We also found orthologs of the Survival Motor Neuron 1 (Smn1) protein, which is essential for the functional development of motor neurons in fish, worm as well as fly [70]. Mutants of all three species show embryonic lethality or development arrest, only for the *C. elegans* mutant an improvement of the phenotype by activating a corresponding small conductance Ca²⁺-activated K⁺ channel (SK channel) is described [70]. An orthologues cluster, holding such an SK channel, is classified in the set of 'Neuron related'.

The *C. elegans* protein UNC-76 and its fly ortholog, present in one cluster, bind to kinesin respectively, which is essential for the transport of membrane bound organelles in neural and nonneural cells, and is required for axonal transport [71] [72]. Loss of *unc-76* leads to severe axonal transport defects [73] [74].

The zebrafish the microtubule-associated protein (tau), representing on orthologues set of the class 'Neuron related', is expressed in the developing central nervous system [75]. *C. elegans* and *D. melanogaster*

tau mutants show changes in the actin cytoskeleton, neuronal dysfunction and perturbed axonal transport of mitochondria [?] [76] [77] [78] [79] and expression is detected for both animals in the neuronal tissue of the developing embryo [?] [80]. Orthologues from two clusters are described in *Drosophila* to interact with tau. Tau and futsch, ortholog of another 'Neuronal related' cluster, both are expressed in neurons of the developing *Drosophila* embryo [80]. During embryogenesis futsch is important for the neuron and synaptic growth, colocalizes with microtubules and additionally is negatively regulated in non-neuronal tissue [81] [82]. It was shown that the fly protein milton, an adaptor protein essential for axonal transport of mitochondria and part of the class of 'Mitochondrial and RNA related' of our set, is interacting with tau during the process of neurogenesis [76].

Mitochondria and RNA-regulation related orthologs. Milton is one of the 19 'mitochondria and RNA-regulation related' orthologs of set $L \cap M$ and the *D. melanogaster* protein holds an corresponding *D. rerio* ortholog, Trak2. In fly Milton connects trafficking mitochondria to the microtubule cytoskeleton and is one mechanism to match mitochondrial distribution with neuronal activity [83], which is essential to regulate the local energy demand in neurons [84].

Expression profiles

Global expression Bilateria-specific genes are highly expressed. As shown in SuppXX Figure, both the mean and median expression of bilaterian-specific genes are higher than the genomic expression levels. This is especially apparent for the median expression. Expression levels are higher during embryonic and pupae phases in *D.melanogaster*, especially going up during and shortly after gastrulation in *D.rerio*, and at late embryonic stages in *C.elegans*.

To understand if the higher expression levels are due to the specific functions of bilaterian genes, we focus on the expression profiles of the genes contained in set $L \cap M$. First, for each of the three model species, we pooled genes from all clusters in set $L \cap M$ and computed the mean of the expression levels for each of the available developmental stages. To compare this profile to genes which are not bilaterian-specific, we generated two backgrounds. The first one consists of random samples of genes from the respective species. The distribution shown as white boxplots in Figure 2 is the distribution of means across these samples. The second background contains randomized samples of genes with similar GO annotations as those found in set $L \cap M$ (background of 'matched functions'; green boxplots in Figure 2).

For all three model organisms we find striking similarities: expression levels of the matched-function genes are higher than those of the random genes throughout the life cycle. Bilateria-specific genes, however, show a clearly distinct pattern. They tend to be significantly less expressed during early development or more expressed at later stages than genes in the matched-function background.

Bilateria-specific profiles In order to obtain a more fine-grained picture of bilaterian expression patterns, we analysed individual expression profiles for the model species, applying several statistic tests in order to detect profiles that were significantly enriched among Bilateria-specific genes.

For each species, we find several over-represented expression profiles. By clustering these profiles (see methods), specific patterns emerge for the mean expression of the clusters. We identified a very limited number of highly characteristic profiles in *D. melanogaster*, *D. rerio*, and *C. elegans* (Figure 2).

It is apparent that there are three distinct sets of *D. melanogaster* profiles. Two sets of profiles are double-peaked and clearly interrelated: the first set is peaked around early embryonic and late larval/early pupal stages, while the second set is peaked at late embryos/early larval and at pupal stages, in correspondence to decreasing or low expression levels of the first set.

The presence of two peaks is not shared with the other two model species, since almost all the statistically significant profiles of *D. rerio* and *C. elegans* are single-peaked. The two peaks corresponds to the two transitions in *Drosophila* - embryo/larvae and larvae/pupae - compared with the single transition in the other two species.

The third set is associated to ribosomal proteins, and therefore it is modulated by ribosome production rates across different stages.

For the other species, in *D. rerio* there are two significant clusters of profiles, one for MCOs and one for all genes. Again, the two profiles are clearly interrelated, since the Bilateria MCOs tend to be highly expressed until gastrulation, when the other profile begins to rise until a peak around hatching (the embryo-larvae transition), coherently with the picture from *Drosophila*. In *C. elegans*, there are many significantly enriched profiles (Supplementary Figure E), most of them monotonically increasing or decreasing across the embryonic stages, a few of them peaked around stage 4 or stages 6-7-8.

There is no clear pattern of shared profiles across orthologs in different species. Even in a single species, different paralogs of the same gene can have widely different expression patterns.

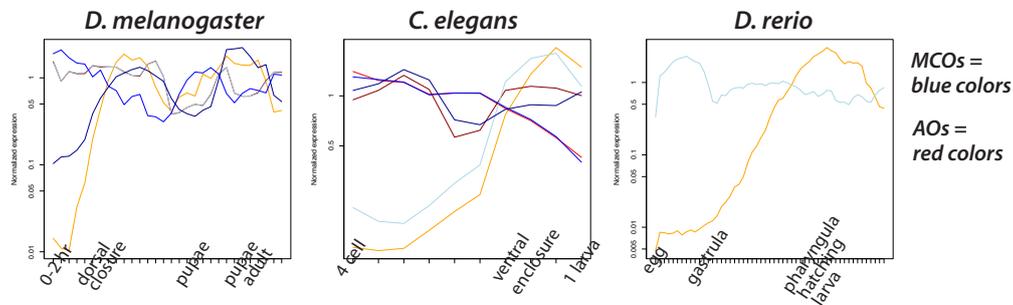


Figure 3. Significance testing for obtained expression profiles. We retrieved different numbers of significant clusters for each of the species. The ones found do confirm the pattern of up- and downregulation observed in the expression profiles for our functional classes. The up-regulation of MCOs early in development in some clusters might be influenced by signalling genes in *D. rerio*.

Expression versus function Different expression profiles can be driven by differences in function among bilaterian genes. We separated the genes in set $L \cap M$ according to the different functional categories described above, then for each category we performed the matched-function randomisation analysis described before for set $L \cap M$.

Similarly to the double-peaked profiles in Figure 2 for *D. melanogaster*, the orthologous genes related to neuron development show significant decrease in expression from the first till the 12-14 hr embryo stage, followed by an immediate increase and a peak around the late embryo phase. During the larvae phase the expression drops to a minimum at L3 larvae phase, followed by an constant decrease of expression till the first pupae stage. Morphology related proteins follow a similar pattern, but with a more pronounced peak around late pupae stages. The expression pattern for the muscle related orthologous proteins in *D. rerio* follows essentially the late-peak pattern in figure 2 with a very low expression rate from zygote till the last gastrulation stage. From that stage onwards, the expression increases constantly during the segmentation phase till the last hatching stage, where the expression does not drop any more and instead stays on a significantly high level. The pattern for the signalling related proteins of *D. rerio* mirrors the early-peak pattern from figure 2. Expression profiles for all three model organisms in all six functional classes are provided in the Supplement.

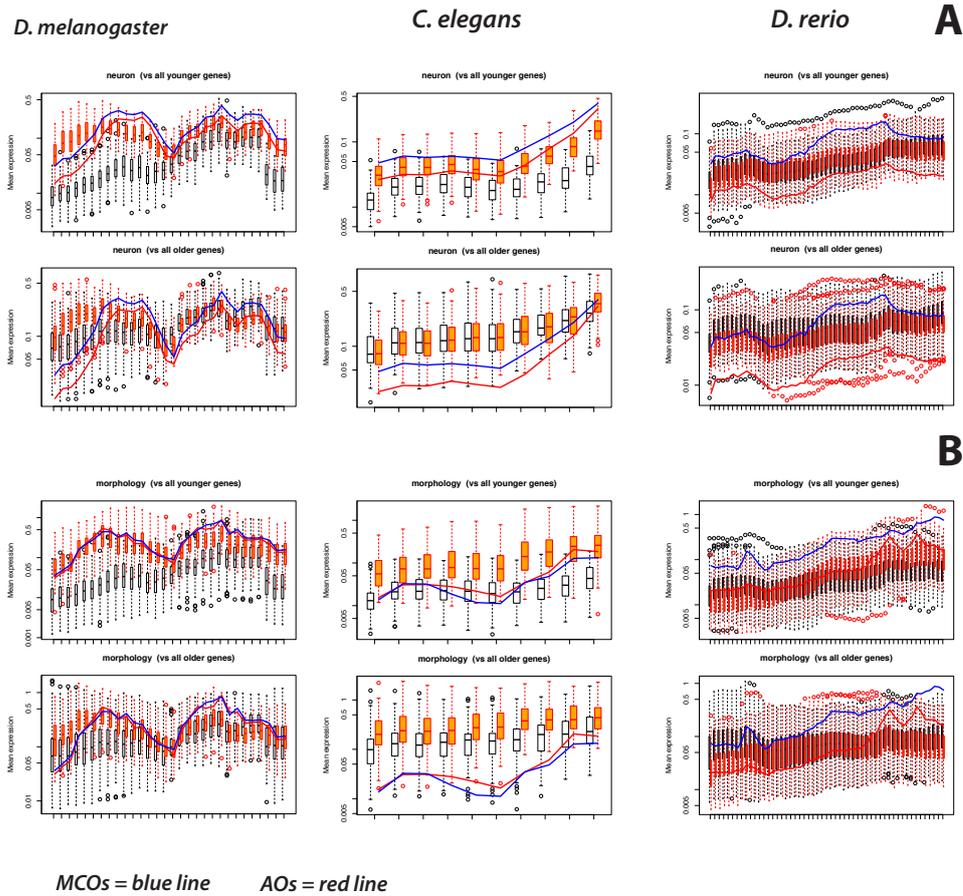


Figure 4. Comparing genes in our functional groups to previously defined phylostrata for **A** “neuron related” and **B** “morphology related”. All Orthologs (AO) are displayed by the red, Most Conserved Orthologs (MCOs) by the blue line. Orange and white boxplots represent the distribution of the means of expression levels in background-samples (see text). Orange: samples of matched-function. White: random samples. In both classes expression of our genes deviates from the genomic background in all three species, showing the peaked structure described in the main text.

Expression and function versus Age Index

We were interested to gain a deeper understanding of how our proteins compare to proteins that originated before and after the last common ancestor of all Bilateria. Therefore, we reanalysed matched-function gene expression with older and younger proteins according to previously defined transcriptomic age indices (phylostrata) [85] as background. Again we used a random set of proteins with similar GO terms as control, but restricted this to proteins that belong to either a younger or an older phylostratum with respect to Bilateria. For proteins classified as acting in neuron development, we see a striking correlation with proteins of the phylostrata younger than the emergence of Bilateria. This is most obvious in *D. melanogaster* where the low expression from the very first till the 12-14 hour embryo stage described above is mirrored. The same striking pattern can be seen in proteins related to morphology and partly for RNA and mitochondrial related proteins. For “muscle related” and signalling related proteins, the pattern is more contrasted, but the match of the expression pattern with phylostratigraphically younger genes is often better than with older proteins. For example, in *D. rerio* both the Bilateria profile and the one of younger proteins do not increase significantly during the gastrulation and early segmentation stages, but during the mid and late segmentation stages. In *D. melanogaster* and *C. elegans* the bilaterian expression patterns follow the ones of the younger proteins, but do not match them.

Discussion

Data & Cluster

Our assay critically hinged on a reliable identification of proteins retained across divergent Bilateria species after more than 540 Ma of evolution. Such an approach can only succeed when complete genomic data is available for the organisms included. Thus, we accepted the possibility of independent loss in phyla, where only fragmentary genomic data is available, and based our orthology screen preferably on sequenced, completed draft genomes. The *A. californica* data, which we had to exclude from downstream analyses, showed that even this is not without problems.

As a first criterium we required that our pooled proteins from different bilaterian species do not have an apparent direct ortholog in non-bilaterian species. A major caveat here is the scarcity of genomic data from non-bilaterian metazoan taxa. Thus, it is clear that despite our sequential strategy of first

using stringent reciprocal blast searches including 7 non-bilaterian species, followed by orthology inference with OrthoMCL, which has been shown to have a very high specificity [86,87], some non-bilaterian genes might have been missed inside and outside of our clusters. However, the stringent pipeline should ensure that proteins in our clusters in most cases have no, or only highly divergent orthologs at least in the non-bilaterian species included. Focusing on the crown clades by analysing the derived model organisms, especially in our functional analyses, we are confident to have captured a set of proteins that depict the vast genetic diversity of bilaterians.

By filtering our clusters we arrived at several hierarchically ordered sets. We decided to focus only on the intersection of sets L and set M for manually curated downstream analyses. These 85 clusters allowed for only one secondary loss within the Lophotrochozoa, Ecdysozoa, and Deuterostomia, and includes all 3 model organisms. It is thus evolutionarily conserved across a broad array of species, while errors through missing single proteins, due to incompleteness of genomic data, should have no great effect. Hence, it should be the evolutionarily most informative of our sets. We are thus confident that our proteins are not only bilaterian specific, but due to their orthology across the taxon they must have been present in the last common bilaterian ancestor (LCBA). They are also likely to conduct crucial functions in the present day bilaterian animals, as otherwise we would assume more loss events in 540Ma of evolution. This set contains also house-keeping proteins which we expect to be represented by set A, which expression is driven by their correlation with the growth rate.

GO terms reveal a link to development

Using model organisms is a powerful approach to retrieve detailed information about the obtained orthologs. It gives the possibility to evaluate the proteins in their developmental role due to the availability of expression data and allows for integration and comparison of several descriptions from functional (lab-based) assays.

We used GO annotations as a first - although approximate - source of information towards functions conserved in our clusters. In contrast to usual GO analyses, we were strongly interested in retained functions of orthologs across different species. We took advantage of the early splits in between almost all bilaterian phyla - the phylogeny is essentially starlike and the lineages to the model organisms evolved independently shortly after the split from the LCBA - to build a new Multi-Species Gene Enrichment Analysis (MSGEA).

As far as we know, our method is the only existing test that is sensitive to coherent enrichment of Gene

Ontologies across different species, and it is able to detect terms that were enriched in the ancestor of the Bilateria. Standard single-species Fisher's exact tests for enrichment cannot capture coherent signals of moderate enrichment across different species, which represent a considerable fraction (30-70%) of the significant terms detected by our method. As far as we know, MSGEA is the first method for enrichment analysis that is able to detect these signals. Our method computes an exact p -value, grounded in the same statistics as the usual GO enrichment analyses, and in fact it reduces to the Fisher's exact test once applied to a single species, thereby ensuring consistency with existing approaches.

Since the test is based on the hypothesis of independent evolution of each lineage, it can be applied only to the simplest, starlike phylogenies. Real phylogenies are rarely starlike, so this could look like a limitation of the method. However, early branching of major clades is common in large phylogenies []. Hence, phylogenies of a few species in crown clades are often well approximated by a starlike one: this is well illustrated by the case of our model species. For this reason, the starlike assumption is not particularly restrictive. However, the exact method is computationally intensive for large sets of genes. For more than 3-4 species, further numerical approximations are required to overcome this limitation.

Relying on the MSGEA method to predict significance of enrichment across species, we were able to create a list of terms most likely to be associated to retained functions for our set of genes. Most terms are associated with development in general, as well as neuron formation, muscles, and cell-cell-communication in particular. Finding these terms is not outright surprising, as these functional categories seem obviously important for Bilateria and have been described as characteristic for them [?]. However, it is now known that rudiments of these described characteristics have been existing already before the LCBA [?, ?, ?, ?, ?]. Nevertheless, as orthology is a strictly phylogenetic classification and not one reflecting function, it appears intriguing that we found most of the proteins in our sets are strongly connected to development and function in processes in this phase of life.

Functional data show the connectivity in our sets of orthologous clusters

For many of the proteins in set $L \cap M$, experimentally validated data (e.g. In-Situ hybridisations and/or Antibody stainings connected to gene knock out assays with RNAi or mutants) were available. These data confirm a strong connection of the proteins to development and in particular to neurogenesis, muscle formation, cell-cell interactions, and the general process of creating a (species specific) adult morphology (see Supplementary Table GENES).

Due to the complex interplay of developmental processes our own functional classes cannot be mutually exclusive. For example, within the “Morphology related” classification, some proteins are also acting in neuron formation. As one example, we classified the *Drosophila* gene Hr46 in “Morphology related”. It acts in the building of the ventral nerve cord in this organisms, while the *C. elegans* ortholog Nhr-23 acts in hypodermis and cuticle formation, which are processes during ventral enclosure of the roundworm Schiffer 2013. This stage has been described as possibly phylotypic for *Caenorhabditis* species Levin.

Similarly *Disabled* orthologs found in another cluster and classified as “Morphology related” are acting in epidermal cells during ventral enclosure of *C. elegans* [?] and epithelia morphogenesis during dorsal closure of *D. melanogaster*, where it is an essential of cell migration [?]. However, the orthologs are also connected to neuroblast development in *C. elegans* [?] and associated with proper development of neural crest progenitors in *D. rerio* [?], which again shows how intertwined developmental processes at this stage are.

Due to its importance for the differentiation of neurons and its presence in the developing brain we classified the *Drosophila* transcription factor Disabled [?] [?] [?] and the *C. elegans* ortholog Ceh-26 [?] [?] as “Neuron related”. Interestingly both transcription factors are additionally employed within the development process of tubular network formation during embryogenesis [?] [?] which is, similar to the creation of a complex neuronal network, a fundamental step towards an functional animal body.

Note also that in some cases bilaterian proteins can be recruited for different functions, related or unrelated to development. For example, the excretory canals in *C. elegans* are not conserved across Nematoda, therefore Ceh-26 is very likely to be an example of neo-functionalisation.

Expression patterns indicate a distinct phase in development as unifying for Bilateria

We performed several different analyses of the most relevant expression profiles for Bilateria-specific proteins. We both characterized some characteristic bilaterian profiles and outlined Bilateria-specific differences in average expression at each stage with respect to genes with the same function.

Capturing the gene expression patterns of bilaterian genes is crucial to understand their actual function. At present, there are methods for detection of tissue-specific profiles, or differential expression analysis, but we are not aware of methods for detection of enriched profiles across a set of genes. We developed two sets of methods: one aimed at studying the impact of function on the average expression levels, and

another aimed at finding Bilateria-specific expression profiles. The first method is based on randomization and is similar in spirit to [], while the second set is a combination of basic, simple but robust statistical methods (correlations, Mann-Whitney and Fisher’s exact tests).

Other choices of methods to detect Bilateria-specific profiles would have been possible, using either clustering or supervised learning/classification. For example, an immediate approach to the search for enriched expression profiles would be (i) clustering the expression profiles of all genes of a species, then (ii) apply Fisher’s exact test to the frequency of Bilateria profiles in expression clusters. However, the results would then depend on the clustering method chosen and its parameters (results not shown). The choice of the correct clustering method is not trivial in itself and depends on the purpose []. On the other hand, our tests are simple, robust (being based on standard, well-studied statistics), statistically grounded, and deliver an exact p -value. Most importantly, they were able to capture the interesting Bilateria-specific patterns in the data.

Bilaterian genes follow some definite patterns during development. These profiles are not exclusive of bilaterian genes, as could well be expected since these genes are connected to genes of different ages in the regulatory networks. More interestingly, these pattern are very different but interrelated, contrasting with the simple idea of a single profile describing a whole phylostratum. Here we observe a different, although not opposite, picture: Bilateria are characterized by several profiles interacting with each other. In particular, the two most important profiles for each species are roughly anticorrelated.

The expression profiles of the proteins in our sets peak towards the end of development in all species (clearly seen in *C. elegans*). In *D. melanogaster* an additionally up-regulation is apparent during and after gastrulation when dorsal-closure happens and the larval morphology is build. This single and double peak pattern is in line with description from modEncode, which arrived at a set of orthologs acting in comparable stages of *D. melanogaster* and *C. elegans* development starting from a co-expression analyses [88,89]. The double peaked pattern in the expression profile of *D. melanogaster* development reflects two large waves of cell proliferation and differentiation [88]. The simpler pattern peaking towards the later stages of embryogenesis in *C. elegans*, which is also present in our *D. rerio* analysis, is congruent with up-regulation of expression from the start of ventral enclosure in the worm, see Schiffer 2013. Similarly the pharyngula stage of *D. rerio* has been described as phylotypic and the ensuing hatching phase is characterized by morphogenesis [?]. Thus between these organisms expression of our retrieved orthologs appears stage congruent, pointing towards a life-cycle phase where the final adult body-plan is build.

In the holometabolous fly the situation is more complex; essentially a body morphology is build twice, for the larva and then for the adult. Almost consequently the two expression peaks we found for our proteins reflect this developmental processes. In this way the highly derived crown-clade *D. melanogaster* [] resembles what has been described as ancestral to Bilateria, a maximal indirect development [?] including a type of primary larva, which is morphologically very different from the adult [?]. Such a bona-fide Type I embryogenesis [?], is still found in many marine invertebrates like for example *Hydroides elegans* [?]. It has been argued that stem-group bilaterians were minute planktotrophic creatures similar to such primary larva and the evolution of development into adult forms was a key asset for early Bilateria [?, ?]. Thus, regulatory processes underlying the morphogenic organization of an adult bauplan might have led to evolution of modern animals [?]. Crucial genes and gene regulatory networks [?] orchestrating developmental mechanisms during the transition into an adult could be preserved today as the backbone of bilaterian diversification [?]. The expression patterns found for our set of genes and by the expression primed analysis of Li24985912 appear to fit with these theories, as do the enriched GO terms and functional data for single proteins. Both our and the modEncode data found proteins acting in a variety of functions in development [89], without the sets of orthologs containing previously assumed key-genes (e.g. HOX genes). This could point towards the theory that differences in GRNs and their deployment being more important than the conservation of single key genes [?] for the evolution of Bilateria. While non-Bilateria have a similar complex distribution and abundance of gene regulatory elements and systems [?] our data indicate that the Bilateria are unified by a set of developmental processes with only a small set of conserved genes. This inherent plasticity of the molecular backbone on the basis of flexible GRNs could allow for the exchange of most of the genes, while function of crucial processes is retained. Thus evolution by tinkering Jacob:1977tf in a system liable to change would be underlying the huge morphological divergence in Bilateria, which is not seen in non-bilaterian metazoa. A link between evolutionary innovations and the emergence of novel genes has been suggested [?]. This is in support of our data, where expression on later stages, when adult morphology is build, is similar to phylo-stratigraphically younger genes used as background in our analysis. Our most conserved orthologs, which are likely to have the more ancestral sequence, are on the other hand more similar to phylo-stratigraphically older genes in their expression.

Our data do not directly argue for a phylotypic stage

Thus, although at least in *D. rerio* and *C. elegans* the developmental stages where we found expression to increase have been described as potentially phylotypic [?, ?] our data do not directly support the idea of a phylotypic stage encoded in the gene expression of Bilateria [?]. However, they could support the idea of a phylotypic stage is a key target for evolutionary tinkering [?] through the re-shuffling of GRNs. The observation that from the gene set described as potentially phylotypic for Nematoda based on an analysis of the crown group *Caenorhabditis* alone, mainly those genes that shape adult morphology are retained in phylogenetically remote species in the roundworms [?] would be in line with this. Thus, the process of forming an adult body could be seen as ancient and conserved [?], with some orthologs retained in their function across large evolutionary time spans and the huge diversity of Bilateria. It is striking to note that this diversity is most apparent in adult morphology. The taxon thus seems to be united by the process actually preceding the formation of its diverse body plans, which allow bilaterian species to exploit almost all habitats on earth.

Stem Bilateria might have been similar to planktotrophic larva

While our data lent some support to the theory of minute planktotrophic stem species in Bilateria we are yet no able to propose reasons for the formation of adult body plans. Hypothesis arguing for a change in ecological conditions appear interesting in this respect [?], but more data is needed to find a link to genes found by us.

Meanwhile, it appears important to note that measures of divergence and molecular clocks are based on snps and indels, e.g. [?, ?, ?], but like morphological characters the exchange (loss and gain) of full genes might have played a much more crucial role in evolution, see [?]. This type of genomic diversity is yet not comprehensively captured in studies on the evolution of Bilateria and its sub-taxa and should be regarded as an avenue to follow. Such studies should be coupled with assays to unravel the functional connections in GRNs and their re-shuffling in divergent taxa.

Materials and Methods

Using a comparative genomic approach, we identified proteins shared among Bilateria. To represent the vast divergence of the different bilaterian clades, subsets of species of the three major clades -

Table 2. Species considered to contrast bilaterian and non-bilaterian protein repertoires

species	source	no. proteins in database	no. proteins in all COP*s	no. COPs represented	no. proteins per COP
Bilateria					
Deuterostomia					
<i>D. rerio</i>	Ensembl	41693	1190	428	2.78
<i>S. purpuratus</i>	NCBI	42420	652	260	2.51
<i>M. musculus</i>	Ensembl	40732	867	357	2.43
Ecdysozoa					
<i>A. gambiae</i>	Ensembl	13133	380	307	1.24
<i>D. melanogaster</i>	FlyBase	23849	989	392	2.52
<i>C. elegans</i>	WormBase	25634	602	291	2.07
Lophotrochozoa					
<i>A. californica</i>	NCBI	1093	15	7	2.14
<i>Capitella spI</i>	JGI	32415	642	376	1.71
<i>H. robusta</i>	JGI	23432	420	325	1.29
<i>L. gigantea</i>	JGI	23851	492	337	1.46
Non-Bilateria					
<i>Amphimedon queenslandica</i>	NCBI	9908			
<i>Arabidopsis thaliana</i>	TAIR	28952			
<i>Dictyostelium discoideum</i>	Dictybase	13426			
<i>Hydra magnipapillata</i>	NCBI	17563			
<i>Nematostella vectensis</i>	JGI	27273			
<i>Saccharomyces cerevisiae</i>	SGD	6692			
<i>Trichoplax adhaerens</i>	JGI	11520			

*COP: cluster of orthologous proteins

Deuterostomia, Ecdysozoa, Lophotrochozoa - have have been chosen (Table 2). We included more species from Lophotrochozoa than from the other two clades, since well-annotated complete genome sequences from Lophotrochozoa model-organisms were still under-represented in public databases at the begin of our studies. To obtain bilaterian specific proteins, a selection of seven non-bilaterian species derived from a previous study [90] have been used as an exclusion criterion for the further analysis. Proteomes (see Table 2), and data from InterProScan [91] and the Gene Ontology consortium [92] were downloaded from the corresponding online addresses.

Blast and Clustering

To find orthologs we set up a stringent analysis pipeline. First, we searched for homologs in bilaterians using the stand-alone Blast version 2.2.25+. The downloaded proteins sequences were grouped and again reciprocally blasted with a cut-off E-value of $\leq 10^{-5}$. We excluded all proteins for which a homolog was identified in at least one non-bilaterian. Second, we verified the potential bilaterian-specific proteins using two different ortholog finders, InParanoid and OrthoMCL [93,94]. Both were rated highly in benchmarking studies which analyzed the performance of orthology-prediction methods [86,87]. For the purpose of detecting orthologs of very diverged species, we chose OrthoMCL version 2.0.2 and the associated MCL version 11-335, as they were shown to be more robust [86,87]. Third, we grouped the ortholog clusters into the following four sets (Fig 1)

- C** : containing a set of 506 clusters of orthologous proteins; each cluster contains at least one representative from all three clades, Lophotrochozoa, Ecdysozoa and Deuterostomia
- M** : containing a set of 160 clusters of orthologous proteins; in each cluster all the model organisms, *D. rerio*, *D. melanogaster* and *C. elegans*, and at least one species from each of the three clades are represented
- L** : containing 125 clusters of orthologous proteins; each cluster contains all bilaterian species listed in Table 2, except for those whose absence can be explained by at most one loss event along the species tree (Fig 1A), but which is not leading to the complete absence of any of the three major clades
- A** : containing 34 clusters of orthologues proteins; each cluster contains representatives of all species (but not necessarily from *A. californica*).

Due to its currently still poor sequence quality, we treated the Lophotrochozoon *A. californica* differently: where available we included *A. californica* orthologs, but we did not require them to be present in set

.

The above data sets are ordered by degree of confidence in the bilaterian specificity of the protein. At the same time they reflect the level of conservation of the proteins across different species. Seeking to be conservative regarding bilaterian specificity on the one hand (set

being most conservative), but also to analyze a dataset which is as comprehensive as possible, we concentrate in our further analysis on the intersection of sets

and

(Figure 1; called $L \cap M$). This set comprises 85 clusters.

In order to account for paralogs in model species, for each of them (*D. rerio*, *D. melanogaster*, *C. elegans*) we extracted the UCSC tracks of the base-by-base PhastCons conservation scores [95] across insects, fishes and *Caenorabdhitis* (*CHECK*) respectively. We used these scores to rank all the genes in a given cluster according to the fraction of very conserved sites, i.e. sites with PhastCons score > 0.99 . We selected the highest-ranking one for each cluster as the 'Most Conserved Ortholog' (MCO). These genes are likely to be 'true' orthologs, based on the idea that strong conservation reflects long-term evolutionary (and functional) constraint and that paralogs tend to diverge faster. We used the fraction of very conserved sites, instead of the average conservation score, since we are interested in the degree of conservation in function, not in sequence. Functional conservation is related to high conservation of alleles at functional sites, while the other (not strictly functional) bases could evolve fast. However, the two measures are highly correlated genome-wide ($r = 0.91$ for *Drosophila*, 0.84 for *Danio* and 0.71 for *C.elegans* across all exons, see Figure [conservationscores.pdf]).

We performed all subsequent analyses both on the complete set of all orthologs/paralogs (AOs) in a given cluster and on MCOs only.

Multi-Species Gene Enrichment Analysis

To examine the global function of the potentially Bilateria-specific proteins, the gene ontology (GO) terms [92] were obtained from the Gene Ontology Database for all genes in the above sets and for all model species *Drosophila*, *Danio* and *C.elegans*.

Our aim was to find terms enriched in Bilateria-specific genes in comparison to the genomic background. For this enrichment analysis, we included the counts of all GO graph descendants for each GO term, we then computed the p -values for all terms present in each species considered and corrected for multiple testing using the Benjamini-Hochberg algorithm [96]. Since our orthologs originate from multiple species, we tested for single-species enrichment for all terms in each of the three species, as well as for multi-species GO enrichment. For this purpose, we developed a novel method for Multi-Species Gene Enrichment Analysis (MSGEA), described below.

Our Multi-Species Gene Enrichment Analysis method assumes that all lineages splitted approximately at the same time (i.e. starlike phylogeny) and evolved independently after that. The null hypothesis of the MSGEA test is that a given GO term was not overrepresented in the ancestor of all lineages. Therefore, a significant p -value shows that the genome of the ancestor was enriched in the GO term considered.

We denote by $n_{go,s}$ the count (i.e., number of occurrences) of the GO term go in species s contained in our set and by $N_{go,s}$ the count for the whole genome. Furthermore, we define $n_s = \sum_{go} n_{go,s}$ and $N_s = \sum_{go} N_{go,s}$. Since we assume that all species evolved independently and that go was not overrepresented in the ancestor, a conservative assumption for the null distribution of the $n_{go,s}$ is given by independent hypergeometric distributions for the counts in each species with parameters $N_{go,s}$, n_s and N_s . The enrichment statistics X_{go} is then defined as the sum of the normalized enrichments of $n_{go,s}$ across species, i.e. the sum of the z -scores:

$$X_{go} = \sum_s \frac{n_{go,s} - E(n_{go,s})}{SD(n_{go,s})} \sim \sum_s \frac{n_{go,s} - N_{go,s}n_s/N_s}{\sqrt{N_{go,s}n_s/N_s}} \quad (1)$$

where the Poisson approximation is used to define the score.

The p -value is the probability

$$p_{go} = \text{Prob}(x \geq X_{go} | N_{go,s}, n_s, N_s),$$

where the distribution of x follows from the hypergeometric distributions of the $n_{go,s}$ with the above parameters. For a single species, this test reduces to the standard one-tail Fisher's Exact Test for GO enrichment. The exact estimation of p -values is computationally intensive; an optimized code developed in C is available from the authors upon request.

Functional categorisation of genes from model organisms

To gain a deeper functional understanding, we employed biomaRt to mine literature databases for functional studies in model organisms with a focus on development. We used the Bioconductor module [97] biomaRt 2.19.3 to extract information for proteins present in our cluster set $L \cap M = M \cap L$. Wormbase release WS220 was queried for *C. elegans* proteins and ENSEMBL 75 for *D. melanogaster* and *D. rerio* proteins. We then queried extensively the literature for experimentally established proofs of protein

function. We could attribute the results to six major distinct categories, described in detail below and summarized in Table refbig table. These categories represent prominent molecular functions during the processes of embryogenesis and development (Table ??). Based on the retrieved annotations and GO terms, we grouped our proteins in set $L \cap Mintothesecategories$.

Expression

We retrieved expression data from different developmental stages in *D. melanogaster* [98], *D. rerio* [99] and *C. elegans* [100]. *D. rerio* data for adult stages were not used. The expression profiles were normalized by taking the logarithm in base 10 of the expression levels and then by subtracting the $\log_1 0$ of the mean expression at each stage.

We developed some tests to find characteristic expression patterns for Bilateria. First, we computed the Pearson correlation between expression profiles for different genes. For each profile in each set of Bilateria-specific genes, we performed two types of tests: (i) a Mann-Whitney test on the distribution of correlations, comparing the correlation of the profile with other bilaterian genes versus the correlation with genome-wide profiles, in order to detect profiles that are more correlated to the ones of other bilaterian genes than to the rest of the genome; (ii) for each profile, we classified the remaining genes as “highly correlated” or “not highly correlated” in expression, using correlation thresholds of $r = 0.5, 0.7$ and 0.9 ; then, we tested for enrichment of correlated profiles among Bilateria by Fisher’s Exact test.

A Benjamini-Hochberg correction for multiple testing was applied to the p -values resulting from these tests. All the significant profiles were clustered based on their correlations r by complete-linkage hierarchical clustering implemented in the R statistics software package, using $1 - r$ as distance measure and selecting clusters at height $h = 0.75$.

Expression versus function

We performed a randomisation test to understand if the mean (logarithmic) expression level of Bilateria-specific genes at each stage could be explained by their function. We performed this analysis on the set $L \cap Monly$.

For this purpose, we randomly sampled from the whole genome 250 sets of genes of the same size as our set, in two different ways. One was a random sampling of genes from the genome. The other was a matched-function sampling: first, for each GO term in the original set, we listed all genes with the same

annotation and extracted from this list a number of random genes equal to the number of occurrences of the term; second, we pooled together all the resulting lists; third, we extracted from this list a number of random genes equal to the number of genes in the original set. This way, we obtained sets of the same size and function (on average) as our original set.

We repeated the same analysis for the whole set $L \cap M$ and for all subsets corresponding to the six functional categories described above.

Age index of bilaterian specific proteins

We used the phylostratigraphic classification [?] for *Drosophila*, *Danio* and *C.elegans* [data obtained by Domasev-Lošo] to obtain two groups from the proteins included in this classification: (i) old proteins, i.e. proteins with inferred origin that predates than Bilateria according to the phylostratigraphy, equivalent to phylostrata 1-7; (ii) young proteins, i.e. proteins appeared after the Bilateria split according to the phylostratigraphy, equivalent to phylostrata 7 and higher. Proteins that appear to be Bilateria-specific according to BLAST search (i.e. the ones in our sets A, L, M, C) were included in both groups, irrespective of the original phylostratigraphic classification.

We repeated all the analyses described in the previous sections, comparing our Bilateria-specific genes with the genome-wide background of older or younger genes only.

Acknowledgments

Financial support by DFG-SFB680 and Volkswagen Stiftung in its Initiative for Evolutionary Biology (to PHS). We thank Georgios Koutsovoulos and Sujai Kumar from the University of Edinburgh for support with orthology inference and annotation.

References

1. Marschall CR (2006) Explaining the cambrian "explosion" of animals. Annual Review of Earth and Planetary Sciences 34: 355-384.
2. Erwin D (2009) Early origin of the bilaterian developmental toolkit. Philos Trans R Soc Lond B Biol Sci 364: 2253-2261.

3. Carter F, Frankson T, Pintard J, Edgecombe B (2011) Seroprevalence of helicobacter pylori infection in adults in the bahamas. *West Indian Med J* 60: 662-665.
4. Hejnal A, Martindale M (2008) Acoel development indicates the independent evolution of the bilaterian mouth and anus. *Nature* 456: 382-386.
5. Knoll A, Carroll S (1999) Early animal evolution: emerging views from comparative biology and geology. *Science* 284: 2129-2137.
6. De Robertis E (2008) The molecular ancestry of segmentation mechanisms. *Proc Natl Acad Sci U S A* 105: 16411-16412.
7. Dunn C, Hejnal A, Matus D, Pang K, Browne W, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452: 745-749.
8. Raven P, Johnson G (2001) *Biology*. McGraw-Hill Science/Engineering/Math; 6 edition.
9. Averof M, Cohen S (1997) Evolutionary origin of insect wings from ancestral gills. *Nature* 385: 627-630.
10. Mitreva M, Jasmer D, Zarlenga D, Wang Z, Abubucker S, et al. (2011) The draft genome of the parasitic nematode trichinella spiralis. *Nat Genet* 43: 228-235.
11. Ikuta T (2011) Evolution of invertebrate deuterostomes and hox/parahox genes. *Genomics Proteomics Bioinformatics* 9: 77-96.
12. Schiffer P, Kroiher M, Kraus C, Koutsovoulos G, Kumar S, et al. (2013) The genome of romanomermis culicivorax: revealing fundamental changes in the core developmental genetic toolkit in nematoda. *BMC Genomics* 14: 923.
13. Ryan J, Burton P, Mazza M, Kwong G, Mullikin J, et al. (2006) The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, nematostella vectensis. *Genome Biol* 7: R64.
14. Zhang X, Shu D, Han J, Zhang Z, Liu J, et al. (2014) Triggers for the cambrian explosion: Hypotheses and problems. *Gondwana Research, ISSN 1342937X* 25: 896-909.

15. Boyle A, Araya C, Brdlik C, Cayting P, Cheng C, et al. (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature* 512: 453-456.
16. Sperling E, Frieder C, Raman A, Girguis P, Levin L, et al. (2013) Oxygen, ecology, and the cambrian radiation of animals. *Proc Natl Acad Sci U S A* 110: 13446-13451.
17. Valentine J, Jablonski D, Erwin D (1999) Fossils, molecules and embryos: new perspectives on the cambrian explosion. *Development* 126: 851-859.
18. Davidson E, Peterson K, Cameron R (1995) Origin of bilaterian body plans: evolution of developmental regulatory mechanisms. *Science* 270: 1319-1325.
19. Davidson E, Erwin D (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311: 796-800.
20. Kortschak R, Samuel G, Saint R, Miller D (2003) EST analysis of the cnidarian acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Curr Biol* 13: 2190-2195.
21. Taylor J, Braasch I, Frickey T, Meyer A, Van de Peer Y (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res* 13: 382-390.
22. Philipp E, Kraemer L, Melzner F, Poustka A, Thieme S, et al. (2012) Massively parallel RNA sequencing identifies a complex immune gene repertoire in the lophotrochozoan mytilus edulis. *PLoS One* 7: e33091.
23. Ruaud A, Lam G, Thummel C (2010) The drosophila nuclear receptors DHR3 and betaFTZ-f1 control overlapping developmental responses in late embryos. *Development* 137: 123-131.
24. Kostrouchova M, Krause M, Kostrouch Z, Rall J (2001) Nuclear hormone receptor CHR3 is a critical regulator of all four larval molts of the nematode caenorhabditis elegans. *Proc Natl Acad Sci U S A* 98: 7360-7365.
25. Kouns N, Nakielna J, Behensky F, Krause M, Kostrouch Z, et al. (2011) NHR-23 dependent collagen and hedgehog-related genes required for molting. *Biochem Biophys Res Commun* 413: 515-520.

26. Frand A, Russel S, Ruvkun G (2005) Functional genomic analysis of *c. elegans* molting. *PLoS Biol* 3: e312.
27. Craig T, Sommer S, Sussman C, Grande J, Kumar R (2008) Expression and regulation of the vitamin d receptor in the zebrafish, *danio rerio*. *J Bone Miner Res* 23: 1486-1496.
28. Song J, Kannan R, Merdes G, Singh J, Mlodzik M, et al. (2010) Disabled is a bona fide component of the *abl* signaling network. *Development* 137: 3719-3727.
29. Gertler F, Hill K, Clark M, Hoffmann F (1993) Dosage-sensitive modifiers of *drosophila* *abl* tyrosine kinase function: prospero, a regulator of axonal outgrowth, and disabled, a novel tyrosine kinase substrate. *Genes Dev* 7: 441-453.
30. Cheng H, Govindan J, Greenstein D (2008) Regulated trafficking of the MSP/*eph* receptor during oocyte meiotic maturation in *c. elegans*. *Curr Biol* 18: 705-714.
31. George S, Simokat K, Hardin J, Chisholm A (1998) The VAB-1 *eph* receptor tyrosine kinase functions in neural and epithelial morphogenesis in *c. elegans*. *Cell* 92: 633-643.
32. Kim J, Kang H, Larrivee B, Lee M, Mettlen M, et al. (2012) Context-dependent proangiogenic function of bone morphogenetic protein signaling is mediated by disabled homolog 2. *Dev Cell* 23: 441-448.
33. Nguyen V, Schmid B, Trout J, Connors S, Ekker M, et al. (1998) Ventral and lateral regions of the zebrafish gastrula, including the neural crest progenitors, are established by a *bmp2b*/*swirl* pathway of genes. *Dev Biol* 199: 93-110.
34. Schwarzbauer J, Spencer C (1993) The *caenorhabditis elegans* homologue of the extracellular calcium binding protein SPARC/osteonectin affects nematode body morphology and mobility. *Mol Biol Cell* 4: 941-952.
35. Fitzgerald M, Schwarzbauer J (1998) Importance of the basement membrane protein SPARC for viability and fertility in *caenorhabditis elegans*. *Curr Biol* 8: 1285-1288.
36. Ceinos R, Torres-Nunez E, Chamorro R, Novoa B, Figueras A, et al. (2013) Critical role of the matricellular protein SPARC in mediating erythroid progenitor cell development in zebrafish. *Cells Tissues Organs* 197: 196-208.

37. Torres Nunez E, Sobrino C, Neale P, Ceinos R, Du S, et al. (2012) Molecular response to ultraviolet radiation exposure in fish embryos: implications for survival and morphological development. *Photochem Photobiol* 88: 701-707.
38. Martinek N, Shahab J, Saathoff M, Ringuette M (2008) Haemocyte-derived SPARC is required for collagen-IV-dependent stability of basal laminae in drosophila embryos. *J Cell Sci* 121: 1671-1680.
39. Portela M, Casas-Tinto S, Rhiner C, Lopez-Gay J, Dominguez O, et al. (2010) Drosophila SPARC is a self-protective signal expressed by loser cells during cell competition. *Dev Cell* 19: 562-573.
40. Rhiner C, Lopez-Gay J, Soldini D, Casas-Tinto S, Martin F, et al. (2010) Flower forms an extracellular code that reveals the fitness of a cell to its neighbors in drosophila. *Dev Cell* 18: 985-998.
41. Bauer M, Bridgham J, Langenau D, Johnson A, Goetz F (2000) Conservation of steroidogenic acute regulatory (StAR) protein structure and expression in vertebrates. *Mol Cell Endocrinol* 168: 119-125.
42. Roth G, Gierl M, Vollborn L, Meise M, Lintermann R, et al. (2004) The drosophila gene *start1*: a putative cholesterol transporter and key regulator of ecdysteroid synthesis. *Proc Natl Acad Sci U S A* 101: 1601-1606.
43. Smith E, Winter B, Eissenberg J, Shilatifard A (2008) Regulation of the transcriptional activity of poised RNA polymerase II by the elongation factor ELL. *Proc Natl Acad Sci U S A* 105: 8575-8579.
44. Cai L, Phong B, Fisher A, Wang Z (2011) Regulation of fertility, survival, and cuticle collagen function by the *caenorhabditis elegans* *eaf-1* and *ell-1* genes. *J Biol Chem* 286: 35915-35921.
45. Cai L, Wang D, Fisher A, Wang Z (2014) Identification of a genetic interaction between the tumor suppressor EAF2 and the retinoblastoma protein (rb) signaling pathway in *c. elegans* and prostate cancer cells. *Biochem Biophys Res Commun* 447: 292-298.
46. Ma X, Liu J (2013) Eafs control erythroid cell fate by regulating *c-myb* expression through *wnt* signaling. *PLoS One* 8: e64576.

47. Claperon A, Therrien M (2007) KSR and CNK: two scaffolds regulating RAS-mediated RAF activation. *Oncogene* 26: 3143-3158.
48. Therrien M, Wong A, Rubin G (1998) CNK, a RAF-binding multidomain protein required for RAS signaling. *Cell* 95: 343-353.
49. Cabernard C, Affolter M (2005) Distinct roles for two receptor tyrosine kinases in epithelial branching morphogenesis in drosophila. *Dev Cell* 9: 831-842.
50. Rocheleau C, Ronnlund A, Tuck S, Sundaram M (2005) *Caenorhabditis elegans* CNK-1 promotes raf activation but is not essential for ras/raf signaling. *Proc Natl Acad Sci U S A* 102: 11757-11762.
51. Chiodin M, Achatz J, Wanninger A, Martinez P (2011) Molecular architecture of muscles in an acoel and its evolutionary implications. *J Exp Zool B Mol Dev Evol* 316: 427-439.
52. Steinmetz P, Kraus J, Larroux C, Hammel J, Amon-Hassenzahl A, et al. (2012) Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* 487: 231-234.
53. Gordon A, Homsher E, Regnier M (2000) Regulation of contraction in striated muscle. *Physiol Rev* 80: 853-924.
54. Herranz R, Mateos J, Mas J, Garcia-Zaragoza E, Cervera M, et al. (2005) The coevolution of insect muscle tpnt and tpni gene isoforms. *Mol Biol Evol* 22: 2231-2242.
55. Nongthomba U, Ansari M, Thimmaiya D, Stark M, Sparrow J (2007) Aberrant splicing of an alternative exon in the drosophila troponin-t gene affects flight muscle development. *Genetics* 177: 295-306.
56. Mas J, Garcia-Zaragoza E, Cervera M (2004) Two functionally identical modular enhancers in drosophila troponin t gene establish the correct protein levels in different muscle types. *Mol Biol Cell* 15: 1931-1945.
57. Marin M, Rodriguez J, Ferrus A (2004) Transcription of drosophila troponin i gene is regulated by two conserved, functionally identical, synergistic elements. *Mol Biol Cell* 15: 1185-1196.
58. Naimi B, Harrison A, Cummins M, Nongthomba U, Clark S, et al. (2001) A tropomyosin-2 mutation suppresses a troponin i myopathy in drosophila. *Mol Biol Cell* 12: 1529-1539.

59. Hsiao C, Tsai W, Horng L, Tsai H (2003) Molecular structure and developmental expression of three muscle-type troponin t genes in zebrafish. *Dev Dyn* 227: 266-279.
60. Fu C, Lee H, Tsai H (2009) The molecular structures and expression patterns of zebrafish troponin i genes. *Gene Expr Patterns* 9: 348-356.
61. Takashima Y, Kitaoka S, Bando T, Kagawa H (2012) Expression profiles and unc-27 mutation rescue of the striated muscle type troponin i isoform-3 in *caenorhabditis elegans*. *Genes Genet Syst* 87: 243-251.
62. Myers C, Goh P, Allen T, Bucher E, Bogaert T (1996) Developmental genetic analysis of troponin t mutations in striated and nonstriated muscle cells of *caenorhabditis elegans*. *J Cell Biol* 132: 1061-1077.
63. Andrikou C, Iovene E, Rizzo F, Oliveri P, Arnone M (2013) Myogenesis in the sea urchin embryo: the molecular fingerprint of the myoblast precursors. *Evodevo* 4: 33.
64. Fukushige T, Krause M (2005) The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early *c. elegans* embryos. *Development* 132: 1795-1805.
65. Doe C, Chu-LaGraff Q, Wright D, Scott M (1991) The prospero gene specifies cell fates in the *drosophila* central nervous system. *Cell* 65: 451-464.
66. Oliver G, Sosa-Pineda B, Geisendorf S, Spana E, Doe C, et al. (1993) Prox 1, a prospero-related homeobox gene expressed during mouse development. *Mech Dev* 44: 3-16.
67. Araya C, Kawli T, Kundaje A, Jiang L, Wu B, et al. (2014) Regulatory analysis of the *c. elegans* genome with spatiotemporal resolution. *Nature* 512: 400-405.
68. Yu H, Pretot R, Burglin T, Sternberg P (2003) Distinct roles of transcription factors EGL-46 and DAF-19 in specifying the functionality of a polycystin-expressing sensory neuron necessary for *c. elegans* male vulva location behavior. *Development* 130: 5217-5227.
69. Kolotuev I, Hyenne V, Schwab Y, Rodriguez D, Labouesse M (2013) A pathway for unicellular tube extension depending on the lymphatic vessel determinant prox1 and on osmoregulation. *Nat Cell Biol* 15: 157-168.

70. Edens B, Ajroud-Driss S, Ma L, Ma Y (2014) Molecular mechanisms and animal models of spinal muscular atrophy. *Biochim Biophys Acta* .
71. Barsi-Rhyne B, Miller K, Vargas C, Thomas A, Park J, et al. (2013) Kinesin-1 acts with netrin and DCC to maintain sensory neuron position in *caenorhabditis elegans*. *Genetics* 194: 175-187.
72. Gindhart J, Chen J, Faulkner M, Gandhi R, Doerner K, et al. (2003) The kinesin-associated protein UNC-76 is required for axonal transport in the *drosophila* nervous system. *Mol Biol Cell* 14: 3356-3365.
73. Toda H, Mochizuki H, Flores R 3rd, Josowitz R, Krasieva T, et al. (2008) UNC-51/ATG1 kinase regulates axonal transport by mediating motor-cargo assembly. *Genes Dev* 22: 3292-3307.
74. Chua J, Butkevich E, Worseck J, Kittelmann M, Gronborg M, et al. (2012) Phosphorylation-regulated axonal dependent transport of syntaxin 1 is mediated by a kinesin-1 adapter. *Proc Natl Acad Sci U S A* 109: 5862-5867.
75. Chen M, Martins R, Lardelli M (2009) Complex splicing and neural expression of duplicated tau genes in zebrafish embryos. *J Alzheimers Dis* 18: 305-317.
76. Iijima-Ando K, Sekiya M, Maruko-Otake A, Ohtake Y, Suzuki E, et al. (2012) Loss of axonal mitochondria promotes tau-mediated neurodegeneration and alzheimer's disease-related tau phosphorylation via PAR-1. *PLoS Genet* 8: e1002918.
77. Bolkan B, Kretschmar D (2014) Loss of tau results in defects in photoreceptor development and progressive neuronal degeneration in *drosophila*. *Dev Neurobiol* 74: 1210-1225.
78. Fulga T, Elson-Schwab I, Khurana V, Steinhilb M, Spires T, et al. (2007) Abnormal bundling and accumulation of f-actin mediates tau-induced neuronal degeneration in vivo. *Nat Cell Biol* 9: 139-148.
79. Kraemer B, Schellenberg G (2007) SUT-1 enables tau-induced neurotoxicity in *c. elegans*. *Hum Mol Genet* 16: 1959-1971.
80. Heidary G, Fortini M (2001) Identification and characterization of the *drosophila* tau homolog. *Mech Dev* 108: 171-178.

81. Hummel T, Krukkert K, Roos J, Davis G, Klambt C (2000) *Drosophila* futsch/22c10 is a MAP1B-like protein required for dendritic and axonal development. *Neuron* 26: 357-370.
82. Roos J, Hummel T, Ng N, Klambt C, Davis G (2000) *Drosophila* futsch regulates synaptic microtubule organization and is necessary for synaptic growth. *Neuron* 26: 371-382.
83. Lee K, Lu B (2014) The myriad roles of miro in the nervous system: axonal transport of mitochondria and beyond. *Front Cell Neurosci* 8: 330.
84. van Spronsen M, Mikhaylova M, Lipka J, Schlager M, van den Heuvel D, et al. (2013) TRAK/milton motor-adaptor proteins steer mitochondrial trafficking to axons and dendrites. *Neuron* 77: 485-502.
85. Domazet-Loso T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468: 815-818.
86. Shaye D, Greenwald I (2011) Ortholist: a compendium of *c. elegans* genes with human orthologs. *PLoS One* 6: e20085.
87. Hulsen T, Huynen M, de Vlieg J, Groenen P (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 7: R31.
88. Li J, Huang H, Bickel P, Brenner S (2014) Comparison of *d. melanogaster* and *c. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res* 24: 1086-1101.
89. Gerstein M, Rozowsky J, Yan K, Wang D, Cheng C, et al. (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512: 445-448.
90. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T (2012) The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc Natl Acad Sci U S A* 109: 17507-17512.
91. Zdobnov E, Apweiler R (2001) Interproscan—an integration platform for the signature-recognition methods in interpro. *Bioinformatics* 17: 847-848.
92. Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 25: 25-29.
93. Li L, Stoeckert C Jr, Roos D (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13: 2178-2189.

94. Alexeyenko A, Tamas I, Liu G, Sonnhammer E (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 22: e9-15.
95. Siepel A, Bejerano G, Pedersen J, Hinrichs A, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034-1050.
96. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* 125: 279-284.
97. Guberman J, Ai J, Arnaiz O, Baran J, Blake A, et al. (2011) Biomart central portal: an open database network for the biological community. *Database (Oxford)* 2011: bar041.
98. Cherbas L, Willingham A, Zhang D, Yang L, Zou Y, et al. (2011) The transcriptional diversity of 25 drosophila cell lines. *Genome Res* 21: 301-314.
99. Domazet-Lošo T, Tautz D (2010) Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol* 8: 66.
100. Silver D, Levin M, Yanai I (2012) Identifying functional links between genes by evolutionary transcriptomics. *Mol Biosyst* 8: 2585-2592.

RESEARCH ARTICLE

Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages

Peter Sarkies^{1*}, Murray E. Selkirk², John T. Jones³, Vivian Blok³, Thomas Boothby⁴, Bob Goldstein⁴, Ben Hanelt⁵, Alex Ardila-Garcia⁶, Naomi M. Fast⁶, Phillip M. Schiffer⁷, Christopher Kraus⁷, Mark J. Taylor⁸, Georgios Koutsovoulos⁹, Mark L. Blaxter⁹, Eric A. Miska^{10*}

1 MRC Clinical Sciences Centre, Imperial College London, London, United Kingdom, **2** Department of Life Sciences, Imperial College London, London, United Kingdom, **3** The James Hutton Institute, Invergowrie, Dundee, United Kingdom, **4** Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America, **5** Center for Evolutionary and Theoretical Immunology, Department of Biology, University of New Mexico, Albuquerque, New Mexico, United States of America, **6** Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada, **7** Zoologisches Institut, Universität zu Köln, Cologne, NRW, Germany, **8** Molecular and Biochemical Parasitology, Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **9** Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom, **10** Wellcome Trust/Cancer Research UK Gurdon Institute, Cambridge, United Kingdom

* psarkies@imperial.ac.uk (PS); eam29@cam.ac.uk (EAM)

Abstract

Small RNA pathways act at the front line of defence against transposable elements across the Eukaryota. In animals, Piwi interacting small RNAs (piRNAs) are a crucial arm of this defence. However, the evolutionary relationships among piRNAs and other small RNA pathways targeting transposable elements are poorly resolved. To address this question we sequenced small RNAs from multiple, diverse nematode species, producing the first phylum-wide analysis of how small RNA pathways evolve. Surprisingly, despite their prominence in *Caenorhabditis elegans* and closely related nematodes, piRNAs are absent in all other nematode lineages. We found that there are at least two evolutionarily distinct mechanisms that compensate for the absence of piRNAs, both involving RNA-dependent RNA polymerases (RdRPs). Whilst one pathway is unique to nematodes, the second involves Dicer-dependent RNA-directed DNA methylation, hitherto unknown in animals, and bears striking similarity to transposon-control mechanisms in fungi and plants. Our results highlight the rapid, context-dependent evolution of small RNA pathways and suggest piRNAs in animals may have replaced an ancient eukaryotic RNA-dependent RNA polymerase pathway to control transposable elements.



OPEN ACCESS

Citation: Sarkies P, Selkirk ME, Jones JT, Blok V, Boothby T, Goldstein B, et al. (2015) Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. *PLoS Biol* 13(2): e1002061. doi:10.1371/journal.pbio.1002061

Academic Editor: Laurence D Hurst, University of Bath, UNITED KINGDOM

Received: September 1, 2014

Accepted: January 2, 2015

Published: February 10, 2015

Copyright: © 2015 Sarkies et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Small RNA sequencing data has been deposited at GEO (GSE56651).

Funding: PS salary was funded by a Research Fellowship from Gonville and Caius College Cambridge and by a Career Development Award from the Medical Research Council. EAM received funding from Cancer Research UK (RG57329, <http://www.cancerresearchuk.org>) and an ERC Starting Grant (RG58558, <http://erc.europa.eu>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: EAM's spouse is a member of the *PLOS Biology* editorial staff; in accordance with the PLOS policy on competing interests she has been excluded from all stages of the review process for this article.

Abbreviations: dsRNA, double-stranded RNA; miRNA, micro RNA; nt, nucleotide; piRNA, Piwi interacting RNA; RdRP, RNA dependent RNA polymerase; RNAi, RNA interference; siRNA, small interfering RNA.

Author Summary

Transposable elements are segments of DNA that have the ability to copy themselves independently of the host genome and thus pose a severe threat to the integrity of the genome. Organisms have evolved mechanisms to restrict the spread of transposable elements, with small RNA molecules being one of the most important defense mechanisms. In animals, the predominant small RNA transposon-silencing mechanism is the piRNA pathway, which appears to be widely conserved. However, little is known about how small RNA pathways that target transposons evolve. In order to study this question we investigated small RNA pathways across the nematode phylum, using a well-studied model organism—the nematode *Caenorhabditis elegans*—as the starting point. Surprisingly we found that the piRNA pathway has been completely lost in all groups of nematodes bar those most closely related to *C. elegans*. This finding raises the intriguing question of how these nematodes are able to control transposable element mobilization without piRNAs. We discovered that there are other small RNA pathways that target transposable elements in these nematodes, employing RNA-dependent RNA polymerases in order to make small RNAs antisense to transposable elements. Intriguingly, the most ancient of these mechanisms, found in the most basal nematodes, is a Dicer-dependent RNA-directed DNA methylation pathway. This pathway shares strong similarity to transposon-silencing mechanisms in plants and fungi, suggesting that it might have been present in an ancient common ancestor of all eukaryotes. Our results highlight the rapid evolution of small RNA pathways and demonstrate the importance of examining molecular pathways in detail across a range of evolutionary distances.

Introduction

Transposable elements are found in almost all eukaryotic genomes, and present a severe threat to the integrity of the germline and the survival of the species. Consequently, organisms have evolved robust pathways to silence the expression of transposable elements and restrict their spread [1–3]. Small (18–35 nucleotide [nt]) RNAs are amongst the most important of these pathways and several different transposon-control mechanisms involving small RNAs are found across the eukaryotic domain. In animals, transposon silencing is controlled by Piwi-interacting small RNAs (piRNAs), which associate with the conserved Piwi subfamily of Argonaute proteins and are essential for fertility in *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, and the nematode *C. elegans* [4,5]. Along with microRNAs, which associate with the Ago subfamily of Argonautes, piRNAs are widely conserved across the animal kingdom [4,6,7]. However, other small RNA pathways are restricted to specific phyla, and the evolutionary and functional relationships between them are unclear, particularly because the majority of information available relates to a few very distantly related model organisms. Thus to understand how small RNA pathways evolve, a range of organisms over a variety of different evolutionary distances need to be studied. To carry out such an analysis we chose to study their evolution across the phylum Nematoda.

The best understood nematode is the model organism *C. elegans*, in which extensive studies of small RNAs have been undertaken. *C. elegans* possesses several classes of small RNAs, most of which are conserved. In *C. elegans* as in other organisms microRNAs (miRNAs) are transcribed from individual genomic loci to form hairpins that are processed by the activity of Dicer to produce mature miRNAs. The sequences of many *C. elegans* miRNAs are highly

conserved all the way to humans and they have important functions in regulating key developmental transitions [8].

The *C. elegans* genome encodes two members of the Piwi subfamily of Argonautes, PRG-1 and PRG-2. The *prg-2* gene is the result of a recent gene duplication and does not appear to function directly in the piRNA pathway, but *prg-1* encodes a functional protein, which is expressed in the germline and binds to piRNAs [9,10]. However, in contrast to the high conservation of miRNAs and miRNA-processing, *C. elegans* piRNAs have some important differences to piRNAs in other animals. In *C. elegans* the 5' U bias common to most animal piRNAs is conserved; however, they are only 21 nt long as opposed to the 26–30 nt more common in *M. musculus* and *D. melanogaster* [9,10]. In addition, piRNAs are produced from individual loci that are transcribed to produce short (26–30 nt) precursors that are processed to give rise to mature piRNAs [11,12], as opposed to the long piRNA precursor transcripts produced in *M. musculus* and *D. melanogaster* that are processed to give rise to multiple piRNAs per genomic locus [7,13]. The majority of piRNA loci in *C. elegans* are associated with an upstream sequence motif [9,10,14].

C. elegans piRNAs also differ from *D. melanogaster* and *M. musculus* piRNAs owing to their different mechanism of target silencing. piRNA-mediated silencing does not involve the direct cleavage of targets by the Slicer endonuclease domain of PRG-1. Instead, piRNAs silence their targets by initiating the synthesis of an abundant class of small interfering RNAs (siRNAs) through an RNA-dependent RNA polymerase (RdRP) [10,15]. These siRNAs align predominantly antisense to targets, are ~22 nt, and start with a guanine (G), thus are also referred to as 22G-RNAs [16,17]. Importantly, because each 22G-RNA is produced by a RdRP, they carry a 5' triphosphate, whilst both piRNAs and miRNAs possess a 5' monophosphate [16,17]. The *C. elegans* RdRPs RRF-1 and EGO-1 are required for 22G-RNA biogenesis, with the RRF-2 RdRP being dispensable [18]. The fourth *C. elegans* RdRP, RRF-3 is required instead for the production of another class of small RNAs, the 26G-RNAs, which have a 5' monophosphate [19]. It remains unclear whether RRF-3's catalytic activity is required for 26G-RNA production. 22G-RNAs associate with multiple “worm”-specific Argonaute proteins (WAGOs) [5] to bring about target silencing, and in addition to being produced downstream of piRNA targeting, 22G-RNAs are also produced downstream of target recognition by other classes of endogenous small RNAs and RNA interference induced by exposure to double-stranded RNA (dsRNA) [20].

Despite divergence in biogenesis and silencing mechanisms, *C. elegans* piRNAs have a similar function to those in other organisms, as they target transposable elements for silencing [10,15]. Additionally, *C. elegans prg-1* mutants show fertility defects [5], and become sterile over many generations [21], meaning that an important role for the piRNA pathway in protecting the function of the germline is conserved across animal species. Thus the *C. elegans* piRNA pathway represents an interesting example of where a conserved central core (the Piwi/piRNA complex) has acquired different upstream and downstream components whilst retaining its ancestral function.

In order to gain further insight into how piRNAs evolve in the context of other small RNA pathways, we used *C. elegans* as a basis to guide an examination of small RNAs and the proteins that bind to them across the known diversity of the phylum Nematoda (Fig. 1A). Our analysis reveals that, surprisingly, piRNAs have been lost several times independently across the phylum. Instead, we find that in the Chromadoria group of nematodes (clades III–V), 22G-RNAs produced by RNA dependent RNA polymerase operate in the absence of piRNAs to target transposable elements. In the Dorylamia group of nematodes, more distant to *C. elegans*, (clades I/II), small RNAs targeting transposons are produced by a different RNA dependent RNA polymerase pathway, in this case acting processively to generate dsRNA that is then

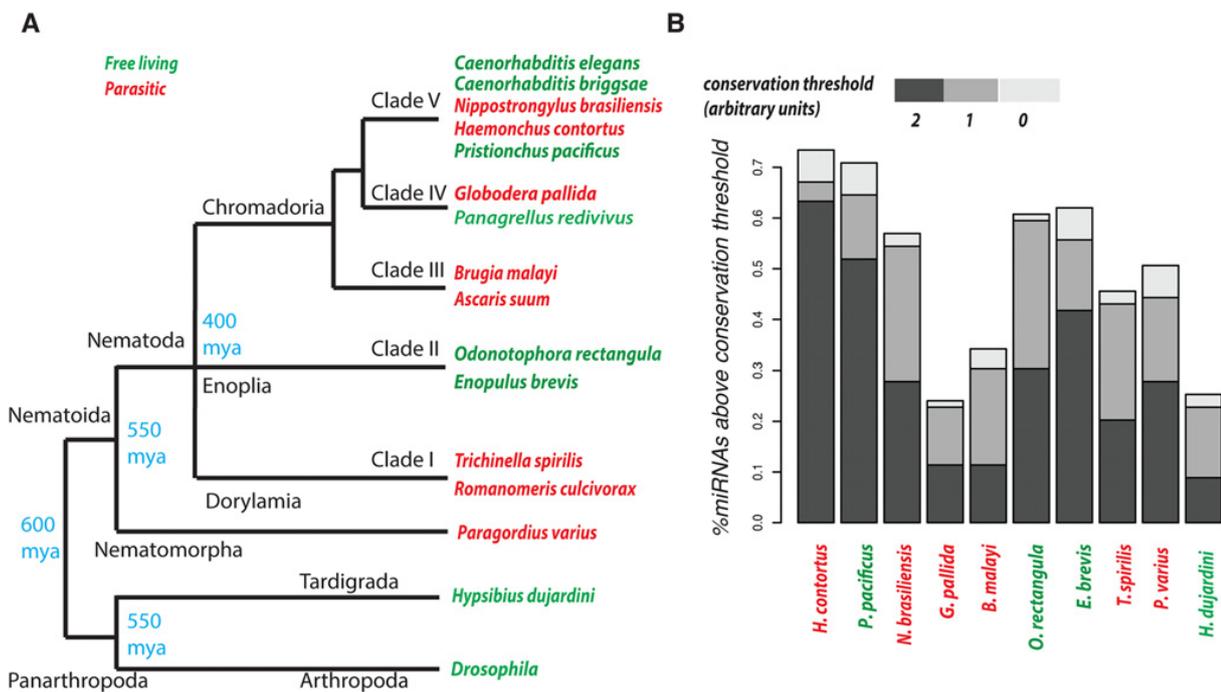


Fig 1. Conservation of miRNAs in the phylum Nematoda. (A) Cladogram of the phylogenetic systematics of Nematoda, using the clades I–V as defined by Blaxter and colleagues [22]. Numbers in blue represent approximate divergence times for key nodes [23]. Red and green represent parasitic and free-living species respectively. (B) Percentage of *C. elegans* miRNAs conserved in nematode species coloured according to using three different stringency cutoffs (see S1 Text for detailed description).

doi:10.1371/journal.pbio.1002061.g001

processed by Dicer into small RNAs. This pathway also involves DNA methylation of transposable elements thus displays similarity to the RNA directed DNA methylation pathway found in plants and fungi. Our results provide a clear example of the rapid diversification of molecular pathways involved in silencing repetitive elements but at the same time identify hitherto unknown conservation at the core of the eukaryotic small RNA machinery.

Results and Discussion

To investigate small RNA pathway evolution, we sequenced small RNAs from 11 species across the phylum Nematoda, and analysed previously published data from two more, spanning the known diversity (Fig. 1A; S1 Text). We selected at least one species from each of the five clades (I–V; *C. elegans* is in clade V) [22]. Our sample included parasites of animals and plants, and free-living nematodes, and included two representatives of the neglected clade II marine species. Additionally, where possible, we collected samples from at least two life cycle stages (S1 Text), always including adults, as some small RNA pathways are enriched in or exclusive to the germline in *C. elegans* [9,10]. We first analysed small RNAs with a 5' monophosphate, which in *C. elegans* include miRNAs and piRNAs (S1 Fig.). To assess miRNA conservation we used the annotated *C. elegans* miRNAs as a reference and calculated a conservation score based on sequence conservation and relative expression levels (see Materials and Methods). By this measure all nematode species showed conservation of at least 20% of *C. elegans* miRNAs (Fig. 1B), with the proportion of miRNAs conserved rising with decreasing phylogenetic distance to *C. elegans* (Spearman's Rho = 0.47; $p = 0.02$). We focussed on miRNA families that are

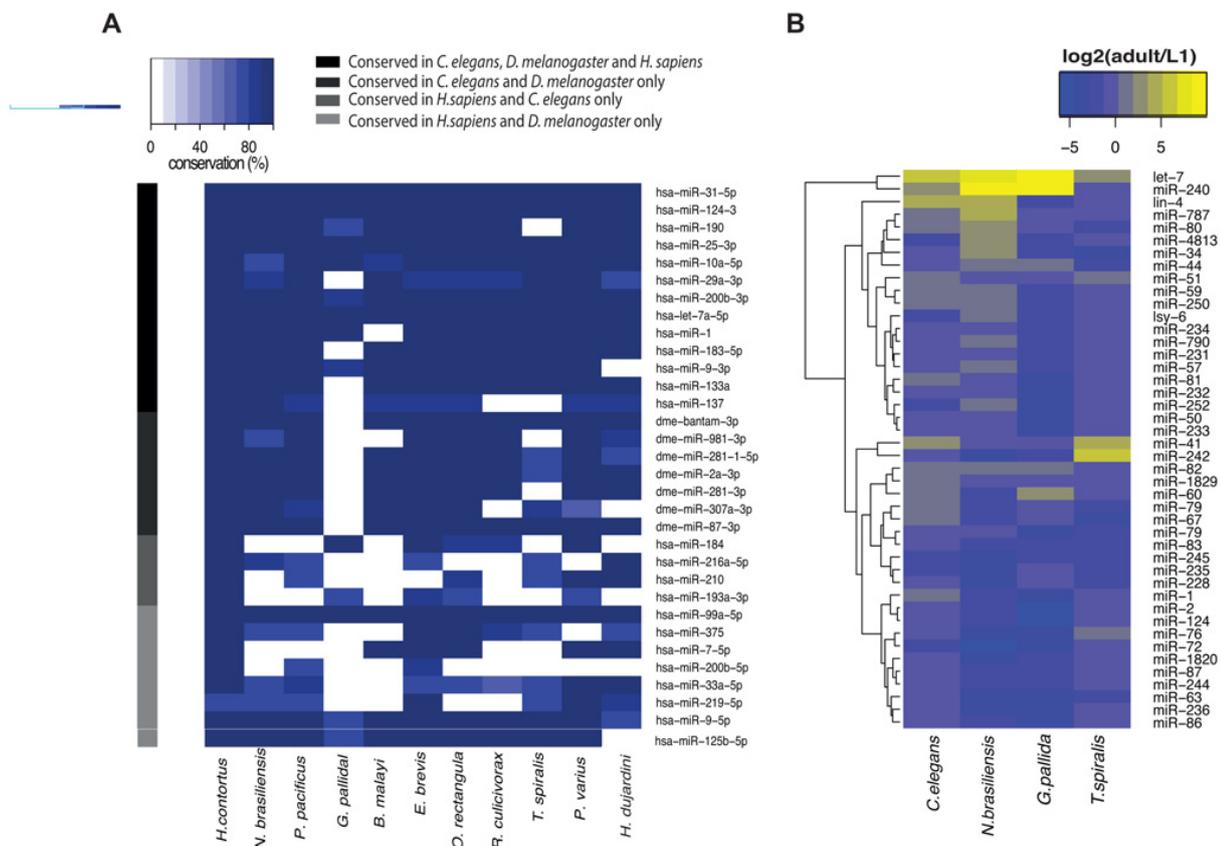


Fig 2. miRNA families and expression in the phylum Nematoda. (A) Conservation of conserved miRNA families, grouped by seed in nematodes. The family is named after the most abundant *H. sapiens* family member, or the human member if no *H. sapiens* member exists. The colour bar indicates conservation of a miRNA in *C. elegans*, *D. melanogaster*, *H. sapiens* or a combination thereof. (B) The developmental profile of miRNAs is conserved across Nematoda. All *C. elegans* miRNAs present in all of the species profiled are shown.

doi:10.1371/journal.pbio.1002061.g002

conserved in at least two out of *C. elegans*, *D. melanogaster*, and *Homo sapiens* as likely ancestral bilaterian miRNAs. The majority were conserved across all the species we examined, and only two of these miRNAs were apparently lost in all nematodes profiled; however, there were isolated examples of species-specific losses of miRNAs, suggesting that their conservation amongst bilaterians is not universal (Fig. 2A). Additionally, we compared small RNA levels for miRNAs at early larval (L1 or L2) and adult stages for *C. elegans*, *Trichinella spiralis* (clade I), *Globodera pallida* (clade IV), and *Nippostrongylus brasiliensis* (clade V), as the life cycles of these species permitted collection of the required life stages. Broadly, these nematodes all showed similar developmental miRNA expression dynamics to *C. elegans* with correlation between expression changes in *C. elegans* and other nematodes better for more closely related nematodes (Figs. 2B and S2). Taken together, these data show that most miRNA sequences and their developmental regulation are highly conserved across the Nematoda.

We next investigated piRNAs. In *C. elegans*, piRNAs are 21 nt long and start with a uracil (U) also referred to as 21U-RNAs [9,10,14]. Most piRNAs in *C. elegans* are clustered in the genome and are associated with a defined upstream sequence motif [9,10]. Previous analysis has shown that similar upstream motifs are found upstream of 21U-RNAs in other *Caenorhabditis* nematodes and in *Pristionchus pacificus* [1,3]. Since the motif may have diverged in other

nematodes we aligned small RNAs to their source genomes and searched for motifs upstream of putative piRNA loci. Using this approach we were able to recover a motif with strong similarity to the *C. elegans* core motif upstream of 21U-RNAs in *Haemonchus contortus* [24] (clade V) and *N. brasiliensis* (Fig. 3A–3C). Thus the 21U-RNAs and the *C. elegans* core motif are highly conserved across the clade V nematodes. However, we were unable to identify a motif for any nematode species outside of clade V. This is unlikely to be due to poor genome assembly as the genome of *N. brasiliensis* is the least complete of any of the genomes in our study (S1 Text). Moreover, when predicted miRNAs were excluded from the analysis, we did not detect small RNAs of any size with a 5' U bias in non-clade V species (Figs. 4 and S4). Thus we did not detect 21U-RNAs or 25–32 nt long piRNAs in nematodes outside clade V. Consistent with our analysis, piRNAs were previously reported to be absent in the parasitic nematode *Ascaris suum* (clade III) [25] and the free-living nematode *Panagrellus redivivus* (clade IV) [26]. Given the topology of the phylogenetic tree (Fig. 1A), these data suggest that the piRNA pathway has been lost independently in multiple nematode lineages.

To confirm the loss of piRNAs, we explored the conservation of proteins involved in the RNA interference (RNAi) pathway, as defined in *C. elegans* (Fig. 3D). We found high conservation of the proteins involved in the miRNA pathway as well as some proteins known to be involved in other endogenous small RNA pathways (see below). However the *C. elegans* Piwi orthologue PRG-1 was absent from all non-clade V nematodes [28] despite being highly conserved in *D. melanogaster* (Fig. 3D). Strikingly, the only other protein showing this pattern of conservation was HENN-1 (Fig. 3D), which stabilises piRNAs by adding a 2' O methyl at the 3' end and is thus an important component of the piRNA pathway in animals including *C. elegans* [29–33]. Furthermore, we also found no PRG-1 orthologue in other clade III and IV nematodes with genome assemblies, nor analysing draft transcriptome data from *E. brevis* (clade II) (S1 Text). This distribution of conservation is unlikely to be due to a sampling error due to incomplete genome assemblies because such a distribution of conservation, placing *D. melanogaster* closer to *C. elegans* than to any non-clade V nematodes was very unusual within the *C. elegans* proteome (estimated Jack-knife $p < 10^{-4}$; see Materials and Methods). Moreover, this loss is unlikely to be an effect of rapid evolution of PRG-1 because it evolves more slowly relative to the median rate of all *C. elegans* proteins in all the species in which we detect it (S5A Fig.). To test this hypothesis further, we simulated evolving PRG-1 to the distance between PRG-1 and *D. melanogaster* Piwi 1,000 times and spiked the simulated protein into the *T. spiralis* gene set. In 1,000 simulations the maximum e-value we observed, corresponding to the weakest hit, between the simulated protein and PRG-1 was still 10^{10} lower than the best hit to PRG-1 in the true *T. spiralis* set (S5B Fig.).

In order to test whether an independent approach could reproduce the loss of PRG-1 in non-clade V nematodes we used phmmer to find the highest scoring homologue of *D. melanogaster* Piwi in all the nematode species we tested. We then constructed a maximum-likelihood phylogenetic tree of these proteins, spiking in the known Piwi proteins from mammals as well as *C. elegans* ALG-1 and *D. melanogaster* AGO as examples of members of the Ago subfamily, responsible for Dicer-dependent small RNA binding [6]. Whilst the best homologues to Piwi in every clade V nematode clustered with the other Piwi proteins, the best homologues of Piwi in every non-clade V nematode were clearly members of the Ago subfamily (Fig. 5). Taken together these data strongly support the loss of PRG-1, and with it the piRNAs that depend on Piwi proteins for their stability, independently in several nematode lineages outside of clade V.

As the repeated loss of piRNAs in such a speciose phylum was unexpected, we decided to examine other related ecdysozoan phyla. We first profiled small RNAs in tardigrades, which are Panarthropods, and thus more closely related to *D. melanogaster* than to nematodes [34]. We identified small RNAs in the tardigrade *Hypsibius dujardini* showing a strong 5' U bias and

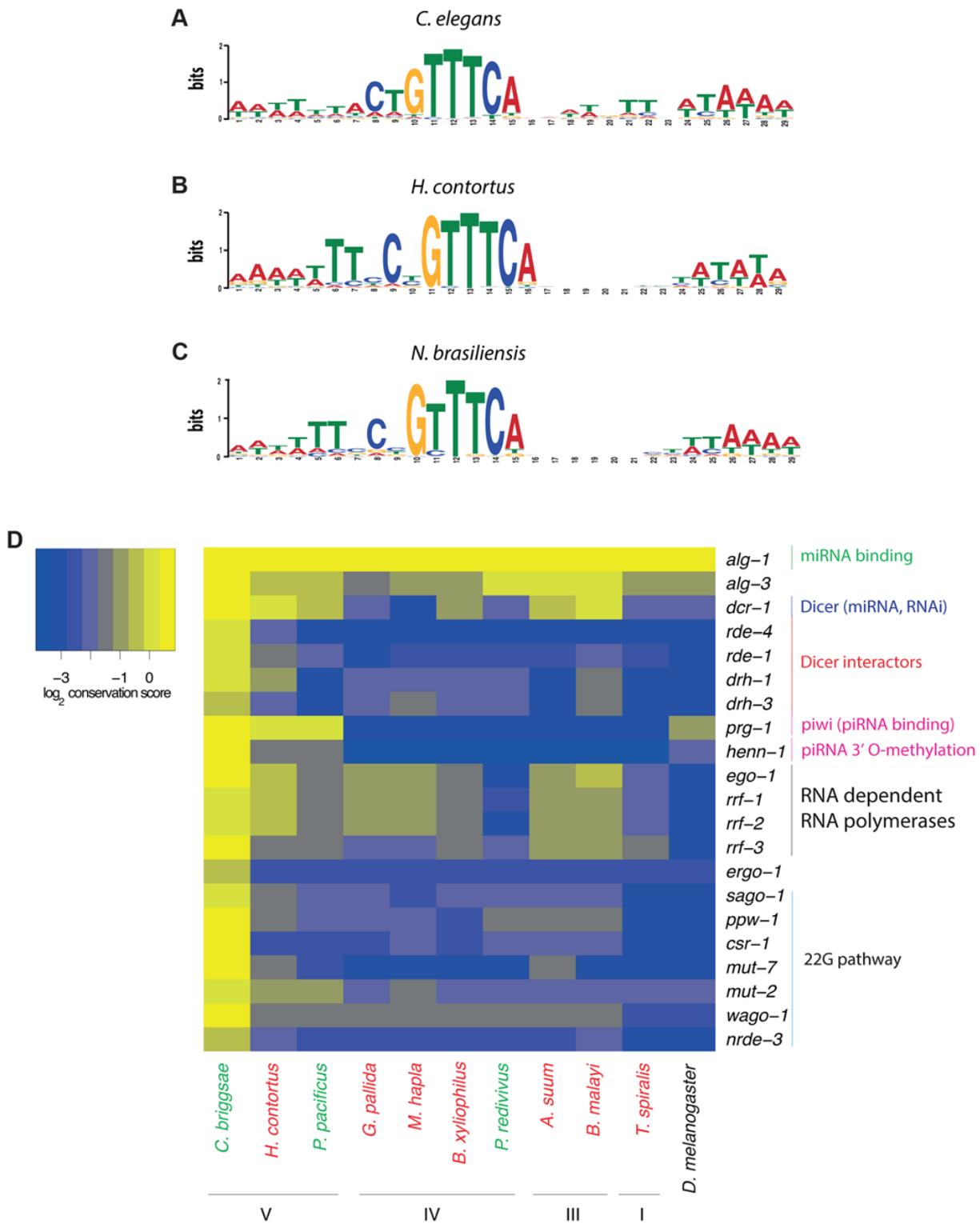


Fig 3. Loss of the piRNA pathway in nematodes outside of clade V. (A–C) Sequence motifs were *de novo* predicted in the upstream regions of aligned 21U-RNA sequences in clade V nematodes: (A) *C. elegans*; (B) *H. contortus*; (C) *N. brasiliensis*. The motifs are presented as sequence logos generated using the MEME program [27]. The upstream sequences for *C. elegans* were taken from [9]. The upstream sequences for *H. contortus* and *N. brasiliensis* 21U-RNAs are in the S4 Data. (D) Conservation of selected *C. elegans* small RNA pathway proteins. Conservation is calculated as the logarithm of the score of the best blast hit in bits normalized to the length of the protein (see S1 Text).

doi:10.1371/journal.pbio.1002061.g003

with a modal length of 26 nt (S3A Fig.). Moreover, we identified a Piwi homologue from draft transcriptome data, suggesting that these small RNAs could be piRNAs (S1 Text). *H. dujardini* piRNAs were clustered on the genome, and showed a signature suggestive of ping-pong amplification as characterized in *D. melanogaster* [13], whereby the tenth nucleotide of overlapping transposons shows a bias towards adenine (A) (S3B Fig.). piRNAs in the tardigrade are thus more similar to those of *D. melanogaster* than to *C. elegans*.

We next examined *Paragordius varius*, member of the phylum Nematomorpha, which is the sister phylum of Nematoda [34]. Interestingly, although we clearly identified miRNAs in *P. varius*, including many that are widely conserved in nematodes (Fig. 2A), we were unable to identify longer piRNAs similar to those seen in tardigrades (S3C Fig.). Thus we concluded that the longer 25–32 nt piRNAs found in many animal phyla may have been lost in the common ancestor of Nematoda and Nematomorpha. Consistently, we were unable to identify a Piwi orthologue analysing draft transcriptome data from *P. varius* (S1 Text).

Given the essential role of piRNAs in targeting transposable elements for silencing, we wondered whether nematodes that lack piRNAs might use other small RNA pathway to target transposable elements. In *C. elegans*, piRNAs act upstream of the generation of 22G-RNAs by RNA dependent RNA polymerase [10]. However, 22G-RNA-mediated silencing can in some cases persist for some time independently of piRNA activity and can also compensate for the absence of piRNAs in some circumstances [21]. We therefore tested whether 22G-RNAs are able to compensate for the absence of piRNAs in nematodes outside of clade V. Importantly, this class of RdRP-derived small RNAs retain their 5' triphosphate [16,17] allowing us to use comparison between small RNA sequencing of 5' monophosphate only species with 5' triphosphate and 5' monophosphate containing species (S1 Fig.) to detect 5' triphosphorylated small RNAs with a 5' G in *Brugia malayi* (clade III) and *G. pallida* (clade IV). These small RNAs had a modal length of 22 nt in *B. malayi* and 22–26 nt in *G. pallida* (Fig. 4C, 4D, 4G, and 4H). However, we did not detect any 5' triphosphorylated small RNAs in *T. spiralis* (clade I) (Fig. 4E and 4F), *Romanomeris culicivox* (clade I), *Enoplus brevis* (clade II), or *Odontophora rectangula* (clade II) (S4 Fig.). This suggests that RdRP-derived 22G-RNAs evolved in the last common ancestor of nematode clades III–V.

To examine this analysis further we examined the conservation of RdRPs in nematodes. In *C. elegans*, the RdRPs RRF-1, EGO-1, and RRF-2 generate 22G-RNAs whereas the RdRP RRF-3 is required for a separate small RNA pathway involving Dicer [35]. We identified RdRPs in all nematode clades. However RRF-1, RRF-2, and EGO-1 were only found in clades III–V (Fig. 3D; S1 Text), whilst RRF-3-like genes were found in all clades (Fig. 3D), including clade II (S1 Text). Thus the RRF-1 RdRP family likely arose in the last common ancestor of clades III–V. To investigate this further we performed multiple sequence alignment and phylogenetic analysis of eukaryotic RdRPs (Fig. 6). This analysis suggested that nematode RNA dependent RNA polymerases fall into two groups, the RRF-3 family RNA dependent RNA polymerases, conserved across the whole nematode phylum, and the RRF-1/EGO-1 family that is only conserved in clades III–V (Fig. 6A; S1 Text). Interestingly the RRF-1/EGO-1 family can be distinguished from RRF-3 on the basis of a conserved insertion, containing two proline residues and a tryptophan residue, which is not present in RRF-3 or any RdRPs from other organisms (Fig. 6B). In viruses, a key difference between *de novo* initiating RdRPs and those that initiate

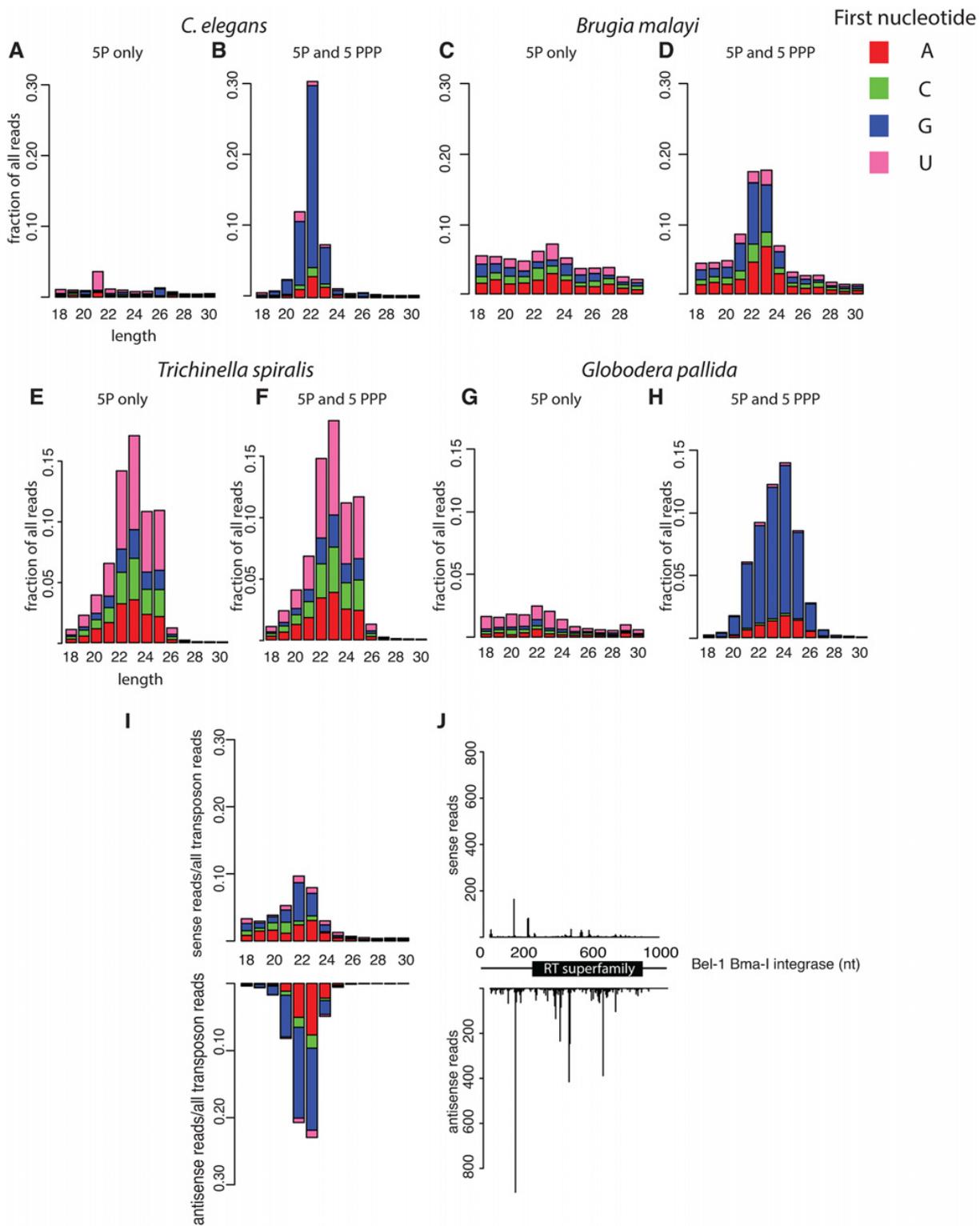


Fig 4. siRNAs combat transposons across the phylum Nematoda. (A–H) Non-processive RdRP activity evolved in the ancestor of clade III–V nematodes. Comparison between libraries prepared from 5' monophosphate only and 5' monophosphate and 5' triphosphate material demonstrates that 5' triphosphate small RNAs are present in *C. elegans* (A, B), *G. pallida* (G, H), and *B. malayi* (C, D) but not found in *T. spiralis* (E, F). (I) 22G-5' triphosphate small RNAs target transposons in *B. malayi*. (J) An example *B. malayi* transposon showing the distribution of small RNAs across the predicted transposon sequence. The transposon shown is a member of the BEL/PAO retrotransposon family. Complete Repeatmasker annotations of *B. malayi* are in [S4 Data](#).

doi:10.1371/journal.pbio.1002061.g004

in a primer-dependent manner is the presence of an extra loop in de novo polymerases, often containing aromatic residues, which enables the initiating G nucleotide to be fixed in the active site despite weak interactions with the template [36]. Although eukaryotic RdRPs are very distantly related to viral RdRPs, and possess different catalytically active residues [37], there are some similarities between the overall architecture of the only eukaryotic RdRP with a solved structure, QDE-1 from *Neospora crassa*, and viral RdRPs [38]. In vitro QDE-1 uses looped-back RNA from the template as a primer [39,40], although the enzyme may also initiate de novo at the 3' end of certain templates [39]. It is therefore plausible that, analogous to viral RdRPs, the conserved extra loop in RRF-1 family RdRPs is important for de novo initiation, whilst the more ancient RRF-3 family polymerases utilise primer-dependent initiation similarly to QDE-1. This would be consistent with the different products of the two groups of nematode RdRPs: short, 5' triphosphorylated G-RNAs arise from the activity of the de novo initiating RRF-1/EGO-1 family polymerases whilst longer RNAs are made by the RRF-3 polymerases (Table 1).

Having established that 22G-RNA RdRP products are conserved in clades III–V we tested whether they could be important in defence against transposable elements in nematodes lacking piRNAs. We identified potential transposable element sequences in the nematode genomes with RepeatMasker, and aligned small RNAs to them to identify potential siRNAs. *B. malayi*

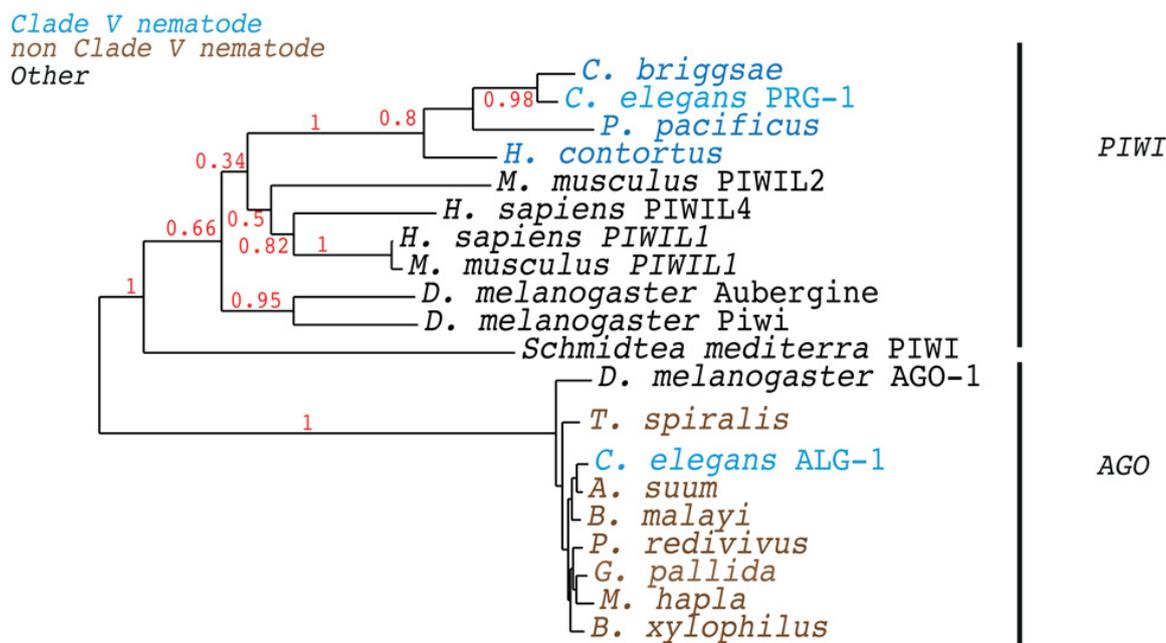


Fig 5. Loss of the Piwi protein in nematodes outside of clade V. Maximum likelihood phylogenetic tree of the best homologues of *D. melanogaster* Piwi found in the analysed nematode genomes. We included a subset of known Piwi proteins and Ago proteins from other organisms for comparison. The alignment is in [S1 Data](#). Bootstrap branch support values from 100 bootstraps are shown. The tree file is in [S4 Data](#).

doi:10.1371/journal.pbio.1002061.g005

and *G. pallida* both had abundant 5' triphosphate RNAs aligning to transposons (Fig. 4I and 4J; S6 Fig.). There were more 22G siRNA sequences mapping antisense than sense (for both: $p < 2.2e-16$, Chi-squared test against a uniform distribution). In *B. malayi* the antisense bias was greater for 22G-RNAs than for non 22G-RNA sequences, whilst in *G. pallida* the antisense bias was seen for 20–26 nt long 5' triphosphorylated G-RNAs ($p = 3.4e-7$ and $p = 7.7e-4$ respectively, Chi-squared test with Yates continuity correction). This finding suggests that transposon silencing is mediated by RdRPs in species other than *C. elegans*, even though these species lack piRNAs.

We wondered what might initiate the formation of RdRP siRNAs against transposons in the absence of piRNAs. Interestingly, we detected small quantities of 22–23 nt 5' monophosphate RNAs aligning to both strands of transposons in *B. malayi* and *G. pallida* (Figs. 4 and S6). These are likely siRNAs generated by Dicer acting on dsRNA and may therefore act to recruit RdRP activity to transposons without the activity of piRNAs (S7 Fig.). It is not clear why *C. elegans* and other clade V nematodes have retained the piRNA pathway in addition to the 22G-RNA pathway. PRG-1 has functions independent of initiating 22G-RNA mediated silencing [21], thus one possibility is that clade V-specific transposable elements exist that cannot be silenced by 22G-RNAs but require PRG-1 for their control. Loss of these elements in other lineages might have enabled subsequent loss of the Piwi protein.

T. spiralis has neither piRNAs nor 5' triphosphate small RNAs. Nevertheless, aligning small RNAs from *T. spiralis* to transposon sequences showed abundant 23–25 nt RNAs enriched antisense to transposons (Fig. 7A; $p < 2e-16$, Chi-squared test against a uniform distribution). Several features of these sequences suggest that they are of an evolutionary distinct origin to the piRNAs found in either *C. elegans* or *D. melanogaster*. First, piRNAs are made by Dicer-independent mechanisms [7]. However, overlapping sense-antisense pairs of *T. spiralis* small RNAs showed a two nt 3' overhang consistent with their being the product of Dicer activity on dsRNA (Figs. 7B and S1). Second, piRNAs in *C. elegans* and *D. melanogaster* are produced by transcription from clusters containing exceptionally high numbers of piRNA sequences [13,14]; *T. spiralis* 23–25 nt small RNAs, however, are distributed more evenly throughout the genome ($p < 1e-7$ to *D. melanogaster*, $p = 0.01$ to *C. elegans*, Kolmogorov-Smirnov test for different distributions), such that clusters with >10 times the mean density of small RNAs genome-wide are virtually absent (S8 Fig.). Third, piRNAs are typically enriched for 2' O-methylation. We tested for this modification using protection against sodium periodate and whilst we readily detected specific protection of *C. elegans* 21U-RNAs, *T. spiralis* 23–25 nt small RNAs were not protected by the treatment (S8 Fig.), consistent with our finding that the HENN-1 protein responsible for 2' O-methylation of piRNAs is not conserved in *T. spiralis* (Fig. 3D).

Given that the *T. spiralis* transposon-silencing small RNAs were produced by Dicer, and that RNA dependent RNA polymerase is conserved in *T. spiralis*, we wondered whether RdRP could produce dsRNA as a substrate for Dicer using the transposon mRNA as a template (S6 Fig.). Such a pathway is found in plants and fungi, and acts upstream of DNA methylation and/or histone modification to silence transposons, a process known as RNA-directed DNA methylation [41–43]. While absent in *C. elegans* and related nematodes, DNA methylation is found in *T. spiralis* [44] and the DNA methylation machinery is present in *R. culicivora* [45]. We analysed CG methylation using genome-wide bisulfite sequencing of *T. spiralis* adults [44] and found a highly significant increase in meCpG density in LTR retrotransposons and DNA transposons compared to genome-wide meCpG (Fig. 7C). Moreover we observed a striking correlation between methylated regions of the genome and enrichment for small RNAs (Fig. 7D), implying that RNA-directed DNA methylation may occur in *T. spiralis*. Furthermore one of the three Dicer paralogues in *T. spiralis* contains a high-scoring bipartite nuclear

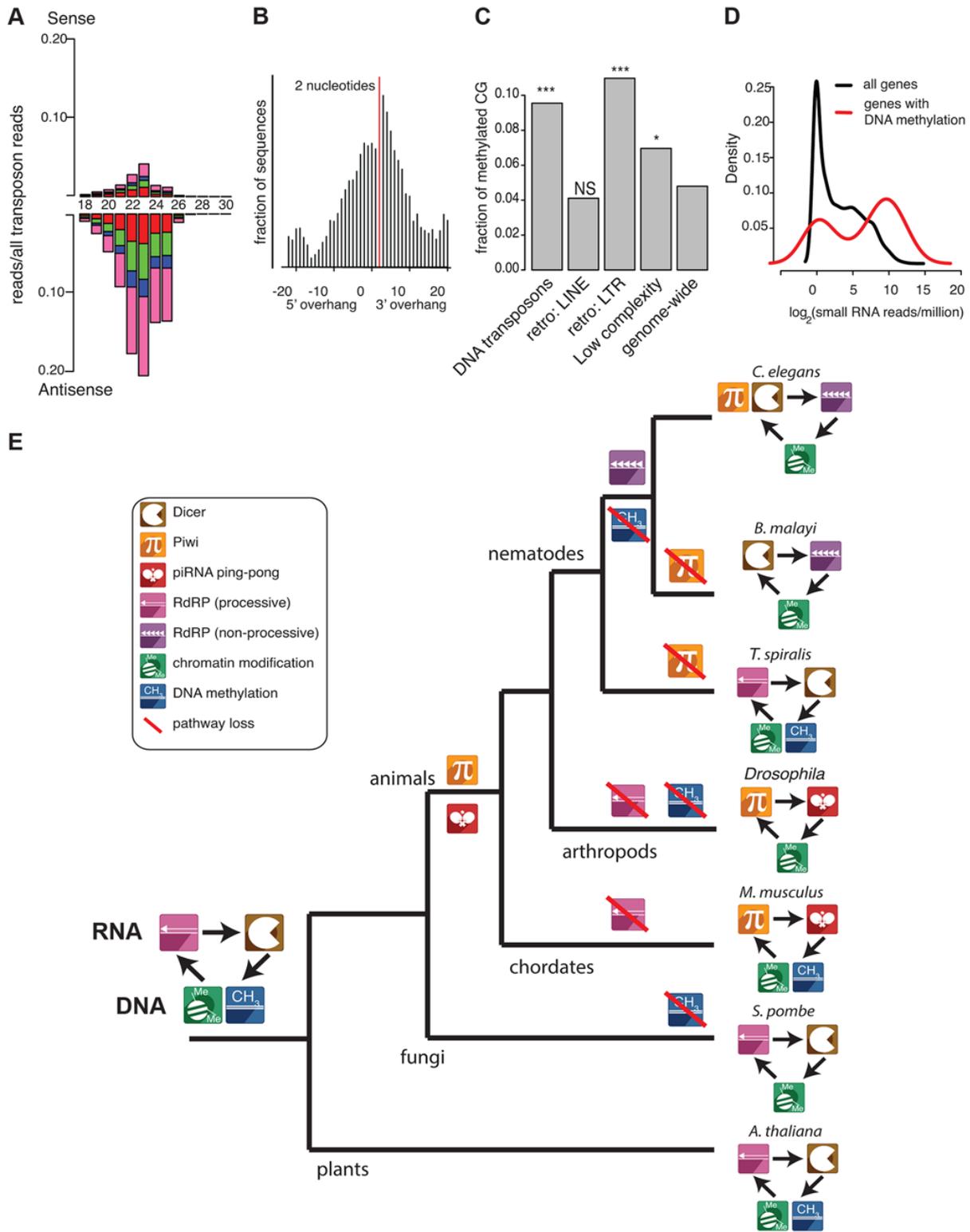


Fig 7. Conservation of eukaryotic transposon-defence mechanisms. (A) *T. spiralis* endo siRNAs align predominantly antisense to transposon sequence. (B) Anti-transposon endo siRNAs in *T. spiralis* have the characteristic 2 nt overhang of Dicer products. (C) CpG DNA methylation in *T. spiralis* is enriched at LTR retrotransposons and DNA transposons relative to genome-wide CG DNA methylation levels. Categories shown are as defined by Repeatmasker. ***chi-squared $p < 1e-40$; *chi-squared $p < 1e-5$. *T. spiralis* Repeatmasker annotations are in the supporting information. DNA methylation data for the analysis was taken from [41]. (D) Evidence for RNA directed DNA methylation in *T. spiralis*. The density plot shows small RNA reads for all genes (black) and genes with DNA methylation (red). A shift towards higher levels of small RNA reads is seen for DNA methylated genes. (E) A model for the evolution of transposon silencing pathways. Expression of transposable elements is recognized at the level of RNA through structural or sequence features. Consequently, transposable elements are silenced post-transcriptionally through cleavage (RNA) or transcriptionally through histone modification and/or DNA methylation (DNA). This interplay between the RNA and DNA level through different pathways is shown here. Symbols indicate pathways used for defence against transposons in a species, not conservation of individual protein factors. Note that RdRPs can act upstream or downstream of Dicer. Branch lengths are for illustration only.

doi:10.1371/journal.pbio.1002061.g007

localization motif (S3 Data). *Schizosaccharomyces pombe* Dicer also has a bipartite nuclear localization motif and enzymes involved in plant RNA-directed DNA methylation are nuclear localized [46,47]. Thus we propose that RNA-directed DNA methylation mediated by RdRPs and nuclear Dicer acts to silence transposon sequences in *T. spiralis*.

We suggest that RNA-directed DNA methylation involving Dicer and RdRP activity is an ancestral transposon silencing mechanism for eukaryotes (Fig. 7E). This system provides target recognition, amplification of small RNAs, and transcriptional repression thus ensuring robust transposon silencing. During the evolution of animals, this central module diversified to utilise piRNAs for target recognition. In some cases, as in mammals or *D. melanogaster*, piRNAs can operate independently of RdRP activity by using a different amplification strategy (the ping-pong mechanism). In surveying the diversity of small RNAs within Nematoda, we found evidence of both the ancestral Dicer/RdRP pathway and the piRNA pathway. In addition a unique non-processive RdRP activity evolved, responsible for generation of a novel class of small RNAs—the 22G-RNAs. We predict that similar diversity is likely to be evident across other animal phyla—indeed, evidence exists for RdRPs in some arthropods [37]—underlining the need to survey wide groups of organisms beyond a few well characterised model systems in order to understand how molecular pathways have evolved.

Materials and Methods

Sample Collection

C. elegans was grown according to standard procedures [48]. *G. pallida* were collected as described [49]. *T. spiralis* adults and L1 were isolated according to standard procedures. *N. brasiliensis* adults were isolated from rodents and L1 larvae were collected from fecal matter according to standard procedures. *E. brevis* were collected from soil filtrates in Sylt, Germany with assistance from Werner Armonies (Alfred Wegener Research Station) and *O. rectangula* were collected from sand filtrates in Vancouver, Canada. In both cases collected nematodes were stored for transport in 1 ml RNA Later. *P. varius* were collected by placing infected crickets in water and frozen at -80°C for RNA extraction.

RNA Extraction and Small RNA Library Preparation

RNA was extracted using Trizol according to standard procedures. RNA Following RNA isolation, RNA was treated with 20U 5' polyphosphatase (Epicenter) for 30 min to remove 5' triphosphates and allow 5' independent library construction or used directly for 5' dependent library construction (S1 Fig.). To test for protection against 3' end oxidation, we treated RNA resuspended in sodium borate buffer (pH 8.6) either with sodium periodate at a final concentration of 25 mM or with an equal volume of water as a control sample. We incubated for 10

min at room temperature and quenched the reaction with glycogen for 10 min before desalting the RNA with a G25 column (GE healthcare) and precipitating with Ethanol.

All small RNA libraries were made with the Illumina Truseq method according to the manufacturer's instructions.

Analysis of Small RNA Sequencing Data

Fasta files with one sequence for each read were generated by using Cutadapt (v1) to trim adapters and a custom Perl script to convert fastq into fasta files. Small RNAs of between 15 and 33 nt were then selected using a custom script written in Perl. In the absence of genome sequence, a custom script written in Perl was used to collect the length and first nucleotide of all species and tabulate this information. The output was read into R to generate barplots. To carry out alignments, to various genomes or transposon sequences from RepeatMasker predictions, we used Bowtie [50] with parameters—v 0—k 1—best, to select only for reads that matched perfectly to the genome. This step generated sam files of the alignments, which were converted into bam files using samtools [51] and bed files using Bedtools [52] before reading into R for further analysis. Detailed description of the analysis of specific classes of small RNAs is available in [S1 Text](#).

Protein Evolution Analysis

In order to interrogate protein evolution we independently developed a method similar to those used for similar analyses [53]. To test for the conservation of *C. elegans* proteins across the nematode phylum, the predicted proteome from each species analysed was downloaded as a fasta file from the sources indicated in [S1 Text](#). As a non-nematode species we downloaded the predicted *D. melanogaster* proteome from Flybase (www.flybase.org). Formatdb was then run on each species separately to create individual databases. The *C. elegans* predicted proteome was used as input and blastp was run separately against each database. This generated output files that were filtered using a custom Perl script running in the BioPerl environment to select the best blast hit for each *C. elegans* protein for each database tested. In order to select the best hit, the bit score was divided by the length of the input sequence, as this method is more reliable than using the e-value [54]. To draw the heatmap in [Fig. 2](#), a curated list of small RNA related proteins was selected from the complete table. In order to test the statistical significance of the incongruity of the PRG-1 loss to the rest of the proteome we replaced the conservation table with a binary matrix representing presence absence calls at the 0.18 threshold, which is average length normalized bit score seen for PRG-1 in non-clade V nematodes. We then used a Jack-knife method to test how likely the pattern of conservation seen in PRG-1 is to occur amongst the proteome. We randomly removed 10% of the data and counted the number of proteins remaining above this threshold in each species, and repeated this to give 10,000 total jack-knives, and the number of times that *D. melanogaster* appeared to have lost fewer proteins than all other non-clade V nematodes. This did not occur once in any of the simulations, thus giving an estimated *p*-value of $<1e-4$. For identification of selected RNAi proteins in *E. brevis* and *H. dujardini*, transcriptome databases from the draft transcriptomes were built using Formatdb and tblastn was used with *C. elegans* proteins. The best hit for selected proteins was then tested by reciprocal blast against *C. elegans* proteins to test for orthology.

To construct phylogenetic trees of Piwi and RdRP proteins we used separate runs of phmmer with *D. melanogaster* Piwi, *C. elegans* RRF-1 and *C. elegans* RRF-3 to identify the best hit of each in the nematode genomes. The sequences were aligned using Muscle using 16 iterations and the alignment was refined with Gblocks 0.91b using default parameters. Phylogeny was obtained using PhyML using the Blosum62 substitution matrix [55].

DNA Methylation Analysis

To test for DNA methylation in our strain of *T. spiralis*, which may have differences to the isolate used by Gao and colleagues [44], we purified genomic DNA from the interphase of a Trizol extraction and used the Bisulfite-gold kit to convert these. We used Taq PCR and the primers from Gao and colleagues [44] and sequenced the PCR products directly. This confirmed 50% methylation at the CG sites identified [44].

We took DNA methylated genes (including transposon coding sequences) as defined [44] and identified the number of small RNAs mapping to these genes and compared to all genes. To estimate the density of DNA methylation at transposons, we used a custom script written in Perl to extract all CHG and CG potential methylation sites within each repetitive element predicted by Repeatmasker and interrogated the methylation status. CHG methylation genome-wide was around 4-fold lower than CG methylation. We grouped repetitive sequences by classes and calculated the percent methylation at each site as the number of converted reads/total reads and then counted the number of bases showing greater than 5% methylation. A chi-squared test was used to calculate the significance using the genome-wide percentage methylation to predict the expected value.

Supporting Information

S1 Data. Multiple sequence alignment for PRG-1/Piwi homologues.

(TXT)

S2 Data. Multiple sequence alignment for RdRPs.

(TXT)

S3 Data. Nuclear localization sequence in *T. spiralis* Dicer.

(DOCX)

S4 Data. Processed data underlying figures and charts.

(ZIP)

S1 Fig. Small RNA sequencing methodology. (A) 5' dependent library preparation only allows RNAs with 5' monophosphates to be ligated to adapters. It thus allows sequencing of Dicer products (both miRNAs and siRNAs) and mature piRNAs. (B) 5' independent library preparation allows both 5' triphosphate and 5' monophosphate species to be ligated to adapters and thus enables sequencing of RNA dependent RNA polymerase products (22G-RNAs in *C. elegans*). (PDF)

S2 Fig. Further analysis of conservation of miRNA sequences and regulation. (A) Mean miRNA conservation across eight nematode species is plotted against expression of the miRNA in adult *C. elegans*. (B–D) Developmental expression changes of miRNAs in nematode species compared to that of the homologous miRNA in *C. elegans*, with the Spearman's rank correlation coefficient shown for each species. (PDF)

S3 Fig. Small RNA sequencing of putative nematode outgroups. (A) 5' monophosphate small RNAs in the tardigrade *H. dujardini*, showing longer sequences with a 5' U bias, putative piRNAs. (B) Putative *H. dujardini* piRNAs show a prominent ten nucleotide overlap, with the tenth nucleotide of the 5'-most piRNA showing a bias towards (A), consistent with ping-pong amplification. (C, D) 5' monophosphate and 5' triphosphate small RNA sequencing from *P. varius* (Nematomorpha) showing absence of longer 5' U species and no evidence of 5' triphosphorylated small RNAs. (PDF)

S4 Fig. Small RNA sequencing of clade V, clade II, and clade I nematodes confirms absence of 5' triphosphorylated small RNA populations and piRNAs outside of clade V. (A, B) 5' monophosphate only (A) and 5' mono and triphosphate (B) sequencing of small RNAs from *P. pacificus* (clade V). (C) 5' mono and triphosphate sequencing of small RNAs from *N. brasiliensis* (clade V). (D, E) 5' monophosphate only (D) and 5' mono and triphosphate (E) sequencing of small RNAs from *E. brevis* (clade II) collected from Sylt in Germany. (F) 5' mono and triphosphate sequencing of small RNAs from *O. rectangula* (clade II) collected from Vancouver in Canada. (G, H) 5' monophosphate only (G) and 5' mono and triphosphate (H) sequencing of small RNAs from *R. culicivorax* (clade I). (I) Collapsing the *R. culicivorax* sequences so that only unique sequences are represented removes a prominent 26T peak; this represents one abundant sequence and is thus not likely to be a piRNA, consistent with the absence of PRG-1/Piwi in this species. (PDF)

S5 Fig. Supplemental analysis of the evolution of PRG-1 in nematodes. (A) Blastp score in bits/length for the best hit to *C. elegans* PRG-1 compared to the median and interquartile range of the best hit for all *C. elegans* proteins. Members of the Piwi subfamily are shown in red (see Fig. 2E) and members of the Ago subfamily shown in purple (see Fig. 2E). (B) Histogram showing the result of 1,000 simulations the evolution of PRG-1 to the distance between *C. elegans* PRG-1 and *D. melanogaster* Piwi. xAxis is the e-value found after spiking the evolved protein into the *T. spiralis* genome; the red line represents the e-value of the best hit to PRG-1 within the true *T. spiralis* genome. (PDF)

S6 Fig. Small RNA sequencing of transposon-matching 5' triphosphorylated siRNAs in clade III and clade IV nematodes. (A) 5' mono and 5' triphosphate sequencing demonstrates that 22–26 nt triphosphorylated small RNAs align predominantly antisense to transposons in *G. pallida*. (B) Collapsing to unique sequences retains the bias towards antisense orientation in both *G. pallida* and *B. malayi*. (C, D) 5' monophosphate only sequencing shows evidence of 23 nt 5' monophosphate small RNAs aligning antisense to transposon sequences in both *B. malayi* (C) and *G. pallida* (D), indicating that Dicer recognises transposons in these organisms. (PDF)

S7 Fig. Potential mechanisms for generation of siRNAs from transposons by Dicer. (A) Dicer cleavage of dsRNA originating from transcription of transposon sequences could feed into the small RNA pathway. (B) RNA dependent RNA polymerase could generate long dsRNA using the transposon sequence as a template, which would then be processed by Dicer to generate siRNAs. (PDF)

S8 Fig. Further analysis of *T. spiralis* siRNAs. (A) Sequencing of small RNAs following treatment with 200 mM sodium periodate compared to control samples shows that *C. elegans* 21U-RNAs are specifically protected against oxidation whilst *T. spiralis* 23–25 nt siRNAs are lost following oxidation. The peak at 28–30 nt in *T. spiralis* reflects two abundant ribosomal RNA sequences as shown by its loss upon collapsing the sequence data to unique sequences (far right hand panel). (B, C) Cluster analysis across the genome shows that regions with high density of piRNAs found in *C. elegans* and *D. melanogaster* are not found for *T. spiralis* 23–25 nt siRNAs. (B) Shows genome-wide distribution of *T. spiralis* 23–25 nt siRNAs, *C. elegans* piRNAs, and *D. melanogaster* piRNAs. Reads are binned in 100 kb windows across the genome and coloured by contigs or chromosomes according to the genome assembly, with the contigs

or chromosomes sorted in order of the total number of small RNAs mapping to them. (C) Shows the cumulative fraction of sequences in (B) that are found in regions with greater than or equal to the density indicated on the x -axis. *C. elegans* and *D. melanogaster* both have more sequences mapping to higher density regions than *T. spiralis* does.
(PDF)

S1 Text. Supplementary tables 1 and 2 and extended experimental procedures.
(DOCX)

Acknowledgments

We thank Sylviane Moss for high-throughput sequencing support. We thank Charles Bradshaw for help with computation and IT. We thank Marie-Anne Felix and Frank Jiggins for critical comments on the manuscript. We thank Matt Berriman (Wellcome Trust Sanger Centre, Hinxton, Cambridge, UK) for allowing us to use unpublished genomic sequencing data for *N. brasiliensis*. We thank Einhardt Schierenberg (University of Cologne, Germany) and Werner Armonies (Alfred Wegener Institute, Sylt, Germany) for help with collection of *E. brevis*.

Author Contributions

Conceived and designed the experiments: PS MES MLB EAM. Performed the experiments: PS. Analyzed the data: PS EAM. Contributed reagents/materials/analysis tools: PS MES JTJ VB TB BG AA NMF CK PMS ES MJT BH GK MLB EAM. Wrote the paper: PS MES JTJ VB TB BG NMF CK PMS MJT BH MLB EAM.

References

1. Shi Z, Montgomery TA, Qi Y, Ruvkun G (2013) High-throughput sequencing reveals extraordinary fluidity of miRNA, piRNA, and siRNA pathways in nematodes. *Genome Res* 23: 497–508. doi: [10.1101/gr.149112.112](https://doi.org/10.1101/gr.149112.112) PMID: [23363624](https://pubmed.ncbi.nlm.nih.gov/23363624/)
2. McCue AD, Slotkin RK (2012) Transposable element small RNAs as regulators of gene expression. *Trends Genet* 28: 616–623. doi: [10.1016/j.tig.2012.09.001](https://doi.org/10.1016/j.tig.2012.09.001) PMID: [23040327](https://pubmed.ncbi.nlm.nih.gov/23040327/)
3. de Wit E, Linsen SEV, Cuppen E, Berezikov E (2009) Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res* 19: 2064–2074. doi: [10.1101/gr.093781.109](https://doi.org/10.1101/gr.093781.109) PMID: [19755563](https://pubmed.ncbi.nlm.nih.gov/19755563/)
4. Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12: 246–258. doi: [10.1038/nrm3089](https://doi.org/10.1038/nrm3089) PMID: [21427766](https://pubmed.ncbi.nlm.nih.gov/21427766/)
5. Yigit E, Batista PJ, Bei Y, Pang KM, Chen C-CG, et al. (2006) Analysis of the *C. elegans* Argonaute Family Reveals that Distinct Argonautes Act Sequentially during RNAi. *Cell* 127: 747–757. PMID: [17110334](https://pubmed.ncbi.nlm.nih.gov/17110334/)
6. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, et al. (2014) The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 21: 743–753. doi: [10.1038/nsmb.2879](https://doi.org/10.1038/nsmb.2879) PMID: [25192263](https://pubmed.ncbi.nlm.nih.gov/25192263/)
7. Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318: 761–764. PMID: [17975059](https://pubmed.ncbi.nlm.nih.gov/17975059/)
8. Kaufman EJ, Miska EA (2010) The microRNAs of *Caenorhabditis elegans*. *Semin Cell Dev Biol* 21: 728–737. doi: [10.1016/j.semcdb.2010.07.001](https://doi.org/10.1016/j.semcdb.2010.07.001) PMID: [20637886](https://pubmed.ncbi.nlm.nih.gov/20637886/)
9. Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, et al. (2008) PRG-1 and 21U-RNAs Interact to Form the piRNA Complex Required for Fertility in *C. elegans*. *Mol Cell* 31: 67–78. doi: [10.1016/j.molcel.2008.06.002](https://doi.org/10.1016/j.molcel.2008.06.002) PMID: [18571452](https://pubmed.ncbi.nlm.nih.gov/18571452/)
10. Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, et al. (2008) Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol Cell* 31: 79–90. doi: [10.1016/j.molcel.2008.06.003](https://doi.org/10.1016/j.molcel.2008.06.003) PMID: [18571451](https://pubmed.ncbi.nlm.nih.gov/18571451/)
11. Gu W, Lee H-C, Chaves D, Youngman EM, Pazour GJ, et al. (2012) CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped SmallRNAs as *C. elegans* piRNA Precursors. *Cell* 151: 1488–1500. doi: [10.1016/j.cell.2012.11.023](https://doi.org/10.1016/j.cell.2012.11.023) PMID: [23260138](https://pubmed.ncbi.nlm.nih.gov/23260138/)

12. Weick E-M, Sarkies P, Silva N, Chen RA, Moss SMM, et al. (2014) PRDE-1 is a nuclear factor essential for the biogenesis of Ruby motif-dependent piRNAs in *C. elegans*. *Genes Dev* 28: 783–796. doi: [10.1101/gad.238105.114](https://doi.org/10.1101/gad.238105.114) PMID: [24696457](https://pubmed.ncbi.nlm.nih.gov/24696457/)
13. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128: 1089–1103. PMID: [17346786](https://pubmed.ncbi.nlm.nih.gov/17346786/)
14. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, et al. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127: 1193–1207. PMID: [17174894](https://pubmed.ncbi.nlm.nih.gov/17174894/)
15. Bagijn MP, Goldstein LD, Sapetschnig A, Weick EM, Bouasker S, et al. (2012) Function, targets, and evolution of *Caenorhabditis elegans* piRNAs. *Science* 337: 574–578. doi: [10.1126/science.1220952](https://doi.org/10.1126/science.1220952) PMID: [22700655](https://pubmed.ncbi.nlm.nih.gov/22700655/)
16. Sijen T, Steiner FA, Thijssen KL, Plasterk RHA (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* 315: 244–247. PMID: [17158288](https://pubmed.ncbi.nlm.nih.gov/17158288/)
17. Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* 315: 241–244. PMID: [17124291](https://pubmed.ncbi.nlm.nih.gov/17124291/)
18. Sijen T, Fleenor J, Simmer F, Thijssen KL, Parrish S, et al. (2001) On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* 107: 465–476. PMID: [11719187](https://pubmed.ncbi.nlm.nih.gov/11719187/)
19. Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, et al. (2009) 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 106: 18674–18679. doi: [10.1073/pnas.0906378106](https://doi.org/10.1073/pnas.0906378106) PMID: [19846761](https://pubmed.ncbi.nlm.nih.gov/19846761/)
20. Vasale JJ, Gu W, Thivierge C, Batista PJ, Claycomb JM, et al. (2010) Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. *Proc Natl Acad Sci U S A* 107: 3582–3587. doi: [10.1073/pnas.0911908107](https://doi.org/10.1073/pnas.0911908107) PMID: [20133583](https://pubmed.ncbi.nlm.nih.gov/20133583/)
21. Simon M, Sarkies P, Ikegami K, Doebley A-L, Goldstein LD, et al. (2014) Reduced insulin/IGF-1 signaling restores germ cell immortality to *caenorhabditis elegans* Piwi mutants. *Cell Rep* 7: 762–773. doi: [10.1016/j.celrep.2014.03.056](https://doi.org/10.1016/j.celrep.2014.03.056) PMID: [24767993](https://pubmed.ncbi.nlm.nih.gov/24767993/)
22. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, et al. (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* 392: 71–75. PMID: [9510248](https://pubmed.ncbi.nlm.nih.gov/9510248/)
23. Rota-Stabelli O, Daley AC, Pisani D (2013) Molecular timetrees reveal a Cambrian colonization of land and a new scenario for Ecdysozoan evolution. *Curr Biol* 23: 392–398. doi: [10.1016/j.cub.2013.01.026](https://doi.org/10.1016/j.cub.2013.01.026) PMID: [23375891](https://pubmed.ncbi.nlm.nih.gov/23375891/)
24. Winter AD, Weir W, Hunt M, Berriman M, Gilleard JS, et al. (2012) Diversity in parasitic nematode genomes: the microRNAs of *Brugia pahangi* and *Haemonchus contortus* are largely novel. *BMC Genomics* 13: 4. doi: [10.1186/1471-2164-13-4](https://doi.org/10.1186/1471-2164-13-4) PMID: [22216965](https://pubmed.ncbi.nlm.nih.gov/22216965/)
25. Wang J, Czech B, Crunk A, Wallace A, Mitreva M, et al. (2011) Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles. *Genome Res* 21: 1462–1477. doi: [10.1101/gr.121426.111](https://doi.org/10.1101/gr.121426.111) PMID: [21685128](https://pubmed.ncbi.nlm.nih.gov/21685128/)
26. Srinivasan J, Dillman AR, Macchietto MG, Heikkinen L, Lakso M, et al. (2013) The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle. *Genetics* 193: 1279–1295. doi: [10.1534/genetics.112.148809](https://doi.org/10.1534/genetics.112.148809) PMID: [23410827](https://pubmed.ncbi.nlm.nih.gov/23410827/)
27. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202–W208. doi: [10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335) PMID: [19458158](https://pubmed.ncbi.nlm.nih.gov/19458158/)
28. Buck AH, Blaxter M (2013) Functional diversification of Argonautes in nematodes: an expanding universe. *Biochem Soc Trans* 41: 881–886. doi: [10.1042/BST20130086](https://doi.org/10.1042/BST20130086) PMID: [23863149](https://pubmed.ncbi.nlm.nih.gov/23863149/)
29. Billi AC, Alessi AF, Khivansara V, Han T, Freeberg M, et al. (2012) The *Caenorhabditis elegans* HEN1 ortholog, HENN-1, methylates and stabilizes select subclasses of germline small RNAs. *PLoS Genet* 8: e1002617. doi: [10.1371/journal.pgen.1002617](https://doi.org/10.1371/journal.pgen.1002617) PMID: [22548001](https://pubmed.ncbi.nlm.nih.gov/22548001/)
30. Montgomery TA, Rim Y-S, Zhang C, Downen RH, Phillips CM, et al. (2012) PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog henn-1. *PLoS Genet* 8: e1002616. doi: [10.1371/journal.pgen.1002616](https://doi.org/10.1371/journal.pgen.1002616) PMID: [22536158](https://pubmed.ncbi.nlm.nih.gov/22536158/)
31. Kamminga LM, van Wolfswinkel JC, Luteijn MJ, Kaaij LJT, Bagijn MP, et al. (2012) Differential impact of the HEN1 homolog HENN-1 on 21U and 26G RNAs in the germline of *Caenorhabditis elegans*. *PLoS Genet* 8: e1002702. doi: [10.1371/journal.pgen.1002702](https://doi.org/10.1371/journal.pgen.1002702) PMID: [22829772](https://pubmed.ncbi.nlm.nih.gov/22829772/)
32. Horwich MD, Li C, Matranga C, Vagin V, Farley G, et al. (2007) The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* 17: 1265–1272. PMID: [17604629](https://pubmed.ncbi.nlm.nih.gov/17604629/)
33. Saito K, Sakaguchi Y, Suzuki T, Suzuki T, Siomi H, et al. (2007) Pimet, the *Drosophila* homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends. *Genes Dev* 21: 1603–1608. PMID: [17606638](https://pubmed.ncbi.nlm.nih.gov/17606638/)

34. Mallatt JM, Garey JR, Shultz JW (2004) Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Molecular phylogenetics and evolution* 31: 178–191. PMID: [15019618](#)
35. Gent JI, Lamm AT, Pavelec DM, Maniar JM, Parameswaran P, et al. (2010) Distinct phases of siRNA synthesis in an endogenous RNAi pathway in *C. elegans* soma. *Mol Cell* 37: 679–689. doi: [10.1016/j.molcel.2010.01.012](#) PMID: [20116306](#)
36. Ferrerorta C, Arias A, Escarmis C, Verdaguier N (2006) A comparison of viral RNA-dependent RNA polymerases. *Curr Opin Struct Biol* 16: 27–34. PMID: [16364629](#)
37. Zong J, Yao X, Yin J, Zhang D, Ma H (2009) Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* 447: 29–39. doi: [10.1016/j.gene.2009.07.004](#) PMID: [19616606](#)
38. Salgado PS, Koivunen M, Makeyev EV, Bamford DH, Stuart DI (2006) The structure of an RNAi polymerase links RNA silencing and transcription. *PLoS Biol* 4: e434. PMID: [17147473](#)
39. Makeyev EV, Bamford DH (2002) Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. *Mol Cell* 10: 1417–1427. PMID: [12504016](#)
40. Talsky KB, Collins K (2010) Initiation by a eukaryotic RNA-dependent RNA polymerase requires looping of the template end and is influenced by the template-tailing activity of an associated uridylyltransferase. *J Biol Chem* 285: 27614–27623. doi: [10.1074/jbc.M110.142273](#) PMID: [20622019](#)
41. Volpe TA (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297: 1833–1837. PMID: [12193640](#)
42. Chan SWL, Zilberman D, Xie Z, Johansen LK, Carrington JC, et al. (2004) RNA silencing genes control de novo DNA methylation. *Science* 303: 1336. PMID: [14988555](#)
43. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328: 916–919. doi: [10.1126/science.1186366](#) PMID: [20395474](#)
44. Gao F, Liu X, Wu X-P, Wang X-L, Gong D, et al. (2012) Differential DNA methylation in discrete developmental stages of the parasitic nematode *Trichinella spiralis*. *Genome Biol* 13: R100. doi: [10.1186/gb-2012-13-10-r100](#) PMID: [23075480](#)
45. Schiffer PH, Kroiher M, Kraus C, Koutsovoulos GD, Kumar S, et al. (2013) The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda. *BMC Genomics* 14: 923. doi: [10.1186/1471-2164-14-923](#) PMID: [24373391](#)
46. Pontes O, Li CF, Nunes PC, Haag J, Ream T, et al. (2006) The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* 126: 79–92. PMID: [16839878](#)
47. Li CF, Pontes O, El-Shami M, Henderson IR, Bernatavichute YV, et al. (2006) An ARGONAUTE4-containing nuclear processing center colocalized with cajal bodies in *Arabidopsis thaliana*. *Cell* 126: 93–106. PMID: [16839879](#)
48. Brenner S (1974) The genetics of *Caenorhabditis elegans*. *Genetics* 77: 71–94. PMID: [4366476](#)
49. Jones JT, Kumar A, Pylypenko LA, Thirugnanasambandam A, Castelli L, et al. (2009) Identification and functional characterization of effectors in expressed sequence tags from various life cycle stages of the potato cyst nematode *Globodera pallida*. *Molecular Plant Pathology* 10: 815–828. doi: [10.1111/j.1364-3703.2009.00585.x](#) PMID: [19849787](#)
50. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25. doi: [10.1186/gb-2009-10-3-r25](#) PMID: [19261174](#)
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25: 2078–2079. doi: [10.1093/bioinformatics/btp352](#) PMID: [19505943](#)
52. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842. doi: [10.1093/bioinformatics/btq033](#) PMID: [20110278](#)
53. Tabach Y, Billi AC, Hayes GD, Newman MA, Zuk O, et al. (2013) Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* 493: 694–698. doi: [10.1038/nature11779](#) PMID: [23364702](#)
54. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96: 4285–4288. PMID: [10200254](#)
55. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59: 307–321. doi: [10.1093/sysbio/syq010](#) PMID: [20525638](#)

Chapter 3

Discussion

The manuscripts discussed and included in this thesis were presented in chronological order. However, for easier argumentation I will use a different order in the discussion.

Although the six publications included in this thesis are set in a general context, which is analysed on different taxonomic levels, they are of course independent assays. I will therefore shortly discuss the main findings of each of the publications and then outline a uniting framework. In this framework, a special emphasis will be put on the **M1** manuscript, “the *Panagrolaimus* genome project”, as the analysis of the evolution of parthenogenesis in these nematodes was at the core of my thesis.

The thesis further contains the manuscript **M3**, which will only be treated as a sub-section in the discussion of **P2**, as my contribution to this project was mostly limited to generation of data and some editing of the text. It does, however, contribute to the overall framework. Although the Discussion section includes some recapitulation of the main results of each of the 6 manuscripts, it will be necessary for the reader to refer to the manuscripts for figures, tables and a complete description of results. This is also true for references.

3.1 Discussion of **P2**: A clade I genome exemplifies the huge differences in the developmental genetic toolkit of Nematoda

The first paper I will discuss is, **P2**, which compares the molecular toolkits available in a newly sequenced Dorylaimia (clade I, a part of Enoplea consisting of clades I and II) nematode, *Romanomermis culicivorax*, with the model worm *Caenorhabditis elegans* (clade V), a second dorylaim, *Trichinella spiralis*, and the beetle *Tribolium castaneum* as an outgroup species from Arthropoda. Assuming a split between Chromadorea and Enoplea more than 420 Ma ago (Rota-Stabelli et al., 2013), and independent evolution of the lineages leading to *C. elegans* and *R. culicivorax*, the total amount of evolutionary time would be more than 800 Ma. There are two key messages that can be taken from the analysis of the *R. culicivorax* genome sequence:

1. The genome of *T. spiralis*, the other clade I nematode previously sequenced, cannot be seen as typical for this group of nematodes, and
2. the molecular backbone of development in *C. elegans* is very different from clade I nematodes.

It has to be assumed that every contemporary species in every phylum is highly derived in its own right and therefore neither *R. culicivorax*, nor *T. spiralis* can serve as ideal role models for dorylaim genomics in general. But, especially the comparison with *T. castaneum*, indicates that *T. spiralis* has lost many possibly pan-ecdysozoan genes that are retained in *R. culicivorax*. Additionally, the *R. culicivorax* genome has more genes in common with the *C. elegans* genome than has *T. spiralis*. These findings imply that *T. spiralis* is highly derived even within Dorylaimia, possibly due to its parasitic life cycle without any free-living stages. As for the 2nd point, the divergence of the *C. elegans* molecular toolkit for development was not fully appreciated before. Some cellular traits were shown to differ in early development between various nematodes (see introduction), but from the genetic perspective, research so far has mainly focussed on finding orthologous proteins between *C. elegans* and other organisms, e.g. *D. melanogaster* and *H. sapiens* (Hunt, 2011). These genes were then assumed to have similar functions. While we were indeed able to identify some genes active in development that are shared between distantly related nematode species (e.g. *mex-3*), we were surprised to find that our data suggest that approximately half of a core set of ~1,800 genes that are active in early development of *Caenorhabditis* nematodes (described in Levin et al., 2012) are not conserved in clade I nematodes. Some key pathways acting in core *C. elegans* developmental processes are affected by this. Among these, the initial establishment of polarity and the definition of the early anterior-posterior axis are

striking examples. Hence, it seems reasonable to assume that a large part of the interacting protein networks seen in *C. elegans* are considerably different from those in Dorylaimia.

Discussion of M3: Evolutionary dynamics of small RNA pathways demonstrate the genomic plasticity of the nematode systems

In the manuscript **M3** one specific gene set, small RNAs, was analysed in all major groups of the nematode tree, particularly including two clade II species, and a nematomorph, the closest outgroup to Nematoda. Small RNAs, which are present in all eukaryotes, are important in silencing transposons, and thus essential to maintain genome integrity (Siomi et al., 2011). In *C. elegans* several classes of small RNAs have been characterised and extensively analysed. Among these, especially micro RNAs (miRNAs) have been found to be important in development and highly conserved across different phyla (Ambros et al., 2003). It is also known that in Bilateria miRNA gene families have undergone massive inflations in some branches and secondary loss in others (Erwin et al., 2011). In *C. elegans* the high conservation of miRNAs and their processing machinery is thwarted by divergence in its Piwi-interacting small RNAs (piRNAs). These are, as in other organisms, silencing transposons in the germline, but are much shorter in *C. elegans* (21 in comparison to 26–30 nucleotides, Das et al., 2008). In the **M3** manuscript it is reported that these piRNAs have been lost independently in several clades of Nematoda. Intriguingly, these losses have been compensated for by different mechanisms to silence transposons. In Chromadorea species outside of clade V 22G-RNAs are produced by a RNA dependent RNA polymerase, while in Enoplea a different RNA dependent RNA polymerase pathway first generates dsRNA that is then processed by the Dicer protein into small RNAs (see **M3** for details). The Piwi proteins, which interact with the piRNAs, are a subfamily of the Argonautes, a conserved group of proteins. PRG-1 is the *C. elegans* orthologue of *Drosophila* Piwi. In the light of finding the divergent pathways for transposon silencing outside of clade V, it is intriguing that also no orthologue of the protein PRG-1 was detected in any of the species from the remaining clades which were analysed in this study¹. The results from **M3** thus indicate the presence of divergent pathways for transposon silencing in Nematoda, which is important in the *C. elegans* germline. Additionally, in the Enoplea species transposable elements appear to be methylated with the aid of the dsRNA/Dicer silencing pathway, which fits to findings from

¹As explained in the Methods (from page 15) finding orthologous proteins can be difficult and especially hampered by low quality genomes and gene predictions. It was therefore a major concern of several of the reviewers of this manuscript that the Piwi orthologue PRG-1 might have simply been overlooked in species outside of clade V by BLAST+ searches. To demonstrate the power of BLAST+ to detect even remote orthologous a simulation assay was conducted in which PRG-1 was artificially evolved, then placed into the *T. spiralis* gene set and searched for with BLAST+. This showed that BLAST+ found the evolved PRG-1 sequences in the *T. spiralis* gene set at far better similarity thresholds than any of the putative *T. spiralis* candidates. BLAST+ therefore appears capable of identifying even remote sequences as homologs, which is important for our studies of orthology, where BLAST+ is used as the first step in the prediction pipelines. This topic will be further elaborated on in a technical part of the discussion of **M1** on page 155.

the **P2** project, where we found proteins involved in methylation enriched in comparison to *C. elegans*. The dsRNA/Dicer mechanism could be seen as ancestral to eukaryotes, as similar pathways are present in plants and fungi. While this on the one hand demonstrates the plasticity of the genomic backbone of the phylum, the conservation of the small RNA machinery between some Enoplea and remote eukaryotic outgroups on the other hand indicates some links to outgroup taxa in general in Enoplea, which are lost in Chromadorea and particularly the *Caenorhabditis* crown clade. This will be further discussed in the section on **M2** from page 162, and the **Outlook** section from page 169 onwards.

In general, the findings from **P2** and **M3** suggest that huge differences in the developmental toolkit between nematodes from clade I and clade V have evolved, indicating Developmental System Drift (DSD) to play a major role in the evolution of roundworms. In other words, the *C. elegans* molecular toolkit for development is far from being archetypical for Nematoda. These findings prompt a set of follow-up questions:

Given the presence of five clades, it could be assumed that there is a steady loss and gain of some genes and even of whole gene families across the tree. Thus the questions arises if there is an evolutionary trajectory from a ground state to a potentially more derived one, or if evolution is more random across the phylum with independent gains and losses in single clades (or smaller taxa)?

Can the available genetic toolkit, the set of retained and novel genes and their families, be set in correlation with the split between Enoplea and Chromadorea nematodes?

With regard to both questions, it would be interesting to analyse more than just the small RNA pathway from clade II nematodes, especially as these species are phylogenetically closer to the outgroup than clade I (Holterman et al., 2006); confirmed by the latest phylogenomic analysis (G. Koutsovoulos, personal communication).

The number of genes across animal taxa varies, but is not drastically different with many species having around 20,000 genes (Lynch, 2007). While *T. spiralis* has fewer genes than *C. elegans*, we found *R. culicivora*x to potentially have more than the model organism. Given the necessity to govern a plethora of developmental processes it is thus reasonable to assume that Dorylaimia nematodes do not simply lack half of the important *C. elegans* genes acting in early development without having any replacement. On the contrary, they should possess other genes to conduct these roles. We have started to uncover these unknown molecular toolsets available in Dorylaimia and other non-*Caenorhabditis* taxa. This will be shortly explained in the Outlook section on page 169 ff. However, to answer the first question, the analysis of developmental system drift and the re-shuffling of gene regulatory networks was first extended to the genus and then the clade level in the **P3** and **M1** manuscripts, which will be discussed in the following sections.

3.2 Discussion of P3: Closely related taxa show variations in developmental gene expression

The publication **P3** is concerned with evolution on almost the lowest taxonomic levels, i.e., between species and between genera. A group of nematodes in clade IV is called the infraorder Panagrolaimorpha, because of their morphological similarity to species in the genus *Panagrolaimus*. As explained in the included publication many species in Panagrolaimorpha are morphologically very similar and particularly in the genus *Panagrolaimus* a very high degree of uniformity exists. However, in their molecular phylogenetic analysis on the evolutionary origin of parthenogenesis in *Panagrolaimus*, Lewis et al. (2009) found that the genus falls into two distinct sub-taxa, which they termed PI and PII. I observed that species from the Lewis group PII do indeed share some morphological similarities with somewhat longer nematodes named *Propanagrolaimus* (Andrássy, 2005). I used large-scale sequencing data, which I had assembled for my research on the evolution of sex from groups PI and PII, as well as outgroup data, to robustly reconcile the phylogeny of these worms. This was necessary, as Lewis et al. (2009) had only used short fragments of ribosomal genes and some mitochondrial sequences. Our data not only allow us to confirm a genus *Propanagrolaimus* sensu (Andrássy, 2005) on the basis of molecular sequences for a first time, but also indicates that the group PII nematodes are members of this genus. Interestingly, we found that these two closely related and morphologically extremely similar genera differ in several developmental traits. First, the expression pattern of the *skn-1* gene, which is crucial to initiate the endo-/mesoderm specifying gene cascade in *C. elegans* (Maduro, 2006) is very different between nematodes from *Panagrolaimus* and *Propanagrolaimus*. The pattern found in the *Propanagrolaimus* species is actually similar to the one found in species of the *Acrobeloides/Cephalobus* group. Also, with respect to early cell lineage, we observed distinct difference between *Panagrolaimus* and *Propanagrolaimus*. While they all differ from the pattern in *C. elegans*, a closer similarity of *Propanagrolaimus* to the *Acrobeloides/Cephalobus* group was not found. Although, so far only the expression pattern of a single gene and the timing of early embryonic cell divisions has been studied, it seems reasonable to expect more variations on the level of interaction networks controlling early development in these nematodes. From these data it seems that already on low taxonomic levels a substantial amount of plasticity in the molecular underpinnings of developmental systems can exist. Still, the developmental plasticity does not affect the morphology in a way leading to great divergence in the adult forms between *Panagrolaimus* and *Propanagrolaimus*, while e.g. *Panagrolaimus* and *Acrobeloides/Cephalobus* nematodes can be clearly defined on the morphological level. It also seems clear that on a broader phylogenetic scale even more fundamental changes have to be expected, which were explored in the third manuscript, **M1**.

3.3 Discussion of M1: DSD has changed GRNs in clade IV nematodes and might facilitate the evolution of parthenogenesis

As already mentioned in the previous section on **P3**, Panagrolaimid nematodes belong to clade IV of the nematode phylum. Species in the genus *Panagrolaimus* are attractive study objects, because of the presence of closely related sexual (amphimictic) and parthenogenetic species, which are easily cultured in the laboratory. In the included manuscript **M1**, a genomic approach is taken to gain first insights into the evolution of parthenogenesis in the genus *Panagrolaimus*. In total, new genomes for three *Panagrolaimus* and one *Propanagrolaimus* species were sequenced and assembled, and one additional parthenogenetic *Panagrolaimus* species was re-assembled to improve genome quality. These data were complemented by RNASeq of transcriptomes for 4 of these 5 species and by data of two additional parthenogenetic strains. Including outgroup species in *Propanagrolaimus*, *Panagrellus*, an array of other species in clade IV, *C. elegans*, and the clade III species *Ascaris suum* allowed to compare the genomic backbone available in *Panagrolaimus* and clade IV in general, with that of the model organism. The analysis, especially the inference of absence for important genes known from *C. elegans*, in *Panagrolaimus* and the close outgroups *Propanagrolaimus* and *Panagrellus*, crucially hinged upon assembling reliable genomes to predict robust gene sets from these data. The genome assembly pipeline in figure 7 on page 18 depicts the measures implemented to this end. Including several *Panagrolaimus* species (2 amphimictic, 5 parthenogenetic) as well as several outgroup species, ensures that even if a gene is not found in one species owing to deficiencies in assembly and/or annotation, it should be present in one of the other datasets. Thus, the number of false negatives should be greatly reduced.

3.3.1 Early developmental processes and the machinery of sex determination and fertilisation is divergent in clade IV nematodes

I focussed on genetic pathways organising developmental processes related to the evolution of sex, as well as axis specification and endo-/mesoderm formation where expression data had indicated divergence (see previous section). My findings appear to perfectly fit with Davidson's theories of gene regulatory network evolution: in each GRN, some genes are universally conserved, while others are not present at all in any of the clade IV tested species (see Supplementary files to **M1**). These genes must therefore have evolved in the lineage leading to *C. elegans* in clade V. It is remarkable that certain genes conducting key functions in *C. elegans* are missing from the *Panagrolaimus* and other clade IV species. The following list contains examples and their function in *C. elegans*:

- *med-1,2* → involved in endo-mesoderm development
- *end-1,3* → involved in endo-mesoderm development
- *tbx-35* → involved in endo-mesoderm development
- *pie-1* → early axis specification
- *fog-1* → sperm fate
- *fog-2* → F-Box protein; hermaphrodite spermatogenesis
- *xol-1* → master sex pathway switch
- *her-1* → sex determination, male development
- *sdc-1,2,3* → dosage compensation
- *fox-1* → dosage compensation
- *cep-1* → meiotic segregation
- *chk-2* → chromosome pairing

While implications of the absence of these genes are explained in the manuscript, I will focus on one aspect in detail here: the absence of the *med* and *end* genes from the endo-/mesoderm GRN. This was an interesting finding against the backdrop of our previous analysis of expression patterns of the *skn-1* gene (see **P3** and the discussion on page 148). The former result may have led to the conclusion that the function of this gene is shifted only between *Propanagrolaimus* and *Caenorhabditis*, whilst being similar between the model organisms and *Panagrolaimus*. As these genes conducting intermediary steps in the pathway - in *C. elegans* *skn-1* activates *med-1,2*, *med-1,2* then activate *end-1,3*, and *end-1,3* activate *elt-2* - the second result, however, promotes the idea that the whole GRN is different from *C. elegans* in all Panagrolaimids (all of the clade IV species analysed to be precise). It is possible that *skn-1* is not acting in endo-/mesoderm formation at all in these species, which would corroborate theories from Goszczynski (2005) proposing a *skn-1* independent pathway for the process. These authors propose a factor “X” to be directly acting on *end-1,3*. My data, in comparison, indicate that all four of the *end-1,3* and *med-1,2* genes were acquired in the lineage leading to *C. elegans*. *Skn-1* or the unknown factor “X” are either directly acting on the GATA-factor *elt-2*, which in turn is acting on a plethora of further genes (McGhee et al., 2009), or some yet to be discovered intermediate players are active in the clade IV species. Using Orthoinspector, I retrieved an *elt-2* candidate in the Panagrolaimids, raising the possibility for conservation of parts of the *C. elegans* gut-specification pathway. However, that the *C. elegans* protein was not retrieved in a group

of orthologues with the other species in the OrthoMCL pipeline could also point towards a completely different molecular underpinning to the process of endo-mesoderm specification in clade IV. Further studies are needed to understand how much evolutionary plasticity is inherent in this important developmental process. A similar picture is apparent for genes acting in processes like sex determination, germline specification, or DNA repair, which all could play a role in the evolution and maintenance of parthenogenesis. From the list given above it can be concluded that the sex determination pathway in clade IV species must be different from *C. elegans*, in particular with regard to dosage compensation conducted by the *sdc* genes and *fox-1*. In all these processes, some players are universally conserved in the analysed species, while others are restricted to *C. elegans* (potentially the genus *Caenorhabditis*) alone. Thus, to truly understand the molecular changes needed to establish and maintain parthenogenesis it will be necessary to unravel the genetics of these processes in detail in the *Panagrolaimus* and other clade IV species. One strategy to achieve this would be a stage and life-cycle specific RNASeq assay, as will be introduced in the Outlook section shortly (see page 172). The genomic backbone established for *Panagrolaimus* and *Propanagrolaimus* presented here will serve as a basis from which such assays exploring the molecular biology of parthenogenesis can be launched.

These inferences of absence, re-capitulated here for a few genes as examples, but detailed in the discussed manuscript **M1**, have major implications for our understanding of the evolution of molecular processes between wider taxonomic units. At least in Nematoda, but most likely across all Metazoa, a certain process (e.g. endo-/mesoderm formation) might be conserved as a specific phase in the course of development, but genes orchestrating it can apparently easily be exchanged or replaced by developmental system drift. Hence, to understand the genetics of development in divergent taxa (within a phylum), data from model systems can only be used as first approach, and in depth studies of several species per taxon are needed.

3.3.2 Genes acting in anhydrobiosis are found in *Panagrolaimus*

The manuscript **M1** however, does not only analyse the evolution of developmental genes and their possible loss and gain, but also investigates additional intriguing life-history traits in the taxon: *Panagrolaimus* species possess a strong potential to survive desiccation (Shannon et al., 2005) (local samples were for example isolated from the core of dried blackberry branches; Einhard Schierenberg, personal communication) and complete freezing (Wharton and Ferns, 1995) (the parthenogenetic species *P. davidi* inhabits areas of Antarctica, *ibidem*). In the manuscript **M1** we tried to uncover more of the molecular and genomic background of this trait called anhydrobiosis. Using the power of our comparative system with a species from the closely related *Propanagrolaimus* outgroup, which had been shown to be less amenable to anhydrobiosis (Shannon et al., 2005), we extended preliminary work

in *P. davidi* (Thorne et al., 2014). We found that clade IV species possess an additional, divergent set of small Heat Shock Proteins compared to *C. elegans*. These proteins have been described to act as chaperones preventing protein aggregation in species undergoing desiccation and are strongly up-regulated in response to desiccation in several anhydrobiotic taxa, for example tardigrades (Reuner et al., 2010). However, the family of sHSPs in *Panagrolaimus* does not appear to be expanded in comparison to other clade IV species analysed, which argues for a subfunctionalisation of these stress-response proteins into the process of anhydrobiosis, rather than for an adaptive change to the gene family. We did identify several other gene families potentially connected to anhydrobiosis, which appear over-represented in the *Panagrolaimus* species in comparison to *Propanagrolaimus*. Among these were not only the well known factors, which protect proteins from agglomeration, but also several protein domains involved in DNA repair, such as HSP-70 (see M1 for details). This was of special interest, as it also opened a link to Horizontal Gene Transfer.

3.3.3 Horizontal Gene Transfer might have favoured the evolution of anhydrobiosis in *Panagrolaimus*

Horizontal gene transfer (HGT) is rampant among Bacteria (McInerney et al., 2011), especially in the marine environment (McDaniel et al., 2010) and might be a driving force in adaptive processes in prokaryotes (Sankar, Schiffer and Wiehe in preparation). It has also been reported in metazoans (Boto, 2014). Preliminary analyses indicated that in nematodes up to 6% of genes per genome could have been acquired laterally from various donor organisms in the lineages leading to the species seen today (Etienne Danchin, personal communication). Based on these observations we analysed the *Panagrolaimus* species and close outgroups for signatures of HGT and correlated our findings to life-history traits of the species. We could link at least one horizontally acquired gene in *Panagrolaimus* to anhydrobiosis. A photolyase was found in all *Panagrolaimus* species, and a homologue identified in *A. suum*, and *A. nanus*, but not in any other of the clade IV species analysed, nor in *C. elegans*. Photolyases are ancient bacterial enzymes, potentially the earliest DNA repair system to evolve (Todo, 1999), which mediate DNA repair after UV radiation damage. Current theory holds that photolyases were present in early metazoa and then independently lost in other lineages (Lucas-Lledo and Lynch, 2009), in particular also in the one leading to *C. elegans*. However, close inspection of the photolyases in the species described above showed that the *Panagrolaimus* enzyme is distinctly different from the one in *A. suum* and appears also to be divergent from the *A. nanus* homologue. As both outgroup species are positioned on two separate phylogenetic branches and are both possessing a different photolyase from the one in *Panagrolaimus*, our data suggest independent gain events. It is interesting to note that Lucas-Lledo and Lynch (2009) reported gain of photolyases in some bacteria, but did not consider HGT for eukaryotes, illustrating the general negligence of the mechanism outside of

prokaryotes in the past. Strand breaks in DNA are expected to be common when organisms undergo anhydrobiosis and have in turn been associated with the uptake of foreign DNA when genomes are repaired during re-hydration (Hespeels et al., 2014). The acquisition of this photolyase might thus have favoured the evolution of anhydrobiosis in *Panagrolaimus*. While we found active expression of the gene in *Panagrolaimus*, a test for its action in DNA repair after anhydrobiosis would need life-cycle-phase specific expression analysis; either *in situ* hybridisations against the mRNA of the gene or RNASeq of the transcriptome would have to be conducted in re-hydrating animals. Analysing more of the horizontally gained genes we found functional descriptions that can also be brought in connection with DNA stability and genome integrity, such as “methyltransferase activity” and “DNA binding”. Thus, it is possible that the strand breaks occurring during desiccation in turn favour the acquisition of genomic material that enhances the chances for survival after rehydration (see **M1** for further discussions). In contrast to the previously described changes in existing genetic networks, such lateral acquisition of genes illustrates a fundamentally different way how evolutionary novelty can be established.

The enhancement of DNA repair mechanisms could also be favourable under parthenogenesis, where the accumulation of even slightly deleterious mutations will lead to fast extinction (see **Excursus 2** on page 156). It is necessary to re-emphasise at this point that the detected photolyase was acquired well before the evolution of parthenogenesis and is not present in all of the other parthenogenetic species in clade IV in our assay (e.g. present in *A. nanus*, absent in the *Meloidogyne* species). Thus, it cannot be a prerequisite to the evolution of parthenogenesis. According to the presented data on mutation rates (see **Excursus 2** on page 156), the postulated improvement of the DNA repair mechanism did not prevent the evolution of higher rates in the parthenogenetic species.

3.3.4 Parthenogenetic *Panagrolaimus* species appear to be polyploid hybrids

Lastly, the accrued data appear to confirm our hypothesis that the parthenogenetic *Panagrolaimus* species evolved through a hybridisation event between closely related sexual species and became polyploid in the process. Parthenogenesis following hybridisation has previously been reported for many unisexual species (Simon et al., 2003), but only been confirmed in *Meloidogyne* in Nematoda (Lunt et al., 2014). Lunt et al. (2014) even found a complex system of sequential hybridisation events in *Meloidogyne floridiensis*. To confirm our hypothesis of polyploid hybrids, backed by data coming from genome size measurements and karyotypes (see Results in **M1**), I developed a mapping pipeline (detailed in **M1**) to detect signatures of hybridisation based on polymorphism frequencies in the sequencing data. This analysis, as well as a genome coverage based assay, did support the idea of hybridisation in the parthenogenetic *Panagrolaimus* species. With the parthenogenetic *Panagrolaimus*

species being bona fide hybrids, genomic data suggesting the same for the parthenogen *Diploscapter coronatus* and PCR data pointing towards a hybrid origin of *Plectus sambe-sii* and an so far unnamed *Protorhabditis* species (strain JB137), it appears likely that the mating of two closely related sexual species could be an important route to parthenogenesis in Nematoda. While further data are needed to confirm this pattern, the switch to parthenogenesis through hybridisation of closely related species does solve some problems faced by hybrids. In general, hybrids have been considered as less evolutionary fit, but this assumption is currently changing, with some data indicating hybridisation underlying diversification in plants (Arnold et al., 2012) and playing a role in adaptive radiations in butterflies (Consortium, 2012; Martin et al., 2013). Still, hybridisation will often lead to either sterile or lethal, or otherwise less fit genotypes, for examples difficulties in forming bivalents during meiosis between divergent stretches of homologous chromosomes, or gene-dosage problems when alleles evolve different expression levels in the parental species (Maheshwari and Barbash, 2011). Also, the classical Dobzhansky-Muller model of hybrid incompatibility (HI) can reduce fitness via the evolutionary change of interaction patterns of genes in the GRNs of the parental species. HI can for example lead to fitness loss (or death) when lineage specific co-evolution between two genes causes protein structures which either cannot interact, or gain detrimental interaction capability, when brought together as alleles in a hybrid (see page 143 ff. in Nei, 2013, and see figure 2 in Maheshwari and Barbash, 2011, for an illustration). However, if the hybridisation event involves polyploidy such problems between divergent alleles might be overcome, as two copies per allele from one parent will be always available. At the same time the hybrid background could yield a temporary adaptive advantage and buffer against the accumulation of deleterious mutations (see **Excursus 2** on page 156 for more details). Many hybrids are polyploid, especially in plants, but these are very often sterile, which in plants led to the hypothesis that species capable of selfing might be more likely to form successful hybrids (see page 324 ff. in Coyne and Orr, 2004). While amphimictic animals are usually not self-fertile (simply lacking the respective organs of both sexes) parthenogenesis might act as an escape route to successful reproduction. One might argue that the likelihood for successful hybridisation events including polyploidisation and ensuing evolution of parthenogenesis is rather small, but the enormous number of meiofaunal animals (Creer et al., 2010; Fonseca et al., 2010) should allow for a reasonable success rate of inter-species copulation events. The few parthenogenetic species we find would then be the ones that happened to be successful in achieving normal development and viable offspring.

The study **M1** was able to provide some insights into the evolutionary plasticity of developmental gene regulatory networks. They are liable to change as proposed in Davidson's theories of evolution, owing to developmental system drift. However, this plasticity actually hampers finding an answer to the question of convergence (changes in the same gene in different organisms) or parallelism (changes in different genes) in the molecular background

of the evolution of parthenogenesis. It will only be possible to address this question after conducting further experiments to find the genes acting in the parthenogenetic species. Coupled with the finding of a hybrid origin in the parthenogenetic *Panagrolaimus* species and indications for this in other nematodes the question of convergence or parallel evolution on the genomic level becomes even more interesting. As indicated above a possible approach could include RNASeq of specific life-cycle stages, which could also include a comparison of *Panagrolaimus* and *Meloidogyne* species. This will be introduced in the Outlook section 3.6.2 on page 172. The genomic data compiled for the presented manuscript will serve as a backbone for such deeper studies into the evolution of parthenogenesis.

Analysing the evolution of anhydrobiosis we inferred the possibility for subfunctionalisation of existing genes, and (somewhat unsurprisingly) an apparent overrepresentation of proteins already known to be acting in the process. However, the possible link between the evolution of anhydrobiosis in *Panagrolaimus* and HGT illustrates how important the exchange of genetic material across different domains of life could be, in providing species with an adaptive advantage. In this way HGT could be seen as constraint breaking (Nei, 2013) and thus facilitate new evolutionary routes.

3.3.5 The use of two orthology detection pipelines enhances specificity and sensitivity of the approach

For a technical aspect, this study has shown the power of an approach using OrthoMCL in conjunction with Orthoinspector (see Methods on page 15 ff). While OrthoMCL has been shown to be very robust in finding orthologues that are also functionally conserved, Orthoinspector appears to uncover some connections obscured to OrthoMCL (see Methods section). In **M1** for example the *elt-2* gene was found in the clade IV species, but it was only detected by Orthoinspector. This can be taken as indication of great divergence of the protein sequence. Given the absence of major upstream genes as discussed in the first part of this section (page 149), this appears to be at least partly functional divergence. However, without the use of both programs *elt-2* would have been marked as absent. The different but complementary approach of both programs can also be seen from the data on the F-Box protein *fog-2* listed above (page 149 ff). OrthoMCL and Orthoinspector both retrieve no orthologue for this gene in any of the other clade IV species, but in contrast to OrthoMCL Orthoinspector does list the 208 paralogues present in *C. elegans*. It has been previously shown that *C. elegans* has a large inflation of F-Box genes (more than 300), while other animals like *D. melanogaster* or *H. sapiens* possess far less (in the range of 20 – 40) (Kipreos and Pagano, 2000). In the paper **P2**, discussed on page 145 ff, BLAST+ searches were conducted to re-confirm OrthoMCL inferences of absence, but for future studies the approach using OrthoMCL in conjunction with Orthoinspector appears very valuable. However, even this will not be fail-proof and more elaborate measures employing for example hidden Markov

models (HMMs) and searching directly on genomic sequence appear most promising to remedy current shortcomings. Such an approach is for example implemented in the figmap pipeline (Curran et al., 2014), where sets of known orthologous proteins are first scanned for domains, from which HMMs are build, which are then used to directly screen genomic contigs circumventing potential erroneous gene prediction software.

Excursus 2: Mutation Rates in parthenogenetic nematodes

In the following, I will discuss experiments and data which were an essential part of my thesis, but are not yet available in the form of a manuscript. As the topic is parthenogenesis like in M1, an Excursus into these analyses is included at this point of the text.

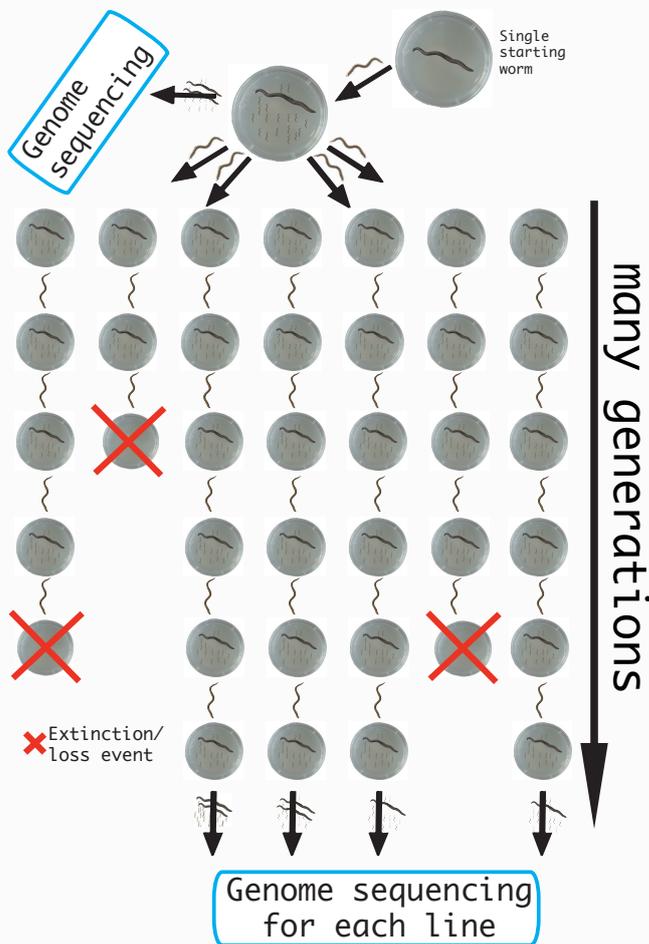
Although there are some “evolutionary scandals” (a term coined by Maynard Smith, 1978) of ancient parthenogenetic and species-rich taxa, like the bdelloid rotifers with an estimated age of 80–100 million years (Welch and Meselson, 2000), parthenogenetic taxa are typically found on phylogenetically terminal branches, i.e. in the vast majority of cases these taxa appear to become extinct in short evolutionary time scales and reversals to amphimixis do not occur (Maynard Smith, 1978; Schon and Martens, 1998). One suspected reason for this fast route to extinction is what Felsenstein (1974) coined *Muller’s Ratchet* (Muller, 1964) under the assumption that most mutations are deleterious (Goyal et al., 2012), the accumulation of mildly deleterious mutations that cannot be purged by recombination and subsequent outcrossing may lower fitness and ultimately lead to an unviable genotype. Experimental evidence to support this idea remains sparse, but it has been shown recently that low-quality genotypes deteriorate in a feedback loop (Sharp and Agrawal, 2012). Some researchers have argued that obligate parthenogenetic species might evolve a lower mutation rate, possibly mediated by enhanced DNA-repair systems (Schon and Martens, 1998) to avoid the ratchet. Low mutation rates and low genetic diversity in some ancient obligate parthenogenetic Crustacea (Darwinulidae; Ostracoda) support this idea (Schon et al., 2003). However, estimates from the parthenogenetic bdelloid rotifers did not reveal lower mutation rates in comparison to their closest sexual relatives (Welch and Meselson, 2001). Still, the estimates for Darwinulidae and the Bdelloids are derived from comparative phylogenetic approaches of very few genes (Schon et al., 2003; Welch and Meselson, 2001), not genome wide analyses. They might thus be strongly biased by the selection of genes. In the age of genomics it is now possible to conduct such mutation rate studies on a genome-wide level. Mutation accumulation (MA) line experiments to measure the effect of accrued slightly deleterious mutations under neutral evolution (excluding selection) have been established long ago, with Michael Lynch likely being the first to apply this technique to *Daphnia* more than 30 years ago (Lynch, 1985). The most famous experiment of such kind, however, is the long-term MA line in *E. coli* conducted by Richard Lenski and his team, which has reached more than 58,000 generations in 25 years and is still on-going (Pennisi, 2013; Bacteria are

3.3 Discussion of M1: DSD has changed GRNs in clade IV nematodes and might facilitate the evolution of parthenogenesis

Excursus 2 (Cont.)

of course not picked as single individuals, but the general process is the same). Similar to the *E. coli* experiment, the first MA lines in *C. elegans* were started before it was easily possible to sequence and re-sequence genomes and thus only phenotypic fitness could be measured (Vassilieva et al., 2000; Vassilieva and Lynch, 1999). Starting first from massive PCR assays and then moving to 2nd generation sequencing MA line experiments have now been conducted in several Metazoa, such as *C. elegans* and *D. melanogaster* (Denver et al., 2004, 2009; Haag-Liautard et al., 2007; Keightley et al., 2009).

To measure for the first time the genomic mutation rate in a parthenogenetic species, which was actually called for by (Baer et al., 2007) in a review on this topic, I used the *C. elegans* experiments as a template to set up an MA line experiment.



The general experimental setup is relatively simple, starting with a single specimen individual offspring are randomly chosen and separately distributed into fresh culture habitats (new agar plates for nematodes). Each individual is then the founder of one line. Every generation a single offspring per line is then again randomly chosen and put into a fresh culture habitat. In amphimictic species every generation two individuals of different sex have to be picked of course, but lines are maintained separately throughout the experiment. The culture habitat is set up to be as beneficial as possible for the tested species. Food, temperature, humidity, and

Figure 8: The general workflow of an MA-lines experiment is depicted here. Details are given in the Excursus text.

other factors are kept constant and possibly at an optimum to prevent selection. The next generation is transferred around the mid of the parents reproduction cycle to not select for early or late offspring. This procedure enforces an artificial population bottleneck keeping the effective population size to 1 in parthenogenetic and 2 in outcrossing species. Only

Excursus 2 (Cont.)

(nearly-)neutral mutations should accrue in the lines and only their fitness effects (if any) will be measured. As unexpected accidents can always happen (e.g. worms crawling of the plate or flies escaping from their tubes) and spontaneous killer-mutations can arise (e.g. terminally stopping reproduction) previous generations are kept to allow to try up to 3 backups per generation per line. If a line fails to propagate even after 3 backup attempts it is deemed extinct due to the accumulation of deleterious mutations. Figure 8 on the preceding page depicts the general procedure.

Starting with 100 parthenogenetic *Panagrolaimus* PS1159 lines and maintaining 71 lines for the hermaphrodite *Propangrolaimus* JU765 as control, my project was run for over 2.5 years in total (initially it was thought these species were both from the genus *Panagrolaimus*, see section 3.2 on page 148). The parthenogenetic lines experienced a drastic drop in number with only 15 remaining (equalling 15% of the starting lines) when the project was terminated. At this time the parthenogenetic lines had reached a maximum of 52 generations, but owing to the increasing need of backups early on in the experiment many lines were only between 30 and 40 generations old. At the same point 30 (42%) of the hermaphrodite lines were still viable, having reached a maximum of 50 but the majority being between 30 and 40 generations. Although it could in theory be possible that more of the PS1159 lines were lost through worms leaving the plate, this appears unlikely and should also have been counteracted by the “3-backups-strategy”. Thus, the observed difference in retained lines also indicates different fitness effects between the modes of reproduction. To further measure this effect I developed a fecundity assay, in which we counted offspring produced by the bottlenecked MA lines in comparison to worms from the founder population. For this, individual worms were placed in a hanging drop of liquid growth medium under the lids of in multi-well plates, a method adapted from Muschiol and Traunspurger, 2007, see there for an illustration. Worms are unable to leave these drops, but produce eggs and offspring hatches. Given that *Panagrolaimus* typical habitats are moist sandy soils and litter fractions of soils (De Goede et al., 1993) and *Propanagrolaimus* was isolated on algal mats in an underwater cave (Muschiol and Traunspurger, 2007) such semi-liquid environments should not affect the nematodes’ reproductive success, at least not in intra-species comparisons. For each line and the control several worms (5 - 10) were tested and all layed eggs and hatching offspring were counted. Preliminary results from the experiment, conducted with support from Magarete Schierenberg, showed a difference in offspring production between the PS1159 founder generation (which had been stored in the anhydrobiotic state, see **M1**, for more than 2.5 years and the bottlenecked MA lines. The founder generation produced on average as many offspring as the control species PS1579 (see **M1** for phylogeny), while the MA lines produced much less, see figure 9 on the facing page. In the hermaphrodite JU765, no difference between MA lines and founder generation could be detected. Some MA lines had more offspring on average than the founder generation, while others were

Excursus 2 (Cont.)

worse. As for example in some Lines 4 out of 5 tested worms failed to produce any offspring, a second iteration of this experiment should be evaluated to increase statistical significance. But already now the data appear to indicate that the loss in fitness was much more pronounced in the parthenogenetic MA lines, than in the hermaphrodite ones.

It was necessary to construct reliable reference genomes to map genomic data obtained from the MA lines to access the mutation rates. The process of reference genome assembly is detailed in the Methods section and also in M1; the PS1159 and JU765 reference genomes for the MA line mapping are the ones to be published in this manuscript. As noted there, the reference genomes are constructed not only from the founder populations (of the, MA

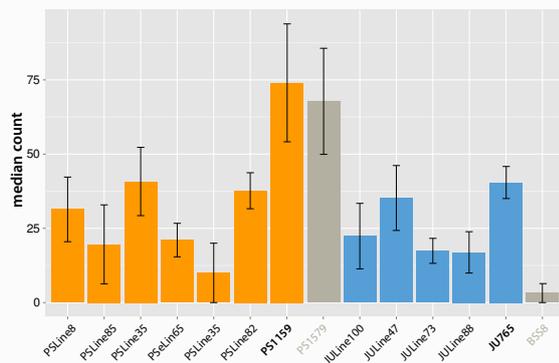


Figure 9: A fitness assay was conducted in the MA lines after many generations of bottlenecking compared with the founder generations and two control species. Shown are the mean numbers of eggs and offspring produced by each of the lines tested, the founder generations and control species. PS1159 founder and lines are in orange, JU765 in blue. Control species black. Error bars show standard error of the mean. Details are given in the Excursus text.

line) but also incorporate the bottlenecked genomes. This is due to the fact that some of the sequencing libraries a large amount of bacterial contamination reducing the available target species reads in them (PS1159), or even had to be discarded for containing eukaryotic contamination (JU765; one library contained *Arabidopsis* another *C. elegans* reads resulting from handling errors in the sequencing centre). Another reason for the need of higher coverage and thus the incorporation of the MA line derived read sets was the larger than initially expected genome size of PS1159.

Inflating the PS1159 populations to large enough numbers of worms for the sequencing was difficult, as the nematodes produced few offspring and these were growing slowly. This can be interpreted as a general indication of fitness loss in the parthenogenetic MA lines. In total six PS1159 and six JU765 MA lines were sequenced. For four each of the PS1159 and JU765 lines technical replicates were sequenced, i.e. these libraries were sequenced twice to allow detecting possible sequencing errors and to increase overall coverage.

To access mutation rates the sequenced MA lines were mapped against the respective reference genomes and mutations then called from these mappings. A variety of tools was used to this end. After cleaning, as described in M1, the reads were mapped against the references using the `clc_mapper` (v.4.2). The resulting file (in the proprietary `cas` format of CLCbio) was then converted into the universal `bam` format, while discarding unmapped reads. The `bam` files were then sorted using `bamtools` (Barnett et al., 2011) and a `mpileup` file containing raw snps and corresponding read and mapping qualities was created for each using `samtools` (Li et al., 2009). Based on these `pileups` mutations were then called using

Excursus 2 (Cont.)

VarScan (v.2.3.6) (Koboldt et al., 2012). The VarScan software detects the most likely 'real' mutations by filtering for ambiguously aligned reads, coverage and quality. To exclude scoring snps that did not arise in the MA process but are a reflection of the hybrid background only mutations that were supported by 99% of reads at a given position with at least 10% coverage were counted with VarScan. The re-sequenced libraries were treated independently and the VarScan output filtered for snps found in both sequencing runs. Each MA line was an independent experiment and consequently the exact same mutation occurring at the same position in two independent lines is unlikely (Saxer et al., 2012). Thus, I used VCFtools (Danecek et al., 2011) to select only those mutations, which were restricted to a single line. The mutation rate μ could then be calculated according to a formula by Keightley et al. (2009):

$$\mu = \frac{\text{no. of mutations called}}{nt \times \text{no. of sites}}$$

Here n stands for the number of MA lines and the mean number of generations in the MA lines is represented by t , **no. of mutations called** are the total number of mutations inferred (here at the 99% threshold) from **no. of sites** that have been considered in the mapping approach, see Keightley et al., 2009 for details. Using this formula I calculated 6.82×10^{-08} mutations per site per generation in the parthenogenetic PS1159 and 6.5×10^{-09} in the hermaphrodite control JU765. These results have to be regarded as preliminary, as a second mapping tool and variant calling software should be tested to validate the results (currently in process) and statistical tests need to be conducted to obtain standard error estimates by evaluation all lines on their own. However, obtaining indications for a much higher mutation rate in the parthenogenetic lines than the hermaphrodite controls, or the published rates for model organisms, is quite intriguing. A first implication could be that Muller's ratchet is indeed clicking in parthenogenetic lineages, driving these to extinction on short evolutionary time scales. Given the fitness effects observed after 30 - 50 generations in the short MA experiments, **how then can parthenogenetic species survive for prolonged evolutionary times?**

One explanation could be the huge population sizes in many invertebrate taxa. The strong effect of the ratchet on small populations was shown in experiments with digital organisms (Misevic et al., 2004) and computational models for structured ("patchy") Bacteria populations (Combadão et al., 2007). Gordo and Charlesworth (2000) estimated that within large populations, thousands of generations could pass before the ratchet causes a considerable decline in fitness. Thus, the very high numbers of individuals per species per location (Barrière and Félix, 2005; Robertson and Freckman, 1995) could buffer against fast extinction in nematodes (and other invertebrates) after the evolution of parthenogenesis. A polyploid hybrid background, as inferred for the *Pangrolaimus* species could even enhance the evolutionary capacity of parthenogenetic species by providing a masking effect through

Excursus 2 (Cont.)

the availability of an extra allele per gene. It is thought that already diploidy at least temporarily masks against deleterious mutations (Lynch et al., 1993). It has been argued that parthenogenesis can be selectively advantageous (i.e. be more evolutionarily fit) under specific environmental conditions and parthenogenetic animal taxa are frequently found in extreme environments, where they seem to have an (at least short-term) evolutionary advantage over outcrossing species (chapter 6 in Schön et al. (2009)). It has already been argued that a hybrid origin could lend parthenogenetic species an adaptive advantage in extreme environments (Kearney, 2005) and I hypothesise that a gain of genetic diversity due to higher mutation rates could act to this effect. These would allow parthenogenetic species to develop a greater adaptive potential on shorter time scales than sexual congeners are capable of, especially given the fast growth rate of parthenogenetic populations. Extinction should still occur faster than in sexual species, but as at least some beneficial mutations will arise even in small sexual populations (Goyal et al., 2012), the ratchet could temporarily speed up adaptation in large populations under a changing extreme environment. An interesting follow up question from this would of course be if more more parthenogenetic taxa are observed during times of climate change - this, as it seems, can be tested in the near future.

While data from my MA line experiments indicate a short evolutionary life time for most parthenogenetic species due to Muller's ratchet, they do not explain the persistence of some ancient unisexual lines. One route to persist for these species could be the abolishment of meiosis and transition to apomictic (mitotic) parthenogenesis, which is at least found in the bdelloid rotifers (Mark Welch et al., 2004a,b), but for example also in many parthenogenetic nematodes of the genus *Meloidogyne* (Lunt et al., 2014), where parthenogenesis is thought to have evolved more than 17 Ma ago (Castagnone-Sereno and Danchin, 2014). Polyploidy has been found in both, rotifers and nematodes Flot et al. (2013); Lunt et al. (2014); Triantaphyllou (1966), and it has recently been shown that ameiotic gene conversion could act as a mechanism to limit the effects of Muller's ratchet in the Bdelloidea *Adineta vaga* (Flot et al., 2013). Similarly, ameiotic recombination has been reported in parthenogenetic *Daphnia* species (Omilian et al., 2006).

Here, I would cautiously suggest that the intrinsic diversity of a hybrid origin will lead to a divergence of alleles in the absence of meiosis.

This has been called the Meselson effect, when described in diploid systems (Butlin, 2002). It has been ruled out for ancient parthenogenetic oribatid mites (Schaefer et al., 2006) and *A. vaga* (Flot et al., 2013). However, we have to consider that each of the two just cited analyses is a snapshot in time, restricted to one or very few species. Thus, in principle one could assume that when meiosis and the repair mechanisms it includes is not acting every generation (as is the case in the mitotic systems) alleles diverge. Some of these alleles will accumulate deleterious mutations, others will gain beneficial mutations. Occasional rounds

Excursus 2 (Cont.)

of ameiotic gene conversion or recombination could weed out the bad alleles in some lineages (especially when multiple copies of the healthy alleles are present in polyploid species). This would keep these lineages evolutionary fit and able to speciate, while the ones where to many deleterious mutations accrue or bad alleles happen to replace good ones in gene conversion will simply be lost. However, this pattern of overwriting bad alleles in ameiotic gene conversion or recombination on the one and extinction of lineages on the other hand would be obscuring the Meselson effect from our view. Only long-term experiments including rounds of re-sequencing in these taxa could reveal its presence.

I would speculate that especially the gene conversion is of interest here. This could be hampered in hybrid polyploid meiotic systems as homologous chromosomes from different parental species might not correctly pair (Maheshwari and Barbash, 2011); for example in triploid systems only one set of chromosomes might exchange material every generation. However, the ameiotic gene conversion might be efficiently working under mitosis, as the rotifer examples seems to indicate (Flot et al., 2013). While speculative the theory just laid out could explain why described ancient parthenogens appear to be ameiotic (Schön et al., 2012).

3.4 Discussion of M2: Genes from the Cambrian Explosion unite Bilateria in a developmental phase

Changing perspective from loss to conservation and from diversity to unifying traits we also analysed a far larger group of animals, which Nematoda are only a small part of, the Bilateria. Nematoda along with Panarthropoda and some less known and less specious groups like Onychophora, Priapulida and Kinorhyncha constitute the Ecdysozoa, the moulting animals, which again are united in Bilateria along with Lophotrochozoa, Vertebrata, Echinodermata, and some smaller groups Dunn et al. (2008). Traditionally, Bilateria have been of high interest to evolutionary biologist, as this group contains the vast majority of animal species living on earth today (humans among them), and was thought to have come into existence in a relatively short time frame, the so called Cambrian Explosion about ~ 540 Ma ago (Erwin et al., 2011). Given that non-Bilateria show complex body organisations (Martindale, 2005), including bilateral symmetry (polyps of Anthozoa, see Berking, 2007), and mesoderm derived tissues (in Ctenophores, see Ryan et al., 2013), the assumed strict distinction between Bilateria and non-Bilateria appears to become somewhat fuzzy. Even more so, recent work on non-bilaterian animals has shown that genomic complexity in these taxa is high (Hudry et al., 2014; Ryan et al., 2013; Schwaiger et al., 2014). Still, the Ctenophore example is intriguing as mesoderm as well as nervous-system specification appear to be orchestrated by an independently evolved genetic system (Moroz et al., 2014). The four manuscripts dis-

cussed in the previous sections illustrate an unexpected amount of change in development and its genomic background on different taxonomic levels in Nematoda, which is in line with these findings. It is clear on the other hand that many processes and genes are conserved among animals, and indeed many such genes have been described (Hunt, 2011).

One major trait of Bilateria might be the hugely divergent morphologies or phenotypes evolved in the taxon (from humans, to roundworms, flies, and mussels), including heavily derived forms like *Sacculina* barnacles, adult Tunicates, or Holothuroidea and other echinoderms, as well as the potential to (rapidly) evolve new phenotypes, for example in the Panarthropoda. Given this tremendous divergence we were interested to know which genes were conserved across Bilateria and to what extent such genes could be conducting conserved bilaterian-specific functions in different organisms even after more than 600 Ma² of evolutionary time. To increase the likelihood of an inferred gene set to act in a bilaterian-specific way, we also tried to exclude genes with non-bilaterian orthologues. We used data from 10 bilaterian species from Ecdysozoa, Lophotrochozoa, and Deuterostomia, as well as 7 non-bilaterian species including the plant *Arabidopsis thaliana* as a remote outgroup in our assay. From this we compiled 4 sets of orthologues, which allowed for different patterns of independent loss of genes along the branches in Bilateria (for example in the most strictest set all tested species had to be present, while in another set only the three model organism where required to be present). By mining literature databases we then functionally characterised 85 groups of orthologues in an intersection of two of these sets, which contained the three model species *C. elegans*, *D. melanogaster*, and *D. rerio*, as well all other species tested, but allowing for maximally one loss along the branches of the bilaterian tree (see figure 1 in **M2**).

From this analysis, as well as a GO-term screen, it appears that almost all of the genes in our set are acting in development of the model organisms. In particular, certain developmental aspects in neuron and muscle formation, as well as the general process of morphogenesis are conserved across Bilateria through these genes. Intriguingly, when analysing patterns of expression of our genes in the three model organisms we also find one conserved pattern: many of the genes are up-regulated in distinct but similar developmental stages between worm (*C. elegans*), fly (*D. melanogaster*) and fish (*D. rerio*). This is particularly true towards the end of embryonic development in all three species and then for a second time in the pupa stage of the fly, when the adult body is constructed. The one expression peak towards the end of embryonic development observed in *C. elegans* and *D. rerio* corresponds to what was described as possible phylotypic stages in *Caenorabditis* (Levin et al., 2012) and the fish (Kimmel et al., 1995), also see the **P2** manuscript. These data are in line with recent findings from the modEncode project where an approach starting from expression

²The Cambrian explosion is currently dated to about 540Ma (Erwin et al., 2011), but the genes must have been present in the stem species of all Bilateria (excluding for rampant Lateral Gene Transfer later on in evolution). Even assuming only a medium length for the fuse of the Cambrian Explosion (Davide Pisani, personal communication) 600 Ma is a conservative estimate for the last common bilaterian ancestor.

data (not groups of orthologues) found similar peaks (Gerstein et al., 2014; Li et al., 2014). Finding two expression peaks in *D. melanogaster* in comparison to one in *C. elegans* and *D. rerio* is of particular interest, as it can support the hypothesis that the genes found by us are important in shaping the body form, a process happening twice in the fly - during the transition from embryo to larva and then again during metamorphosis.

In other words, despite the huge evolutionary distances and the large amount of phenotypic disparity among Bilateria, parts of the genetic machinery for setting up the nervous system, muscles, and building the adult bilaterian body is retained even in very distantly related species. What is more, these genes are also expressed in comparable developmental phases. Still, among the genes in our inferred sets of orthologues we did not find previously described key players (e.g. HOX genes): our finding resembles more a set of small parts which keeps the machinery running. It appears that the process of the building of body form could be more important than the underlying molecular toolkit and could constitute the trait uniting Bilateria. Following this, our findings appear in line with the idea that stem taxa of Bilateria were small animals, similar to larval forms in modern marine taxa, and the invention of the adult body plan was a key event in the history of the taxon, see Peterson and Davidson (2000) and Chen et al. (2000) for an introduction to this hypothesis. It should be noted, that this hypothesis is far from accepted among Palaeontologists and evolutionary Biologists, see for example Budd and Jensen (2000) for arguments in favour of larger animals in the stem group.

The main limitation of our project is the lack of expression data from a variety of taxa outside the model organisms, even more so given the huge amount of DSD and GRN shuffling described in the previous sections. Thus, to be able to propose a truly pan-bilaterian genetic machinery for certain developmental phases it will be necessary to accumulate more data, genomic and functional, from a variety of neglected and elusive taxa, such as the mentioned onychophorans or priapulids (genomes in progress; Georg Mayer, personal communication and <http://genome.wustl.edu/genomes/detail/priapulid-caudatus>), and lophotrochozoans in general. Especially the latter as well as echinoderms are of interest, as these animals (but also many other marine species) retain what has been termed “maximal indirect development” (Peterson et al., 1997). The evolution of this type of constrained development, where the morphology of the usually planktotrophic larva is completely different from the adult, has been proposed as an evolutionary asset of early bilaterians (Blackstone and Ellison, 2000). Similar to *Drosophila*, one would expect to observe a “double peak” in the developmental expression pattern of our orthologues in species with such larval forms; one phase of up-regulation in the transition from embryo to larva, the other when the adult is formed from the larva. Finding this would argue for the hypothesis that the developmental phase during which the adult body form is build is at the core of bilaterian evolution and unites the taxon.

Recent studies of the genomes and developmental transcriptomes from two Comb Jellies

(Ctenophores) revealed a divergent molecular toolkit for mesoderm and neuronal signalling in these animals (Moroz et al., 2014; Ryan et al., 2013), while an analysis of gene regulation in the Cnidarian *Nematostella* illustrated high genomic complexity in this animal (Schwaiger et al., 2014). This hints at the (genomic) diversity in non-bilaterian taxa, where a much smaller number of genomes are available and much less species are studied in the laboratory. Therefore, it will be necessary to extend our genomic database in non-Bilateria to learn more about connections and disparities between these and the Bilateria. For example, the marine cnidarian *Hydractinia echinata* is currently genome and transcriptome sequenced (with my participation) and preliminary analysis of sex related genes in this species showed links to genes known from Bilateria species.

Such genome and transcriptome projects are very likely to succeed when researchers with different areas of expertise collaborate. An example how such collaborations can be fostered is discussed in the next section.

3.5 Discussion of **P1**: We need to sequence more genomes in collaborative efforts

The first four manuscripts (**P2**, **M3**, **P3**, **M1**) presented here indicate considerable Developmental System Drift which re-shuffled Gene Regulatory Networks (including in the system of sexual and unisexual reproduction), while the fifth (**M2**) manuscript presented data from a metazoan-wide analysis which identified a process that could unify the morphologically diverse Bilateria. The emerging picture appears to be that while some biological processes in development might be shared among groups of animals separated by 600 Ma of evolution, the list of universal genes potentially restricted to Bilateria is small. In other words there is a high level of genomic disparity between taxa in Bilateria. The recent descriptions of divergent genetic systems in Comb Jellies (Moroz et al., 2014) (discussed in **M2**) show that this variability is likely present across Metazoa. In Nematoda only ~25 species are currently genome sequenced, with a strong emphasis on Chromadorea, including human parasites in clade III, plant parasites in clade IV and the species around *C. elegans*. Although datasets on some additional Enoplea species have recently been published, these are the vertebrate parasites, e.g. *Trichuris suis* and *Trichuris muris* (Foth et al., 2014; Jex et al., 2014), close to *T. spiralis*. It is clear that a huge diversity within Nematoda still awaits discovery.

We need to assemble more genomic data in an efficient way, i.e. without the waste of time and money resulting from different groups sequencing the same organism. Ideally, groups with overlapping interest should bring their expertise together and collaborate as much as possible. Our paper entitled "959 Nematode Genomes: a semantic wiki for coordinating sequencing projects" published in 2011 may be the most important of the included manuscripts. While not describing scientific discoveries, it presents a platform for researchers to consoli-

date their approaches to collect genomic and transcriptomic data across the nematode phylum. The 959 Nematode Genomes wiki platform and database www.nematodegenomes.org is meant to help to avoid duplicate work and foster collaborations by allowing researchers to flag their species of interest, provide summary details of sequencing strategies, data already acquired, and get in contact with colleagues interested in the same species. Collaborations have already been facilitated through the platform, for example the *Dirofilaria immitis* genome project (Godel et al., 2012). With the ever growing amount of data accrued for each sequenced species and the plethora of different evolutionary questions to explore only such collaborations will help us to gain a deeper understanding of evolution across the whole phylum.

3.6 Conclusions and Outlook

To summarise the publications included and discussed in the light of the introduction I will first draw some major conclusions here, before outlining at least two strings of future research programmes, which will allow us to gain deeper insight into evolutionary processes in Nematoda in general and the evolution of parthenogenesis in particular.

3.6.1 Concluding remarks

In the Introduction some specific questions have been asked, and some at least partial answers can be given based on the publications included. The first set of questions was:

- Is the highly conservative Bauplan model of nematodes (see above) reflected by an equally conservative genetic toolkit for early development or is the apparent plasticity on the cellular level found in some developmental processes reflected by rapid genomic evolution? In other words, how much plasticity in the genetic toolkits, or DSD, is inherent to specific taxonomic levels within the phylum?

It appears that the answer to the second, not the first, question would be “yes”: The 4 included publications mainly analysing nematodes, do suggest that on all taxonomic levels, from between morphologically very uniform *Panagrolaimus* and *Propanagrolaimus* species, to *C. elegans* and *R. culicivora*, the genetic toolkit of early development undergoes considerable change - even on short evolutionary time scales.

- If genes known for their crucial function in *C. elegans* early development (and maybe in outgroup species, e.g. *Drosophila*) are conserved, are the respective interaction partners of these important genes retained as well (on different taxonomic levels)? This question explores the divergence of GRNs.

The emerging pattern is that in all analysed GRNs only some genes are conserved. Interaction partners and thus the whole network is under considerable evolutionary change -

Developmental System Drift is constantly re-shuffling Gene Regulatory Networks. This general pattern was illustrated in the discussion of **M1** with the *skn-1* endo-/mesoderm gene cascade as one example, and the A-P axis system in the discussion of **P2** as another. But examples for a large group of developmental processes are given in the included publications. Davidson's theories of GRN network evolution are therefore supported by our data.

- How is potential plasticity involved in the evolution of parthenogenesis, and are convergent or parallel genomic pathways taken in the process?

Findings in the **M1** indicate that in the GRNs which orchestrate fertilisation, egg-activation and early axis determination, some key genes are conserved while others are rapidly substituted. The liability of genetic systems to rapid changes appears to facilitate the evolution of parthenogenesis by allowing the adaptation of the respective GRNs to "the needs" of a newly evolved unisexual species. However, the ubiquitous change also made it impossible to infer from genomic data alone which genetic pathways are changed. Thus, especially the second part of this question could not be answered yet, but a research plan to do so will be introduced in section 3.6.2 on page 169 ff.

- Assuming drastic changes in the GRNs, as proposed by Davidson: is there any set of universally conserved genes which is functioning in comparable life-cycle processes across the diversity of bilaterian species?

Our comparative study did indeed identify such a set of genes. We could also show that processes in which these genes appear to act in remotely related organisms (fly, worm, and fish) are also conserved. Most intriguingly these are developmental processes, which raises the possibility of shared developmental phases for Bilateria, maybe even laying the foundation for their evolution more than 600 Ma ago.

In more general terms the findings presented here indicate a rapid evolutionary turnover of genes acting in Nematoda development. Molecular evolution was previously found to be very high in the phylum, which is evidenced by long branches found in phylogenetic studies, see e.g. (Holterman et al., 2008, 2006; van Megen et al., 2009). However, the actual base-per-generation mutation rates in *C. elegans* or in the hermaphrodite species JU765 (see **Excursus** 2 on page 156) do not significantly deviate from for example *Drosophila melanogaster* (Denver et al., 2004, 2009; Keightley et al., 2009), or *Arabidopsis thaliana* (Warthmann et al., 2010). At this point it is not clear, which evolutionary forces drive the huge divergence observed on the gene sequence level (i.e. cause large branch lengths).

One idea proposed here is that we observe what is in part an artefact of taxon under-sampling; the phylum could simply be so specious that we are missing many species in-between when comparing a set of species. This would be in line with estimates that species numbers in Nematoda could exceed 1 Million, and may even be as high as 10 Million (see section 1.3 on page 5 ff.). Under-sampling is particularly likely in marine samples, which

are hard to obtain. When analysed in bulk, the marine environment appears to be equally nematode rich as the terrestrial (Danovaro et al., 2009; Fonseca et al., 2010) and thus many species must await discovery there. Conversely, one could assume that many species along the branches are extinct and we see mainly crown species in crown groups - the tips of the phylogenetic tree. This, however, seems unlikely given the constant sampling of new species, for example as mentioned above for *Caenorhabditis* with now 45 recognised species, or the diversity found in the species *P. pacificus* alone (McGaughran et al., 2013).

Still, the processes observed in the included works, the frequent loss and gain of genes, and the evolutionary change leading to divergence of sequences beyond a point where these are recognised as orthologues (thus indicating the possibility of functional change), is a different mutational process than the neutral substitution and fixation of single nucleotides. Ohno had pointed to the evolutionary power of gene duplications (Ohno, 1970) and as Nei recently argued, these kinds of large mutations (i.e. the duplication of a gene, an exon, or even a large chromosomal region) could indeed be seen as the driving forces of evolution, in some case breaking constraints by creating novelty (Nei, 2013). Some of the evolutionary changes we observed in the GRNs will have been adaptive. For example barriers to inter-species mating have evolved rapidly in *Caenorhabditis* nematodes potentially driven through intra-species sperm competition (Ting et al., 2014). The genetic component of the processes in the just cited work, like oocyte degradation, could be mediated for example through sperm-oocyte interacting proteins (the authors give a variety of possible candidates, including Major Sperm Proteins), which we found to differ between *Panagrolaimus* and *C. elegans*, too.

However, Sommer argued that not every trait evolving in development can and should be seen as steered by selection (Sommer, 2009). And it indeed appears unlikely that the frequent change and turnover of many genes we found in developmental systems even between closely related genera is entirely adaptive. Even more so, as nematode embryonic development happens inside the protective environments of eggs, which in many cases are encased in a shells hard enough to survive extended bleach treatments in the laboratory (killing adult worms, bacteria and fungi). Population bottlenecks are also easily imagined when single individuals are transferred to a new environment, for example by hitchhiking on flying Sauropsida. It is therefore necessary to assume drift to play an important role in speciation processes, too.

Another lesson from the presented analysis is that Horizontal Gene Transfer can indeed lead to the gain of diversity and constraint-breaking adaptive potential in and between groups of animals. Especially the photolyases, which have apparently been independently gained in different nematode taxa, do call attention to the public goods hypothesis introduced by McNerney et al. (2011). These authors conjecture that genes in general cannot be seen as entities belonging to a certain taxon (like taxonomist assume), but are frequently exchanged. They state that for example *E. coli* as a species has $\sim 18,000$ genes, but a single

strain only has up to 5,500 genes in its genome. Although bacteria are of course highly diverse, the example of the hyper diverse *C. brenneri* (Dey et al., 2013), mentioned already in the introduction (section 1.3 on page 5) and the large number of ORFan genes found in many species (Srinivasan et al., 2013) indicate the huge genomic plasticity in nematodes (and potentially Metazoa in general). Part of this diversity could indeed be gained through the random re-distribution of genes by HGT.

Initial divergence, speciation and further divergence of developmental GRNs after the formation of new species will thus be driven by both, selection and drift. These rapid evolutionary processes would be largely uncoupled from the neutral mutation rate measured in experiments on laboratory model animals.

In summary, we find a puzzling dualism across all hierarchical levels of the phylum (from genera upwards), on the one hand the strict vermiform Bauplan on the other hand the huge divergence in cellular and molecular patterns in the development of this form. Given the huge range of food and habitats exploited by the worms it is safe to assume that genes underpinning metabolism are also under constant change (see for an example the *B. xylophilus* and *H. contortus* genome reports, Kikuchi et al., 2011; Schwarz et al., 2013). This may well be what makes nematodes so successful in terms of species numbers and diversity:

having arrived at a common, somewhat constrained, but generally working body form early on, genetic resources, instead of being wasted on dead end Bauplan models (e.g. Trilobites, or Ammonites), could have been freed for evolutionary tinkering (Jacob, 1977), allowing nematodes to exploit all available habitats on earth (though worms don't have wings). This last, somewhat informal statement about wings, however illustrates the possible constraints by fixing a Bauplan model early on in evolution. Still, our data, mostly showing absence of *C. elegans* genes in other Nematodes, do not give an answer yet to the the questions how exactly, and when the derived vermiform morphology evolved. The set of genes acting when the worm shape is realised seems to be rather small (see the *Romanormis genome project* and the manuscript on Bilateria) and thus we need to study in more detail the genomes of Enoplea nematodes as well as close outgroups. Such an approach including some preliminary findings, comparing data from these taxa to outgroups across Bilateria will be outlined in the following final section of the presented thesis.

3.6.2 Outlook

Based on the projects presented in this thesis two major, intertwined strings for future research emerge. A substantial amount of work has been devoted to one of these strings in the Schierenberg lab during the last 3 years with my help, already delivering preliminary findings:

Understanding the molecular backbone of development in Enoplea

While we found evidence that DSD is acting at all taxonomic levels in Nematoda in **P3**, **M1**, and **P2**, these works are also limited. In these studies the molecular toolkits of development are explored from the *C. elegans* point of view, describing what is different and what is missing in the other roundworms in comparison. We found all the analysed species to lack a large part of the elaborate *C. elegans* genetic toolkit for early development, but it is reasonable to assume that these nematodes have other genes in place to fulfil the respective crucial functions. Thus, as was already indicated at the very end of the **Introduction** (section 1.6 on page 11), it would be very interesting to explore if Enoplea nematodes, being phylogenetically closer to outgroup taxa than *C. elegans*, also share more of the genes found in other Metazoa (Ecdysozoa in particular and, on a wider scale, Bilateria) than our standard model.

To this end we conducted a study comparing data from all major nematode clades, including the first large-scale data for a clade II species, *Enoplus brevis*, the closest outgroup to Nematoda within Nematoidea, a nematomorph, as well a first genomic data from a Tardigrade, *Hybribius dujardini*, and several outgroup species in Panarthropoda, Lophotrochozoa and Deuterostomia. Although not fully analysed, this assay has already revealed surprising links to outgroups in the molecular toolkits of Enoplea nematodes. In particular, it uncovered some highly interesting genes, important in bilaterian development, that had so far been thought to be lost in nematodes based on the *C. elegans* data. For example the BMP antagonists chordin and noggin were found by us in enoplean nematodes, as were hunchback, an orthologue of knirps, and orthodenticle. So far it can only be speculated about the function of these genes in Enoplea. However, their retainment indicates that processes like

- DV-axis formation and gastrulation (chordin/noggin), Martindale (2005),
- HOX gene regulation and segmentation (hunchback), Marques-Souza et al. (2008),
- and AP-axis building and segmentation (knirps, a *Drosophila* gap gene; orthodenticle, A-P patterning of brains), Hirth et al. (2003); Naggan Perl et al. (2013); Tautz (2004)

although phenotypically modified, might still be orchestrated by similar genetic mechanisms in these species. Especially, the case of the BMP antagonists in gastrulation appears intriguing as the clade II species *Tobrilus diversipapillatus* was found to pass through a gastrulation

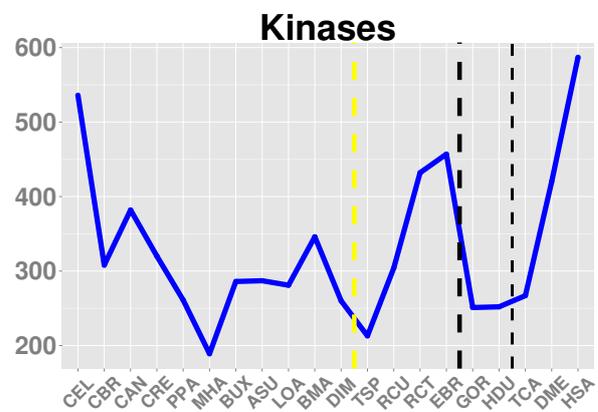


Figure 10: Dynamics of the number of kinases found across Nematoda and outgroups. *C. elegans* is the leftmost species (CEL). The yellow line separates Chromadorea from Enoplea, the first black line Enoplea from Nematopomorpha and Tardigrada, and the second these from *Tribolium*, *Drosophila*, and *Homo sapiens*.

with a large blastocoel, reminiscent of what is observed in classical models like the sea urchin (Schierenberg and Schulze, 2008). Additionally, we also find that some molecular pathways known to be derived in *C. elegans* were already altered in the earliest nematodes, or the nematode stem group. For example our data indicate that the number of wnt genes, which is at least 11 in basally branching Bilateria like *Daphnia* and *Platynereis*, as well as in humans and the non-bilaterian *Nematostella* (Janssen et al., 2010), is already reduced in the nematomorph and Enoplea, with further loss leading to the 5 genes in *C. elegans*. Similarly, our study delivers genomic evidence for a theory that was previously only based on ESTs (Aboobaker and Blaxter, 2010, 2003): No complete set of HOX genes is present either in the nematomorph or the Enoplea species. Thus, a shift in these signalling systems must already have its origin in the stem group of Nematodea.

Another interesting aspect of *C. elegans*' biology is the huge inflation of some gene families, such as F-Box proteins (~520 genes) or Kinases (~438 genes) (Hunt, 2011). Our data will allow us to unravel the dynamics of such gene-family inflations across the phylum (see figure 10 on the facing page for preliminary data on the Kinome). These data make it possible to access the molecular background of the Enoplea/Chromadorea divide in the nematode phylum and to look at evolutionary forces across the phylum. Containing Nematoda, the Ecdysozoa are still a problematic taxon, as ecdysis appears to be the only shared trait uniting a morphologically derived array of animals (Panarthropoda, Kinorhyncha, Loricifera, Priapulida, Tardigrada, Onychophora, Nematomorpha, and Nematoda) (Telford and Littlewood, 2009). Our cross-Bilateria study indentified sets of retained genes in large taxonomic groups, but indicated that the conservation of processes might be more important than these genes. A similar assay focussing on Ecdysozoa could thus refine our knowledge about this group and elucidate its common origin on a molecular basis. One plan for the future is to explore genomic diversity in this taxon at least at the order level, to determine what unites as well as what separates sub-taxa. Such a study should then also incorporate analyses of genomic structure (e.g. micro- and macrosynteny) to provide further details about genome evolution itself.

For Nematoda I developed a hypothesis from our preliminary data why Chromadorea are so distinctly different in their genomic backbone from Enoplea, which in turn are much more similar to outgroup animals:

As the first Chromadorean species to conquer land could have been a parasite (suggested in Rota-Stabelli et al., 2013) **I propose that it possessed a somewhat reduced gene set** (frequently observed in parasites, see our data on *T. spiralis* in **P2**). Later, free-living Chromadorea would have had to evolve new molecular functions from this reduced set of gene families, forcing evolution to tinker (Jacob, 1977) with a divergent molecular toolkit in processes conducted by a conserved machinery in other metazoa. To explore this hypothesis in greater depth and to address the questions raised towards diversity in the evolution of development at the end of the discussion of **P2** and **P1** it will be necessary to **sequence**

more worms, in particular in groups close to the Chromadorea/Enoplea divide and more taxa from the latter named group in general. I started already a genome and transcriptome project for a *Plectus* species, which appears to be a "stepping stone" in terms of nematode embryonic development (Schulze et al., 2012), but it will be equally important to **establish good laboratory models** in Enoplea to conduct functional studies. Thus, the community of Nematologists should follow steps already taken by, for example, researchers working on arthropods, by establishing species like the beetle *Tribolium*, or the bug *Oncopeltus* as laboratory models.

One major problem for predicting genes and their proteins from genomic or transcriptomic data is that two proteins from distantly related species might be clustered as orthologues, but still have a very divergent functional context. Although the OrthoMCL program has been tested and verified to find likely function-retaining orthologues, the cases of *skn-1* discussed in **M1** and *mex-3* in **P2** are good examples for genes where at least the expression domains are shifted (even in closely related species). As it is possible to conduct classical wet lab studies in *R. culicivora*, the species could be an invaluable model for further research into the biology of Enoplea nematodes. When correlated with the well-studied *R. culicivora* embryonic development (Schulze and Schierenberg, 2008), functional assays (e.g. antibody stainings) based on the now available genomic data could demonstrate how conserved or divergent proteins act, for example, in early axis formations.

Understanding the molecular underpinnings of parthenogenesis

The second proposed assay would build on the data from the *Panagrolaimus* genome project (**M1**) to analyse the evolutionary origin of parthenogenesis and the implicated changes to GRNs acting in processes like sex determination, axis formation, and egg activation. As explained in **M1**, I found that only some of the genes acting in these processes, which are known from *C. elegans*, are present in the clade IV worms. Thus, to truly understand which genes are playing a part in the evolution of parthenogenesis and if the genetic mechanisms underlying successful parthenogenesis are under convergent or parallel evolution in related taxa, the genes acting in these species have to be first identified and then compared with other genera where parthenogenesis evolved.

Parthenogenesis evolved independently in several branches of clade IV. As a result of my work for this thesis, genome sequences are now available for amphimictic and parthenogenetic *Panagrolaimus* species. Several *Meloidogyne* species (parthenogenetic and amphimictic) have been sequenced (Abad et al., 2008; Lunt et al., 2014; Opperman et al., 2008). Currently, I am preparing to sequence a parthenogenetic isolate of the *Aphenlenchus avenae* species complex. This is of special interest as in some of the parthenogenetic *Aphenlenchus* species the production of males can be induced through an increase of temperature (Hansen et al., 1973). Having a set of species from three genera at hand it will be possible to iden-

tify gene sets involved in male function and sex determination in a comparative RNASeq assay. For each of the genera *Panagrolaimus*, *Meloidogyne*, and *Aphenlenchus* differential gene expression analyses in different life cycles and sexes can be conducted. By isolating and sequencing RNA from

- juvenile (virgin) female parthenogenetic and amphimictic worms,
- adult virgin female parthenogenetic and amphimictic worms,
- adult egg laying female parthenogenetic and amphimictic worms,
- males,
- and male producing females of the temperature dependent parthenogenetic *A. avenae*

it will be possible to identify genes that are expressed in the adult parthenogenetic females, which are usually contributed by males (in amphimictic species) and likely to act in egg activation, as well as identify genes that appear to determine the male phenotype. These data can be complemented with novel, as well as already existing data on specific developmental stages (for example 1-cell and 8-cell stages) of key species, which have been generated in the laboratory of E. Schierenberg to analyse the axis determination network.

Based on the established phylogeny in clade IV it will then be possible to infer whether or not the candidate genes (or a subset of them) are shared among all parthenogenetic species in this taxon, which would indicate parallel evolution, or if the independent acquisition of the parthenogenetic phenotype is due to independent sub- or neo-functionalisation of different genes; thus, indicating convergence.

Chapter 4

Bibliography

- Abad, P., J. Gouzy, J.-M. Aury, P. Castagnone-Sereno, E. G. J. Danchin, et al. (2008). *Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita*. Nature Biotechnology, 26(8):909–915.
- Aboobaker, A. and M. Blaxter (2010). *The nematode story: Hox gene loss and rapid evolution*. Advances in experimental medicine and biology, 689:101–110.
- Aboobaker, A. A. and M. L. Blaxter (2003). *Hox Gene Loss during Dynamic Evolution of the Nematode Cluster*. Current Biology, 13(1):37–40.
- Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, et al. (1997). *Evidence for a clade of nematodes, arthropods and other moulting animals*. Nature, 387(6632):489–493.
- Ambros, V., R. C. Lee, A. Lavanway, P. T. Williams, and D. Jewell (2003). *MicroRNAs and other tiny endogenous RNAs in C. elegans*. Current Biology, 13(10):807–818.
- Andrássy, I. (2005). *Free-Living Nematodes of Hungary*. Nematoda Errantia. Hungarian History Museum, Budapest, 1st. edition.
- Arnold, M., E. S. Ballerini, and A. N. Brothers (2012). *Hybrid fitness, adaptation and evolutionary diversification: lessons learned from Louisiana Irises*. Heredity, 108(3):159–166.
- Avise, J. (2008). *Clonality: The Genetics, Ecology, and Evolution of Sexual Abstinence in Vertebrate Animals*. The Genetics, Ecology, and Evolution of Sexual Abstinence in Vertebrate Animals. Oxford University Press, 1st. edition.
- Baer, C. F., M. M. Miyamoto, and D. R. Denver (2007). *Mutation rate variation in multicellular eukaryotes: causes and consequences*. Nature Reviews Genetics, 8(8):619–631.
- Baldi, C., S. Cho, and R. E. Ellis (2009). *Mutations in two independent pathways are sufficient to create hermaphroditic nematodes*. Science, 326(5955):1002–1005.

-
- Barnett, D. W., E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth (2011). *BamTools: a C++ API and toolkit for analyzing and managing BAM files*. *Bioinformatics*, 27(12):1691–1692.
- Barrière, A. and M. Félix (2005). *High Local Genetic Diversity and Low Outcrossing Rate in Caenorhabditis elegans Natural Populations*. *Current Biology*, 25(13):1176–1184.
- Becks, L. and A. F. Agrawal (2010). *Higher rates of sex evolve in spatially heterogeneous environments*. *Nature*, 468(7320):89–92.
- Becks, L. and A. F. Agrawal (2012). *The evolution of sex is favoured during adaptation to new environments*. *PLoS Biology*, 10(5):e1001317.
- Bell, G. (1982). *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. University of California Press, Berkeley.
- Berking, S. (2007). *Generation of bilateral symmetry in Anthozoa: A model*. *Journal of Theoretical Biology*, 246(3):477–490.
- Beukeboom, L. and N. Perrin (2014). *The Evolution of Sex Determination*. Oxford University Press, 1st. edition.
- Blackstone, N. W. and A. M. Ellison (2000). *Maximal indirect development, set-aside cells, and levels of selection*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 288(2):99–104.
- Blaxter, M. (2011). *Nematodes: The Worm and Its Relatives*. *PLoS Biology*, 9(4):e1001050.
- Blaxter, M., P. de Ley, J. Garey, L. Liu, P. Scheldeman, et al. (1998). *A molecular evolutionary framework for the phylum Nematoda*. *Nature*, 392(6671):71–75.
- Boto, L. (2014). *Horizontal gene transfer in the acquisition of novel traits by metazoans*. *Proceedings of the Royal Society B: Biological Sciences*, 281(1777):20132450–20132450.
- Boveri, T. (1899). *Die Entwicklung von Ascaris megalocephala mit besonderer Ruecksicht auf die Kernverhaeltnisse*. *Festschrift fuer C. von Kupffer*. Jena: Fischer, pages 383–430.
- Brenner, S. (1974). *Genetics of Caenorhabditis elegans*. *Genetics*, 77(1):71–94.
- Budd, G. E. (2001). *Why are arthropods segmented?* *Evolution and Development*, 3(5):332–342.
- Budd, G. E. and S. Jensen (2000). *A critical reappraisal of the fossil record of the bilaterian phyla*. *Biological reviews of the Cambridge Philosophical Society*, 75(2):253–295.

- Burglin, T. (2006). *Homologs of the Hh signalling network in C. elegans*. WormBook, ed. The C. elegans Research Community, <http://www.wormbook.org>.
- Butlin, R. (2002). *The costs and benefits of sex: new insights from old asexual lineages*. Nature Reviews Genetics, 3(4):311–317.
- Bywaters, J. H., E. L. Lasley, A. Sokoloff, S. R. R., H. Fennessy, et al. (1959). *Tribolium Information Bulletin*.
- C. elegans Sequencing Consortium (1998). *Genome sequence of the nematode C. elegans: a platform for investigating biology*. Science, 282(5396):2012–2018.
- Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, et al. (2008). *MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes*. Genome Research, 18(1):188–196.
- Castagnone-Sereno, P. and E. G. J. Danchin (2014). *Parasitic success without sex - the nematode experience*. Journal of Evolutionary Biology, 27(7):1323–1333.
- Chalfie, M., Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher (1994). *Green fluorescent protein as a marker for gene expression*. Science, 263(5148):802–805.
- Chen, F., A. J. Mackey, J. K. Vermunt, and D. S. Roos (2007). *Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes*. PLoS ONE, 2(4):e383.
- Chen, J. Y., P. Oliveri, C. W. Li, G. Q. Zhou, F. Gao, et al. (2000). *Precambrian animal diversity: putative phosphatized embryos from the Doushantuo Formation of China*. Proceedings of the National Academy of Sciences, 97(9):4457–4462.
- Combadão, J., P. R. A. Campos, F. Dionisio, and I. Gordo (2007). *Small-world networks decrease the speed of Muller’s ratchet*. Genetical research, 89(01):7–18.
- Consortium, T. *Heliconius*. G. (2012). *Butterfly genome reveals promiscuous exchange of mimicry adaptations among species*. Nature, 487(7405):1–5.
- Coyne, J. A. and H. Orr (2004). *Speciation*. Sinaur Associates, Inc., 1st. edition.
- Creer, S., V. G. Fonseca, D. L. Porazinska, R. M. Giblin Davis, W. Sung, et al. (2010). *Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises*. Molecular Ecology, 19(s1):4–20.
- Curran, D. M., J. S. Gilleard, and J. D. Wasmuth (2014). *Figmap: a profile HMM to identify genes and bypass troublesome gene models in draft genomes*. Bioinformatics, 30(22):3266–3267.

-
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, et al. (2011). *The variant call format and VCFtools*. *Bioinformatics*, 27(15):2156–2158.
- Danovaro, R., S. Bianchelli, C. Gambi, and M. Mea (2009). α -, β -, γ -, δ -and ε -diversity of deep-sea nematodes in canyons and open slopes of Northeast Atlantic and Mediterranean margins. *Marine Ecology Progress Series*, 396:197–209.
- Das, P. P., M. P. Bagijn, L. D. Goldstein, J. R. Woolford, N. J. Lehrbach, et al. (2008). *Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline*. *Molecular Cell*, 31(1):79–90.
- Davidson, E. H. (2006). *The Regulatory Genome*. *Gene Regulatory Networks In Development and Evolution*. Academic Press, 1st. edition.
- Davidson, E. H. and D. H. Erwin (2006). *Gene regulatory networks and the evolution of animal body plans*. *Science*, 311(5762):796–800.
- De Goede, R. G. M., B. C. Verschoor, and S. S. Georgieva (1993). *Nematode distribution, trophic structure and biomass in a primary succession of blown-out areas in a drift sand landscape*. *Fundamental and applied Nematology*, 16(6):525–538.
- de Ley, P. (2006). *A quick tour of nematode diversity and the backbone of nematode phylogeny*. *WormBook*, ed. The C. elegans Research Community, <http://www.wormbook.org>.
- Denver, D., K. Morris, M. Lynch, and W. Thomas (2004). *High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome*. *Nature*, 430(7000):679–682.
- Denver, D. R., K. A. Clark, and M. J. Raboin (2011). *Reproductive mode evolution in nematodes: insights from molecular phylogenies and recently discovered species*. *Molecular Phylogenetics And Evolution*, 61(2):584–592.
- Denver, D. R., P. C. Dolan, L. J. Wilhelm, W. Sung, J. I. Lucas-Lledó, et al. (2009). *A genome-wide view of Caenorhabditis elegans base-substitution mutation processes*. *Proceedings of the National Academy of Sciences*, 106(38):16310–16314.
- Deppe, U., E. Schierenberg, T. Cole, C. Krieg, D. Schmitt, et al. (1978). *Cell lineages of the embryo of the nematode Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, 75(1):376–380.
- Dey, A., C. K. W. Chan, C. G. Thomas, and A. D. Cutter (2013). *Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri*. *Proceedings of the National Academy of Sciences*, 110(27):11056–11060.

- Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, et al. (2008). *Broad phylogenomic sampling improves resolution of the animal tree of life*. *Nature*, 452(7188).
- Edgecombe, G. D., G. Giribet, C. W. Dunn, A. Hejnol, R. M. Kristensen, et al. (2011). *Higher-level metazoan relationships: recent progress and remaining questions*. *Organisms Diversity & Evolution*, 11(2):151–172.
- Eisenmann, D. M. (2005). *Wnt signaling*. WormBook, ed. The C. elegans Research Community, WormBook,.
- Elmer, K. R. and A. Meyer (2011). *Adaptation in the age of ecological genomics: insights from parallelism and convergence*. *Trends In Ecology & Evolution*, 26(6):298–306.
- Engelstadter, J. (2008). *Muller’s Ratchet and the Degeneration of Y Chromosomes: A Simulation Study*. *Genetics*, 180(2):957–967.
- Erwin, D. H., M. Laflamme, S. M. Tweedt, E. A. Sperling, D. Pisani, et al. (2011). *The Cambrian conundrum: early divergence and later ecological success in the early history of animals*. *Science*, 334(6059):1091–1097.
- Felsenstein, J. (1974). *The Evolutionary Advantage of Recombination*. *Genetics*, 78:737–757.
- Ferguson, E. L. (1996). *Conservation of dorsal-ventral patterning in arthropods and chordates*. *Current Opinion in Genetics & Development*, 6(4):424–431.
- Ferrarini, M., M. Moretto, J. A. Ward, N. Šurbanovski, V. Stevanović, et al. (2013). *An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome*. *BMC Genomics*, 14(1):670.
- Fire, A., S. Xu, M. Montgomery, S. Kostas, S. Driver, et al. (1998). *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans*. *Nature*, 391(6669):806–811.
- Flot, J.-F., B. Hespeels, X. Li, B. Noel, I. Arkhipova, et al. (2013). *Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga*. *Nature*, 500(7463):453–457.
- Fonseca, V. G., G. R. Carvalho, W. Sung, H. F. Johnson, D. M. Power, et al. (2010). *Second-generation environmental sequencing unmasks marine metazoan biodiversity*. *Nature Communications*, 1(7):1–8.
- Foth, B. J., I. J. Tsai, A. J. Reid, A. J. Bancroft, S. Nichol, et al. (2014). *Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction*. *Nature Genetics*, 46(7):693–700.

-
- Franco, H. L. and H. H.-C. Yao (2012). *Sex and hedgehog: roles of genes in the hedgehog signaling pathway in mammalian sexual differentiation*. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 20(1):247–258.
- Gerstein, M. B., J. Rozowsky, K.-K. Yan, D. Wang, C. Cheng, et al. (2014). *Comparative analysis of the transcriptome across distant species*. Nature, 512(7515):445–448.
- Godel, C., S. Kumar, G. Koutsovoulos, P. Ludin, D. Nilsson, et al. (2012). *The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets*. FASEB journal: official publication of the Federation of American Societies for Experimental Biology, 26(11):4650–4661.
- Goldstein, B., L. Frisse, and W. Thomas (1998). *Embryonic axis specification in nematodes: evolution of the first step in development*. Current Biology, 8(3):157–160.
- Gordo, I. and B. Charlesworth (2000). *On the speed of Muller’s ratchet*. Genetics, 156(4):2137–2140.
- Goszczyński, B. (2005). *Reevaluation of the Role of the med-1 and med-2 Genes in Specifying the *Caenorhabditis elegans* Endoderm*. Genetics, 171(2):545–555.
- Goyal, S., D. J. Balick, E. R. Jerison, R. A. Neher, B. I. Shraiman, et al. (2012). *Dynamic mutation-selection balance as an evolutionary attractor*. Genetics, 191(4):1309–1319.
- Haag-Liautard, C., M. Dorris, X. Maside, S. Macaskill, D. Halligan, et al. (2007). *Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila**. Nature, 445(7123):82–85.
- Halanych, K. M., J. D. Bacheller, A. M. Aguinaldo, S. M. Liva, D. M. Hillis, et al. (1995). *Evidence from 18S ribosomal DNA that the lophophorates are protostome animals*. Science, 267(5204):1641–1643.
- Hansen, E. L., E. J. Buecher, and E. A. Yarwood (1973). *Alteration of sex of *Aphelenchus avenae* in culture*. Nematologica, 19(1):113–116a.
- Heger, P., M. Kroiher, N. Ndifon, and E. Schierenberg (2010). *Conservation of MAP kinase activity and MSP genes in parthenogenetic nematodes*. Developmental Biology, 10(1):51.
- Hespeels, B., M. Knapen, D. Hanot-Mambres, A. C. Heuskin, F. Pineux, et al. (2014). *Gateway to genetic exchange? DNA double-strand breaks in the bdelloid rotifer *Adineta vaga* submitted to desiccation*. Journal of Evolutionary Biology, 27(7):1334–1345.

- Hirth, F., L. Kammermeier, E. Frei, U. Walldorf, M. Noll, et al. (2003). *An urbilaterian origin of the tripartite brain: developmental genetic insights from Drosophila*. *Development*, 130(11):2365–2373.
- Holterman, M., O. Holovachov, S. van den Elsen, H. van Megen, T. Bongers, et al. (2008). *Small subunit ribosomal DNA-based phylogeny of basal Chromadoria (Nematoda) suggests that transitions from marine to terrestrial habitats (and vice versa) require relatively simple adaptations*. *Molecular Phylogenetics And Evolution*, 48(2):758–763.
- Holterman, M., A. van der Wurff, S. van den Elsen, H. van Megen, T. Bongers, et al. (2006). *Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades*. *Molecular Biology And Evolution*, 23(9):1792–1800.
- Hudry, B., M. Thomas-Chollier, Y. Volovik, M. Duffraisse, A. Dard, et al. (2014). *Molecular insights into the origin of the Hox-TALE patterning system*. *eLife*, 3:e01939–e01939.
- Hunt, P. (2011). *OrthoList: a compendium of C. elegans genes with human orthologs*. *PLoS ONE*, 6(5):e20085.
- Jacob, F. (1977). *Evolution and tinkering*. *Science*, 196(4295):1161–1166.
- Janssen, R., M. Le Gouar, M. Pechmann, F. Poulin, R. Bolognesi, et al. (2010). *Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation*. *BMC Evolutionary Biology*, 10(1):374.
- Jex, A. R., P. Nejsum, E. M. Schwarz, L. Hu, N. D. Young, et al. (2014). *Genome and transcriptome of the porcine whipworm Trichuris suis*. *Nature Publishing Group*, 46(7):701–706.
- Kearney, M. (2005). *Hybridization, glaciation and geographical parthenogenesis*. *Trends In Ecology & Evolution*, 20(9):495–502.
- Keightley, P. D., U. Trivedi, M. Thomson, F. Oliver, S. Kumar, et al. (2009). *Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines*. *Genome Research*, 19(7):1195–1201.
- Kikuchi, T., J. A. Cotton, J. J. Dalzell, K. Hasegawa, N. Kanzaki, et al. (2011). *Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen Bursaphelenchus xylophilus*. *PLoS pathogens*, 7(9).
- Kimble, J. and D. Hirsh (1979). *The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans*. *Developmental Biology*, 70(2):396–417.

-
- Kimmel, C. B., W. W. Ballard, S. R. Kimmel, B. Ullmann, and T. F. Schilling (1995). *Stages of embryonic development of the zebrafish*. *Developmental Dynamics*, 203(3):253–310.
- Kiontke, K., A. Barriere, I. Kolotuev, B. Podbilewicz, R. Sommer, et al. (2007). *Trends, stasis, and drift in the evolution of nematode vulva development*. *Current biology : CB*, 17(22):1925–1937.
- Kiontke, K., N. Gavin, Y. Raynes, C. Roehrig, F. Piano, et al. (2004). *Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss*. *Proceedings of the National Academy of Sciences*, 101(24):9003.
- Kiontke, K. C., M.-A. Félix, M. Ailion, M. V. Rockman, C. Braendle, et al. (2011). *A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits*. *BMC Evolutionary Biology*, 11(1):339.
- Kipreos, E. T. and M. Pagano (2000). *The F-box protein family*. *Genome Biology*, 1(5):reviews3002.1–3002.7.
- Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, et al. (2012). *VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing*. *Genome Research*, 22(3):568–576.
- Koonin, E. V. (2005). *Orthologs, paralog, and evolutionary genomics*. *Annual Review of Genetics*, 39:309–338.
- Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter (2013). *Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots*. *Frontiers in Genetics*, 4:237.
- Kumar, S., P. H. Schiffer, and M. Blaxter (2011). *959 Nematode Genomes: a semantic wiki for coordinating sequencing projects*. *Nucleic Acids Research*, 40:D1295–1300.
- Lambshhead, P. J. D. and G. Boucher (2003). *Marine nematode deep-sea biodiversity – hyperdiverse or hype?* *Journal of Biogeography*, 30(4):475–485.
- Lee, D. L., editor (2002). *The Biology of Nematodes*. Taylor & Francis Press, 1st. edition.
- Levin, M., T. Hashimshony, F. Wagner, and I. Yanai (2012). *Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo*. *Developmental Cell*, 22(5):1101–1108.
- Lewis, S., L. Dyal, C. Hilburn, S. Weitz, W. Liao, et al. (2009). *Molecular evolution in Panagrolaimus nematodes: origins of parthenogenesis, hermaphroditism and the Antarctic species P. davidi*. *BMC Evolutionary Biology*, 9(1):15.

- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al. (2009). *The Sequence Alignment/Map format and SAMtools*. *Bioinformatics*, 25(16):2078–2079.
- Li, J. J., H. Huang, P. J. Bickel, and S. E. Brenner (2014). *Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data*. *Genome Research*, 24(7):1086–1101.
- Linard, B., J. D. Thompson, O. Poch, and O. Lecompte (2011). *OrthoInspector: comprehensive orthology analysis and visual exploration*. *BMC Bioinformatics*, 12(1):11.
- Lucas-Lledo, J. I. and M. Lynch (2009). *Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family*. *Molecular Biology And Evolution*, 26(5):1143–1153.
- Lunt, D. H., S. Kumar, G. Koutsovoulos, and M. L. Blaxter (2014). *The complex hybrid origins of the root knot nematodes revealed through comparative genomics*. *PeerJ*, 2(8):e356.
- Lynch, J. A. and S. Roth (2011). *The evolution of dorsal-ventral patterning mechanisms in insects*. *Genes & Development*, 25(2):107–118.
- Lynch, M. (1985). *Spontaneous mutations for life-history characters in an obligate parthenogen*. *Evolution*, pages 804–818.
- Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates Incorporated, 1st. edition.
- Lynch, M., R. Bürger, D. Butcher, and W. Gabriel (1993). *The mutational meltdown in asexual populations*. *Journal Of Heredity*, 84(5):339–344.
- Maduro, M. F. (2006). *Endomesoderm specification in *Caenorhabditis elegans* and other nematodes*. *BioEssays*, 28(10):1010–1022.
- Maheshwari, S. and D. A. Barbash (2011). *The genetics of hybrid incompatibilities*. *Annual Review of Genetics*, 45(1):331–355.
- Mark Welch, D. B., M. P. Cummings, D. M. Hillis, and M. Meselson (2004)a. *Divergent gene copies in the asexual class *Bdelloidea* (*Rotifera*) separated before the bdelloid radiation or within bdelloid families*. *Proceedings of the National Academy of Sciences*, 101(6):1622–1625.
- Mark Welch, J. L., D. B. Mark Welch, and M. Meselson (2004)b. *Cytogenetic evidence for asexual evolution of bdelloid rotifers*. *Proceedings of the National Academy of Sciences*, 101(6):1618–1621.

-
- Marlow, H. (2013). *Ectopic activation of the canonical wnt signaling pathway affects ectodermal patterning along the primary axis during larval development in the anthozoan *Nematostella vectensis**. *Developmental Biology*, 380(2):324–334.
- Marques-Souza, H., M. Aranda, and D. Tautz (2008). *Delimiting the conserved features of hunchback function for the trunk organization of insects*. *Development*, 135(5):881–888.
- Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, et al. (2013). *Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies*. *Genome Research*, 23(11):1817–1828.
- Martindale, M. Q. (2005). *The evolution of metazoan axial properties*. *Nature Reviews Genetics*, 6(12):917–927.
- Maynard Smith, J. (1978). *The Evolution of Sex*. Cambridge University Press, 1st. edition.
- McDaniel, L. D., E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, et al. (2010). *High frequency of horizontal gene transfer in the oceans*. *Science*, 330(6000):50.
- McGaughran, A., K. Morgan, and R. J. Sommer (2013). *Unraveling the evolutionary history of the nematode *Pristionchus pacificus*: from lineage diversification to island colonization*. *Ecology and Evolution*, 3(3):667–675.
- McGhee, G. R. (2011). *Convergent Evolution. Limited Forms Most Beautiful*. MIT Press, 1st. edition.
- McGhee, J. D., T. Fukushige, M. W. Krause, S. E. Minnema, B. Goszczynski, et al. (2009). *ELT-2 is the predominant transcription factor controlling differentiation and function of the *C. elegans* intestine, from embryo to adult*. *Developmental Biology*, 327(2):551–565.
- McGregor, A. P., M. Pechmann, E. E. Schwager, and W. G. Damen (2009). *An ancestral regulatory network for posterior development in arthropods*. *Communicative & integrative biology*, 2(2):174–176.
- McInerney, J. O., D. Pisani, E. Baptiste, and M. J. O’Connell (2011). *The public goods hypothesis for the evolution of life on Earth*. *Biology Direct*, 6(1):41.
- Miller, J. R., S. Koren, and G. Sutton (2010). *Assembly algorithms for next-generation sequencing data*. *Genomics*, 95(6):315–327.
- Misevic, D., R. E. Lenski, and C. Ofria (2004). *Sexual reproduction and muller’s ratchet in digital organisms*. In *Ninth International Conference on Artificial Life, (Boston MA)*, pages 340–345. MIT Press.

- Mitreva, M., D. P. Jasmer, D. S. Zarlenga, Z. Wang, S. Abubucker, et al. (2011). *The draft genome of the parasitic nematode Trichinella spiralis*. *Nature Genetics*, 43(3):228–235.
- Moroz, L. L., K. M. Kocot, M. R. Citarella, S. Dosung, T. P. Norekian, et al. (2014). *The ctenophore genome and the evolutionary origins of neural systems*. *Nature*, 510(7503):109–114.
- Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. *Nature Methods*, 5(7):621–628.
- Müller, H. (1903). *Beitrag zur Embryonalentwicklung von Ascaris megalcephala*. *Zoologica*, 41.
- Muller, H. J. (1964). *The relation of recombination to mutational advance*. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9.
- Muschiol, D. and W. Traunspurger (2007). *Life cycle and calculation of the intrinsic rate of natural increase of two bacterivorous nematodes, Panagrolaimus sp. and Poikilolaimus sp. from chemoautotrophic Movile Cave, Romania*. *Nematology*, 9(2):271–284.
- Naggan Perl, T., B. G. M. Schmid, J. Schwirz, and A. D. Chipman (2013). *The evolution of the knirps family of transcription factors in arthropods*. *Molecular Biology And Evolution*, 30(6):1348–1357.
- Nei, M. (2013). *Mutation-Driven Evolution*. Oxford University Press, 1st. edition.
- Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, 1st. edition.
- Omilian, A. R., M. E. A. Cristescu, J. L. Dudycha, and M. Lynch (2006). *Ameiotic recombination in asexual lineages of Daphnia*. *Proceedings of the National Academy of Sciences*, 103(49):18638–18643.
- Opperman, C. H., D. M. Bird, V. M. Williamson, D. S. Rokhsar, M. Burke, et al. (2008). *Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism*. *Proceedings of the National Academy of Sciences*, 105(39):14802–14807.
- Otto, S. P. and T. Lenormand (2002). *Resolving the paradox of sex and recombination*. *Nature Reviews Genetics*, 3(4):252–261.
- Pennisi, E. (2013). *The man who bottled evolution*. *Science*, 342(6160):790–793.
- Peterson, K. J., R. A. Cameron, and E. H. Davidson (1997). *Set-aside cells in maximal indirect development: evolutionary and developmental significance*. *BioEssays*, 19(7):623–631.

-
- Peterson, K. J. and E. H. Davidson (2000). *Regulatory evolution and the origin of the bilaterians*. Proceedings of the National Academy of Sciences, 97(9):4430–4433.
- Pevzner, P. (2001). *Fragment assembly with double-barreled data*. Bioinformatics, 17 Suppl 1:S225–233.
- Pevzner, P., H. Tang, and M. Waterman (2001). *An Eulerian path approach to DNA fragment assembly*. Proceedings of the National Academy of Sciences, 98(17):9748.
- Piccolo, S., Y. Sasai, B. Lu, and E. M. De Robertis (1996). *Dorsoventral patterning in Xenopus: inhibition of ventral signals by direct binding of chordin to BMP-4*. Cell, 86(4):589–598.
- Pires-daSilva, A. and R. J. Sommer (2003). *The evolution of signalling pathways in animal development*. Nature Reviews Genetics, 4(1):39–49.
- Pisani, D., L. L. Poling, M. Lyons-Weiler, and S. B. Hedges (2004). *The colonization of land by animals: molecular phylogeny and divergence times among arthropods*. BMC Biology, 2:1.
- Poinar, G., H. Kerp, and H. Hass (2008). *Palaeonema phyticum gen. n., sp. n. (Nematoda: Palaeonematidae fam. n.), a Devonian nematode associated with early land plants*. Nematology, 10(1):9–14.
- Reuner, A., S. Hengherr, B. Mali, F. Förster, D. Arndt, et al. (2010). *Stress response in tardigrades: differential gene expression of molecular chaperones*. Cell Stress and Chaperones, 15(4):423–430.
- Robertson, G. P. and D. W. Freckman (1995). *The Spatial Distribution of Nematode Trophic Groups Across a Cultivated Ecosystem*. Ecology, 76(5):1425.
- Rota-Stabelli, O., A. C. Daley, and D. Pisani (2013). *Molecular Timetrees Reveal a Cambrian Colonization of Land and a New Scenario for Ecdysozoan Evolution*. Current biology : CB, 23(5):392–398.
- Ryan, J. F., K. Pang, C. E. Schnitzler, A. D. Nguyen, R. T. Moreland, et al. (2013). *The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution*. Science, 342(6164):1242592–1242592.
- Saxer, G., P. Havlak, S. A. Fox, M. A. Quance, S. Gupta, et al. (2012). *Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba Dictyostelium discoideum*. PLoS ONE, 7(10):e46759.

- Schaefer, I., K. Domes, M. Heethoff, K. Schneider, I. Schon, et al. (2006). *No evidence for the 'Meselson effect' in parthenogenetic oribatid mites (Oribatida, Acari)*. *Journal of Evolutionary Biology*, 19(1):184–193.
- Schatz, M., A. Delcher, and S. Salzberg (2010). *Assembly of large genomes using second-generation sequencing*. *Genome Research*, 20(9):1165–1173.
- Schatz, M. C., J. Witkowski, and W. R. McCombie (2012). *Current challenges in de novo plant genome sequencing and assembly*. *Genome Biology*, 13(4):243.
- Schierenberg, E. and J. Schulze (2008). *Many roads lead to Rome: different ways to construct a nematode*, chapter 14, pages 261–280. *Key Themes in Evolutionary Developmental Biology*. Cambridge University Press.
- Schmidt, J., V. Francois, E. Bier, and D. Kimelman (1995). *Drosophila short gastrulation induces an ectopic axis in Xenopus: evidence for conserved mechanisms of dorsal-ventral patterning*. *Development*, 121(12):4319–4328.
- Schon, I. and K. Martens (1998). *DNA repair in ancient asexuals - a new solution to an old problem?* *Journal of Natural History*, 32(7):943–948.
- Schön, I., K. Martens, P. J. Dijk, and P. van Dijk, editors (2009). *Lost Sex: The Evolutionary Biology of Parthenogenesis*. Springer Science, 1st. edition.
- Schon, I., K. Martens, K. van Doninck, and R. Butlin (2003). *Evolution in the slow lane: molecular rates of evolution in sexual and asexual ostracods (Crustacea : Ostracoda)*. *Biological Journal Of The Linnean Society*, 79(1):93–100.
- Schön, I., R. L. Pinto, S. Halse, A. J. Smith, K. Martens, et al. (2012). *Cryptic Species in Putative Ancient Asexual Darwinulids (Crustacea, Ostracoda)*. *PLoS ONE*, 7(7):e39844.
- Schulze, J., W. Houthoofd, J. Uenk, S. Vangestel, and E. Schierenberg (2012). *Plectus - a stepping stone in embryonic cell lineage evolution of nematodes*. *EvoDevo*, 3(1):13.
- Schulze, J. and E. Schierenberg (2008). *Cellular pattern formation, establishment of polarity and segregation of colored cytoplasm in embryos of the nematode Romanomermis culicivorax*. *Developmental Biology*, 315(2):426–436.
- Schulze, J. and E. Schierenberg (2009). *Embryogenesis of Romanomermis culicivorax: an alternative way to construct a nematode*. *Developmental Biology*, 334(1):10–21.
- Schulze, J. and E. Schierenberg (2011). *Evolution of embryonic development in nematodes*. *EvoDevo*, 2(1):18.

-
- Schwaiger, M., A. Schonauer, A. F. Rendeiro, C. Pribitzer, A. Schauer, et al. (2014). *Evolutionary conservation of the eumetazoan gene regulatory landscape*. *Genome Research*, 24(4):639–650.
- Schwander, T., S. Vuilleumier, J. Dubman, and B. J. Crespi (2010). *Positive feedback in the transition from sexual reproduction to parthenogenesis*. *Proceedings of the Royal Society B: Biological Sciences*, 277(1686):1435–1442.
- Schwarz, E. M., P. K. Korhonen, B. E. Campbell, N. D. Young, A. R. Jex, et al. (2013). *The genome and developmental transcriptome of the strongylid nematode Haemonchus contortus*. *Genome Biology*, 14(8):R89.
- Shannon, A., J. Browne, J. Boyd, D. Fitzpatrick, and A. Burnell (2005). *The anhydrobiotic potential and molecular phylogenetics of species and strains of Panagrolaimus (Nematoda, Panagrolaimidae)*. *Journal Of Experimental Biology*, 208(12):2433–2445.
- Sharp, N. P. and A. F. Agrawal (2012). *Evidence for elevated mutation rates in low-quality genotypes*. *Proceedings of the National Academy of Sciences*, 109(16):6142–6146.
- Simon, J., F. Delmotte, C. Rispe, and T. Crease (2003). *Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals*. *Biological Journal Of The Linnean Society*, 79(1):151–163.
- Simpson, J. T. (2014). *Exploring Genome Characteristics and Sequence Quality Without a Reference*. *Bioinformatics*.
- Siomi, M. C., K. Sato, D. Pezic, and A. A. Aravin (2011). *PIWI-interacting small RNAs: the vanguard of genome defence*. *Nature Reviews Molecular Cell Biology*, 12(4):246–258.
- Siracusa, G., D. G. Whittingham, M. Molinaro, and E. Vivarelli (1978). *Parthenogenetic activation of mouse oocytes induced by inhibitors of protein synthesis*. *Journal of embryology and experimental morphology*, 43:157–166.
- Skiba, F. and E. Schierenberg (1992). *Cell lineages, developmental timing, and spatial pattern formation in embryos of free-living soil nematodes*. *Developmental Biology*, 151(2):597–610.
- Sommer, R. (2006). *Pristionchus pacificus*. *WormBook*, ed. The C. elegans Research Community, <http://www.wormbook.org>.
- Sommer, R. J. (2009). *The future of evo-devo: model systems and evolutionary theory*. *Nature Reviews Genetics*, 10(6):416–422.

- Sommer, R. J., A. Eizinger, K. Z. Lee, B. Jungblut, A. Bubeck, et al. (1998). *The Pristionchus HOX gene Ppa-lin-39 inhibits programmed cell death to specify the vulva equivalence group and is not required during vulval induction*. *Development*, 125(19):3865–3873.
- Srinivasan, J., A. R. Dillman, M. G. Macchietto, L. Heikkinen, M. Lakso, et al. (2013). *The Draft Genome and Transcriptome of Panagrellus redivivus are Shaped by the Harsh Demands of a Free-Living Lifestyle*. *Genetics*, 193(4):1279–1295.
- Stanke, M. and S. Waack (2003). *Gene prediction with a hidden Markov model and a new intron submodel*. *Bioinformatics*, 19(Suppl 2):ii215–ii225.
- Stothard, P. and D. Pilgrim (2003). *Sex-determination gene and pathway evolution in nematodes*. *BioEssays*, 25(3):221–231.
- Sulston, J. E. and H. R. Horvitz (1977). *Post-embryonic cell lineages of the nematode, Caenorhabditis elegans*. *Developmental Biology*, 56(1):110–156.
- Sulston, J. E., E. Schierenberg, J. G. White, and J. N. Thomson (1983). *The embryonic cell lineage of the nematode Caenorhabditis elegans*. *Developmental Biology*, 100(1):64–119.
- Tautz, D. (2004). *Segmentation*. *Developmental Cell*, 7(3):301–312.
- Telford, M. J. and D. T. J. Littlewood, editors (2009). *Animal Evolution. Genomes, Fossils, and Trees*. Oxford University Press, 1st. edition.
- Thorne, M. A. S., H. Kagoshima, M. S. Clark, C. J. Marshall, and D. A. Wharton (2014). *Molecular analysis of the cold tolerant Antarctic nematode, Panagrolaimus davidi*. *PLoS ONE*, 9(8):e104526.
- Tian, H., B. Schlager, H. Xiao, and R. J. Sommer (2008). *Wnt signaling induces vulva development in the nematode Pristionchus pacificus*. *Current Biology*, 18(2):142–146.
- Timmons, L. and A. Fire (1998). *Specific interference by ingested dsRNA*. *Nature*, 395(6705):854–854.
- Ting, J. J., G. C. Woodruff, G. Leung, N.-R. Shin, A. D. Cutter, et al. (2014). *Intense Sperm-Mediated Sexual Conflict Promotes Reproductive Isolation in Caenorhabditis Nematodes*. *PLoS Biology*, 12(7):e1001915.
- Todo, T. (1999). *Functional diversity of the DNA photolyase/blue light receptor family*. *Mutation Research*, 434(2):89–97.
- Triantaphyllou, A. C. (1966). *Polyploidy and reproductive patterns in the root-knot nematode Meloidogyne hapla*. *Journal of Morphology*, 118(3):403–413.

-
- True, J. R. and E. S. Haag (2001). *Developmental system drift and flexibility in evolutionary trajectories*. *Evolution and Development*, 3(2):109–119.
- Valentine, J. W. (1986). *Fossil Record of the Origin of Baupläne and Its Implications*. In D. M. Raup and D. Jablonski, editors, *Patterns and Processes in the History of Life*, pages 209–222. Springer.
- van Megen, H., S. van den Elsen, M. Holterman, G. Karssen, P. Mooyman, et al. (2009). *A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences*. *Nematology*, 11(6):927–950.
- Vangestel, S., W. Houthoofd, W. Bert, and G. Borgonie (2008). *The early embryonic development of the satellite organism *Pristionchus pacificus*: differences and similarities with *Caenorhabditis elegans**. *Nematology*, 10:301–312.
- Vassilieva, L., A. Hook, and M. Lynch (2000). *The fitness effects of spontaneous mutations in *Caenorhabditis elegans**. *Evolution*, 54(4):1234–1246.
- Vassilieva, L. and M. Lynch (1999). *The rate of spontaneous mutation for life-history traits in *Caenorhabditis elegans**. *Genetics*, 151(1):119–129.
- Venter, J. C. (2003). *A part of the human genome sequence*. *Science*, 299(5610):1183–1184.
- Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, et al. (2001). *The sequence of the human genome*. *Science*, 291(5507):1304–1351.
- Warthmann, N., R. Clark, R. Shaw, and D. Weigel (2010). *The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana**. *Science*, 327(5961):92–94.
- Welch, D. and M. Meselson (2000). *Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange*. *Science*, 288(5469):1211–1215.
- Welch, D. and M. Meselson (2001). *Rates of nucleotide substitution in sexual and anciently asexual rotifers*. *Proceedings of the National Academy of Sciences*, 98(12):6720–6724.
- Werren, J. H. and D. W. Loehlin (2009). *The Parasitoid Wasp *Nasonia*: An Emerging Model System With Haploid Male Genetics*. Cold Spring Harbor Protocols, 2009(10):pdb.emo134.
- Wharton, D. and D. Ferns (1995). *Survival of intracellular freezing by the Antarctic nematode *Panagrolaimus davidi**. *Journal Of Experimental Biology*, 198(Pt 6):1381–1387.
- Wiegner, O. and E. Schierenberg (1998). *Specification of gut cell fate differs significantly between the nematodes *Acroboloides nanus* and *Caenorhabditis elegans**. *Developmental Biology*, 204(1):3–14.

- Wiegner, O. and E. Schierenberg (1999). *Regulative development in a nematode embryo: a hierarchy of cell fate transformations*. *Developmental Biology*, 215(1):1–12.
- Woese, C. R., O. Kandler, and M. L. Wheelis (1990). *Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya*. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.
- Wu, T. D. and S. Nacu (2010). *Fast and SNP-tolerant detection of complex variants and splicing in short reads*. *Bioinformatics*, 26(7):873–881.
- Zerbino, D. R. (2009). *Genome assembly and comparison using de Bruijn graphs*. Ph.D. thesis, University of Cambridge, Darwin College.

Zusammenfassung in deutscher Sprache

Das Phylum Nematoda wird durch seine immense Artenvielfalt charakterisiert. Nematoden bewohnen nahezu alle Habitate der Erde. Trotz dieses Vorkommens in einer großen Zahl von Ökosystemen zeichnen sich Nematoden durch einen sehr strikten Bauplan aus, der zwischen zwei Großgruppen innerhalb des Phylums in den letzten 400 Millionen Jahren nur kleinsten Änderungen unterworfen war. Diese Konservierung der adulten Körperform ist unter Tieren etwas Besonderes. Diese außergewöhnliche Konservierung des Bauplans hat im Verbund mit einem in den Modellorganismen *C. elegans*, *P. pacificus* und *Ascaris* sehr ähnlichen Muster in der frühen Entwicklung zu der Annahme geführt, dass dieses Muster der Frühentwicklung beispielhaft für das ganze Phylum sei. Angeführt durch Arbeiten im Schierenberglabor in den letzten 25 Jahren ist allerdings klar geworden, dass es bedeutende Unterschiede in der Frühentwicklung verschiedener Nematoden aus unterschiedlichen Ästen des Phylums gibt. Diese Befunde haben zusammen mit Daten, die in unterschiedlichen Arten der Panarthropoden erhoben wurden, die Frage aufgeworfen, ob das molekulare „Toolkit“ zur Verwirklichung dieser unterschiedlichen Entwicklung einer ebensolchen Variabilität unterworfen ist. In der hier vorgelegten Arbeit wird diese Frage aus einer genomischen Perspektive durch das Kompilieren und Auswerten sehr großer Datenmengen bearbeitet. Diese Daten erlauben einen Vergleich von Arten, die phylogenetisch gesehen an entgegengesetzten Enden des Phylums zu finden sind und somit durch hunderte Millionen Jahre der Evolution voneinander getrennt sind. Es wurden aber auch Untersuchungen auf hierarchisch tiefer liegenden Ebenen der Verwandtschaft, zwischen Arten in einer bestimmten Großgruppe innerhalb der Nematoden und zwischen einzelnen Gattungen durchgeführt. In all diesen Taxa wurden genregulatorische Netzwerke (GRNs) der Frühentwicklung analysiert und mit dem „Entwicklungstoolkit“ von *C. elegans* verglichen. Ich nutzte diesen Ansatz, um aktuelle Theorien zur Evolution von genregulatorischen Netzwerken zu prüfen. Hieraus ergibt sich das Bild, dass Gene, welche als Zwischenstufen in diesen Netzwerken fungieren, häufig durch den Prozess des „Developmental System Drift“ (DSD) ausgetauscht werden, während hierarchisch höher und tiefer liegende Gene erhalten bleiben. Dies steht im Einklang mit den aktuellen Theorien zur Evolution der GRNs. Trotz dieser Divergenz innerhalb der Nematoden zeigte eine Analyse der extrem diversen Bilateria, zu denen Nematoden gehören, auch einen Entwicklungsprozess auf, der über große evolutionäre Distanzen konserviert ist. Dies ist jene Phase der Entwicklung, in der die adulte Körperform entsteht.

Parthenogenese evolvierte in mehreren Gattungen in den Nematoden, mit einem potentiellen „Hotspot“ in der „Clade IV“ des Phylums. Ich nutzte die Daten, die ich zur Bearbeitung der oben geschilderten Fragen erhob, um den Ursprung und die zugrundeliegenden molekularen Mechanismen von Parthenogenese in der Gattung *Panagrolaimus* zu erforschen. Während es durch die Veränderungen in den GRNs durch das DSD derzeit noch nicht möglich war, molekulare Mechanismen genau aufzuklären, konnten Belege dafür gefunden werden, dass

parthenogenetische Arten in *Panagrolaimus* polyploide Hybride sind. Dieser Befund unterstützt die Annahme, dass Hybridisierung auch in Nematoden vielfach zur Entstehung von parthenogenetischen Arten führt. Parthenogenetische Arten sind oftmals in neuen und unvorteilhaften Habitaten zu finden. Interessanterweise sind *Panagrolaimus*-Arten, im Unterschied zu *C. elegans* und den meisten anderen Nematoden, in der Lage, komplett auszutrocknen und wieder zum Leben zu erwachen - ein Prozess, der Anhydrobiose genannt wird. Wir untersuchten den molekularen Hintergrund von Anhydrobiose in *Panagrolaimus* und fanden einige Gene, die bereits zuvor mit diesem Prozess in Verbindung gebracht wurden. Interessanterweise fanden wir auch eine direkte Verknüpfung zu Genen, die durch horizontalen Gentransfer (HGT) gewonnen wurden. Diese können *Panagrolaimus*-Arten während der Rehydrierungsphase ihrer DNA reparieren und dadurch einen adaptiven Vorteil liefern. Dies zeigt, wie wichtig der Prozess des HGT auch in Tieren sein kann, ein Aspekt, der bisher in wissenschaftlichen Untersuchungen vernachlässigt wurde. Die genomischen und transkriptomischen Daten, die im Verlauf dieser Doktorarbeit gewonnen und analysiert wurden, können zukünftig als Basis für Projekte dienen, welche Entwicklungssysteme im Hinblick auf GRNs und DSD untersuchen, sowie Anhydrobiose und den molekularbiologischen Hintergrund von Parthenogenese analysieren.

Beteiligung an den angeführten Publikationen

1. Zur Publikation (**P1**) mit dem Titel “959 Nematode Genomes: a semantic wiki for coordinating sequencing projects” wurden von mir hauptsächlich Projektideen beigebracht. Ich stellte das Projekt auf mehreren Konferenzen vor und war am Verfassen der Publikation beteiligt.
2. Die Publikation (**P2**) mit dem Titel “The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda” wurde in weiten Teilen von mir selbst verfasst. Ich führte die meisten der vorgestellten Analysen durch und war an den restlichen direkt beteiligt. Weiterhin leistete ich große Beiträge zum Design der Experimente und der Planung des Projekts. Ich leitete und organisierte die internationale Kollaboration, die zum Erstellen dieser Publikation nötig war.
3. Die Publikation (**P3**) mit dem Titel “Developmental variations among Panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit” wurde in weiten Teilen von mir verfasst. Die Publikation wurde von mir geplant und die bioinformatischen Analysen wurden von mir durchgeführt. Ich betreute eine Studentin, welche Daten für die Publikation erhob und analysierte.
4. Das im Manuskript **M1** dargestellte Projekt mit dem Arbeitstitel “How to survive the extreme: a multi genome analysis reveals evolutionary traits of cryptobiosis and routes to parthenogenesis” wurde zu weiten Teilen von mir erdacht und durchgeführt. Von den bereits in dieser Arbeit abgedruckten Teilen des Manuskripts wurde nur der Methoden- und Ergebnisteil über Horizontal Gene Transfer sowie der Teil der Einleitung, der sich mit Anhydrobiose befasst, nicht von mir verfasst. Jedoch habe ich auch diese Teile editiert. Die Analysen zum Horizontal Gene Transfer wurden von Etienne Danchin durchgeführt und die Analyse des Anhydrobioseteils wurde teilweise von Ann Burnell durchgeführt. Simon Wong half bei der Genvorhersage in den beiden Transkriptomen und dem Interproscan. Ich leitete und organisierte die internationale Kollaboration, die zum Erstellen dieser Publikation nötig war.
5. Zu dem Manuskript **M2** mit dem Titel “Genes from the Cambrian Explosion” trug ich auf mehreren Ebenen bei. Ich war maßgeblich an der Ausarbeitung der Methodik und der Fragestellungen beteiligt. Ich führte einige der gemachten Analysen durch und entwickelte Scripte hierfür. Die im Manuskript dargelegten Interpretationen sind größtenteils von mir. Ich schrieb den Abstract, die Einleitung und weite Teile der Diskussion und war an der Aus- und Überarbeitung sämtlicher anderer Teile ebenso beteiligt.

6. Für die im Manuskript (**M3**) mit dem Titel “Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages” gemachten Analysen stellte ich von mir errechnete Daten bereit. Weiterhin half ich bei der Erstellung der Endfassung des Manuskripts und diskutierte mit dem Hauptautor, Peter Sarkies, die von den Gutachtern gemachten Kommentare und Änderungsvorschläge.

Erklärung und Lebenslauf

Erklärung (entsprechend §4 Abs. 1 Nr. 9 der Promotionsordnung vom 02. Februar 2006, mit Änderungen vom 10. Mai 2012)

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie abgesehen von unten angegebenen Teilpublikationen noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Einhard Schierenberg betreut worden.

Datum

Unterschrift

Veröffentlichte Teilpublikationen

1. Kumar, S., Schiffer, P. H., Blaxter, M., (2012), *959 Nematode Genomes: a semantic wiki for coordinating sequencing projects*, Nucleic Acid Research.
2. Schiffer et al., (2013). *The genome of Romanomermis culicivorax: revealing fundamental changes in the core developmental genetic toolkit in Nematoda*, BMC Genomics.
3. Schiffer et al., (2014), *Developmental variations among Panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit*, Genes, Development and Evolution.
4. Die Manuskripte **M1**, **M2** und **M3** (s.o.) sind bereits zur Veröffentlichung eingereicht, bzw. werden in Kürze eingereicht.

ZU MEINER PERSON

Philipp Schiffer

Auf dem Bachacker 2

50354 Hürth

Tel.: 02233 - 398083

philipp.schiffer@gmx.de

geboren am 8. November 1980

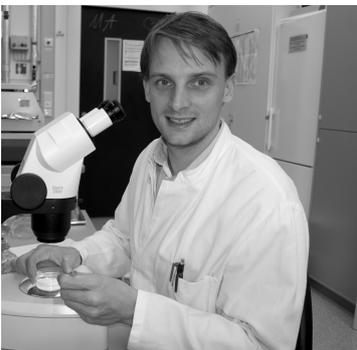
in Köln

Staatsangehörigkeit: deutsch

verheiratet, 2 Kinder

Ausbildungsgrad

Diplombiologe



Philipp Schiffer, Auf dem Bachacker 2, 50354 Hürth, Tel.: 02233 - 398083

LEBENS LAUF

DIPL OM

Januar 2009 Diplom in Biologie
Bewertung: sehr gut - mit Auszeichnung -
Diplomarbeit an der Universität zu Köln:
*„Cryptic speciation and Morphological Plasticity in the le-
padomorph barnacle Lepas anatifera and related species“*

STUDIUM

2001 – 2008 Diplomstudium in Biologie
an der Universität zu Köln
Hauptfach: Zoologie
Nebenfächer: Genetik, Paläontologie

WEHRDIENST

2000 – 2001 Wehrdienst beim Heer, Falkensteinkaserne Koblenz

SCHULBILDUNG

Juni 2000 Abiturprüfung
1991 – 2000 Albert Schweitzer Gymnasium Hürth

BERUFLICHE TÄTIGKEITEN

seit 2014/05 Wissenschaftlicher Mitarbeiter, Institut für Genetik,
Universität zu Köln (UzK)
seit 2014/04 Akad. Rat im Hochschuldienst, Zoologisches Institut,
UzK
Jan 2014 – März 2014 Wissenschaftliche Hilfskraft, Zoologisches Institut, UzK
Aug. 2013 – Dez. 2013 Wissenschaftlicher Mitarbeiter, Institut für Genetik,
UzK
Aug. 2009 – Juli 2013 Wissenschaftlicher Mitarbeiter, Zoologisches Institut,
UzK

Philipp Schiffer, Auf dem Bachacker 2, 50354 Hürth, Tel.: 02233 - 398083

März 2009 – Juli 2009 Studentische Hilfskraft, Institut für Geologie und Mineralogie, UzK
Feb. 2002 – Okt. 2006 Studentische Aushilfe, Bike & Outdoor Company, Hürth

BESONDERE KENNTNISSE

Sprachen Englisch: sehr gut in Schrift und Sprache
 Französisch: Grundkenntnisse
 Latein: Abschluss mit Latinum