The Nuanced Evolutionary Consequences of Duplicated Genes

Emily Anna Baker

St Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2023

The Nuanced Evolutionary Consequences of Duplicated Genes

Emily Anna Baker
St Peter's College

Submitted for the degree of D.Phil.
Hilary Term 2023

Abstract

Gene duplication is postulated to facilitate evolutionary change by providing new genetic material which can gain or lose functions over evolutionary time. There are many instances throughout metazoan evolution which implicate gene duplications as forerunners of innovation. Though often these exist in the literature as implications only. Remarkably, there are still many facets of paralogous gene divergence which are poorly understood. These facets form the basis of questions to be addressed in this thesis. How do new genes acquire new functions? What is the reason for the expansion of certain gene families in certain taxa? Why do some paralogous genes remain redundant for millions of years of evolution? Are paralogues destined to become less essential than the genes from which they were derived? *Caenorhabditis elegans* and its close relatives represent powerful and tractable models for answering these questions. With their unrivalled suite of genomic and molecular genetic resources, the *Caenorhabditis* genus will be the stage on which this quest to understand the evolutionary consequences of duplicated genes will play out. But *Caenorhabditis* nematodes are not the main characters in this story. Our characters, as it were, will be four very different gene families.

Here, the fates of duplicated genes are explored in light of the various complexities of paralogue diversification. While gene family size will provide an overall focus for this study, insights garnered from paralogue dynamics within families are expected to illuminate the mechanisms by which gene duplications can be a vehicle for evolutionary change. Factors such as the evolution of domain architecture following tandem duplication, as well as asymmetric gene diversification, overlapping functionality between paralogues, and the role of rapid paralogue divergence in major evolutionary events, such as the formation of new species, will all be explored.

In Chapter 3, the taxon-specific dynamism of the T-box family of transcription factors in the *Caenorhabditis* genus is investigated, with particular focus on the functional evolution of the *tbx-35* and *tbx-36* gene pair only found in *C. elegans*. Despite the accumulation of severe deleterious mutations in gene pair among various wild populations, it is shown that the *tbx-35* and *tbx-36* have each acquired very different roles in the early embryo. Chapter 4 characterises the pervasiveness of redundancy in a slightly smaller, more conserved, gene family. The redundancy relationships in the Warthog family in *C. elegans* are reconciled with the evolutionary fates members therein have adopted, following which it is seen that particular clades of the family specialise in particular aspects of postembryonic development. Chapter 5 focusses on the consequences (and underpinnings) of asymmetric paralogue divergence for gene family dynamics in the Drd family of oxidoreductases, finding that asymmetric paralogue divergence can be underpinned by the loss of biological function, as well as the gain. Lastly, in Chapter 6, changes to domain architecture following gene duplication — and the associated developmental consequences thereof — are explored through the Myrf family of TFs. It is found that subfunctionalisation, while shown elsewhere in this thesis to occur via regulatory change, can also occur via the complementary degeneration of domain architecture. Taken together, these results characterise hitherto poorly understood aspects of duplicated gene evolution.

But this is not to say that this thesis is a collection of just-so stories to try and persuade you of the importance of gene duplication to complex, multifaceted evolutionary processes. It is more than that. Rather, it will be argued that only with the kind of careful genetic approach that is employed in this thesis can we truly understand not only the fates that duplicated genes adopt, but meaningfully comprehend the roles they play in evolution. Together, these findings will provide a new framework for furthering our understanding of the very many evolutionary consequences of duplicated genes.

Thesis supervisor: Professor Alison Woollard

Acknowledgements

It is a truth universally acknowledged that although only one person's name appears on the front of a thesis, in actuality, it represents the efforts and influences of many. I have been surrounded by incredible support, both academic and personal, throughout my DPhil — the lion's share of which coming from Alison herself. It seems a lifetime ago since I first wrote to Professor Alison Woollard asking to work in her group as an undergraduate project student, and another lifetime entirely since I watched her present the Royal Institution Christmas lectures in the midst of my A-levels; these being what galvanised me to pursue science in the first instance (from among my other subjects, all of which were my favourite). If ever there was an example to be cited for the transformative power of engaging children and young people with scientific research, I suppose it could well be this one. There is not the space to list all the ways in which Alison has supported me over the years by way of gratitude, and it would simply feel inadequate to do so. Suffice to say, for these purposes, I am grateful to her for being a supervisor, mentor, teacher, colleague, and friend. We make a great team.

My thanks extend to members of the Woollard group and Biochemistry department whom I have known over the years; friendships are indeed forged in the crucible of long days and nights in the lab. But it is no exaggeration to say that I would not have been in a position to write this thesis were it not for Chloe whose best friendship I rely on every day. The same is just as true for Roddy and Craig, who never fail to make me laugh until my sides ache.

I am under no illusion that without being able to stand on the shoulders of some intellectual giants, I would not have been able to produce the work in this thesis. Being inspired by conversations, had over a coffee or more often a pint of ale, with fantastic scientists over my DPhil years has been a privilege for which I will always be grateful. And so, it is for this reason that I thank: Professor Peter Holland FRS, Professor Jonathan Hodgkin FRS, Professor Catherine Pears, Professor Sebastian Shimeld, and Dr Pete Appleford.

One person whose support cannot go unacknowledged is Aidan Casey KC; without his advice and wisdom I would have felt at sea. Thank you for teaching me to either get yourself a good lawyer or become one.

My DPhil has been a truly formative experience. I would like to thank Francis, Robert, and Paul for showing me the value of good decision-making and that having power is not nearly as important as how you choose to use it.

Lastly, though undoubtedly of most importance, I thank my parents. From taking me to my first open day at Oxford when I was just 13 (at my tiresome insistence), to reading every essay, paper, poem, and chapter I write, your support of me is unwavering and your enthusiasm for all my endeavours indefatigable. I am the luckiest daughter. I dedicate this thesis to you both, as well as to Chloe, Roddy, and Craig.

Declaration of authorship and publications arising from this thesis

Several portions of this thesis are reproduced from publications of which I am the first author, and so all text and data presented here are my own work. In particular, much of the introduction is based on one review article and an essay (Baker and Woollard 2019; Baker and Woollard 2022), and Chapter 4 is based on a recent publication (Baker et al. 2021).

All transcriptional/translational reporter and RNAi constructs — including lines made with those constructs — were made by me over the course of this thesis with the express intention of providing answers to questions I ask in it. The only exceptions to this are: *ajm-1::gfp* used in Chapter 4; the Warthog RNAi clones used in Chapter 4 which were retrieved from the Source Bioscience/Ahringer RNAi library; and *rab-3p::3XFLAG::wrmScarlet::unc-54* used in Chapter 6.

E.A. Baker
14th April 2023

TABLE OF CONTENTS

.

List of abbreviations

3'UTR, 3' Untranslated Region

5'UTR, 5'Untranslated Region

AJM-1::GFP, Apical Junction Marker -1 GFP

ATP, adenosine triphosphate

cDNA, complementary DNA

CeNDR, Caenorhabditis elegans natural diversity resource

CGC, Caenorhabditis Genetics Center

cM, centimorgans

CRISPR/Cas9, Clustered Regularly Interspaced Short Palindromic Repeats

DAPI, 4',6-diamidino-2-phenylindole

DDC, Duplication Degeneration Complementation

Dhh, Desert Hedgehog

DIC, Differential Interference Contrast

DMSO, Dimethyl sulphoxide

DNA, Deoxyribonucleic acid

DRD, Dietary Restriction Downregulated

DP, Dimerisation Partner

DPL-1, DP-like

DTC, Distal Tip Cell

EFL-1, E2F-like

FAXDC2, fatty acid hydroxylase domain containing 2

FEM-1, FEMinization of XX and XO animals -1

FPKM, Fragments Per Kilobase of transcript per Million mapped reads

GFP, green fluorescent protein

GLP-1, abnormal GermLine Proliferation

GRD, GRounDhog

GRL, Ground-Like

Hh-r, Hedgehog-related

HIM, High Incidence of Males

Ihh, Indian Hedgehog

IPTG, Isopropyl β-D-1-thiogalactopyranoside

kb, kilobase(s)

LET, lethal

LG, Linkage Group

LIN, LINeage defective

LR, Left-right

M, Molar

Mab, Male ABnormal

Mb, megabase

modENCODE, Model Organism ENCyclopedia Of DNA Elements

mM, milli Molar

mRNA, messenger RNA

Muv, Multivulva phenotype

MYA, Million Years Ago

MYRF, Myelin Gene Regulatory Factor

NLS, nuclear localisation signal

NGM, Nematode Growth Medium

OMA, Orthologous Matrix Algorithm

ORF, Open Reading Frame

PAML, Phylogenetic Analysis by Maximum Likelihood

PAR, PARtitioning-defective polarity proteins

PAT, Paralysed Arrest at Twofold

PCR, Polymerase Chain Reaction

Pt-r, Patched-related

QUA, QUAhog

RAB, Ras-associated binding

Ras, Rat sarcoma

Rb, Retinoblastoma

RFP, Red Fluorescent Protein

RNA, Ribonucleic acid

RNAi, RNA interference

RRF-3, RNA-dependent RNA polymerase Family

RTK-Ras-MAPK, Receptor-Tyrosine-Kinase / Mitogen-Activated-Protein-Kinase

Runx, Runt-related transcription factor

S.D. Standard Deviation

S.E.M. Standard Error of the Mean

SH-aLRT, Shimodaira–Hasegawa - approximate Likelihood Ratio Test

Shh, Sonic Hedgehog

SR, Stable Redundancy

Tbx, T-box gene (T-domain containing TF)

TDAR, Transient Duplication Associated Redundancy

TF, Transcription Factor

UNC, uncoordinated

WGD, Whole Genome Duplication

Wnt, Wingless Integrated

WRT, (wrt) WaRThog

WT, Wild-type

μm, microlitre

μm, micrometer

The Road Not Taken

By Robert Frost (1874 - 1963)

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

CHAPTER 1

General Introduction

Gene duplications in evolution

Preface

A question that has occupied the minds of biologists for centuries concerns what drives the origin of novelty across the tree of life. Indeed, for many decades now, duplicated genes have sat at the heart of answers to this question, having long been appreciated as both the substrates and catalysts of evolutionary processes. In many ways, it simply stands to reason that new genetic material is a prerequisite for phenotypic innovation. A substantial body of empirical evidence now suggests that from the simplest cell to complex multicellular animals and plants, gene duplications have, irrefutably, made immeasurable contributions to the phenotypic evolution of all life on Earth. Although, not merely drivers of morphological innovation and speciation events, duplicated genes sculpt the evolution of genetic architecture in ways we are only just coming to understand now we have the experimental tools to do so.

Traditionally, there exists the notion that following their formation, duplicated genes can evolve in one of three ways. First, the adoption of new functionality (via changes to either the regulatory or coding sequence or both) by neofunctionalisation. Second, the degeneration of one of the paralogues resulting in its non-functionality by the process of pseudogenisation. And third, the partial, though necessarily complementary, degeneration of both paralogues such that the expression of the single ancestral gene becomes divided between the new duplicates in a process known as subfunctionalisation.

These three modes of duplicated gene evolution have provided the framework for the empirical assessment of paralogue diversification for some twenty years despite not originating from empirical studies. However, it is arguable that no framework (empirically derived or otherwise), should remain unchallenged for such a long length of scientific time given the emergence of new evidence that is hard to fit within it (Baker and Woollard 2022). In any case, it is certainly worth exploring whether 'pigeonholing' paralogues into the classical framework inevitably leads to an underestimation of their complexity (Baker and Woollard 2022).

It goes without saying that our current ability to study the structure and function of genes has eclipsed that of previous decades, providing insights into the complexity of genetic architecture that were unknowable when the traditional ideas of paralogue evolution were being developed. So, in view of such emerging complexities, it is worth scrutinising our understanding of duplicated gene evolution before building on it with the aim of moving towards a more nuanced understanding of the evolutionary consequences of duplicated genes.

But before elaborating on our current understanding of the ways in which duplicated genes evolve, it is prudent, for our investigative purposes, to consider the origin and development of ideas surrounding the role of gene duplications in evolutionary processes.

Origin of an idea: duplicated genes as substrates and catalysts

Even prior to the structure of genes being characterised, Susumu Ohno posited the evolutionary potential of gene duplications in his seminal work '*Evolution by Gene Duplication'* (Ohno 1970). In

his book, Ohno emphasised that evolution by natural selection was an inherently conservative process and that only with new genetic material – generated by gene duplication events – could novelty emerge. Although, even prior to Ohno's work, Ed Lewis had tentatively put forward the idea that new genes arising from duplication, or "higher repetition" could go on to perform *similar* or new functions relative to the pre-existing gene from which they were derived (Lewis 1951). Though erroneously termed "pseudoalleles" by Lewis (adapting the term from Barbara McClintock's original coinage of it (McClintock 1944)), he still hit upon the same idea that would, decades later, prove to be so transformative and conceptually critical to the fields of evolutionary biology and genetics as we know them today.

Beyond its establishment as a revolutionary idea, the concept of gene duplications as a powerful force in evolution remained mechanistically vague until the late 1990s when it was outlined exactly how they might actually catalyse complex evolutionary processes (Force et al 1999; Lynch and Connery 2000). With the aim of addressing a paradox in Ohno's thinking, pioneers of the field Michael Lynch and Allan Force sought to explain why so many duplicated genes persist over the course of evolution (not undergoing pseudogenisation) if neofunctionalisation only occurs in the rarest of circumstances. Their answer came in the form of the Duplication-Degeneration-Complementation (DDC) model, known more commonly as subfunctionalisation (Force et al 1999). Force and Lynch rooted this third mode of paralogue evolution in regulatory divergence as *the* means of functional differentiation and in so doing set the framework for studies of duplicated gene evolution for the twenty years that followed.

With great power — duplicated genes as major drivers of evolutionary change

Gene duplications are a unique class of mutations in that they act as both substrates and catalysts for evolutionary change. Whilst point mutations, indels and other molecular genetic changes may be acted upon by selection if they affect the fitness of the organism, they cannot do so without altering the pre-existing structure and function of the respective gene. Indeed for the most part, changes to protein-coding genes are deleterious as they alter, and therefore compromise the integrity of, the already functional gene product. By comparison, gene duplications provide raw material upon which selection can act, making new evolutionary opportunities possible. In this way, *ceteris paribus*, gene duplication can significantly speed up evolution by providing new — and crucially redundant — genetic material that has no constraints and can freely evolve new functions (Ohno 1970; Naseeb, Ames et al. 2017).

But this is not to say that duplicated genes, upon their generation by the initial duplication event, are always under relaxed selection. Any model of duplicated gene evolution is predicated on their retention following genesis, and in most scenarios, this is unlikely to be the case (Force et al 1999; Lynch and Connery 2000; Baker and Woollard 2022). The reason being that, aside from regulatory sequence changes, gene duplication is the mechanism by which gene dosage is permanently increased. So, gene duplications may well be retained (i.e., by positive selection) should this increase in gene dosage confer a fitness advantage (or even be neutral in terms of fitness consequence), but in many cases it will not. This is to say that intrinsic to the outcome of gene duplication events is a disruption to stoichiometry (that is, the relative quantitative balance of substrates in biological processes), which, conceivably, causes the vast majority of gene duplications to be selected against (Ohno 1970; Lynch and Connery 2000). Thus, as quickly as

duplicated genes arise in a population, they may well disappear. It is for this reason that the frequency of gene duplications as a mutational event were grossly underestimated for a considerable portion of the field's history, and remains an elusive aspect of genome evolution still today (Moore and Purugganan 2003).

Once thought to occur at a rate of 0.2% per gene per million years, the probability of a gene duplication event is actually thought to be closer to 2% per gene per million years, at least in eukaryotes (Moore and Purugganan 2003; Fernando et al. 2021). Of course, this, in many ways, is a meaningless figure — it would be misleading to suggest that any 'given' gene has a 2% probability of duplicating within a 'given' million year timespan. As we shall go on to explore the reasons for in later sections in this chapter, no two paralogue pairs are entirely comparable, and the inherent likelihood of their initial duplication is only one respect in which they are not (Baker and Woollard 2022). No two genes have precisely the same genomic environment, and so do not have precisely the same propensity to duplicate. The genetic and cellular motifs associated with gene duplication mechanisms will be set out in the next section of this chapter, but for now, it is to be stressed that only a tiny proportion of duplicated genes have the kind of evolutionary potential that all are routinely associated with. And for those that are retained beyond the early phase, fewer still evolve to have the kind of roles that lead to major evolutionary transitions.

There are key examples frequently called upon to illustrate how gene duplications facilitate the origin of evolutionary innovations in metazoans (Wagner 2011). From the duplications of the Hox cluster preceding the rise of vertebrates (Holland and Garcia-Fernàndez 1996; Lemons and McGinnis 2006), to the duplication of multiple transcription factors recruited for cardiac development in amniotes (Olsen 2006), and the radiation of the opsin gene family giving rise to

trichromatic colour vision in primates (Dulai et al. 1999); the importance of certain gene duplications to the history of animal evolution is hard to overstate. However, the *direct* evidence linking gene duplication with speciation is scarce. The logic, though, is as sound as it is appealing, and goes as follows: should a gene involved in development or reproduction duplicate to yield paralogues which subsequently diverge in function, those functional, and necessarily biologically irreconcilable differences would inevitably lead to the formation of inviable hybrid progeny, or reproductive incompatibility, respectively.

One such partially characterised example is the hybrid male sterility gene, Odysseus, conserved among members of the *Drosophila* genus (Ting et al. 2004). Odysseus is highly divergent both in terms of gene sequence and expression between *Drosophila* species (while its paralogue, *unc-4*, is highly conserved in both respects). Odysseus exhibits highly variable expression in the germline and soma of different *Drosophila* species, and even between individuals within them, displays a degree of dynamism. The authors of the work imply that it is the inherent instability of Odysseus that make it a candidate for speciation catalysis (Ting et al. 2004), but this was not irrefutably *proven*. This is because the role of any gene duplication, or indeed any class of mutation, is hard to causatively tie down to any one speciation event in complex multicellular organisms. Speciation is a process that in complex systems can only be studied in retrospect, and so investigating the role of one particular genetic variant in such a process is nigh on impossible — inevitably, a myriad of other genomic differences confounds our ability to answer the question. But perhaps a more reliable instigator of speciation is the duplication, not just of a single gene, but an entire genome.

Plants are notorious for the alacrity with which their genomes are duplicated compared to animals. It is, on any view, a mutation of enormous consequence, yet has occurred at least once in upwards

of 70% of all angiosperm lineages (Pozo and Ramirez-Parra 2015). By contrast, only a handful of animal lineages show evidence of whole genome duplication (ray-finned fish (Blomme et al. 2006; Brunet et al. 2006), salmonid fish (Macqueen and Johnston 2014), *Xenopus laevis* (Session et al. 2016), the common house spider (Schwager et al. 2017), horseshoe crabs (Kenny et al. 2016), and acipenseriformes such as paddlefish (Redmond et al. 2022) to name almost all that are known). Underpinning this fundamental difference in genomic behaviour is that plants are forced to withstand extreme environmental perturbation in a way animals are not. As such, the advantages that whole genome duplication (WGD) events have conferred to plants over the course of their evolution has been attributed to a number of factors that accelerate it, such as mutation buffering, dosage effects, increased heterozygosity, as well as the sub- or neofunctionalisation of duplicated genes, all resulting in greater phenotypic variation that could facilitate adaptation to otherwise inhospitable environments (Pozo and Ramirez-Parra 2015; Qiao et al. 2019).

Whole genome duplication has been directly linked to the recently formed polyploid hybrid monkeyflower species, *Mimulus peregrinus,* that arose, remarkably, twice (independently) within the last 140 years — once on the Scottish mainland and a second time on the Orkney Islands (Vallejo-Marín et al. 2015). The parental origin of *M. peregrinus* is thought to be two local populations of *M. robertsii*. And while the limited genetic diversity exhibited by *M. peregrinus* indicates that it may not have a long-term evolutionary trajectory, it is undeniably a species in its own right, becoming instantly genomically distinct from its extant precursor, and by the same token, reproductively isolated from it. In this way, the precise mechanistic context of the gene duplication event in question is critical to consider when making an assessment of the evolutionary consequences of paralogue generation — be this a catapult to speciation by a dramatic WGD, or something altogether less significant.

How are new genes born?

As alluded to in the previous section, gene duplication events come in different flavours, but can broadly be categorised in two ways: small-scale and whole genome duplication events. Taking the first, and most common gene duplication mechanism, small-scale events include tandem duplications which may arise in a variety of ways, namely – non-allelic homologous recombination associated with unequal exchange between sister chromatids (Taylor et al. 1957); unequal crossing over in meiosis prophase I (Smithies 1964); or non-homologous recombination repairing chromosomal breakages caused by replication going awry (Kozul et al. 2004). Small-scale duplication events of this sort are mediated by repetitive elements, such as tandem or inverted repeats, which inadvertently provides regions of homology thereby facilitating recombination. Therefore, depending on the point of recombination, and by extension the location of the repeat, part of a gene, a whole gene or several genes can be duplicated at once.

Small-scale duplications, however, do not necessarily yield paralogues in tandem. Transposition, or more commonly retrotransposition, also have the potential to duplicate entire genes. Retrotransposition occurs when a messenger RNA (mRNA) is retrotranscribed into complementary DNA (cDNA) and subsequently inserted into the genome. The hallmarks of retrotransposed duplicates are traditionally considered to be a lack of introns and the absence of regulatory sequences such as a promoter. It is for the latter reason that paralogues generated via retrotransposition were long thought to merely pseudogenise shortly after their genesis (Vanin 1985). However, it has been reported that up to 12.5% of all human retrogenes – of which there are ~8,000 – are in fact transcribed (Zhang et al. 2003; Vinckenbosch et al. 2006), suggesting that some may escape their fate by co-opting the regulatory apparatus of neighbouring genes or by

generating their own de novo, the latter proven to take place in as little as 15 generations in prokaryotic systems (Yona et al. 2018). In contrast, other studies suggest that as little as 4% of human retrogenes are ever expressed (Harrison et al. 2015). Despite the lack of consensus regarding their transcriptional potential, it is known that paralogues derived from retrotransposition can create introns from internal exonic sequences, a phenomenon known as 'intronization' (Irimia et al. 2008). For this reason, the presence of introns alone cannot be taken as definitive evidence to reject retrotransposition as the mechanism of paralogue generation in any given instance. Thus, it may be the case that the number of functional paralogues derived from retrotransposition has been historically underestimated.

In contrast to its small-scale counterpart, the mechanism of WGD remains elusive. What is known is that there are two sorts, autopolyploidisation and allopolyploidisation. The first involves a WGD event within a species, which may feasibly occur if cytokinesis fails during early development, or if prior to this at fertilisation there is pronuclear fusion involving an unreduced gamete (Spring 2017). The latter might arise if there is a failure in meiosis II, or an inability to extrude the polar bodies at the end of meiosis. The second – allopolyploidisation – is the mutational product of an interspecies cross which almost without exception is thought to lead to sterility due to inevitable mitotic failures when attempting to pair similar, but crucially non-homologous, chromosomes (Spring 2017).

The distinction between the possible mechanisms of WGD is not trivial. This is because the main advantage to deriving paralogues through WGD as opposed to small-scale events is considered to be the maintenance of stoichiometry – all subunits of protein complexes, genes and their regulators, and so on, are simultaneously duplicated such that there are no subsequent dosage imbalances (Ohno 1970; Kondrashov and Kondrashov 2006; Birchler and Veitia 2012). This is

heralded as WGD's major advantage over tandem duplication and explains why paralogues derived from WGD persist over longer periods of evolutionary time compared to any other sort (Birchler and Veitia 2012; Veitia 2017; Birchler and Veitia 2021). In other words, what these events wreak in mutational consequence they make up for in stoichiometric integrity (Ohno 1970). However, this is surely only the case for paralogues derived from autopolyploidisation as those that originate from allopolyploidisation – depending on the extent of the divergence between the two species crossing to generate the initial polyploid – are not necessarily able to interact faithfully. Of course, this would wholly depend on the interaction in question as some interactions (for example certain transcription factor (TF) binding preferences), are inherently promiscuous such that any sequence differences between new paralogues may in fact be inconsequential or could even generate novel targets.

Alternative fates of duplicated genes

The conventional wisdom regarding the fates of duplicated genes was bound to be too simple. Indeed, no author has ever claimed those are the *only* foreseeable ways in which duplicated genes can evolve. And nor could they, because for as long as the three classical models have existed it has been appreciated that paralogues can maintain genetic redundancy with one another for considerably long periods of evolutionary time (Tischler et al. 2006). This is a state of affairs that should not be permitted in the classical view because one paralogue would inevitably be under relaxed selection making true redundancy evolutionarily unstable and thus transient (Lynch and Connery 2000). On the contrary, redundancy is thought to be a pervasive phenomenon among

paralogues generally (Tischler et al. 2006), but this lacks any explanation in the classical framework.

One such explanation, though, may be found in an alternative fate known as hypofunctionalisation (Qian et al. 2010). Hypofunctionalisation involves the reduction in the expression of both duplicated genes such that *both* must be retained by necessity and so are unable to diverge over time (Qian et al. 2010; Gout et al. 2015). However, if paralogues acquire the ability to compensate for one another (e.g., by upregulation of expression), then neither paralogue would have a phenotype if it were to be mutated – redundancy would therefore ensue. This is not subfunctionalisation in the classical sense because functionality (in terms of expression domains) has not been lost – paralogue expression is not spatially distinct. A recently characterised example illuminates hypofunctionalisation dynamics.

T-box TFs, the subjects of Chapter 3, are an uncharacteristically large and dynamic family in the *Caenorhabditis* genus of nematodes. Among two of the 21 paralogues in *Caenorhabditis elegans* are *tbx-37* and *tbx-38*, known for their redundant role in mesodermal induction in the early embryo (Good et al. 2004). Originally put forward as a case of true redundancy, recent work suggests all is not what it first seems when it comes to the interaction between these two paralogues. When endogenous reporter constructs of *tbx-37* and *tbx-38* were made, each was only observable in the early embryo when the other paralogue was knocked out implying each has the capacity to be upregulated when its counterpart is compromised (Charest et al. 2020). And so, it would seem that *tbx-37* and *tbx-38* are paralogues locked into a hypofunctionalised relationship – ordinarily (i.e., in the wildtype scenario) expressed at low levels individually that combine to produce the full expression potential required for normal development. Presumably this combined expression recapitulates the level of expression associated with the ancestral single-copy orthologue, and the

regulatory interaction between *tbx-37* and *tbx-38* has evolved as a compensation mechanism if one paralogue were to deteriorate. It is hard to conceive of another reason for maintaining duplicated genes in such a state as this other than for the purpose of instilling robustness in a gene regulatory network. Indeed, this may be of paramount importance for paralogues involved in critical aspects of development that belong to rapidly evolving, mutationally vulnerable, gene families (both statements that apply to *tbx-37* and *tbx-38*). But hypofunctionalisation is unlikely to always explain instances of persistent redundancy between paralogues.

Another explanation for the maintenance of redundant duplicated genes over long periods of evolutionary time may lie in pleiotropy, i.e., the ability of certain genes to perform multiple functions. The redundancy between *erd-2.1* and *erd-2.2* in *C. elegans* is reminiscent of many such relationships between other paralogous KDEL receptors in the animal kingdom (Tischler et al. 2006; Matthews et al. 2021). Simultaneous knockdown of *erd-2.1* and *erd-2.2* is lethal, yet when abolished individually there are no obvious phenotypic consequences. While not characterised as such in *C. elegans*, work in budding yeast (Semenza et al. 1990) and *Drosophila* (Abrams et al. 2013) suggests the role of the ERD family is to retrieve proteins that have been trafficked to the Golgi apparatus. The Erd paralogues in *C. elegans* are 84% similar in amino acid sequence, and exactly why they have remained as such for approximately 80 million years of *Caenorhabditis* evolution remained mysterious until it became known that they are both able to moonlight as (equally capable) suppressors of deleterious mutations in the acetylcholine transporter, UNC-17 (Matthews et al. 2021). It was deduced that while certain mutations compromised the primary function of the Erd paralogues as receptors, those same mutations, surprisingly, made them able to suppress certain deleterious *unc-17* alleles. And while this kind of allele-specific suppression may at first seem highly coincidental and artifactual, it may on closer inspection lend a suitable explanation for the persistence of an additional Erd paralogue as a genetic 'spare tyre'.

In permitting survival of otherwise deleterious mutants, the preservation of both *erd-2.1* and *erd-2.2* as almost identical genes actually creates their functional potential as regulators of neurotransmission. So more than simply instilling robustness in their basic role as cellular traffickers, maintaining redundant Erd paralogues permits both paralogues the opportunity to flexibly adopt a new function which would not be accessible to a single-copy orthologue, and nor to two divergent Erd paralogues where the roles of the receptor and suppressor were permanently divided between two different genes. In this way, these kinds of longstanding redundancies may, paradoxically, open up striking adaptive opportunities. This differs from neofunctionalisation and subfunctionalisation in that latent pleiotropy maintains the redundancy over evolutionary timescales. But what about moonlighting more broadly? Surely gene duplication offers the perfect solution to single-copy genes that are required to multitask in seemingly unrelated biological scenarios.

Unlike the latent pleiotropy exhibited by the Erd paralogues, some gene products in the natural world display moonlighting behaviour (which can be defined as the ability of a single gene product to perform two or more biochemically unrelated, independent functions (Jeffrey 2018)). In such a scenario, it might seem as though subfunctionalisation (though necessarily with respect to coding sequence) would be the obvious route for newly duplicated moonlighters to take. However, this only rarely occurs (Espinosa-Cantú et al. 2015). Stoichiometric limitations may be relevant (among others) if the functions of the gene product are associated with the same biological context such as a response to a particular environmental perturbation. Such a limitation might have led to the fates of the partially redundant pair of paralogous moonlighters Hxk1 and Hxk2 in budding yeast (Gancedo et al. 2014). These two genes catalyse the phosphorylation of hexoses but both moonlight as transcriptional regulators of Glk1 required for the metabolism of aldohexoses (Rodriguez et al. 2001; Gancedo et al. 2014). Hxk1 and Hxk2 are expressed under different

nutritional conditions yet can fully compensate for the loss of their paralogue by upregulating their expression irrespective of the nutrients present. Their compensatory dynamic is not dissimilar to the interaction between *tbx-37* and *tbx-38* in *C. elegans*, only more drastically so, because Hxk1 and Hxk2 are not ordinarily co-expressed at all. This is to say that Hxk1 and Hxk2 are *specialised* for moonlighting under different circumstances.

Specialisation is more accurately defined as a form of asymmetric paralogue divergence where one paralogue becomes adept at a distinct aspect of the ancestral gene's function, possibly even elaborating on it, while the other retains a broad association with the ancestor. Recent studies have found specialisation to be a phenomenon in large multigene families, the eutherian ETCHbox genes being one such gene family (Royall et al. 2019; Lewin et al. 2021).

ETCHbox genes are fast-evolving members of the homeobox gene superfamily of TFs present only in eutherian mammals, originating by duplication and divergence from *Cone-rod homeobox* (*CRX)* (Royall et al. 2019). As its name suggests, CRX expression is restricted to would-be photoreceptor cells in the developing eye, while ETCHbox genes are expressed only for a short period in preimplantation embryos. The authors compared the repertoire of target genes regulated by metatherian CRX, eutherian CRX, and the ETCHbox genes and found that the metatherian and eutherian CRX orthologues strongly resembled one another in their capacity to activate the same suite of genes required for the adoption of a photoreceptor cell fate (Royall et al. 2019; Lewin et al. 2021). And, while metatherian CRX can regulate a number of genes in eutherian preimplantation embryos, it could not activate the vast majority of ETCHbox targets upon its overexpression. This suggests that, following their derivation, eutherian ETCHbox paralogues assumed a functional niche that went above and beyond the role of their progenitor in the early embryo of the ancestral mammal. However, caution is advised when making inferences about gene evolution from studies

such as this where the analysis of gene function in embryogenesis is restricted to overexpression experiments in cell culture (Royall et al. 2019; Lewin et al. 2021). But nevertheless, at this stage, it may be hypothesised more generally that specialisation is an evolutionary strategy that enables members of gene families to adopt a divide and conquer strategy of functional innovation not available to subfunctionalised duplicates.

Regulatory coevolution among coevolving paralogues

A conceptual flaw in the field-defining notion that newly duplicated genes (derived from small-scale events) have an untapped evolutionary potential is that genes are not merely their promoters, introns, and exons. For a gene to give rise to its phenotype, it is reliant on regulatory elements often located far away from its coding region. Thus, it is axiomatic that new, initially identical, duplicates are not free to assume any imaginable fate given they are inevitably limited by demands on their unduplicated regulatory elements. In this next section we shall explore what is already known about the ways in which paralogues resolve such regulatory conflicts to delineate their expression from one another and how these shape the fates of the duplicates in question.

Two recent illustrations of how this regulatory crossroads can be resolved come from studies in *Drosophila melanogaster*. The first to provide insights into this is the *bric-a-brac* (*bab*) locus, comprising the tandemly-duplicated genes *bab1* and *bab2* that share a single enhancer (Bourbon et al. 2022). While the enhancer regulates overlapping *bab1* and *bab2* expression along the proximodistal axis of the developing leg disc, *bab2*-specific (and more restricted) expression in the leg disc was also shown to be under the control of the same enhancer; but how? It was postulated that divergence in the *bab1 and bab2* promoters facilitates unique interactions with the TF Rotund at the *bab2* locus (Bourbon et al. 2022). The shared enhancer can respond to Rotund, affecting both *bab1* and *bab2* expression, but it is thought the additional interaction with the *bab2* promoter is

responsible for the subsequent recruitment of additional repressors that directly act on *bab2* to delimit its expression (de Celis Ibeas and Bray 2003; Greenberg et al. 2009).

Second, and mechanistically redolent of the action of *bab1 and bab2* in the leg disc, is the expression of the functionally redundant *pdm2* and *nub* paralogues in the developing wing disc (Loker and Mann 2022). Again, the expression of both paralogues relies on a shared enhancer, but their ability to respond to it differs – *nub* responds in all wing progenitor cells and *pdm2* only in a small subset. This differential response is a result of a *pdm2*-specific silencer element in the *pdm2* promoter that receives repressive input from Rotund. And crucially, the repression by Rotund depends on *nub*, allowing *pdm2* to fully respond to the wing enhancer when *nub* expression is compromised thereby enabling functional compensation to occur.

Elaboration on pre-existing regulatory programmes is the most well-trodden route (that we know of) to refining the expression of duplicated genes. However, as the cases of *bab1/bab2 and pdm2/nub* demonstrate, still being linked and under the command of the same *cis*-regulatory element is clearly restrictive with respect to the ability of paralogues to functionally diverge from one another. This is just one reason why the fates of paralogous genes are ultimately constrained by the particular duplication event from which they were derived.

Context-dependence and the limitations on duplicated genes

It seems intuitive that both the mechanism and age of duplication limit the ability of paralogues to persist over evolutionary time because it makes them more or less prone to pseudogenisation (by being more or less able to become functionally distinct). It was not, however, until relatively recently that these ideas were supported by empirical evidence from studies of plant genome

evolution (Rody et al. 2017). By analysing the genomes of 25 taxonomically diverse plant species, the ability of duplicated genes to be retained over the course of evolution in light of their duplication mechanism, gene function, and age of duplication were all investigated (Rody et al. 2017). It was found that tandem paralogues tend to be much younger compared to those that originate from WGD events, but that regardless of how they are derived, TFs are overwhelmingly retained following their duplication. An explanation for why this might be could lie in the behaviour of TFs compared to other genes. The highly interactive nature of TFs means that their stoichiometry is critical (Kuzmin et al. 2020; Ascencio et al. 2021). Thus, it could be argued that TFs should either be lost almost immediately following their duplication or forced to rapidly assume a new function to avoid such a loss.

A hypothesis for why it is that TFs are disproportionately retained following their duplication relates to expression attenuation (Ascencio et al. 2021). This is thought to protect against the particularly harmful short-term consequences of duplicating TFs which act as the major regulators of organismal development. However, it may be that following their downregulation, paralogous TFs are required to evolve much slower as they hypofunctionalise and become unable to diversify – effectively trapped into working as one gene because the stakes are simply too high to subsequently gain or lose functions (Kuzmin et al. 2020; Ascencio et al. 2021; Birchler and Veitia 2021). This would certainly seem an idea supported by the examples introduced thus far, and may explain why redundancy is so prevalent among duplicated TFs especially, flying in the face of the narrative that suggests acquiring new transcriptional regulators is always a path to new evolutionary opportunities.

Nothing in genetics makes sense except in the light of gene duplication — the case for studying a gene in the context of its family

Aside from setting out the terminology and concepts fundamental to the understanding of this thesis, thus far, we have introduced several ideas that will be required for the interpretation of the results in it. What's more, with an eye on pragmatism, it is worth spelling out that, as illustrated by the insights gleaned from the aforementioned recent studies, a more rigorously genetic approach is required to uncover the nuances of duplicated gene evolution. Quite simply, the historic over reliance on using expression pattern as a proxy for gene function, and thus deduce the evolutionary fates that genes have adopted, conspires with the tripartite framework to produce a reductionist (and therefore overly simplistic) picture of gene family evolution.

With an undeniable emphasis on *cis*-regulatory evolution, the mode of gene evolution described by the DDC model has historically been deduced in investigations relying heavily on expression pattern determination techniques such as *in situ* hybridisation (Force et al 1999; Lynch and Connery 2000; Baker and Woollard 2022). In these studies, if the expression of the individual paralogues each has a more restricted expression than an unduplicated outgroup, but their summation does not, this is taken as evidence of subfunctionalisation, a logic which still pervades the field of evolutionary developmental biology today (Baker and Woollard 2019). Not merely conflating the concept of 'gene expression' with 'phenotype' (which may be more or less true for certain genes, such as TFs, but is unlikely to be the case for many others), this approach overlooks coding sequence evolution and, by its nature, neglects to assess the *functionalisation* of the genes in question but merely characterises exactly what it is designed to — their *cis*-regulatory divergence.

It follows that in order to systematically elucidate the functions of duplicate genes, it is necessary to perform genetic analyses, including loss-of-function studies.

Inversely, in the field of genetics, the role of a particular gene is often not studied in concert with its paralogues. Genes are commonly studied in isolation and taken out of their evolutionary context with regards to both duplication and loss events. Whilst this may fail to capture some aspect of their biology or bypass an explanation for a particular genetic phenomenon, it also misses the opportunity to understand how genes function as products of their evolutionary history. That is to say, nothing in genetics makes sense except in the light of gene duplication.

The Nematode Phylum: Beyond 'The Worm'

C. elegans as a powerhouse of developmental biology and genetics

Caenorhabditis elegans is a well-established model organism for the study of development, genetics and genomics as well as seemingly ever-expanding fields of related research. A number of its inherent qualities enabled it to thrive as a genetic and developmental model system from the mid-1970s onwards (Brenner 1974), including its small size, relatively simple anatomy and nervous system, as well as its rapid lifecycle and transparency from egg-to-adult. In addition, *C. elegans* populations consist mostly of self-fertilising hermaphrodites with the occasional spontaneous male, facilitating genetic analysis.

As *C. elegans* became increasingly utilised by biologists, a plethora of resources became available that made 'The Worm', as it is affectionately known, an even more convenient system with which

to answer biological questions. Most notably, the entire invariant cell lineage was traced throughout development (Sulston and Horvitz 1977) and later the whole genome was sequenced, the first complete genome of a multicellular animal (*C. elegans* Sequencing Consortium 1998). Consisting of 100 Mb and encoding 20,190 protein-coding genes (WormBase Release WS270), the worm's compact genome represents an invaluable tool for tackling the complex, multifaceted nature of biological problems. Augmenting the utility of the *C. elegans* paradigm from the point of view of the developmental biologist is the ease with which gene function can be interrogated using a whole gamut of reverse genetic approaches (both knockdowns and knockouts). And complementing the analysis of gene function is the readiness with which gene expression patterns can be ascertained in the worm using both transcriptional and translational reporters.

Beyond 'The Worm'

Following the post-genomic era, high-quality genome assemblies became available for many species from across the nematode phylum (Figure 1.1A) (Osche 1952; Blaxter et al. 1998; Fierst et al. 2015; Kanzaki et al. 2018; Smythe 2019). Nematoda is further divided into three main lineages, namely, Enoplea, Dorylaimia, and Chromadorea, although orders are commonly organised into five major clades that do not correspond to the divisions of classical taxonomy (Blaxter et al. 1998).

Clade V nematodes, including *C. elegans* and other Rhabditina, have been sequenced the most extensively. Non-parasitic, soil-dwelling bacterivorous members of this suborder offer many of the same advantages as *C. elegans* in terms of being tractable experimental systems (Fierst et al. 2015; Smythe 2019). Figure 1.1B summarises the phylogenetic relationship between particular Clade V species. The *Caenorhabditis* genus therein is broadly divided into two supergroups, *Drosophilae* and *Elegans*, a distinction with more of a genomic than a morphological basis (Kiontke and Félix et al.

2011). For the approximately 50 species in this genus, a genome sequence is either available or currently being determined.

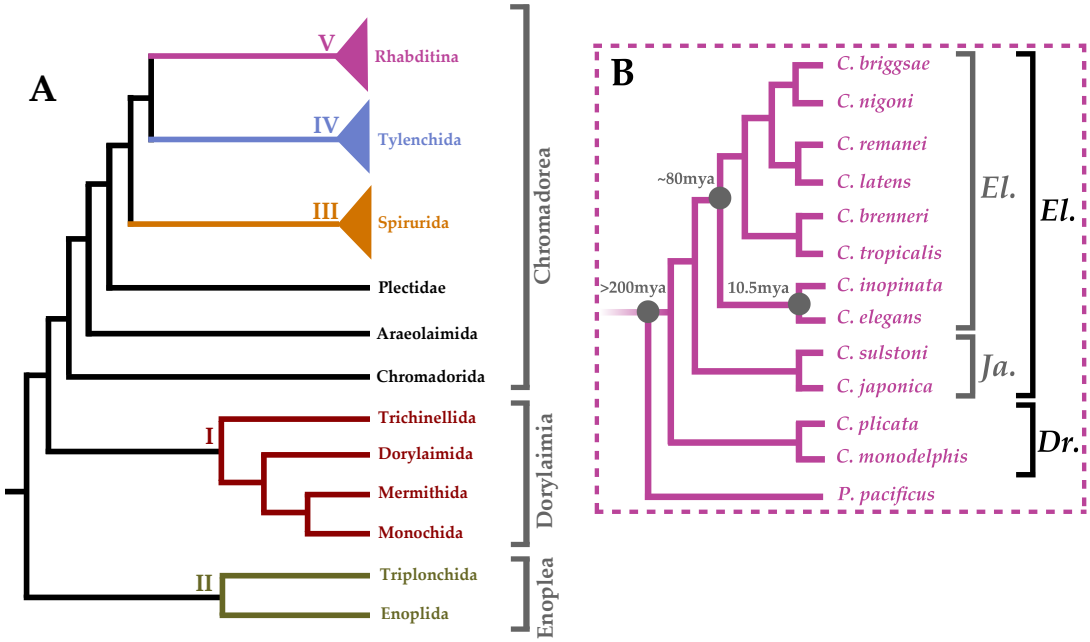


Figure 1.1. Phylogenetic Summary of Nematoda. (A) shows a dendrogram summarising the results of Blaxter et al.'s maximum parsimony analysis as well as the maximum likelihood analysis performed by Smythe et al. on nematode orders [18,19]. Clades I–V are colour-coded: I (burgundy), II (olive), III (orange), IV (light blue), and V (purple). (B) shows a summary dendrogram of the *Caenorhabditis* genus adapted from Bayesian analysis by Stevens et al. [22]. All species are in the *Caenorhabditis* genus and rooted with a *Pristionchus pacificus* outgroup. The *Elegans* (*El.*) and *Drosophilae* (*Dr.*) supergroups are depicted (black) as well as the *Elegans* (*El.*) and *Japonica* (*Ja.*) groups (grey) therein. Figure taken from Baker and Woollard 2019.

Caenorhabditis inopinata is the recently discovered sister species of *C. elegans* (Smythe 2019). Its genome is larger than that of *C. elegans* at 123 Mb, which is thought to be due to highly expanded transposable elements, yet it has lost many protein-coding genes such as those in chemoreceptor gene families. The phylogenetic position of these species, together with their evolutionarily conserved morphology, and the ease with which they can be cultured under laboratory conditions, make *Pristionchus pacificus*, *C. japonica*, *C. inopinata*, *C. brenneri*, *C. remanei*, and *C. briggsae* prime comparative systems for the *C. elegans* model. However, it is worth bearing in mind that it is not

yet established to what extent these species are as amenable to transgenesis as *C. elegans* is known to be. This will, though, be naturally ascertained over the course of this thesis.

C. elegans anatomy and development

It is serendipitous for experimentalists that the *C. elegans* life history strategy resembles that of a pioneer organism — defined by an ability to reproduce rapidly and fruitfully when conditions are favourable (that is when food is abundant and competition and predation are minimal). Central to the life history of a pioneer, though, is one's ability to withstand extremely adverse conditions. When the going gets tough — when food is exiguous — a solution employed by most if not all pioneers is sheer physiological resilience (with tardigrades perhaps being the most archetypal example of this). In *C. elegans*, physiological resilience manifests as the diversion of an entire developmental programme (Byerly and Cassada 1976), modifying its lifecycle to enable it to withstand prolonged periods of starvation and desiccation.

Under favourable conditions, eggs hatch within ten hours of being laid, giving rise to larvae which undergo four moults during the transition to adulthood where the shedding of each cuticle delineates each larval stage, designated L1, L2, L3, and L4 — a process that takes three-and-a-half days at 20 °C though is highly variable at other temperatures (as depicted in Figure 1.2A) (Byerly and Cassada 1976). But when deprived of food, the worm has two diversionary developmental routes (Cassada and Russell 1975). If eggs hatch in the absence of food, the resultant L1 animals arrest thereafter, entering what is known as L1 diapause — morphologically indistinguishable from L1 larvae hatched onto an abundance of food but are marked by their increased resistance to stress. Alternatively, upon starvation or the detection of high population density (denoting intense

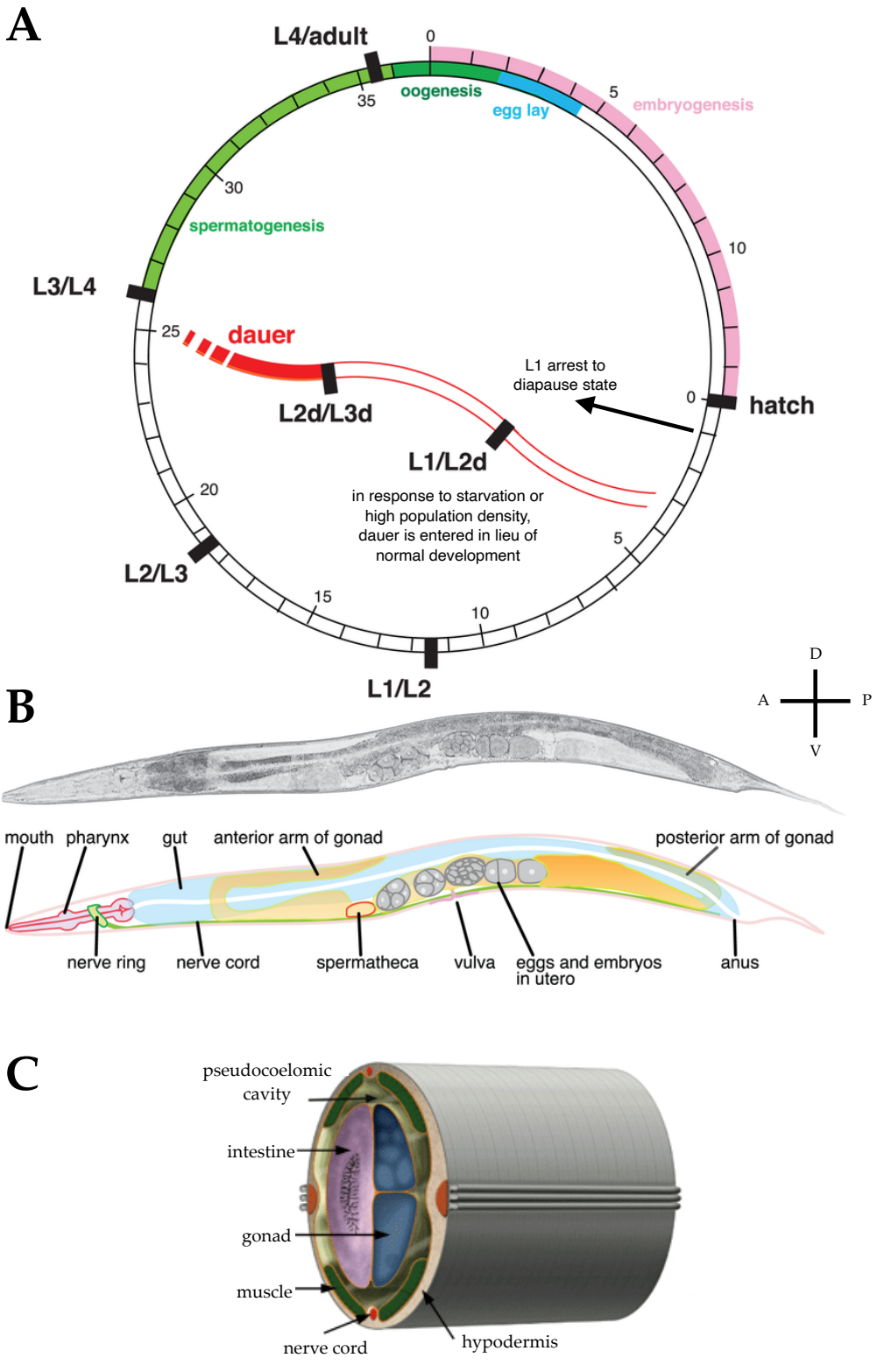


Figure 1.2. Fundamental of *C. elegans* anatomy and development. (A) Schematic showing the *C. elegans* lifecycle with the time taken to reach each developmental stage (at 25°C) shown in hours; numbers on the inside the cycle represent hours from hatching while numbers on the outside the cycle represent hours from laying. Each larval stage (L1, L2, L3, and L4) are denoted by a thick black line and alternative developmental programmes entered in response to starvation (L1

diapause and dauer) are marked up. (B) Nomarski photomicrograph (top) and schematic depiction (bottom) of the worm are shown. The anteroposterior and dorsoventral axes are shown. (C) Cross-section of an adult hermaphrodite with various anatomical features displayed. Panels (A) and (B) are modified from Blaxter 2011 and panel (C) is modified from Altun and Hall 2009.

competition), L2 animals enter a different quiescent state known as 'dauer' (Cassada and Russell 1975). The behavioural and physiological adaptations unique to dauers imparts on them a competency to survive extreme conditions. By way of example, dauers possess a modified cuticle to enhance water-retention, and their mouths become plugged while they cease pharyngeal pumping (along with general motility) almost completely as a means of conserving energy (Cassada and Russell 1975; Byerly and Cassada 1976). So well-adapted are dauers to environmental adversity, they can remain viable in this state for approximately four months, all the while retaining the ability to re-enter regular development to adulthood (Baugh and Sternberg 2006).

Like all roundworms, *C. elegans* is an unsegmented pseudocoelomate. Adult hermaphrodites comprise of 959 somatic cells and are approximately 1 mm long and 50 μm wide though taper in at both ends (Figure 1.2B). In a very crude sense, *C. elegans* is made up of two tubes – one nested inside the other – with a cavity between them filled with pseudocoelomic fluid which acts to maintain the vermiform shape through the hydrostatic pressure it imposes on the surrounding tissue (Figure 1.2C) (Altun and Hall 2009). On the outer side of the pseudocoelomic cavity are the: muscle, hypodermis, excretory and neuronal systems and, at the interface with the outside world, a collagenous cuticle. The inner tube, meanwhile, is home to a less diverse array of tissues, consisting of an alimentary system (punctuated by a pharynx and associated grinder at the anterior end followed by an intestine which runs along the worm for almost its entire length), sitting back-to-back with the reproductive system (Altun and Hall 2009).

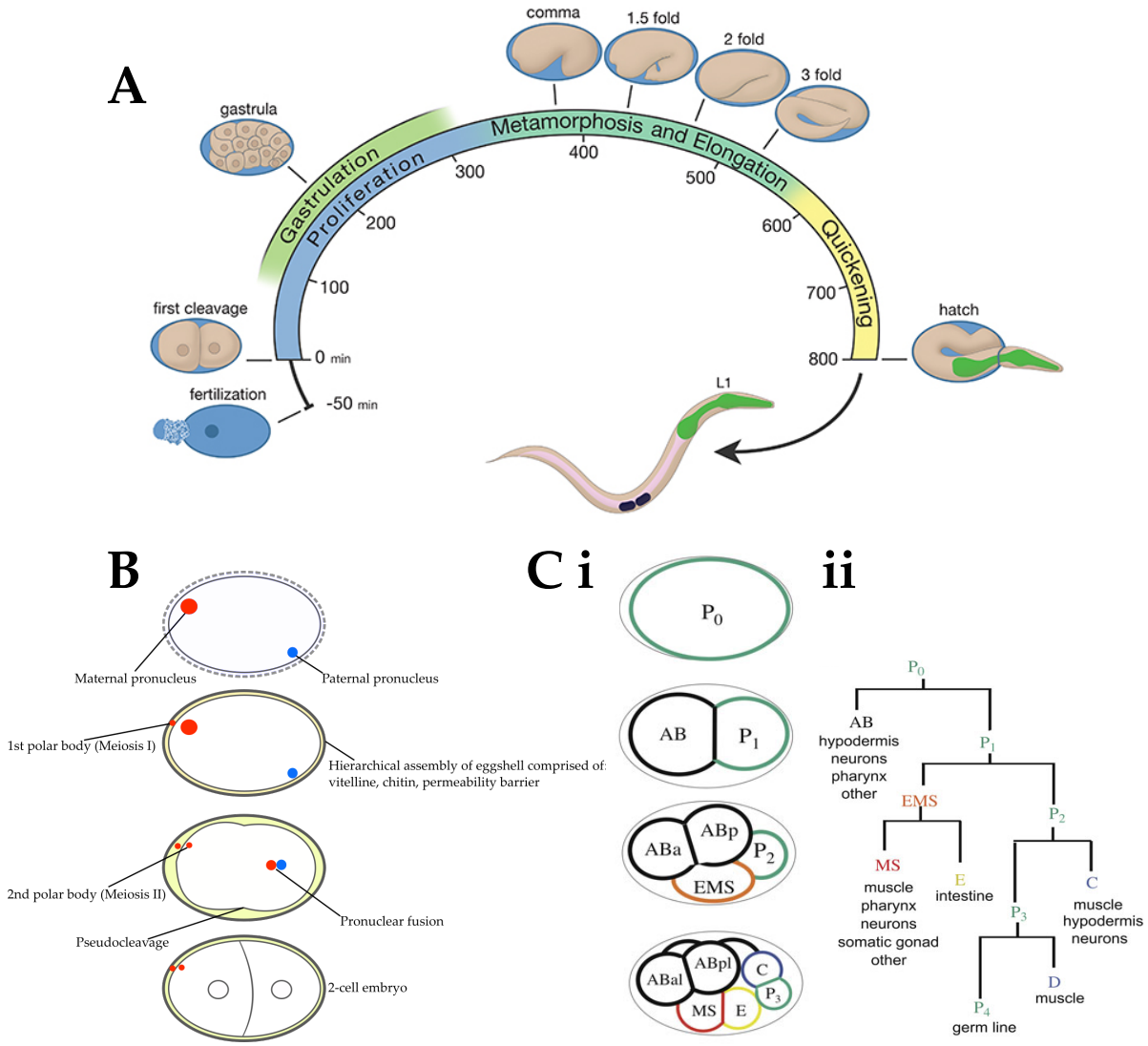


Figure 1.3. Embryonic development in *C. elegans*. (A) Schematised time course of embryogenesis illustrating key embryonic milestones that occur at times (shown in mins) post-fertilisation. (B) Schematic detailing the temporal significance of polar body extrusion simultaneous with the formation of the eggshell. (C) (i) Axial specification in the early embryo is reliably achieved by rounds of asymmetric cell division generating a defined lineage of cells depicted in (ii). Panel (A) is modified from WormBook and panel (C) is taken from Rose and Gonzcy 2014.

During their approximately three week-long life, a hermaphrodite may lay a brood of between 200 to 300 eggs (Byerly and Cassada 1976). Oocytes are self-fertilised as they traverse the spermatheca and post-fertilisation, the zygote undergoes a series of rapid cell divisions before entering gastrulation (the rearrangement of a one-dimensional layer of cells into three germ layers), around

which time the egg is laid (Figure 1.3A). The onset of gastrulation, and indeed the rest of development, hinges on the specification of three principal axes in the early embryo: the anteroposterior axis; the dorsoventral axis; and the left-right axis, and these in turn are contingent on the pronuclear fusion of the maternal and paternal genomes and the proper extrusion of polar bodies from the one-cell embryo. Below we will set out the cellular basis of precisely how these axes are determined, but first let us consider how this latter, and more mechanistically elusive process, is accomplished.

It is a little appreciated fact that the polar bodies (these being the unused products of meiosis) are extruded only after sperm entry into the oocyte (depicted in Figure 1.3B) (Johnston et al. 2006). The first polar body, the diploid product of meiosis I, is extruded prior to the formation of the eggshell proper and so rests on the underside of the outermost layer — the protective chitin interface with the outside world — shrouded in the extra-embryonic matrix surrounding the developing embryo where it remains throughout embryogenesis though presumably disperses after hatching (Johnston et al. 2006). The second polar body, the haploid product of meiosis II, is extruded only after the eggshell and the extra-embryonic matrix are fully formed such that it remains trapped in the filamentous peri-embryonic layer that directly envelopes the embryo. In worms, just as in mammals, this second polar body is destroyed by programmed cell death to avoid any lethal consequences for the ploidy of the developing embryo which, in *C. elegans*, is achieved by its engulfment and subsequent destruction by the blastomeres (Schlientz and Bowerman 2020). The temporal coordination of these processes is a high-wire act with disastrous consequences should it, at any point, go awry; and it is not yet known how pronuclear fusion is delayed to allow the maternally derived polar bodies to be extruded before embryogenesis gets underway. What is

certain, however, is that correct assembly of the eggshell is required for meiotic fidelity, polar body extrusion and polarisation of the *C. elegans* embryo (shown in Figure 1.3Ci) (Johnston et al. 2006).

At fertilisation, sperm entry (specifically the sperm-derived centrosome) defines the posterior end of the embryo (Goldstein and Hird 1996; O'Connell, Maxwell et al. 2000; Sadler and Shakes 2000; Wallenfang and Seydoux 2000; Hamill, Severson et al. 2002; Cuenca, Schetter et al. 2003; Cowan and Hyman 2004; Bienkowska and Cowan 2012) and kickstarts embryogenesis, a defining feature of which is asymmetric cell division (creating a cell lineage which is shown in Figure 1.3Cii). Asymmetrical divisions (these being divisions in which each daughter cell differs in fate) are instrumental not just in the definition of the three axes in the developing animal, but also to creating cellular diversity therein. The arrival of the sperm centrosomal proteins initiates movements in the cortical actomyosin network that cause the polarised separation of PAR proteins (PARtitioning defective) (Kemphues, Priess et al. 1988; Morton, Roos et al. 1992; Watts, Etemad-Moghadam et al. 1996; Munro, Nance et al. 2004) which in turn creates an asymmetric distribution of fate determinants (Kirby, Kusch et al. 1990; Hird and White 1993; Hird, Paulsen et al. 1996, Cheeks, Canman et al. 2004). Consequently the embryo divides asymmetrically along the newly established anteroposterior axis to produce, through asymmetric spindle positioning (Grill, Gonczy et al. 2001; Grill, Howard et al. 2003, Labbe, McCarthy et al. 2004), the large anterior AB and the smaller posterior P_1 blastomeres.

Following the generation of AB and P_1, their respective nucleosomes migrate around each nucleus to sit on a lateral axis. While the nucleosomes of AB do not move from this position to ensure AB divides perpendicular to the first division, the P_1 nucleus-centrosomal complex rotates 90o meaning its division takes place in parallel relative to the first. The division of AB gives rise to ABa

(anterior daughter) and ABp (posterior daughter), named for their final positions at the four cell stage rather than for the division plane of their division; and concomitantly, the division of P_1 gives rise to EMS and P_2. Although the generation of ABa and ABp occurs on a lateral plane, they are known to be functionally equivalent, (Priess and Thomson 1987) and so cannot be said to break the dorsoventral symmetry of the embryo. In contrast, during late prophase of the P_1 division, the spindle rotates once more by 90o and EMS is nudged ventrally to sit beneath ABp, which moves to the dorsal side of the embryo, meaning EMS now defines the ventral side.

It is a natural consequence of having bilateral axes that an animal will have a left and a right side to their body. In this way, the final axis to emerge in development, the left-right axis, is defined on its breaking which, during worm embryogenesis, is on the division of ABa and ABp. These cells divide to give rise to left and right daughters, whereby ABal (left-hand daughter of ABa) and ABpl (left-hand daughter of ABp) are positioned slightly more anteriorly than their right-hand counterparts (Deppe, Schierenberg et al. 1978). Yet, the bilateral pairs of cells on the left and right side of the embryo are functionally equivalent until their fates are instructed otherwise on the inductive signals by MS at the 12- and 24-cell stages (Sulston, Schierenberg et al. 1983; Wood 1991; Hutter and Schnabel 1994; Hutter and Schnabel 1995).

Post-gastrulation, morphogenesis ensues, wherein the embryo becomes enclosed by the epidermis and subsequently elongates, becoming increasingly vermiform, eventually shaping into a distinctive 3-fold, and by this point motile, embryo. After hatching and in the presence of food, post-embryonic development will commence.

Caenorhabditis nematodes and the reverse evo-devo approach

Comparative genomics is a hugely powerful approach for recognising the functional regions of genomes, be this at level of operons, genes, TF binding sites, or other *cis*-regulatory elements. The approach is based on the premise that, over the course of evolution, the genome changes and the rate of this change depends on the strength of stabilising selection acting on a given region; but where divergence is observed, it may be underpinned by change which is under positive selection (though it is of course true that for the most part, divergence is due to elements under relaxed selection). And with respect to the former, the comparative genomic approach has been incredibly fruitful in yielding conserved genomic elements between taxa that play important roles in conserved aspects of organismal development (as found, for example, in Brabin et al. 2011).

Conversely, in more recent years, comparative genomics has proven equally insightful for revealing the significance of genomic differences between organisms, allowing biologists to glimpse into the molecular basis of speciation and evolutionary innovation (reviewed in Baker and Woollard 2019). Unsurprisingly, gene duplications and losses have been at the epicentre of these investigations; the balance between these two dynamic processes not only accounting for copy number differences between species, but also aspects of their morphological diversity as found in studies of moths, molluscs and mammals alike (Holland et al. 2017). By way of analogy to reverse genetics, this approach — termed 'reverse evo-devo' — searches for differences in genes or genetic organisation between species and then asks what effect these differences have on the phenotypes of organisms (Holland et al. 2017).

Prior to this investigation, the reverse evo-devo approach had not been systematically applied to members of the nematode phylum. This was a missed opportunity, because using the reverse evo-devo approach in *Caenorhabditis* nematodes provides not only a chance to assess what the genomic differences are between closely related species, but crucially understand what significance these differences bear on organismal development in greater molecular and physiological detail as compared to other commonly used, though less genetically tractable, systems. It is only too tempting to ponder the potential of such a strategy. Aside from revealing insights regarding the behaviour of genes over the course of evolution, it may be that investigations of this kind could improve our understanding of aspects of developmental biology that have proven evasive in the hands of those relentlessly pursuing evolutionary conservation between systems (thinking particularly of phenomena such as developmental systems drift, or robustness by way of examples here). And so, it could well be that from work such as this, evolutionary developmental biology will find its next set of answers — by asking questions not about what is conserved between animal groups, but the importance of genetic differences observed between them.

But if we are to investigate the evolutionary dynamics of specific gene families in the *Caenorhabditis* genus, it is worth first considering the broader context of genomic divergence in which they sit. Indeed, if the premise of this work is that gene duplications play a significant role in organismal evolution, it is essential to clarify their presence in, nay, their importance to, the models used as part of it. To establish this as a first principle, the results of an analysis that explores the prevalence of gene duplications in certain *Caenorhabditis* species of interest to this investigation (as compared our outgroup, *P. pacificus*) is shown in Figure 1.4. In each case, the number of species-specific paralogues is shown for each gene class (Figure 1.4Ai), where G protein-coupled receptors and F-box domain containing paralogues are provided separately (Figure 1.4Aii) to avoid confounding

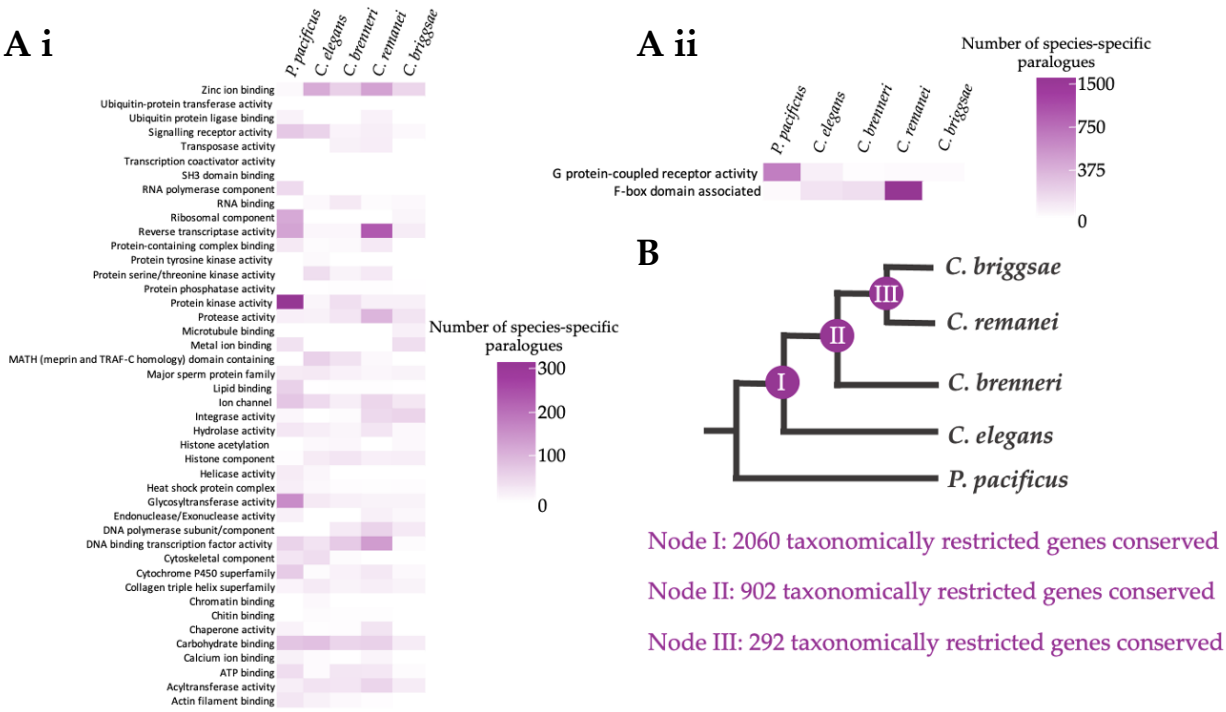


Figure 1.4. Reverse evo-devo in the *Caenorhabditis* genus. (Ai) Comparative genomic OrthoFinder analysis among members of the *Caenorhabditis* genus as compared an outgroup, *P. pacificus*, where only the number of species-specific paralogues belonging to different gene classes are shown (i.e., those not found in any other species represented here). One-to-one orthologues were searched for between all species in OrthoFinder and subsequently discounted from the dataset leaving only paralogues in 'one-to-many' relationships behind. (Aii) The results of the same analysis but exclusively representing highly expanded gene classes: GPCRs and F-box domain encoding genes. (B) Taxon-restricted duplicates at three nodes in the *Caenorhabditis* species tree are shown. At each node (I, II, and III), the number of newly emerged paralogues conserved among those species therein is displayed, these being 'one-to-many' relationships as defined by OrthoFinder where the 'many' paralogues are all conserved between species which fall within that node.

the rest of the analysis with the relative extreme abundance. On a similar vein, Figure 1.4B shows the number of taxon-specific duplicates (shared between all species) at three nodes in the *Caenorhabditis* phylogeny. Although some would argue that the preliminary analysis presented here would be better placed, in the context of a thesis, in a results section, the analysis shown in Figure 1.4 is considered so foundational to the questions posed in subsequent chapters that it has been presented in Chapter 1.

What is clear from this analysis is that there is indeed precedent for investigating the role of gene duplications as an evolutionary force in the *Caenorhabditis* genus because they have so obviously made a mark on the genomes of species therein. Even with respect to entire gene classes, there are tantalisingly foreseeable ways in which species-specific radiations like the ones shown could feasibly lead to species-specific innovations, with perhaps the expansion of the major sperm protein family being the most overtly plausible (these being nematode-specific proteins required for sperm motility and general function).

However, evolutionary innovation and speciation in any taxon is unlikely to be fuelled by duplications in just one gene class — evolution is rarely so uncomplicated — not least because such processes are underpinned by phenotypes dependent on entire gene regulatory networks which encompass many different kinds of gene products. It is also paramount to highlight that attempting to draw out the significance of any one gene duplication by studying them in this strictly bioinformatic way is impractical and biologically ignorant — the role of the marginal kinase paralogue is not comparable to the role of the marginal transcription factor, nor even to another, kinase paralogue. Nevertheless, it is clear that paralogue generation is a feature, if not an engine, of evolution in the *Caenorhabditis* genus — harbouring a wealth of new gene families, inevitably with their own dynamics, that will help us understand the complexities of duplicated gene evolution more generally.

Summary

Our understanding of how duplicated genes evolve has been shaped by a framework – based on theoretical considerations – in which duplicated gene fates are categorised in one of three ways:

neofunctionalisation, pseudogenisation, and subfunctionalisation. This classical framework has defined the scope and interpretation of empirical assessments of paralogue evolution for two decades such that the theory has not only lagged behind the insights derived from empirical studies, but has actually limited them. I have, thus far, drawn attention to recently characterised examples of paralogue dynamics that fall outside the classical framework, suggesting there are in fact many more ways that duplicated genes can evolve than just three. Indeed it is arguable that, due to the unique combinations of paralogue age, mechanistic origin, regulatory constraints, and interacting partners (among others), no paralogues could ever be said to adopt a canonical evolutionary fate – this is to say that there are as many fates as there are paralogous genes.

Aims of this work

This investigation aims to shed more light on the dynamics of, and relationships between, diverging paralogous genes — specifically with respect to a number of factors that go largely unconsidered in most studies of gene family evolution. Nuanced considerations of paralogue evolution, such as the changes to domain architecture following duplication, as well as asymmetric gene diversification, overlapping functionality between paralogues, and the role of rapid paralogue divergence in the formation of new species, will all be explored in turn.

In Chapter 3, the taxon-specific dynamism of the T-box family of TFs in the *Caenorhabditis* genus is investigated, with particular focus on the functional evolution of the *tbx-35* and *tbx-36* gene pair only found in *C. elegans*. Unlike their paralogous counterparts, *tbx-35* and *tbx-36* are expressed in different cells — and are then shown to have divergent functions — in the very early embryo. Subsequent analysis showed, however, that they share a role in mid-embryogenesis which is only revealed when environmental conditions are changed. Like many of the *C. elegans* specific T-box repertoire, *tbx-35* and *tbx-36* have accumulated deleterious mutations in wild populations of *C. elegans* whereby some isolates that have a degenerate copy of *tbx-35* have retained an 'intact' *tbx-36*, meanwhile another, geographically distinct, cohort of isolates exists for which the inverse holds true. The role of *tbx-36* in the early embryo is revealed as a neofunctionalisation underpinned by its acquisition of new regulation upon its transposition to a different chromosome (from its progenitor). And so, when reconciling their developmental roles with their radical patterns of divergence in extant populations of *C. elegans*, the implication is that the *tbx-35* and *tbx-36* gene pair may be putative 'speciation paralogues' operating today.

The *tbx-35* and *tbx-36* gene pair is a prime example of just how impactful paralogue divergence can be for evolutionary processes, but as an example through which to study paralogue dynamics *per se*, they are laden with confounding complexity (being only two of so many other species-specific T-box genes involved in embryogenesis). As such, subsequent chapters explore singular phenomena of paralogue evolution displayed in the T-box family but in smaller, less dynamic, gene families — with each exhibiting a facet of the complexity observed in T-box family evolution.

Chapter 4 characterises the pervasiveness of redundancy in a slightly smaller, more conserved, gene family. The redundancy relationships in the Warthog family in *C. elegans* are reconciled with the evolutionary fates members therein have adopted. The evolutionary origin and the roles of the Warthog family were investigated and it was found that these genes have adopted new functions in aspects of post-embryonic development, including left-right asymmetry and cell fate determination (akin to the functions of their vertebrate counterparts). Analysis of various double and triple mutants of the Warthog family reveals that more recently derived paralogues are not redundant with one another, while a pair of divergent Warthogs do display redundancy with respect to their function in cuticle biosynthesis. Probing the genetic relationships between members of the Warthog family reveals that newer members of taxon-restricted gene families are not always functionally redundant despite their recent inception, whereas much older paralogues can be: a phenomenon considered paradoxical according to the current framework in gene evolution.

Chapter 5 focusses on the consequences (and underpinnings) of asymmetric paralogue divergence for gene family dynamics in the Drd family of oxidoreductases. On cursory inspection, one family member of three, *drd-1.2*, exhibits a radical departure from form in its role in the starvation

response while *drd-1.1* and *drd-1.3* are both involved in male tail development and hermaphrodite gonad development. *drd-1.1* and *drd-1.3* have specialised in male tail development, becoming particularly involved in the specification of different cell lineages therein. However, it is shown that when *drd-1.2* is compromised, both *drd-1.1* and *drd-1.3* are able to compensate and become upregulated upon starvation. As *drd-1.2* is not found to play a role in any aspect of morphological development (of the male tail, hermaphrodite gonad, or otherwise), *drd-1.2* is actually thought to exhibit an extreme rate of divergence because of its loss of functionality, as opposed to that which it has gained.

Finally, in Chapter 6, changes to domain architecture following gene duplication — and the associated developmental consequences thereof — are explored through the Myrf family of TFs. A paradigm for evolution at the sub-gene level following duplication, the Myrf family in the *Caenorhabditis* genus gives unprecedented insight into the impact of paralogue generation when orthologues are seldom duplicated in other taxa (with a single Myrf gene found in all other animals). Redundancy is observed between *myrf-1* and *myrf-2* in *C. elegans*, and so quite why they are both maintained at all is somewhat open-ended. But all Myrf family genes encode one gene product which is post-translationally cleaved into two functional units, and through studying additional Myrf paralogues in other *Caenorhabditis* species, it is implied that one way to lessen the potentially damaging impact of paralogue generation — in lieu of pseudogenisation or hypofunctionalisation — is to divide protein domains between new paralogues (thus removing the need to post-translationally cleave gene products altogether). In this way, additional Myrf paralogues have, ironically, no more or less potential to diversify than a single gene. Conclusions from studying the Myrf family give credence to the idea that not only are some families more

prone to duplication than others, but offers an explanation for why this might be so, and how diverse solutions are arrived upon to cope with them in the event they do.

Taken together, these results characterise hitherto poorly understood aspects of duplicated gene evolution. In particular, it is shown that paralogues cannot be assigned any one label for the fate they adopt, because they do not adopt just any *one* fate; genes contain multitudes when it comes to their pleiotropic functionality in organismal development and homeostasis. As such, the utility of any *model* of duplicated gene evolution is thrown into question by this work. And while it goes without saying that the complex consequences of duplicated genes cannot be fully uncovered by the study of just four gene families (nor by five, twenty, or two hundred), it is undeniable that the investigation to follow moves us closer to having a more nuanced understanding of paralogue dynamics.

CHAPTER 2

Materials and Methods

Bioinformatic approaches

Paralogy detection using Orthologous Matrix Algorithm (OMA)

Orthology was detected in a pairwise manner between *Pristionchus pacificus* and members of the *Caenorhabditis* genus using the OMA web interface — an algorithm for the inference of orthologues among complete genomes (Altenhoff 2018). *P. pacificus* was always set as 'Species 1' and the *Caenorhabditis* species in question set as 'Species 2'. The results from these five sets of analyses were obtained as FASTA files then reformatted and converted to .txt files before import into RStudio.

There are four orthology relationships obtained from the pairwise analyses outlined above: '1:1', '1:many', 'none:1' and 'none:many'. RStudio was used to construct data frames for all the instances of duplication in these five data sets, i.e., rows of '1:many'. A complete R markdown script detailing the code that was used to derive the total number of duplications in these data sets as well as other functions can be found in Appendix I.

Putative orthology relationships were then verified using BLASTp searches. Expect thresholds were set to 1 and all sequences were retrieved unless they were splice variants in which case only the longest isoform was chosen. InterProScan was used to predict the gene ontology profiles of all lineage-specific duplicates.

Molecular phylogenetic and other bioinformatic analyses

Caenorhabditis elegans sequences were obtained from WormBase (http://wormbase.org) and (PSI-)BLAST searched (Altschul et al. 1997) against the genomes of selected nematode species (using the web service default parameters). Representatives from the phylum Nematoda were selected on the basis of genome quality and completeness. Consequently, the sequences are either from major parasites, including: *Trichinella spiralis*, *Brugia malayi*, *Ascaris suum* and *Toxocara canis*; or model organisms, including: *Pristionchus pacificus*, *Caenorhabditis briggsae*, *Caenorhabditis remanei*, *Caenorhabditis japonica*, *Caenorhabditis brenneri*. Multiple sequence alignments were carried out using SeaView software version 4.6.2 (Gouy et al. 2010) and the CLUSTAL Omega programme (default parameters) was used to locally improve the alignment, which was further adjusted by eye.

For phylogenetic tree construction, one of two pipelines was chosen, either Bayesian or maximum likelihood. Taking the first, phylogenetic tree construction for the Drd and Warthog families were achieved using the Bayesian algorithm in MrBayes version 3.2 (Ronquist and Huelsenbeck, 2003). Bayesian inference was performed using the Markov chain Monte Carlo method. Two independent Markov chains were run, each with 1 million iterations with default heating parameters. The first 25% of the trees were discarded as burn-in before compiling consensus trees and summary statistics. Posterior probabilities at each internal node were taken as a measure of support.

Alternatively, for the the Myrf and T-box families, a maximum likelihood pipeline was used to build phylogenies. Maximum likelihood phylograms were constructed using IQ-TREE (Trifinopoulos et al. 2016) and its built-in ModelFinder software (Kalyaanamoorthy et al. 2017).

Branch support was calculated running 10,000 replicates of the SH-like approximate likelihood ratio test and ultrafast bootstrap (10,000 replicates). All tree figures were rendered with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

To map synteny and paralogy relationships, genes directly adjacent to loci in *C. elegans* were searched by tBLASTn of their predicted proteins across the other genomes. If an orthologue was present/detected, their genomic location in a given species was compared to the position of the orthologue in the *C. elegans* reference genome.

Testing for positive selection using Phylogenetic Analysis by Maximum Likelihood (PAML)

The signatures of positive selection were tested for in the Drd family using CodeML implemented in PAML (Yang 2007), using a branch-model to estimate the dN/dS ratio by assigning two independent ratios, specifying the branch leading to each Drd family clade (model = 2, NSsites = 0). Each analysis compared the null model (M0) to a two-ratio branch model (specifying either the *drd-1.1*, *drd-1.2*, *drd-1.3*, or C_bri_drd-1.3 lineage as foreground). Consensus sequences were first compiled for *drd-1.1*, *drd-1.2*, and *drd-1.3* lineages such that selection was assessed for all orthologues in the *Caenorhabditis* genus to remove effects of anomalous rates of divergence in a particular species.

Positive selection was also tested for in the Warthog family, namely for *wrt-6* and *wrt-10* which could not be assessed functionally. Two analyses were ran (one testing for positive selection in the

wrt-6 clade and another testing for positive selection in the *wrt-10* clade) and compared the null model (M0) to a two-ratio branch model (specifying the *C. elegans* lineage as foreground in both).

Assessing mutation accumulation of T-box genes in wild isolates in *C. elegans*

Wild isolate sequences were obtained from the *C. elegans* Natural Diversity Resource Centre (CeNDR (Cook et al. 2017)). In all 21 T-box paralogues, high and moderate effect mutations were identified by calling premature stop codons, missing start codons, frameshift mutations, or splice variants according to Calculated Variant Consequence score implemented in CeNDR.

Ancestral Sequence Reconstruction in the Myrf family using FastML

The Myrf family single copy ancestor amino acid sequence was reconstructed using FastML (Moshe and Pupko 2019) because the rate of evolution was considered relatively low for the whole family (as revealed by the maximum likelihood phylogram) such that a relatively accurate sequence could be reconstructed. The FastML web server (default parameters) was used to generate an ancestral sequence for the Myrf family using the maximum likelihood phylogram and the multiple sequence alignment.

Transcription factor binding site analysis

Transcription factor binding analysis was performed using the JASPAR vertebrate 2022 database (Castro-Mondragon et al. 2022). Putative binding sites were then investigated either experimentally (by their deletion as in the case of the E2F binding site in the promoter of *tbx-36*), or verified using publicly available ChIP-seq data from modENCODE (Araya et al. 2014) mapped onto the genome browser in WormBase.

Worm methods

Strains and maintenance of worms

All strains of *C. elegans* used as part of this investigation were derived from the wild type N2 Bristol strain unless stated otherwise (i.e., from another naturally occurring population or species). Manipulations and maintenance of strains were performed as previously described (Sulston and Hodgkin 1988). Strains used are detailed in Appendix II.

Genotyping

Where it was not possible to follow mutant alleles phenotypically, PCR was used for genotyping. In the case of deletion alleles, PCR primers were designed to anneal either side of the deletion, resulting in a single PCR product in wild type animals, a single, smaller PCR product in individuals homozygous for the deletion, and two differently sized products in heterozygotes. In the case of mutants carrying point mutations, PCR genotyping was performed using the tetra arms PCR approach (Ye, Dhillon et al. 2001). Unless otherwise specified, PCR performed for genotyping purposes was performed using NEB Taq polymerase, using variable annealing temperatures and all other parameters as standard.

***myrf-1(ybq6)* genotyping**

Primers EB22, EB23, EB24 and EB25 were used, with an annealing temperature of 58 °C and an extension time of 1 minute, 20 seconds.

***myrf-2(ybq42)* genotyping**

Primers EB18, EB19, EB20 and EB21 were used, with an annealing temperature of 57 °C and an extension time of 1 minute, 10 seconds.

***wrt-1(tm1417)* genotyping**

Primers EB08 and EB09 were used, with an annealing temperature of 60 °C and an extension time of 3 minutes.

***wrt-2(ok2810)* genotyping**

Primers EB03 and EB04 were used, with an annealing temperature of 60 °C and an extension time of 2 minutes, 30 seconds.

***wrt-3(ok2608)* genotyping**

Primers EB05 and EB06 were used, with an annealing temperature of 60 °C and an extension time of 2 minutes, 30 seconds.

***wrt-4(tm1911)* genotyping**

Primers EB10 and EB11 were used, with an annealing temperature of 59 °C and an extension time of 2 minutes, 30 seconds.

***wrt-5(ok670)* genotyping**

Primers EB12 and EB13 were used, with an annealing temperature of 59 °C and an extension time of 3 minutes.

***wrt-7(ok3271)* genotyping**

Primers EB14 and EB15 were used, with an annealing temperature of 57 °C and an extension time of 2 minutes, 45 seconds.

***wrt-8(tm1585)* genotyping**

Primers EB01 and EB02 were used, with an annealing temperature of 56 °C and an extension time of 3 minutes.

***wrt-9(ok2732)* genotyping**

Primers EB16 and EB17 were used, with an annealing temperature of 59 °C and an extension time of 2 minutes.

RNA interference (RNAi)

Gene knockdown by RNAi was achieved in one of two ways, either by feeding or by injection. Where RNAi was delivered by feeding, it was performed exactly as described in Fire et al. 1998. For RNAi knockdown of: *wrt-1; wrt-2; wrt-3; wrt-4; wrt-5; wrt-6; wrt-7; wrt-8; wrt-9; wrt-10; drd-1.1*; *drd-1.2*; and *drd-1.3,* feeding clones from the Ahringer library were checked and sequenced prior to performing all experiments. Where the RNAi knockdown of genes was performed by injection of dsRNA directly into worms (where specified), an *in vitro* transcription kit (supplied by Promega RiboMAX, protocol performed as per manufacturers' instructions) was used to first synthesise dsRNA that was diluted to ~1 mg/ml for injection.

In the case of RNAi knockdown of *tbx-36*, *efl-1*, *dpl-1*, *myrf-1*, *myrf-2*, and *rec-8*, feeding clones were made.

Single worm lysis

Genomic DNA was obtained by single worm lysis (Williams, Schrank et al. 1992). Two adult hermaphrodites were placed into 2.5 µl of lysis mix (100 µl lysis buffer consisting of 50 µM KCl, 2.5 µM MgCl2, 10 µM Tris-HCl pH 8.3, 0.45% NP40, 0.45% Tween 20, 0.01% gelatine plus 1 µl proteinase K (10 mg/ml)), and the solution frozen at -80 °C, in a thin-walled 0.2 ml PCR tube, for 1 hour. The tube was incubated for 1 hour at 60 °C and subsequently for 15 minutes at 95 °C.

Construction of transgenic worms

Broadly speaking, transgenic animals were made in one of two ways. The first being injections performed as described previously (Mello and Fire 1995) using the *unc-119*$^{+}$ (pDP#MM016β) transformation marker (Maduro and Pilgrim 1995). Injections performed in this way could only be done in an N2 background (wherein an *unc-119* knockout was available to rescue). The second way was only necessary in those cases where transgenic lines were made from wild populations of *C. elegans* and also *P. pacificus* (where, in both instances, knockouts of *unc-119* are not available). For these, a co-injection marker of *myo-2::mCherry* was used and injected at less than 1 ng/µl; successfully transformed progeny were signalled by a red pharynx. As a general principle, where transcriptional reporter lines have been made, constructs were injected at 10-20 ng/µl; where translational reporter lines have been made, constructs were injected at 0.5-5 ng/µl and in the case

of the latter, co-injected in a mix with ScaI-digested gDNA (from the native isolate in question). All plasmids and other injection constructs were verified by sequencing prior to use.

Synthesising circular or linear constructs for bacterial or worm expression

Details of plasmids and PCR products constructed as part of this investigation for expression pattern determination or RNAi-induced knockdown are detailed below. Where PCR products were generated from gDNA, N2 worms (unless specified otherwise) were lysed and the entire 2.5 µl resultant lysis mix used as a template for the reaction. Where PCR products were generated from cDNA, worm lysate was used as a template for reverse transcription using the Thermo Fisher Scientific Maxima H Minus Reverse Transcription Kit as per manufacturers' instructions. All PCR reactions for making the constructs below used either Phusion or Q5 polymerase as high fidelity alternatives to Taq polymerases used elsewhere in this investigation.

Where expression constructs were made, and the expression of a gene was being characterised for the first time in the absence of prior knowledge about its regulatory elements, the putative promoter element (referred to henceforth as the promoter for the sake of brevity) was taken as the region spanning the start codon upstream all the way to the next gene.

tbx-35::gfp for expression in N2, ECA1185, and ECA1191

The C-terminal translational reporters for *tbx-35* were made by the overlap-extension PCR of two distinct products flanked with regions of homology, a) *tbx-35* derived from a particular population of *C. elegans* (either: N2; ECA1185; or ECA1191) and, b) the coding sequence of *gfp* including the

3'UTR of *unc-54*. Taking the first, the open reading frame of *tbx-35,* with its native promoter in tow, was amplified and a 20 nt sequence (complementary to the reverse strand of the start of *gfp*) was added to the 3' end. Second, *gfp*, with its associated 3' UTR, was amplified from pPD107.94 with a 20 nt homology region (corresponding to the 3' end of the *tbx-35* fragment) added to the 5' end. Then, these PCR products were fused together and the final product injected straight into animals (due to its large size) without first being cloned into a vector.

***tbx-36::gfp* for expression in N2, ECA1185, and ECA1191**

The C-terminal translational reporters for *tbx-36* were made in much the same way as the equivalent set of constructs for *tbx-35* detailed above, that is, by the overlap-extension PCR of two distinct products flanked with regions of homology, a) *tbx-36* derived from a particular population of *C. elegans* (either: N2; ECA1185; or ECA1191) and, b) the coding sequence of *gfp* including the 3'UTR of *unc-54*. The only distinction between the two methodologies was the initial amplification of, and the addition of *gfp* sequence homology to, *tbx-36* — here taken in the forward direction to account for the difference in genomic orientation as compared *tbx-35.*

ΔE2F site *tbx-36::gfp*

The *tbx-36::gfp* translational reporter derived from the above was used to make an additional construct which was identical in every way *except* for the deletion of a 28 bp sequence, a putative binding site for the E2F TF. The putative E2F binding site was deleted by amplifying the *tbx-36::gfp* construct in two portions, the first being the so-called 'upstream portion': the promoter of *tbx-36* (minus the 28 bp sequence and 213 bp upstream at the very start of the promoter), and the second being everything else, the so-called 'downstream portion': the small, remaining section of the *tbx-36* promoter downstream of the putative E2F binding site, the open reading frame of *tbx-36*, the

coding sequence of *gfp*, and the 3'UTR of *unc-54*. Crucially, to both PCR products, a NotI restriction site was added (3' in the case of the former and 5' in the case of the latter). Both products were digested individually, ligated together, run on an agarose gel and the correct final product extracted (i.e., the ligated 'upstream' and 'downstream' portions, as opposed to two 'upstream' portions, and so on) and verified by sequencing. The final product was cloned into pCR®-XL-TOPO vector to ensure it existed in sufficient quantities to inject. This final plasmid product generated pAW986.

drd-1.1p::gfp

pAW972 (*drd-1.1* transcriptional reporter) was made in two stages. Firstly, a 2.5 kb region of the *drd-1.1* promoter was amplified and cloned in the forward direction into pCR®-XL-TOPO vector. This construct was then digested with HindIII-HF and PstI-HF; the promoter-containing fragment was ligated into pPD95.75, which had been cut with the same enzymes.

drd-1.2p::mCherry

pAW974 (*drd-1.2* transcriptional reporter) was made in two stages. Firstly, a 1.7 kb region of the *drd-1.2* promoter was amplified and cloned in the reverse direction into pCR®-XL-TOPO vector. This construct was then digested with SphI-HF and PstI-HF; the promoter-containing fragment was ligated into pPD95.67, which had been cut with the same enzymes.

drd-1.3p::gfp

pAW976 (*drd-1.2* transcriptional reporter) was made in two stages. Firstly, a 2.1 kb region of the *drd-1.3* promoter was amplified and cloned in the reverse direction into pCR®-XL-TOPO vector.

This construct was then digested with HindIII-HF and PstI-HF; the promoter-containing fragment was ligated into pPD95.75, which had been cut with the same enzymes.

Ppdrdp::gfp — P. pacificus

pAW990 (the transcriptional reporter construct for the single Drd family orthologue present in *P. pacificus* with the WormBase identifier, WBGene00106782) was made by the amplification of a 1.6 kb region of the *Ppdrd* promoter from the purified gDNA of *P. pacificus (*with the addition of HindIII and SalI restriction sites). The PCR product was digested (with HindIII-HF and SalI-HF), without first being subcloned, and ligated into pPD107.94 in the forward direction, which had been cut with the same enzymes.

myrf-1p::gfp

pAW982 (*myrf-1* transcriptional reporter) was made in two stages. Firstly, a 4 kb region of the *myrf-1* promoter was amplified and cloned in the reverse direction into pCR®-XL-TOPO vector. This construct was then digested with HindIII-HF and SalI-HF; the promoter-containing fragment was ligated into pPD107.94, which had been cut with the same enzymes.

myrf-2p::gfp

pAW983 (*myrf-2* transcriptional reporter) was made in two stages. Firstly, a 2.5 kb region of the *myrf-2* promoter was amplified and cloned in the reverse direction into pCR®-XL-TOPO vector. This construct was then digested with HindIII-HF and SalI-HF; the promoter-containing fragment was ligated into pPD107.94, which had been cut with the same enzymes.

myrf-2.1::mCherry — C. brenneri

The *myrf-2.1::mCherry* C-terminal translational reporter derived from *C. brenneri* was made by the initial overlap-extension PCR of two distinct products flanked with regions of homology, a) *myrf-2.1* derived from *C. brenneri* and, b) the coding sequence of *mCherry* including the 3'UTR of *unc-54*. Taking the first, the open reading frame of *myrf-2.1,* with its native promoter in tow, was amplified and a 20 nt sequence (corresponding to that which follows from the start codon of *mCherry*) was added to the 3' end. Second, *mCherry,* with its associated 3' UTR, was amplified from pPD95.67 with a 20 nt homology region (corresponding to the 3' end of the *myrf-2.1* fragment) added to the 5' end. Then, these PCR products were fused together and the final product injected straight into animals (due to its large size) without first being cloned into a vector.

myrf-2.2::gfp — C. brenneri

The *myrf-2.2::gfp* C-terminal translational reporter derived from *C. brenneri* was made by the initial overlap-extension PCR of two distinct products flanked with regions of homology, a) *myrf-2.2* derived from *C. brenneri* and, b) the coding sequence of *gfp* including the 3'UTR of *unc-54*. Taking the first, the open reading frame of *myrf-2.2,* with its native promoter in tow, was amplified and a 20 nt sequence (corresponding to that which follows from the start codon of *gfp*) was added to the 3' end. Second, *gfp,* with its associated 3' UTR, was amplified from pPD95.75 with a 20 nt homology region (corresponding to the 3' end of the *myrf-2.2* fragment) added to the 5' end with the addition, too, of an SV40 nuclear localisation sequence (NLS) directly after the region of homology to *myrf-2.2,* and so lies adjacent to GFP in the final product. Then, these PCR products were fused together and the final product injected straight into animals (due to its large size) without first being cloned into a vector.

***tbx-36* feeding RNAi clone**

The *tbx-36* RNAi feeding construct was made by first amplifying a 850 bp fragment of *tbx-36* gDNA and cloning it into the pCR®-XL-TOPO vector. From here, the insert was cut using HindIII-HF and SalI-HF and inserted into the same restriction enzyme sites in the RNAi feeding vector L4440, generating pAW988.

***efl-1* feeding RNAi clone**

The *efl-1* RNAi feeding construct was made by first amplifying a 750 bp fragment of *efl-1* cDNA and cloning it into the pCR®-XL-TOPO vector. From here, the insert was cut using SpeI-HF and SalI-HF and inserted into the same restriction enzyme sites in the RNAi feeding vector L4440, generating pAW984.

***dpl-1* feeding RNAi clone**

The *dpl-1* RNAi feeding construct was made by first amplifying a 1000 bp fragment of *dpl-1* cDNA and cloning it into the pCR®-XL-TOPO vector. From here, the insert was cut using SpeI-HF and SalI-HF and inserted into the same restriction enzyme sites in the RNAi feeding vector L4440, generating pAW985.

***myrf-1* feeding RNAi clone**

The *myrf-1* RNAi feeding construct was made by first amplifying a 875 bp fragment of *myrf-1* cDNA and cloning it into the pCR®-XL-TOPO vector. From here, the insert was cut using HindIII-HF and SphI-HF and inserted into the same restriction enzyme sites in the RNAi feeding vector L4440, generating pAW987.

***myrf-2* feeding RNAi clone**

The *myrf-2* RNAi feeding construct was made by first amplifying a 720 bp fragment of *myrf-2* cDNA and cloning it into the pCR®-XL-TOPO vector. From here, the insert was cut using HindIII-HF and XbaI and inserted into the same restriction enzyme sites in the RNAi feeding vector L4440, generating pAW989.

RT-PCR analysis to verify RNAi knockdown

RT-PCR was performed on individual animals obtained from RNAi knockdown or L4440 control plates. One animal was lysed (x 5 animals per treatment) and worm lysate containing total RNA was used as a template for reverse transcription using the Thermo Fisher Scientific Maxima H Minus Reverse Transcription Kit as per manufacturers' instructions (with prior treatment of RNase-free DNase on the lysate directly). The product from this reaction was used as a template for a small PCR reaction (consisting of only 10 cycles) and the levels of cDNA (both of the gene of interest and a positive control, *act-1*) were assessed visually on an agarose gel where the absence of a band in the RNAi-treated samples was taken as evidence of effective knockdown.

DAPI staining of whole worms

Young adult worms were picked into a watch glass containing 10 µl of M9 buffer. 200 µl of 150 nM DAPI dissolved in 100% ethanol was added, followed by incubation in a dark chamber for two hours. Worms were then washed three times in M9 each time and left to soak in 1 ml of M9 overnight at 4 °C in a humid chamber.

DAPI staining of dissected gonads

DAPI staining of isolated germlines allows for the detection of a variety of morphological features, including: nuclear size; shape; and distribution, together with chromosomal structure and spatial organisation in greater resolution than that which can be observed from the fixation and staining of whole animals.

To achieve the DAPI staining of isolated germlines, clean worms were picked into 180 µl of 1x PBS containing 25 mM of levamisole and 10 mM of EDTA in a watch glass and decapitated using a round-edge scalpel blade. Then, 19 µl of 37% formaldehyde was added to the watch glass (to make a final concentration of 3.7%), and fixation was allowed to proceed for 60 minutes at room temperature in a humid chamber. Most of the liquid was then removed and to the tiny volume that remained, 500 µl of 100% methanol (pre-chilled to -20 °C) was added and incubated at room temperature for a further five minutes. Approximately 250 µl of the solution was then removed, followed by three washes with 1 ml of 1x PBS. DAPI was then added to a final concentration of 100 ng/ml and samples incubated for ten minutes, washed a further three times in 1 ml of 1x PBS, then mounted for observation on an agarose pad. 10 µl of VECTASHIELD anti-fade mounting medium was immediately added to the pad and then topped with a cover slip.

Brood size assays

Brood sizes were measured using 20 synchronised L4 worms of each strain. Worms were picked onto individual 55 mm NGM plates seeded with OP50 and transferred to a fresh plate every day

until egg laying had stopped. In all cases, plates were seeded only the day prior to ensure the bacterial lawn was not so thick as to miss any eggs when attempting to score the brood. Counting of the brood (unhatched eggs and young larvae) was performed on the plate which the mother had been on the previous day. Eggs were scored as dead if they remained unhatched 24 hours after the mother was removed, while larval death was recorded based solely on morphological and behavioural characterisation (immobility and translucence). All plates were scored in triplicate, and the mean and standard deviation was calculated for each.

Cuticle permeability assays

Cuticle permeability to DAPI was assayed as described (Xiong et al. 2017). In brief, L4 larvae were washed from plates with M9 buffer prior to staining with DAPI (5 μg/ml each in M9 buffer) for 15 minutes at room temperature with gentle agitation. Subsequently, worms were washed three times with M9 buffer, followed by fluorescence imaging. For microscopy, worms were mounted onto 2% agarose pads, anaesthetised with 3 μl of 20 mM levamisole and sealed with a coverslip before imaging on a Zeiss Axioplan 2 microscope. Samples were observed with a Zeiss Plan Neofluar 20×/0.50 Ph2 objective, images captured using a Zeiss AxioCam and the software AxioVision 4.8. DAPI accumulation was imaged at 100 msec exposure time.

Starvation viability assays

Members of the Drd family, either *drd-1.1*, *drd-1.2*, or *drd-1.3* (or a combination of them together), were knocked down by the injection of dsRNA into individual worms as described above. These animals were then placed on starvation NGM (made without bacto-peptone to prevent the growth of a bacterial lawn) to recover in lieu of regular NGM. The number of dauers and non-dauers per

plate were counted and the Dauer Formation Index (DFI) per plate calculated using the following formula: *DFI = No. of Dauers — No. of Non-dauers / No. of Dauers + No. of Non-dauers*. A more positive index value indicates a greater tendency to form dauers while a more negative index value denotes a preference for entering the reproductive developmental trajectory.

Heat-shocking for males

In the instance of not being able to obtain enough males to perform a genetic cross, perhaps if a high incidence of males (him) family mutation was not present in a particular strain, males were generated by 'heat-shocking' L4 hermaphrodites.

After picking six L4 hermaphrodites onto a plate (x 6 plates), animals were placed at 30 °C for 6 hours before recovery at 20 °C; the resultant progeny were examined for males.

Acknowledgements

The bioinformatics which formed the basis of much of this work relied on the use of a server owned and maintained by Peter Holland and Sebastian Shimeld (both at the Biology Dept., University of Oxford); without whom this project would, therefore, not have been feasible. Many strains used during this project were supplied by the *Caenorhabditis* Genetic Center (CGC) or were the kind gifts of Jonathan Hodgkin (Biochemistry Dept., University of Oxford).

CHAPTER 3

When a new gene learns old tricks: how a species-specific T-box gene came to play a role in the very early embryo

Introduction

T-box genes are a large family of TFs with diverse roles in animal development, and are often the linchpin of processes in vertebrate and invertebrate embryogenesis. Aspects of development in which their roles are now well-characterised include heart organogenesis (Plageman and Yutzey 2005), limb specification (Gibson-Brown et al. 1998), and the control of gastrulation (Showell et al. 2004). As one might expect from genes as critical to organismal development as T-box genes, they (as a family) exhibit minimal copy number variation throughout the animal kingdom, a statement which holds true, as far as is known, for all metazoan taxa (Sebé-Pedrós et al. 2013; Papaioannou 2014) with the exception of the *Caenorhabditis* genus. In the *Caenorhabditis* genus, characterised as part of this work, T-box genes are gained and lost at an unprecedented scale, indicative of their rapid evolution. With a view to understanding how such extreme dynamism in a developmentally important gene family is not merely tolerated, but apparently actively selected for, this chapter sets out to probe the developmental and evolutionary consequences — and drivers — of the taxon-specific radiation of T-box genes.

While known for orchestrating the transcriptional regulation of a range of developmental processes, T-box genes from across animal taxa are unified not only by a conserved DNA-binding domain but, in the case of many family members, their manifest predilection for posterior

patterning (Kispert et al. 1994; Singer et al. 1996; Kusch and Reuter 1999; Pocock et al. 2004). Their speciality in specifying similar posterior cell fates in invertebrates and vertebrates (meaning worm and mouse tails alike) cannot, intriguingly, be a product of evolutionary conservation because non-orthologous T-box genes regulate these processes across systems. However, what is known is that idiosyncratic modes of T-box evolution in primate lineages led to the loss of a tail in the ancestral hominoid that gave rise to modern humans thus leading to a major morphological, and developmental, difference between humans and non-hominoid primates today (Xia and Zhang 2021). In this ancestral hominoid, an *Alu* element inserted into an intron of the *TBXT* gene, also known as *Brachyury*, resulting in a splice isoform of the gene that was unable to specify a tail proper. It is impossible to conclude if this singular genetic change directly resulted in the split between these two now well-established primate lineages, but it nevertheless remains an illustrative example of the power of T-box evolution for shaping the diversity of life, and particularly the role of non-coding sequence evolution as a vehicle for that process.

Back in the realm of invertebrates, two paralogues, *tbx-8* and *tbx-9* — unique to the *Caenorhabditis* genus — are functionally redundant for their roles in the intercalation of posterior dorsal hypodermal cells, in muscle cell positioning, and in intestinal development in *C. elegans* (Pocock et al. 2004). However, excluding the roles of a highly conserved couple of T-box orthologues (Woollard and Hodgkin 2000; Kostas and Fire 2002), and with the obvious exception of *tbx-37* and *tbx-38* discussed in Chapter 1, little is known about the roles, regulation, and targets of other *Caenorhabditis*-specific T-box family members. Indeed, it is a point of note that investigative efforts in the worm have previously focussed on T-box genes found more broadly in other taxa. But like in any system, a mechanistic understanding of how T-box genes act requires an analysis of how their expression is controlled, the identification of their target genes, and an insight into how

different family members exert different effects (with respect to their regulation and targets). From this point of view, T-box paralogues in *C. elegans* provide only too many investigative avenues to pursue given their taxon-specific proliferation. Furthermore, although T-box paralogue pairs in *C. elegans* had not been subject to detailed investigation prior to this, it was appreciated all were expressed in the early embryo (Tintori et al. 2016).

Their early embryonic expression, combined with their phyletic dynamism, make T-box paralogues prime candidates as 'speciation genes' in the *Caenorhabditis* genus. But what precisely is meant by the term? For lack of a formal, agreed upon, definition, here they are defined as any genes which, in their divergence, contribute to the cessation of gene flow between populations over time. Causative loci that lead to speciation have been seldom characterised; this is due, in large part, to how speciation can only be studied retrospectively in most systems. However, naturally occurring populations of *C. elegans*, or wild isolates, diverging for only thousands of years, present a unique opportunity to capture processes which may precede new species divergence, thereby uncovering the genetic hallmarks associated with its onset.

The duplication and diversification of speciation genes is rarely considered, probably because the possibility of redundancy relationships existing between genes facilitating the formation of new species seems counterintuitive. This is exactly the problem which will be addressed in this chapter. Here, we will consider the potential of T-box paralogue divergence to directly catalyse the formation of new species in *C. elegans* populations today. In so doing, we will settle the case of precisely why there are so many T-box genes in a simple roundworm (as compared with more complex animals), and whether this is purely an instance of genetic profligacy, or bona fide developmental innovation.

Results

The taxon-specific proliferation of T-box genes in the *Caenorhabditis* genus is underpinned by their rapid evolution

It is axiomatic that the functional consequences of the particular evolutionary paths gene families take in certain lineages cannot be probed without first characterising those paths using phylogenetics. It is in such a vein that the maximum likelihood phylogram (built from all T-box domain-encoding proteins (Figure 3.1A) mined from the *Caenorhabditis* genus and suitable outgroups) in Figure 3.1B was built — and it draws two key conclusions. First, that the radiation of T-box genes is exclusive to the *Caenorhabditis* genus and not present in neighbouring lineages of nematodes (illustrated here using *P. pacificus*). Second, that the T-box proliferation observed within *Caenorhabditis* nematodes is irregular, that is to say it is not found in other genuses of similar size and age (illustrated here using the *Drosophila* genus).

As many as 51 T-box genes are found in *C. remanei*, with 40 and 36 being observed in *C. briggsae* and *C. brenneri*, respectively. As such, this makes *C. elegans* (of fame for its 21 T-box genes), something of an ordinary, if not below average, member of the genus. By contrast, a mere 9 T-box genes are present in *P. pacificus*, just slightly more than the three members of the *Drosophila* genus presented here. The relative stasis of T-box evolution outside of the *Caenorhabditis* genus is unsurprising, but this begs the question of whether the burgeoning of the T-box family in *Caenorhabditis* nematodes is based on their propensity not just to duplicate, but to diversify more generally. In any case, with a view to eventually drawing conclusions about the consequences of such mutational behaviour, this phylogenetic analysis merely demonstrates that T-box genes have duplicated significantly within the genus and because of that they *may* have historically acted as speciation genes. But their evolutionary potential in this regard is by no means proven — a myriad

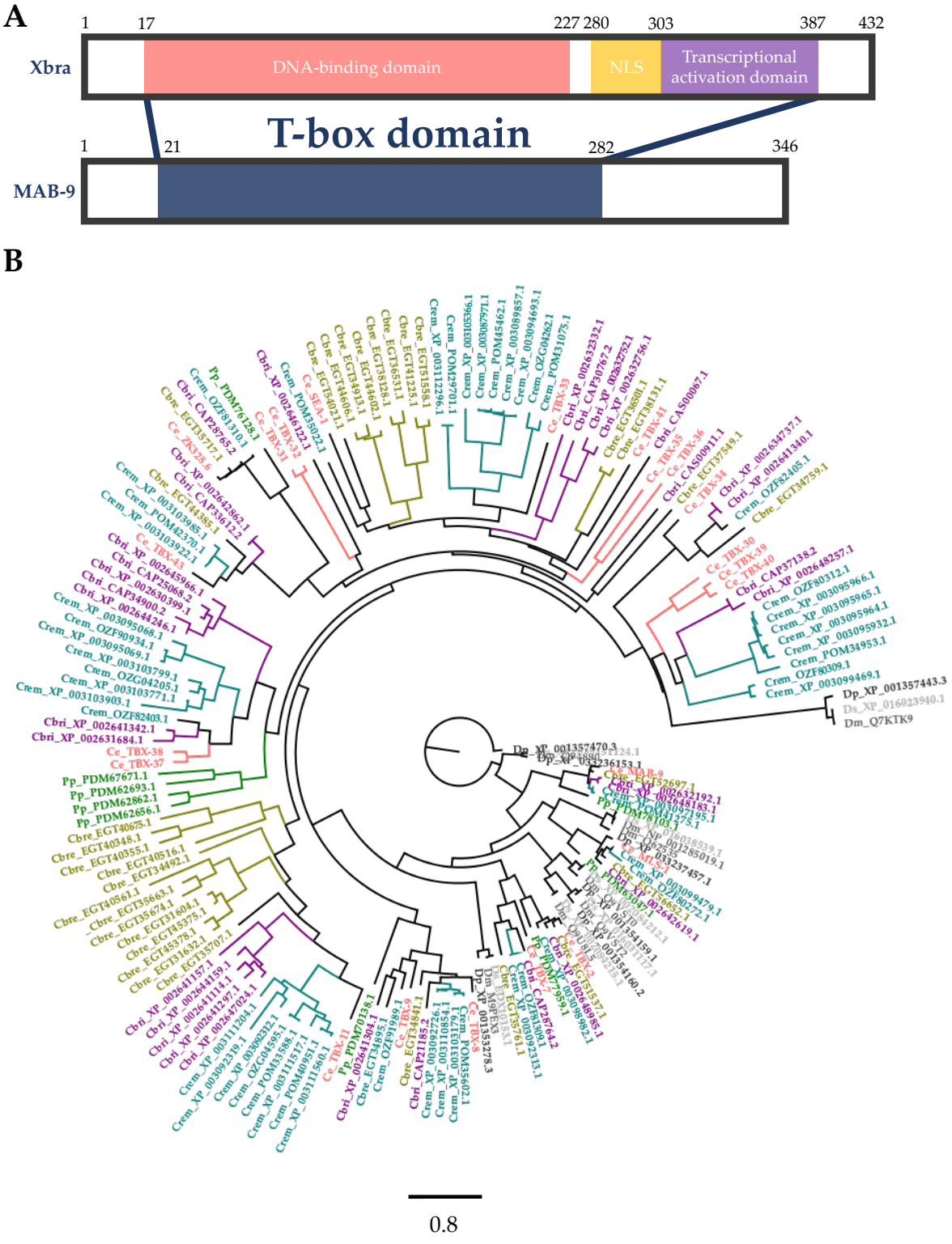


Figure 3.1. Phylogenetic analysis of the T-box family in the *Caenorhabditis* genus. (A) Schematic of T-box domain proteins, Xbra (*Xenopus laevis*) and MAB-9 (*C. elegans*). The domain itself is tripartite; consisting of the DNA-binding domain; a nuclear localisation sequence, and a transcriptional activation domain. Numbers correspond to amino acids. (B) Maximum likelihood phylogram of all T-box domain (predicted) proteins from the genomes of: *C. elegans* (Ce, pink), *C. brenneri* (Cbre, olive green), *C. remanei* (Crem, teal), *C. briggsae* (Cbri, purple) with a *P. pacificus* (Pp, bright green) outgroup, and further outgroups from the *Drosophila* genus: *D. melanogaster* (Dm), *D. pseudoobscura* (Dp), and *D. simulans* (Ds), in various shades of grey. Scale bar is substitutions per

site per million years. Because of the great size of this maximum phylogram, node values have been excluded, but a version of the tree is included in Appendix III which provides these values which may be referred to when evaluating the confidence level in any node or branch placement in the tree.

of genomic divergence (including copy number variants of some genes) is apparent between these species, most of which will have emerged after the speciation events took place. It follows, therefore, that we need to assess their evolvability in a more meaningful way.

To assess this properly, we need to systematically analyse the evolvability of the T-box family among geographically distant, diverging populations today. Figure 3.2 shows the results of exactly this exercise, illustrating the striking patterns of mutation accumulation in extant wild populations, or wild isolates, of *C. elegans*. All available isolate genomes were mined for this exercise, where the quality criteria of genome completeness was met. This amounted to approximately 300 wild isolates. When high and moderate effect mutations (these being: premature stop codons; missing start codons; frameshifts (both out-of-frame insertions and deletions); in-frame insertions and deletions; and substitutions which result in a change of amino acid to one of a different class) are mapped on to the gene tree of *C. elegans* T-box genes, it is revealed that while the family are prone to mutation, this is by no means true for all its members, to an equal extent. In fact, some quite striking patterns emerge.

The bottom portion of the tree contains the more conserved T-box genes (those found in other nematodes and / or animals), and in these, on the whole, high effect mutations have not been accumulated. Meanwhile, in the top portion of the tree which details the paralogy relationships between the T-box genes only found in *C. elegans*, not only are high effect mutations accumulated at a much higher rate generally, but they separate out reciprocally among members of a paralogue

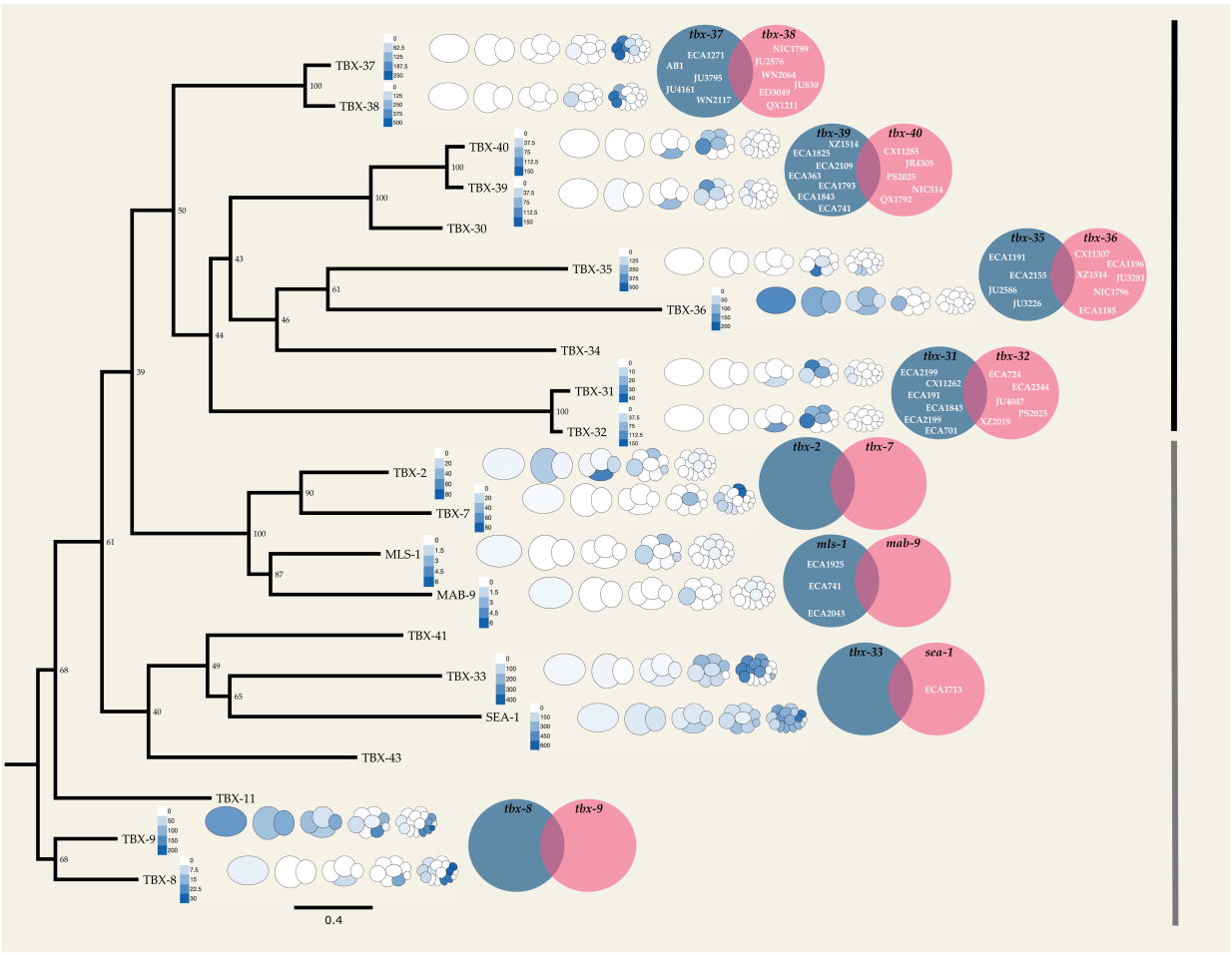


Figure 3.2. Divergence in the T-box repertoire of *C. elegans*. Gene tree of *C. elegans* T-box domains with the transcript abundances of those in paralogue pairs heat-mapped on to pictograms of the developing embryo (up to the 16-cell stage). Transcript abundance here is the mean FPKM (log2 (Fold Change)) where each has its own relative scale. All scRNA-seq data are derived from a dataset generated as part of Tintori et al. 2016. Beside the pictograms are Venn diagrams of the respective paralogue pair showing instances of High effect mutation accumulation in the pairs among wild isolates. The name of the isolate is provided and where it features in the circle indicates the presence of a High effect mutation (premature stop codon, frameshift, missing start codon, or splice isoform variant). So-called moderate effect mutations (small in-frame deletions or missense mutations that change amino acid class) are not mapped on to the Venn diagrams due to their abundance yet are included in Appendix IV. A total of 304 wild isolates were analysed in this experiment — the full library which had a fully sequenced genome which was publicly made available on CeNDR. All isolates which are not in the above diagram, nor in the extended supplementary version, may be taken to contain intact wildtype coding sequences of both paralogues. This is the majority of the 304 isolates known and characterised. Black line highlights the *C. elegans* -specific T-box genes, the grey line covers the more widely conserved T-box genes (among nematodes and other animal groups).

pair with one or other genes in the pair accumulating loss-of-function mutations, but *never* both.

Moderate effect mutations accumulated in this way between paralogous T-box genes are included

in Appendix IV as they proved too abundant (that is they arise too frequently) to include in the clear diagrammatical representation in Figure 3.2. For the following paralogue pairs: *tbx-31*/*tbx-32*; *tbx-35*/*tbx-36*; *tbx-37*/*tbx-38*; and *tbx-39*/*tbx-40*, many high and moderate effect mutations are accumulated in only one of the two members in the pair (albeit in different isolates between the pairs themselves), and all are expressed during early embryogenesis (defined here as up until the completion of the 16-cell stage).

In three of the aforementioned four T-box paralogue pairs, visualisation of publicly available scRNA-seq data (pre-analysed by authors in Tintori et al. 2016) implies that, due to being expressed in the same cells at the same time, they likely exhibit overlapping functionality (or possibly even functional redundancy) in embryonic development. Though long suspected to occur over some length of evolutionary time, reciprocal paralogue mutation accumulation in the way described has never before been shown. It is, therefore, foreseen how one or other of the paralogues — but not both — may be lost in a population; this is the very scenario facilitated by genetic redundancy. However, the exception is the *tbx-35*/*tbx-36* gene pair; because, while *tbx-35* is expressed in MS and descendants thereof, *tbx-36* is present far earlier in the one-cell embryo and in both AB and P_1 (though at lower levels by comparison). Of course, levels of *tbx-36* mRNA detected at such an early stage of development is inevitably maternally contributed as zygotic gene expression does not commence until the 4-cell stage. In any event, it is at this point simply important to acknowledge that *tbx-35* and *tbx-36* are expressed at very different stages of embryogenesis, though appear to have accumulated loss-of-function mutations in a reciprocal fashion in wild populations — two conclusions that are, as yet, hard to reconcile.

Exploring the expression patterns and the roles of the *tbx-35*/*tbx-36* gene pair

Although, this is only a problem to address if both *tbx-35* and *tbx-36* are actually involved in the development of the early embryo in the ways their expression patterns seem to imply. Only if they are will their mutational accumulation patterns among wild isolates require directly addressing. In order to assess this, the expression of both genes was confirmed *in vivo* by building C-terminal translational reporter constructs and their functions probed using knockout (for *tbx-35*) and knockdown (for *tbx-36*) approaches (Figure 3.3). The translational reporter constructs seem, to an approximation, to faithfully recapitulate the expression patterns of *tbx-35* and *tbx-36* gleaned from the scRNA-seq dataset. In Figure 3.3Ai and ii, TBX-35::GFP is present in EMS (the cell which divides to give rise to MS in which *tbx-35* was mRNA was detected previously) and its descendants. In Figure 3.3Aiv and v, TBX-36::GFP is observed in the 2-cell embryo (though was not detectable at other stages of early embryogenesis). However, it should be noted that despite being extrachromosomal arrays (i.e., existing in multicopy and thus overexpressed to variable extents between individual embryos), the level of TBX-36 is lower than one might expect from the scRNA-seq data, but this is most likely due to the general silencing of transgene arrays in the germline (Kelly et al. 1998; Reuben and Lin 2002).

It should not go unnoticed that the 2-cell embryo depicted (in Figure 3.3Av) that is expressing *tbx-36::gfp* (and thus overexpressing *tbx-36* with respect to wildtype levels) appears defective, notably in the ploidy of AB where evidence of an additional nuclear structure is seen. This we will return to below. With regards to the expression of *tbx-35* and *tbx-36* observed later in embryogenesis as depicted in Figure 3.3A, both TBX-35::GFP (iii) and TBX-36::GFP (vi) are detected during gastrulation, at around the 24-cell stage. Of course, this was not apparent from the scRNA-seq data because the transcriptome lineage was not traced beyond the 16-cell embryo.

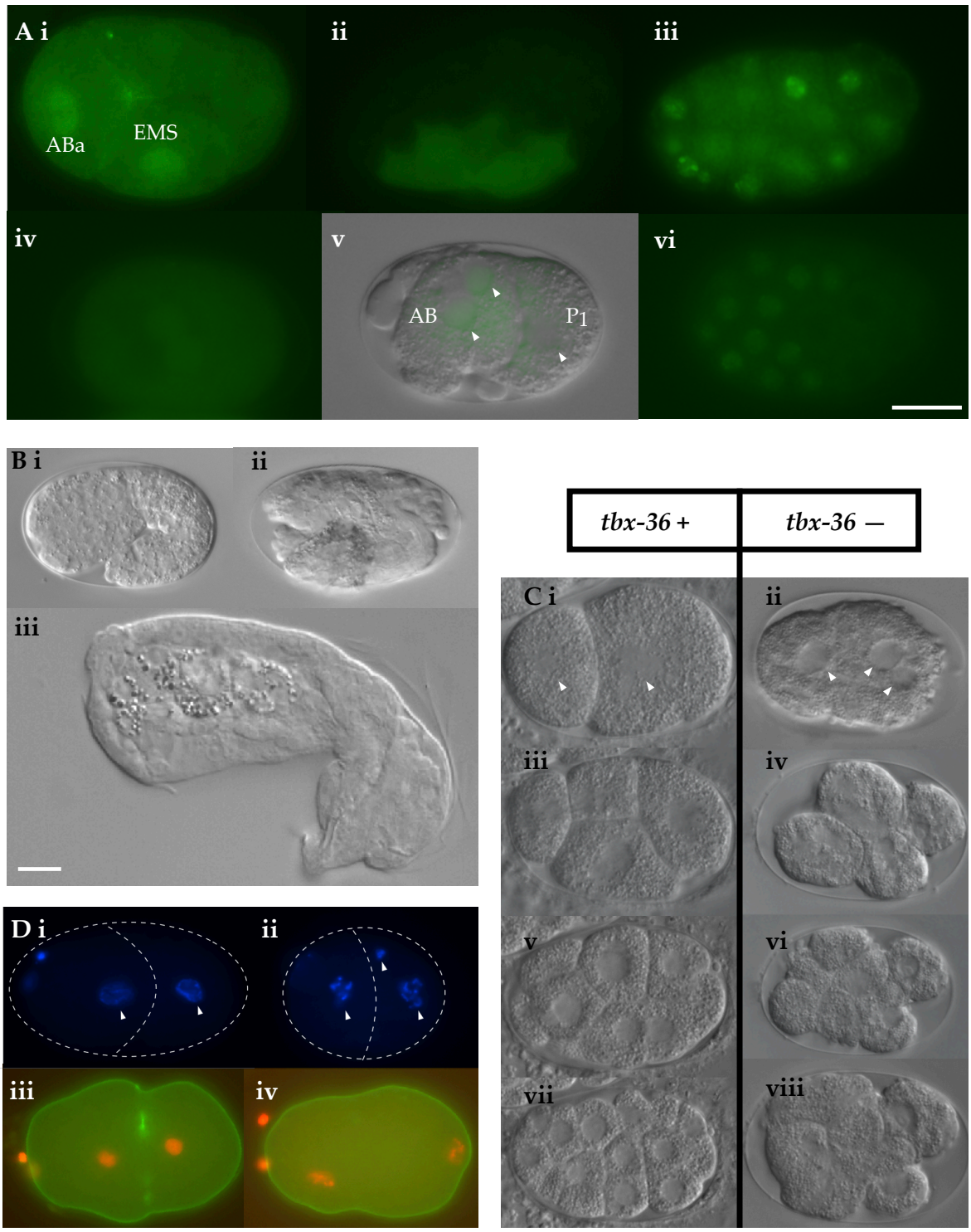


Figure 3.3. Expression and phenotypes of the *tbx-35*/*tbx-36* paralogue pair. (A) Expression patterns of *tbx-35* and *tbx-36* during early and mid embryogenesis using C-terminal translational reporter fusion constructs where the coding sequence *gfp* and the 3'UTR of *unc-54* were fused to *tbx-35* and *tbx-36* just prior to the stop codon by overlap-extension PCR. These constructs were injected as PCR projects into N2 and stable lines were generated and image. Panel (A i, ii, iii) show the expression of TBX-35::GFP in a 4-cell stage embryo, an 8-cell stage embryo, and a 24-cell stage embryo, respectively; meanwhile, (iv, v, vi) show the expression of TBX-36::GFP in a 1-cell stage embryo just prior to the first division, a 2-cell stage embryo where expression is visible in two nuclei of AB (two arrowheads), as well as in EMS (single arrowhead) and a 24-cell stage embryo,

respectively. (B) Phenotype of *tbx-35(tm1789)* mutant embryos, where (i) shows a WT and (ii) shows a *tbx-35(tm1789)* mutant embryo both at the 1.5-fold stage of mid-embryogenesis (during morphogenesis) approx. 425 minutes post-fertilisation — note the abnormality in the *tbx-35(tm1789)* mutant embryo which first becomes apparent at this stage, having completed gastrulation appearing as WT; (iii) reveals the final consequences of a lack of functional *tbx-35*, hatching as an inviable L1 with no muscle. (C) Phenotype of embryos in which *tbx-36* has been knocked down by RNAi via microinjection of dsRNA in the young adult mother, where (i, iii, v, vii) depict an embryo treated with no *tbx-36* RNAi (i.e., EV control) and (ii, iv, vi, viii) depict an embryo treated with *tbx-36* RNAi, all at either 9 (i and ii), 18 (iii and iv), 30 (v and vi), or 45 (vii and viii) minutes post-fertilisation. The same two embryos were lineaged in both cases. Panels (i and ii) at 9 minutes post-fertilisation use white arrowheads to mark nuclei in AB and P_1. Note the additional nuclear structure in P_1 in panel (ii). (D) Cellular basis of *tbx-36* KD embryos: (i) EV control embryo at the 2-cell stage, fixed and stained with DAPI — note the two nuclei as two fully formed circular structures with decondensed chromatin and the two polar bodies, one in focus, one out of focus, which is a natural consequence of their proper extrusion as described in Chapter 1. (ii) *tbx-36* KD 2-cell embryo with two nuclei present in AB and P_1 which have failed to decondense, in addition, note the irregular situation of one polar body which appears inside P_1 itself implying the temporal coordination of polar body extrusion is in some way abnormal in the absence of *tbx-36*. (iii) EV control embryo at the 1-cell stage (the pseudocleavage plane visible). (iv) *tbx-36* KD 1-cell embryo noting again the failure to have decondensed the chromosomes and the improper extrusion of the polar bodies, though this time they *both* sit outside the embryo indicating they have *both* been extruded prior to eggshell formation. In both (iii and iv) the outermost surface of the embryo is labelled using *nmy-2* (green) and the genetic material using H2B (red). Both scale bars (Avi and Biii) are 10 μm. All embryos in the above images were traced at 20 °C.

Of considerable assistance to this investigation was the work of Morris Maduro and colleagues who characterised the role of *tbx-35* in muscle development in 2006 (Maduro et al. 2006). So, it was already appreciated, and expected, that *tbx-35* is expressed in EMS and its descendants, and that it is necessary and sufficient for the development of muscle in *C. elegans*. Unique reporter constructs were generated and knockout experiments were performed here, therefore, so as to independently corroborate the conclusions made by Maduro and colleagues in relation to *tbx-35* function, and Figure 3.3A i, ii, iii and B satisfactorily demonstrates their accuracy. In fact, *tbx-35* is so essential to the specification of the MS blastomere that knockouts, like the one presented here, require balancing for the purposes of strain maintenance (in the case of the *tbx-35(tm1789) II* allele used both here and in Maduro et al. 2006). This essentially means that upwards of 90% of embryos display MS blastomere defects (Maduro et al. 2006). While *tbx-35(tm1789)* embryos complete

gastrulation appearing as wildtype, when morphogenesis commences, the absolute failure to specify the muscle cell lineage becomes apparent (Figure 3.3Bi and ii) and they subsequently hatch as inviable L1 animals, devoid of muscle and signs of life (Figure 3.3Biii).

By contrast, nothing was known about the role of *tbx-36* prior to this investigation. Figure 3.3C illustrates that upon its knockdown by RNAi, as might be expected from its presence at the same stage of embryogenesis, defects are apparent as early as the 2-cell stage. When *tbx-36* knockdown embryos are followed over the course of their development, defects in the first division are immediately evident in the ploidy of the daughter cells, where evidence of an additional nuclear structure is seen in P_1 (when comparing Figure 3.3Ci and ii), not unlike the embryos which overexpresses *tbx-36* over and above wildtype levels mentioned previously. With such a severe cellular impediment being present at such an early stage of embryogenesis, it is no wonder that subsequent divisions in *tbx-36* knockdown embryos are unrecognisable relative to the invariant lineage they would otherwise follow (Figure 3.3Ciii — viii). Inevitably, going on to characterise the basis of this phenotype requires the use of markers and the staining of DNA to elucidate the temporal incoordination of developmental processes.

In Figure 3.3Di and ii, it is revealed by the DAPI staining of DNA that *tbx-36* knockdown embryos fail to decondense chromosomes following their first division, and that what is normally a highly regular process of polar body extrusion displays signs of going awry with a definite polar body-like structure observed *inside* P_1 (when it should be resting at the anterior edge of AB prior to its destruction by the very same, as shown in panel (i)). Therefore presumably, it was an improperly extruded polar body that was recorded as an additional nuclear structure in P_1 in the above. Live imaging of labelled embryos (Figure 3.3Diii and iv) shows that it is the case that polar body

extrusion is improper in the absence of *tbx-36*, with the irregular extrusion of the second polar body leading here to its extrusion far earlier than it should have been, evidenced in its situation outside the embryo trapped just beneath the outermost layer of the eggshell — the same place (and focal plane) as the first polar body. And so, as the loss of *tbx-36* from the early embryo imparts a stochasticity on the should-be regular process of polar body extrusion, it is possible to say that the gene has developmental significance for pivotal chromosomal processes. Indeed, one might speculate these are the kind of processes for which divergence therein could lead to irreconcilable differences between populations of the same species.

An illustration of the kinds of high effect mutations accumulated in some wild populations of *C. elegans* are shown in Figure 3.4A where ECA1185 (with a premature termination codon (PTC) in *tbx-35*) and ECA1911 (with a PTC in *tbx-36*) are included owing to their particular interest to this investigation — both having accumulated high effect mutations in a reciprocal fashion yet both originating from tropical Hawaiian islands, albeit different ones. Now, in the knowledge of their developmental importance and of their distinct roles in embryogenesis, it seems even more unlikely that deleterious mutations of this sort would readily accumulate in ***either*** *tbx-35* ***or*** *tbx-36* among animals to no apparent detriment. But the essence of how this problem is addressed is implied in how this 'either or' rule is strictly adhered to by all extant wild populations of *C. elegans*. Of course, and maybe even unsurprisingly given their recent derivation, the answer lies in the overlapping functionality between these two paralogues.

Figure 3.4B characterises the basis for how proper development can reliably be achieved in the absence of true copies of either *tbx-35* or *tbx-36.* It was initially noted that upon knockdown at 20 °C in N2, the penetrance of defects present in *tbx-36* embryos was not complete, that is just less

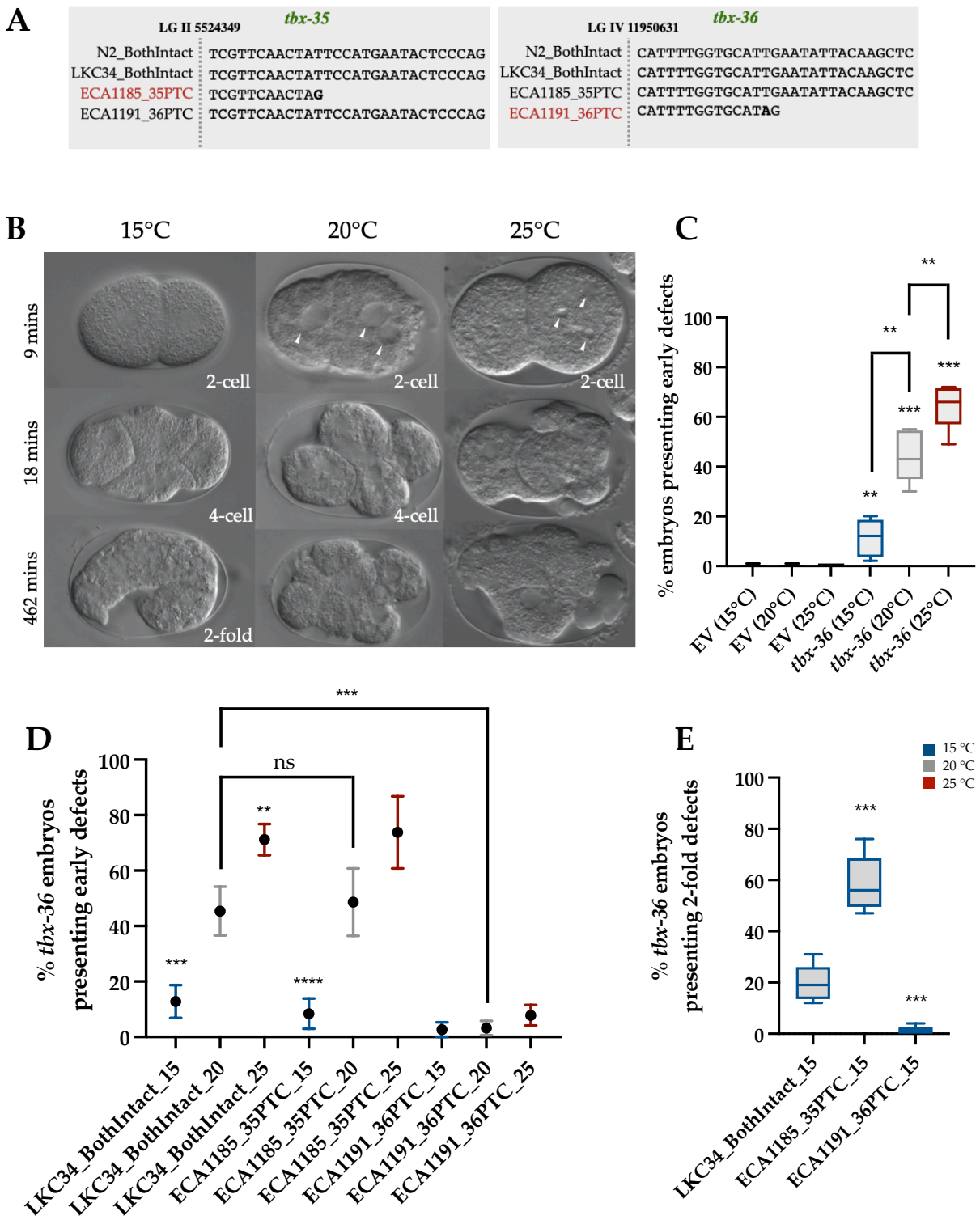


Figure 3.4. Temperature dependency of the *tbx-36* early phenotype. (A) Regions of LGII and LGIV containing the ORFs of *tbx-35* and *tbx-36* are shown in particular wild populations of *C. elegans*. Owing to the mutations they have accumulated (in either *tbx-35* or *tbx-36*) they are of interest to this investigation. LKC34 (isolated from Madagascar) is included another control isolate in addition to N2, having an intact version of both paralogues. ECA1185 has a PTC in *tbx-35* (TAG). ECA1191 has a PTC in *tbx-36* (TAG). The naturally occurring mutations that generate the PTCs in

ECA1185 and ECA1191 (base substitutions in both cases) are shown in bold. (B) Depicts representative images of embryogenesis in *tbx-36* knockdown embryos (in an N2 background) at three different temperatures: 15 °C, 20 °C, and 25 °C. As before, dsRNA corresponding to *tbx-36* was delivered by microinjection into a young adult mother. (C) Quantifies the phenotypic penetrance of the *tbx-36* early embryonic defects shown on knockdown in N2 using box plots (showing the mean with whiskers as the SEM) — upon *tbx-36* knockdown as compared empty vector (EV) control, at different temperatures. Comparisons are pairwise (between EV and *tbx-36* embryos at the same temperature) or as indicated, between temperatures, using bars. (D) Quantifies the phenotypic penetrance of the early embryonic defects shown in LKC34, ECA1185, and ECA1191 at three different temperatures using box plots, again showing the mean with whiskers as the SEM. Comparisons are pairwise as indicated using the bars, or between the same isolate at different temperatures where 15 °C and 25 °C are compared pairwise to the penetrance of defects at 20 °C in that same isolate. (E) Quantifies the number of mid-embryogenesis/ morphogenesis defects in *tbx-36* knockdown embryos that make it through early development (by virtue of being kept at 15 °C), but display defects akin to *tbx-35* knockouts at the 2-fold stage. Quantification is shown in using box plots, again displaying the mean with whiskers as the SEM. Black stars (****$P \leq 0.0001$, ***$P \leq 0.001$, **$P \leq 0.01$, *$P \leq 0.05$, ns$P > 0.05$) show statistically significant differences as determined by Student's *t*-tests. All sample sizes, in each of (C, D, E), exceed 50 different embryos for each isolate at each temperature. In panels (C, D, E), the colour of the box plot indicates the temperature the embryos were exposed to, either 15 °C (blue), 20 °C (grey), or 25 °C (red).

than half displayed the kinds of severe, early defects described above. At first, this was thought to be due to the inherent shortcomings of performing an RNAi experiment — that not all *tbx-36* transcript being was abolished in the process. However, as we will go on to confirm later, the efficacy of *tbx-36* RNAi is impressively high (especially in comparison to *tbx-35* RNAi knockdown which is considerably harder to achieve). And so, the reason behind the partial penetrance of the *tbx-36* phenotype remained a mystery until it was serendipitously discovered that upon a change in temperature to 15 °C, a lower percentage of *tbx-36* knockdown embryos displayed early defects, and that inversely, when the temperature was increased to 25 °C, a higher percentage of *tbx-36* knockdown embryos displayed such defects (quantified in Figure 3.4C). However, even at 25 °C, 20% of embryos treated with *tbx-36* RNAi develop as wildtype and hatch as viable L1 larvae, but this is more likely to be attributable to the incomplete knockdown of *tbx-36* as while knockdown is effective, it is not, unsurprisingly, 100%. We have no reason, based on this work and the work of

others, to suppose that the RNAi machinery is compromised at lower temperatures, or more efficacious at higher temperatures (Fire et al. 1998).

Stated for reasons of clarity and completeness, it is shown embryos treated with empty vector (EV) do not display defects in embryogenesis, at the lower or higher temperatures tested. And it should be noted in passing that no comparable phenotypic temperature dependence was observed in *tbx-35(tm1789)* mutants (in this investigation or in Maduro et al. 2006 (data not shown)).

The next pertinent question to ask is what becomes of the *tbx-36* knockdown embryos that develop as wildtype during early embryogenesis at 15 °C? When followed throughout their embryogenesis, it is seen that the gastrulation of *tbx-36* deficient embryos appears as wildtype but then at morphogenesis, approximately 430 minutes post-fertilisation, defects are displayed at the 2-fold stage (Figure 3.4B) akin to those seen in *tbx-35* knockouts (Figure 3.3B). It is not knowable if the phenotype of *tbx-36* knockdown embryos in mid-embryogenesis is similarly temperature-dependent as the early role appears to be (as at higher temperatures, such a high proportion of effective *tbx-36* knockdowns display early defects that these effectively mask what *could be* going wrong in later development).

Next, apposite to understanding the role that the *tbx-35*/*tbx-36* gene pair play in speciation today, is to access the function of *tbx-36* in mid-embryogenesis in our populations of interest. And so, in Figure 3.4D, it is shown that not only is an isolate with an intact *tbx-36* (and a degenerated *tbx-35*) as dependent on the gene for early embryogenesis as N2 is, but that it also exhibits comparable temperature dependency of the phenotype. This is isolate ECA1185. By the same token, in Figure 3.4E, it is seen that this same isolate with a PTC in *tbx-35*, ECA1185, also requires *tbx-36* later in

development, with morphogenetic defects observable at the 2-fold stage **at an even higher penetrance** than is recorded in N2. This strongly suggests that ECA1185 relies more heavily on *tbx-36* for morphogenesis than N2 does, which of course has *tbx-35* for the very same role. Meanwhile, *tbx-36* knockdown has no effect on either early or mid embryonic processes on the isolate studied here, ECA1191, that has accumulated a PTC in the very same (though has retained a faithful copy of *tbx-35*). The isolates ECA1185 and ECA1191 are compared to LKC34, the latter having an 'intact' or faithful copy of both *tbx-35* and *tbx-36* (with no high effect mutations, such as PTCs, in either paralogue). Though unlike N2, LKC34 has only recently been isolated from the wild (from Madagascar), and so is unlikely to display the kinds of phenotypic artefacts that have been associated with model systems cultivated in a laboratory environment for many decades (Sterken et al. 2015).

The obvious logical progression would be to investigate the role of *tbx-35* in the same trio of isolates as in the above: LKC34, ECA1185, and ECA1191. Alas, the poor efficiency of *tbx-35* knockdown by RNAi, among all wild populations, was the stumbling block for why this could not be done (data not shown). This would be a considerable issue were it not for the previous work reporting the same role for, and expression of, *tbx-35* as described here, and so it is safe to assume that where a population of *C. elegans* has a true copy of *tbx-35* (with no loss-of-function mutations accumulated therein), the gene has the same role in MS blastomere specification as it does in N2. Inversely, where a wild population has a degenerated copy of *tbx-35* (featuring deleterious mutations such as a PTC), it may be assumed — because it simply cannot produce a functional gene product — that *tbx-35* has not the same or similar role in MS blastomere specification and so would not have a phenotype at the 1.5-fold or 2-fold stages of embryogenesis. But for completeness, the expression of *tbx-35* was determined in ECA1185 and ECA1191 as proof of this

principle. In both cases, C-terminal translational reporters were generated from each of ECA1185 and ECA1191 and injected into their respective isolate so as to observe their native expression pattern. Figure 3.5A shows that *tbx-35::gfp* is indeed expressed in ECA1191 during the the early stages of gastrulation, just as in N2, though *tbx-35::gfp* derived from ECA1185 and expressed in the same is not (see Figure 3.5B). The latter, undoubtedly, is due to the presence of the PTC in the ORF of *tbx-35*, meaning that *in vivo,* this C-terminal reporter is not actually expressed.

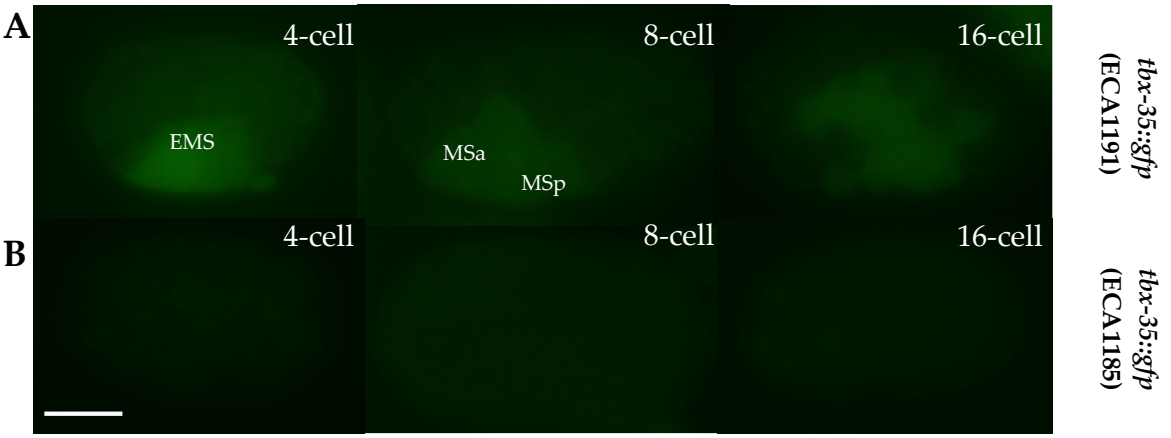


Figure 3.5. Expression of *tbx-35::gfp* in ECA1185 and ECA1191. C-terminal translational fusion reporters of TBX-35::GFP were built for ECA1185 and ECA1191 just as for N2 described previously. In short, the promoter and the coding sequence of tbx-35 — derived from the particular isolate — was fused to the coding sequence *gfp* and the 3'UTR of *unc-54* just prior to the stop codon by overlap-extension PCR. These two constructs were then injected back into their native isolates and the progeny of the stable lines were imaged. (A) Expression of *tbx-35::gfp* (derived from ECA1191) in ECA1191 embryos at the 4-cell (32 mins post-fertilisation), 8-cell (74 mins post-fertilisation), and 16-cell stage (135 mins post-fertilisation). (B) Embryos containing *tbx-35::gfp* (derived from ECA1185) in ECA1185 embryos at the 4-cell (32 mins post-fertilisation), 8-cell (74 mins post-fertilisation), and 16-cell stage (135 mins post-fertilisation). Scale bar is 10 μm. All embryos in the above images were lineage traced at 20 °C.

Returning to the role of *tbx-36* in mid-embryogenesis described above, this would certainly appear to explain its expression during early gastrulation, although, just as for *tbx-35*, it is not known how it is activated at this stage (being switched on again when the embryo comprises of around 24 cells). It is certainly hard to deny the possibility that given the extent to which the expression patterns and roles of the two paralogues appear to resemble one another, phenotypically speaking,

at this stage that they are most likely being activated as part of the same transcriptional response. Though, if this is true, it has to be a distinct regulatory phenomenon to the idiosyncratic expression of *tbx-36* in the very early embryo because the latter is not a feature of *tbx-35* functionality; or, to put it another way, the early embryo is either an expression domain which *tbx-35* has lost or it is one which *tbx-36* has gained as part of its neofunctionalisation. In either event, this must be due to a regulatory interaction unique to the *tbx-36* locus not found to occur in *tbx-35*. It is with this in mind that the regulatory divergence between *tbx-35* and *tbx-36* was investigated.

New genes, new functions, and their new regulators: the regulatory evolution of the *tbx-36* locus differentiates it from its paralogue

A comparative analysis of predicted TF binding sites was performed on the non-coding regions of both *tbx-35* and *tbx-36* (their putative promoter elements and introns) using JASPAR: a database for TF binding profiles. In fact, as included in Appendix V, the same analysis was performed on the other T-box paralogue pairs exclusive to *C. elegans* (the aforementioned *tbx-31*/*tbx-32*; *tbx-37*/*tbx-38*; and *tbx-39*/*tbx-40*), and it was found that *tbx-35* and *tbx-36* shared the fewest predicted TF binding sites in common. In and of itself this does not provide a great deal of insight as to their regulatory divergence because this result, no doubt, is just a proxy for their non-coding sequence dissimilarity, and there is no telling from such a cursory inspection which TF binding sites are of biological relevance without appropriate subsequent verification. Indeed, most predictions of this nature are unlikely to be bona fide TF binding sites. But at the very least, this analysis provides a platform for the interrogation of possible regulatory interactions. One such candidate emerged in this way — when inspecting the hits only found in *tbx-36* and not *tbx-35* — and that is the putative binding site for E2F, or EFL-1 as it is known in *C. elegans*, in the promoter of *tbx-36*.

The modENCODE ChIP-seq database was used to initially verify the legitimacy of the predicted EFL-1 binding site in the promoter of *tbx-36*. In its design, the modENCODE ChIP-seq project profiled a number of TFs critical for *C. elegans* development and performed repeat analyses for each TF at each of the worm's canonical developmental stages, though in reality these represent broad swathes of developmental time encompassing many discrete developmental events. The stages characterised by modENCODE are thus: the early embryo; the late embryo; L1 larvae; L2 larvae; L3 larvae; L4 larvae; and young adult worms. The seven EFL-1 datasets were mapped onto the *tbx-36* locus, and only one yielded a significant hit — a peak stretching 28 bp in the promoter of *tbx-36* in young adult worms (Figure 3.6A). Now, before characterising this regulatory interaction further, it is important to first take a slight digression to explain a peculiarity of E2F transcriptional regulation in *C. elegans*.

E2F transcriptional regulation really needs no introduction. Suffice to say, it forms part of the core transcriptional axis on which the eukaryotic cell cycle is controlled, namely the G1/S transition by mediating the transcriptional activation of many genes. But E2F does not operate in isolation. In all systems in which it is present, E2F functions as a heterodimeric TF with its binding partner, DP (or DPL-1 in *C. elegans*), and is negatively regulated by the binding of Retinoblastoma/Rb (or LIN-35 in *C. elegans*). This tripartite transcriptional regulatory complex could be summarised as being inhibitory of ectopic cell division in somatic tissues, most likely by repressing the expression of target genes. However, there is an additional flavour of E2F signalling in *C. elegans,* one which is totally independent of Rb/LIN-35 (Chi and Reinke 2006). This distinction, it turns out, mirrors a separation in developmental time whereby many genes required for oogenesis and early embryogenesis are largely activated by *efl-1* and *dpl-1* independent of *lin-35* activity, and genes

regulated by the two later in development (that is, the 4-cell stage of embryogenesis onwards), require the kind of repressive transcriptional input imparted by *lin-35* in the way outlined.

The material temporal difference in E2F transcriptional regulation is molecularly determined by the possession of either one of two different TF binding motifs in the *cis*-regulatory regions of target genes themselves. The former suite of E2F-regulated genes is commonly associated with the sequence 5′-TTCGCGCC-3′, and the latter and Rb-independent variety of E2F regulation, is commonly associated with the sequence 5′-TTTTCCAG-3′. It is, however, to be expected that owing to the promiscuous nature of TF binding profiles, total obedience to these motifs is unlikely in all circumstances that E2F binds to its targets as in any case, binding site degeneracy is thought to be the rule, rather than the exception, for TFs more generally (reviewed by Copley 2016). And so, it is for two reasons that we can say with a degree of certainty that *lin-35* is not required for the transcriptional regulation of *tbx-36*, even though E2F/DP, as per the story so far, may very well be. First, the presence of *tbx-36* in the very early embryo, while possibly regulated by *efl-1* and *dpl-1*, cannot logically be associated with the activity of *lin-35* because the involvement of *lin-35* in the E2F transcriptional ensemble occurs only after the formation of the 4-cell embryo (Chi and Reinke 2006), at which point *tbx-36* is no longer expressed. The second and most cogent of the two reasons presented here is that there is no ChIP-seq peak for LIN-35 in the *cis*-regulatory regions of *tbx-36* (including, especially, any that overlay with the putative 28 bp binding site for E2F). This bears out for all developmental stages included in the modENCODE ChIP-seq enterprise (data not shown).

Multiple different approaches were used to empirically assess the putative regulation of *tbx-36* by E2F/DP. The first, and perhaps most convincing, is the deletion of the 28 bp binding site for E2F in the C-terminal reporter construct for *tbx-36*. It is worth spelling out that the deletion introduced

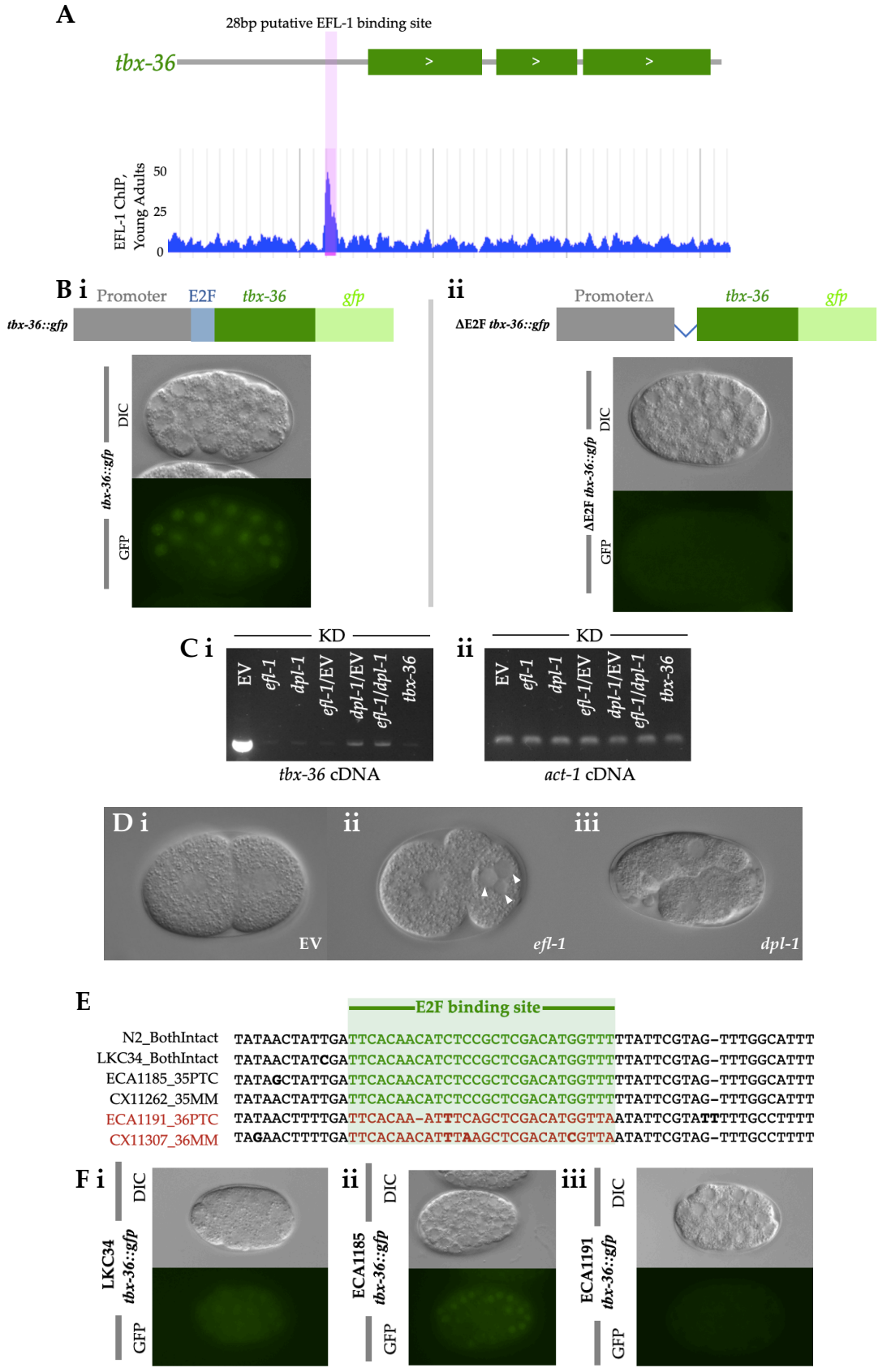


Figure 3.6. Transcriptional regulation of *tbx-36* by E2F/DP. (A) Putative binding site for E2F in the promoter of *tbx-36* as identified by visualising modENCODE ChIP-seq data. E2F ChIP-seq data is available for all life stages ('early' (i.e., 4-cell to 16-cell) embryo, late embryo, L1, L2, L3, L4, young

adult, and then young adult male) — the only peak of significance at the *tbx-36* locus was as shown in young adults. Data were pre-analysed by the modENCODE enterprise such that peaks were just extracted and their notion of significance was taken as true. Vertical pink rectangular band highlights the 28 bp region in the promoter of *tbx-36* that E2F purportedly binds to. (B) (i) The C-terminal *tbx-36::gfp* translational reporter construct (already described) with the putative 28 bp E2F binding site intact — fluorescence is observed at the 24-cell stage. (ii) The mutated C-terminal *tbx-36::gfp* translational reporter construct wherein the putative 28 bp E2F binding site has been deleted by the addition of NotI sites that flank the binding site. Following digestion with NotI (and the deletion of the 28 bp), the ligation of promoter with the coding sequence of *tbx-36* made a construct lacking the 28 bp binding site. This was injected into N2 and the stable lines made were imaged — fluorescence is not observed at the 24-cell stage. (C) RT-PCR to ascertain levels of *tbx-36* following different RNAi treatments (either *efl-1, dpl-1,* and *tbx-36*) as compared Control. All RNAi was delivered by the microinjection of the corresponding dsRNA into the young adult mother. 30 adult worms from each RNAi treatment (themselves containing newly fertilised zygotes and very early embryos) were used in the subsequent RT-PCR analysis. It is important to mention that a period of at least six hours went by before total RNA was extracted (and converted to cDNA) from injected animals, so as to allow time for RNAi knockdown to take full effect. (i) Gel image of the RT-PCR analysis on various knockdown embryos (*efl-1, dpl-1, efl-1/EV, dpl-1/EV, efl-1/dpl-1,* and *tbx-36*) detecting levels of *tbx-36* mRNA relative to EV control. (ii) Levels of *act-1* cDNA were detected on the exact same samples as in (i) as a positive control for mRNA integrity. (D) (i) EV control embryo at the 2-cell stage. (ii) *efl-1* knockdown embryo at the 2-cell stage, note how this image depicts defects appear similar to those in *tbx-36* knockdown embryos with the ploidy of P_1 being irregular (Figure 3.3). (iii) *dpl-1* knockdown embryo at the 2-cell stage. (E) Truncated alignment of a region in the promoter of *tbx-36* containing the 28 bp binding site for E2F among particular wild isolates of interest to this investigation (with mutations accumulated in the ORFs of either *tbx-35* or *tbx-36* — either multiple missense (MM) or PTC mutations). Within the region, a number of mutations (nucleotides in bold) can be observed only in those isolates with mutations in the ORFs of *tbx-36* (ECA1191 which has a PTC in its copy of *tbx-36* and CX11307 which has multiple missense mutations in in its copy of *tbx-36* — both shown in red. The E2F binding site of those isolates with mutations in the ORFs of *tbx-35* (ECA1185 and CX11262) have accumulated no mutations, relative to N2 and LKC34, in their E2F binding sites within the promoter of *tbx-36*. (F) Early gastrulation in particular wild isolates of interest to this investigation expressing their natively derived *tbx-36::gfp* fusion constructs. Note, all these constructs contain both the 28 bp E2F region (whether it possesses a functional binding site or not), and the 213 bp upstream region of the promoter): LKC34 (i) and ECA1185 (ii) visibly contain TBX-36::GFP yet ECA1191 (containing a PTC in the ORF of *tbx-36*), does not (iii). All embryos in the above images were traced at 20 °C.

corresponds *exactly* to the E2F ChIP-seq peak in the promoter of *tbx-36*; that is to say that only those 28 bp were deleted, no more and no less. It will eventually become apparent as to why elaborating on how this construct was made sheds light on the unique evolutionary trajectory taken by *tbx-36*, and so it will be briefly explained below.

The full-length *tbx-36::gfp* translational reporter construct was re-amplified to make two PCR products: A) the 5′ portion of the *tbx-36* promoter that flanks one side of the putative 28 bp E2F binding site and B) everything that sits at the other side of the construct, that is 3′, to the putative 28 bp E2F binding site (being the small remaining portion of the promoter, the ORF of *tbx-36*, *gfp*, and the 3′UTR of *unc-54* which was included in the original construct to ensure the efficient termination of transcription). In their re-amplification, NotI restriction sites were added only to the 3′ and 5' ends of the products A and B, respectively, such that they could both be digested and ligated together to create the full-length *tbx-36::gfp* translational reporter construct again, albeit crucially, this time, without the putative 28 bp E2F binding site. However, product A (containing most of the *tbx-36* promoter), left off 213 bp at the very start of the original full-length construct because within this 213 bp were two NotI restriction sites. As such, there were in total, three varieties of the *tbx-36::gfp* translational reporter generated as part of this investigation: the longest being that which contained the full-length promoter with both the 213 bp upstream region and the putative 28 bp E2F binding site, the second being the same except for the absence of the 213 bp upstream region; and the third being the shortest missing both the 213 bp upstream region and the putative 28 bp E2F binding site. The constructs were sequenced and found to be identical in every other way except for the differences highlighted.

We will return to a comparison of the two longest constructs, but for now, it should be made clear that all images of the *tbx-36* expression shown thus far have been taken of embryos carrying the second construct, containing the putative 28 bp E2F binding site but **not** the 213 bp upstream region. This second construct, in any case, acts as a suitable control for the ΔE2F *tbx-36::gfp* construct, and schematised depictions of the two are shown in Figure 3.6B. At the 24-cell stage, expression of *tbx-36* is clearly shown (in panel (i)) and, by contrast, it is clearly absent in those

embryos carrying the ΔE2F *tbx-36::gfp* construct (shown in panel (ii)). Three independent lines were generated in the case of each, and all appeared as in the images shown in either (i) or (ii), this is to say that the visibility of the constructs (or lack thereof) is not due to the inherent variability in copy number of the extrachromosomal transgene arrays. It is thus possible to say, from this result alone, that the regulatory interaction between E2F and the *tbx-36* locus appears true, nay, more than that, that E2F is actually necessary for the expression of *tbx-36*. This was further verified in two other ways, this time accounting for the role of the binding partner of EFL-1, DPL-1.

Based on insights gleaned thus far, it is to be predicted that in the absence of *efl-1* and / or *dpl-1*, *tbx-36* is not able to be expressed. In a similar vein, perhaps one might expect that in the absence of *efl-1* and / or *dpl-1*, early embryonic defects resembling those observed in *tbx-36* knockdowns are also seen. But there is a major problem with experimentally pursuing this to its logical conclusion, and that is that the RNAi knockdown of either *efl-1* or *dpl-1*, it was known prior to this investigation, has catastrophic phenotypic consequences (Chi and Reinke 2009; Petrella et al. 2011). With *efl-1* and *dpl-1* being required for processes as fundamental as germline development and morphogenesis, let alone the negative regulation of Ras signalling, in their absence, animals are inviable (Chi and Reinke 2009). Especially as knockdown of the complex results in sterility, it was essential to time the delivery of the dsRNAs just right: too early and no eggs would be laid at all, too late and there would be no effect on the very early embryo and therefore the expression of *tbx-36*. As such, the dsRNA was synthesised and delivered by gonadal injection which, irrespective of the demands of this particular experiment, is the most reliable method for RNAi in *C. elegans* and presents the least experiment-to-experiment and animal-to-animal variability (Hammell and Hannon 2016). The dsRNA enabling *efl-1* and / or *dpl-1* knockdown was injected into young adult worms (containing between five to seven retained, fertilised embryos) around the same time as the

delivery of the dsRNA in the case of *tbx-36* knockdown embryos shown previously (though it is worth noting, anecdotally, that *tbx-36* RNAi can be delivered earlier by comparison). It was necessary to use suitable dilution controls where, for instance, the dsRNA for *efl-1* knockdown was diluted to half the concentration to facilitate meaningful comparison with the *efl-1* and *dpl-1* double knockdown embryos.

Performed in the way described above, the effect of *efl-1* and / or *dpl-1* knockdown on the levels of *tbx-36* in young adult animals and early embryos is shown in Figure 3.6C. It is technically impossible to harvest sufficient quantities of mRNA from individual newly fertilised zygotes to perform RT-PCR analysis on and so, in lieu, 30 adult worms from each RNAi treatment (themselves containing newly fertilised zygotes and very early embryos) were used in this analysis. It is important to mention that a period of at least six hours went by before total RNA was extracted from injected animals, so as to allow time for RNAi knockdown to take full effect. Following extraction, total RNA from relevant knockdown samples was converted to cDNA and split into two — on one set levels of *tbx-36* were detected (the results of which feature in panel (i)), and on the other set levels of *act-1* were detected as a normaliser and to ensure that at no point in the process did the RNA become degraded (the results of which feature in panel (ii)). It is seen that levels of *tbx-36* are, as expected, decreased upon *efl-1* and / or *dpl-1* knockdown, this being the case when they are knocked down both singly and together. With regard to the combinatorial knockdowns, it is reassuring that the effect of diluting the dsRNA does not compromise the efficacy of the treatment, as per this readout anyway. In fact, in all the *efl-1* and / or *dpl-1* RNAi scenarios, *tbx-36* decreases to similar levels as it appears to when the gene itself is knocked down, the faint band being just about visible in this lane leading us to infer that *tbx-36* knockdown is not absolute, though levels of transcript are still reliably diminished.

The third and final way in which the regulation of *tbx-36* by E2F/DP was probed was the characterisation of early embryonic phenotypes in the absence of *efl-1* and *dpl-1*, as logically, one might expect they would resemble those in *tbx-36* knockdown embryos because in essence the molecular outcome is the same (i.e., a deficiency of *tbx-36*). However, the caveat with this approach is enormous, because, in the way described, *efl-1* and *dpl-1* are thought to be master regulators of early embryogenesis, and so, in their absence, a whole gamut of early acting genes are not going to be expressed, not merely *tbx-36*. Nevertheless, the early embryonic phenotypes of *efl-1* and *dpl-1* were sought, at this point in the investigation, as a peripheral piece of confirmatory evidence, and so they are included in Figure 3.6D. When comparing Figure 3.6Di and ii, it is seen that *efl-1* RNAi knockdown embryos (like the one depicted) have a ploidy defect in P_1, strikingly similar to the aforementioned 2-cell embryos that develop with insufficient *tbx-36*. The 2-cell embryo in Figure 3.6Diii displays the outcome of *dpl-1* knockdown with what looks more akin to polarity defects reminiscent of those seen in mutants of the PAR family, such as *par-2* and *par-3* (Goldstein and Macara 2007). Indeed, defects of this sort were manifold in both *efl-1* and *dpl-1* knockdown embryos, but further investigation of these fell beyond the scope of this work because they do not pertain to the role, or evolution, of *tbx-36*. Although, the interested reader may care to note that there are significant E2F ChIP-seq peaks (in young adults) in the *cis*-regulatory regions of *par-2*, *par-3, and par-5,* the lattermost gene being, it is valuable to note for later, linked to *tbx-36* on chromosome IV.

But once again, our focus must be brought back to the relevance of this to the role of *tbx-36* in the potential speciation process we see happening before our very eyes. Accordingly, the 28 bp region corresponding to the E2F binding site in the promoter of *tbx-36* (characterised above in N2) was

identified and analysed in the wild populations of *C. elegans* relevant to this investigation. The results of these extracted, aligned sequences are shown in Figure 3.6E. It is seen how the isolates which have accumulated mutations in *tbx-36* (CX11307 and ECA1191) have a highly divergent E2F binding site. Presumably, the extent of the divergence seen in the E2F binding sites of CX11307 and ECA1191 means this 28 bp region no longer facilitates E2F-binding.

To verify this logically plausible scenario — that the naturally occurring mutations in the E2F binding site results in a loss of *tbx-36* expression — translational reporters were each built for LKC34, ECA1185, and ECA1191 in the same fashion as described previously. The former two isolates, it is worth re-mentioning, contain intact *tbx-36* regulatory and coding sequences. The latter isolate does not. Given we have shown that E2F is required to activate *tbx-36* transcription, it is safe to assume that the divergence observed in the 28 bp E2F binding site in the promoter of *tbx-36* in ECA1191 is what causes the lack of TBX-36::GFP expression (made, naturally, from the native ECA1191 *tbx-36* sequence). This is in contrast to those populations with intact *tbx-36* regulatory and coding sequences, LKC34 and ECA1185, which show visible levels of TBX-36::GFP during gastrulation. It is important to note that all three of these constructs were built using the full-length promoter, i.e., including the 213 bp upstream region. Why this matters pertains to the nature of this 213 bp upstream region, and so too the regulatory evolution of *tbx-36*, so it will now be dealt with.

The nuanced regulatory requirements for *tbx-36* relate to *how* it emerged as a principal modulator of the early embryo

The early defects observed in the event of *tbx-36* overexpression bear a striking resemblance to embryos that develop in an absence of *tbx-36*, implying that having too much or too little of the

gene is detrimental to the fidelity of developmental processes. In itself this is not surprising; any transcriptional regulator is itself exposed to regulation, and in turn those regulators are regulated, and so on. This is how developmental processes are coordinated with total accuracy. And so, while the analogy is often made between TFs and simple on/off switches, in reality, their action is more like that of a dial with an ability to fine-tune the expression of downstream genes dependent on their levels in space and time. This, in turn, is dependent on their interactions with activators, as well as repressors. In this way, the gene regulatory networks that underpin development can be thought of as hanging in a very delicate balance indeed, reliant on an orchestra of transcriptional regulators that, in the event any one should become dysregulated, could lead to the whole symphony of development becoming out of tune.

This is an intrinsic property of TFs it is true, but how does it relate to the evolution of *tbx-36* as compared with *tbx-35*? We have already found that *tbx-36* has acquired the regulation of the EFL-1/DPL-1 heterodimer leading to its expression in the very early embryo. And, as *tbx-35* is not under the regulatory command of E2F signalling (shown in the absence of an E2F binding site from its regulatory regions), it is not expressed at such an early stage — so far, so established. But it was previously outlined how, in the processes of generating the ΔE2F *tbx-36::gfp* construct, a slightly longer *tbx-36::gfp* construct was generated as a control (with the E2F binding site still present), but that these were both derived from a longer *tbx-36::gfp* construct with an additional 213 bp at the very start of the promoter which did not feature in either of the other two (this being schematised in Figure 3.7A). The longest version was the very first *tbx-36::gfp* construct made, and all transgenic lines carrying it as an extrachromosomal array did not show expression in the early embryo, and displayed only minimal fluorescence, by comparison to the second *tbx-36::gfp* construct, at the 24-cell stage of embryogenesis (the latter depicted in Figure 3.7B). It is therefore surmised that within

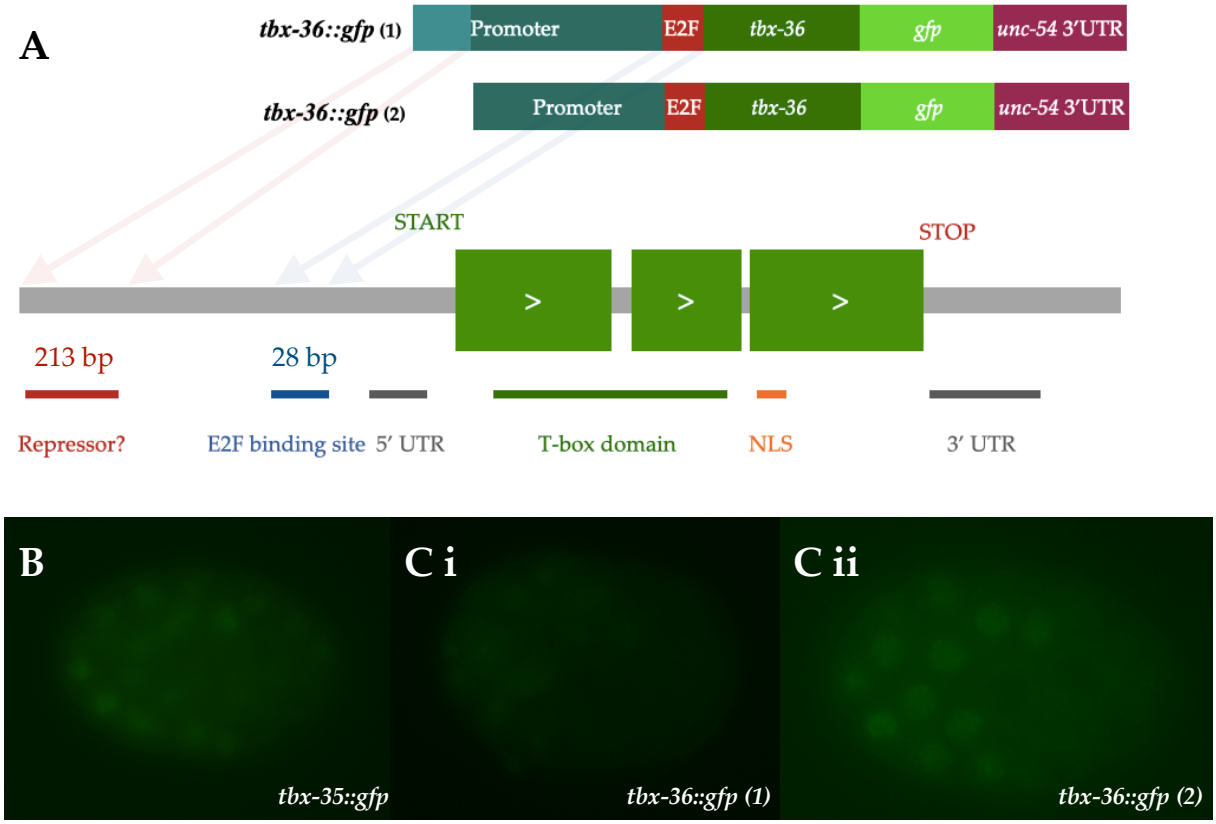


Figure 3.7. Expression modulation of *tbx-36* in embryogenesis (A) Schematic representations of the two C-terminal translational reporters of *tbx-36::gfp* generated as part of this investigation. In the first, there are 213 bp at the 5′ end of the construct (the very start of the promoter) that are not present in the second and are therefore hypothesised to contain some kind of repressive element (for *tbx-36* itself) owing to the brightness of the same construct (the second) when these 213 bp are not present. Schematics of the constructs are not to scale with respect to region size. (B) Depicts a typical 24-cell stage embryo expressing *tbx-35::gfp* for comparison with those presented in (C) where (i) is typical 24-cell stage embryo expressing *tbx-36::gfp* (version 1) and (ii) is a typical 24-cell stage embryo expressing *tbx-36::gfp* (version 2). All embryos in the above images were kept at 20 °C.

this 213 bp region lies a repressor for *tbx-36* itself. Attempts were made using both the JASPAR database and the modENCODE ChIP-seq library to determine what the identity of this repressor could be, but the reality is that less is known (or at the very least curated) about the repressive and suppressive inputs into transcriptional processes as compared their activation. That being said, it is still possible to conclude that there is a response element for some kind of repressor in the promoter of *tbx-36*; but why should it *need* one? The answer to this lies in the origins of *tbx-36*, and this of course relates back to the gene duplication event from which it was derived.

It is a casual observation that all the T-box paralogues in *C. elegans* are small genes (even according to the standards set by the worm genome itself) all sitting at less than 6 kb each. But many, including *tbx-35* and *tbx-36*, are smaller still at under 4 kb, owing to their lack of long introns. This is a more than subtle clue that at some point in their evolutionary history, retrotransposition was a driving force in T-box paralogue generation. Although, as depicted in Figure 3.8A, many paralogue pairs are tightly linked (albeit the pairs themselves distributed) on the various chromosomes, indicating that such a pair were duplicated in tandem. Notably, this is the case for all the T-box paralogues exclusively found in *C. elegans*, all except, that is, for *tbx-35* and *tbx-36.* It is seen that *tbx-35* sits on LG II, while *tbx-36* sits on LG IV. It is tempting to suggest that residing in an entirely different chromosomal location means that *tbx-35* and *tbx-36* reside in unique regulatory environments, and this we will come back to. For now, it is first necessary to establish which of the two locations (if either) is the location of the progenitor of the pair before formulating conclusions about which may have landed among new, unfamiliar, regulatory regions.

In order to settle this, small-scale synteny analysis was performed for those regions around *tbx-35* and *tbx-36* in *C. elegans*. While used more extensively elsewhere in this thesis in subsequent chapters, it should be acknowledged that on the whole, synteny analysis is often unhelpful when it comes to deciphering gene evolution among nematode species because their genomes are notoriously changeable even over relatively short periods of evolutionary time (Stevens et al. 2019). Therefore, only very closely linked genes to *tbx-35* and *tbx-36*, with clear orthologues in other *Caenorhabditis* species, were used in this analysis. Their locations were compared to the locations of the *tbx-35*/*tbx-36 -like* species-specific paralogues in *C. brenneri* and *C. briggsae* (there being no clear single-copy 'ancestral' gene in any of the species analysed as part of the

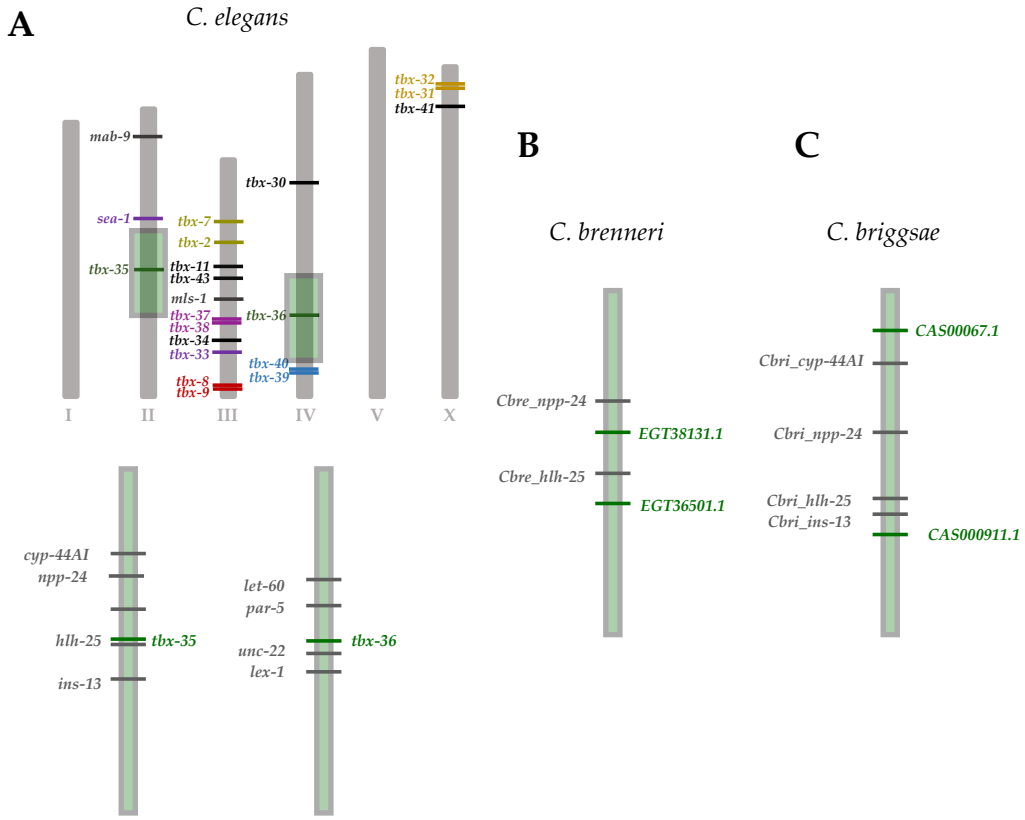


Figure 3.8. Synteny among T-box paralogues in three *Caenorhabditis* species. (A) The chromosomal location of all 21 T-box paralogues are shown in *C. elegans*. Paralogue pairs are colour-coded making it apparent that a considerable number are still closely linked tandem duplicates. The regions of LG II and LG IV on which *tbx-35* and *tbx-36* reside are expanded with syntenic markers which provided genes by which to ascertain syntenic conservation in the two other species. (B) Expanded region of *C. brenneri* genome with two species-specific paralogues that were the closest possible *C. brenneri* homologues to the *tbx-35*/*tbx-36* as per the phylogenetic analysis in Figure 1B. The two depicted are shown to be in a region syntenic with LG II in *C. elegans* on which *tbx-35* resides. (C) Expanded region of *C. briggsae* genome with two species-specific paralogues that were the closest possible *C. briggsae* homologues to the *tbx-35*/*tbx-36* as per the phylogenetic analysis in Figure 1B. The two depicted are shown to be in a region syntenic with LG II in *C. elegans* on which *tbx-35* resides, just as for *C. brenneri*.

phylogenetic analysis shown previously, suggesting that this gene has been lost in other species or that it has given rise to a host of divergent lineage-specific paralogues therein). Figure 3.8B and C portray these *tbx-35*/*tbx-36* *-like* species-specific paralogues in *C. brenneri* and *C. briggsae*, respectively, and it is seen that synteny is somewhat conserved between the regions these paralogues lie in and LG II in *C. elegans*, the home of *tbx-35*. No syntenic markers, close or distant,

were matched between the region housing *tbx-36* on LG IV and these regions in *C. brenneri* and *C. briggsae*. It is, therefore, most likely that the location of *tbx-35* in *C. elegans* corresponds to the original chromosomal location of the progenitor of the *tbx-35*/*tbx-36* gene pair in a now extinct species, meaning that it was *tbx-36*, not *tbx-35*, that found itself amidst new *cis*-regulatory apparatus, or at least as far as has been proven thus, in a new chromosomal location.

It is not possible, given the comparable size of all T-box paralogues in *C. elegans*, to discern whether *tbx-36* was derived from retrotransposition or was tandemly duplicated and then subsequently transposed to LG IV. But the precise fact of the event here is not of particular significance; rather, what is of importance to assess is whether the act of finding itself on a new chromosome enabled *tbx-36* to follow a distinct evolutionary path not accessible to *tbx-35*.

The coevolution of genes and their *cis*-regulators remains something of an enigma, and what resembles a 'chicken or egg' phenomenon. It is plausible that one way a gene's existing regulatory commitments (and, it follows, its function) can be overhauled is by entering into a foreign genomic environment where a host of unfamiliar regulatory elements are already present. And, the more of any one kind of regulatory element there is in a given region of the genome, the more likely the gene entering into such an environment will assume that kind of regulation. It was with this thinking that the possibility that *tbx-36* entered into a so-called E2F 'hotspot' on LG IV was investigated.

Using the modENCODE ChIP-seq database across all life stages, the number and location of the putative E2F binding sites were mapped onto regions of chromosome II and IV, equal distances either side of *tbx-35* and *tbx-36*, respectively. The results of this analysis are shown in Figure 3.9.

Figure 3.9. Characterising the E2F regulatory landscape around the *tbx-36* locus. All modENCODE ChIP-seq peaks for E2F binding are mapped onto the *tbx-36* (top) and the *tbx-35* (bottom) loci, on LGIV and LGII, respectively. Both *tbx-35* and *tbx-36* are shown centrally, circled, in green. The two types of E2F transcriptional regulation are separated out, with the E2F regulation of post-embryonic genes denoted by binding sites in grey (Rb-dependent E2F activity), and the unique binding sites — containing a subtly different binding motif — for the E2F regulation of genes required for oogenesis and early embryogenesis (which is Rb-independent) in yellow. The results of this analysis are quantified in the table on the right, where the factor of gene density is taken into account when calculating the frequency of sites (given by the quotient: net sites / gene density) at the two comparably sized loci. Gene density is essential to discount for as presumably the presence of more genes in a region of the same size increases the likelihood that any binding site, not just for E2F, will be found.

Owing to their unique binding motifs corresponding to their temporal distinction in development, it was essential to differentiate between E2F regulation associated with LIN-35 and that which is not. The quantification of the E2F binding sites in the regions surveyed, even when taking into account gene density in the two regions of the same size, reveals that around the *tbx-36* locus on chromosome IV, there is an abundance of E2F response elements relative to the analogous region on chromosome II around the *tbx-35* locus. Indeed, there seems to be a particularly tight cluster of E2F binding sites, both those that are associated with LIN-35 and those that are not, directly

adjacent to *tbx-36*. It is noteworthy that genes which lie in this selfsame region include *par-5* — an essential gene required for the first asymmetric cell division during embryogenesis, and *let-60* — an essential Ras gene controlling the switch between vulval and hypodermal cell fates during vulval organogenesis.

In the knowledge amassed above, it is possible to propose, with a degree of confidence, that in landing on LG IV, by whatever mechanism, *tbx-36* acquired E2F regulation. And, given we now know E2F/DP are required for the expression of *tbx-36*, it is likely that had *tbx-36* landed in a different region of the genome altogether, bereft of E2F regulation, *tbx-36* would have probably not become a key regulator in the very early embryo.

Discussion

The premise of this chapter was to assess the significance of the extreme dynamism exhibited by the T-box family in the *Caenorhabditis* genus, and in doing so, investigate the potential role of copy number variation more broadly in the generation of new species by understanding its consequences for fundamental biological processes. Irrefutably, it is impossible to close such a complex, multifaceted, matter with even the most careful examination of just one gene family. But that is not to say that insights cannot be provided on the issue by the kind of analysis performed here on the T-box family.

T-box genes, it was casually noticed prior to this investigation, are prone to duplication in the *Caenorhabditis* genus. But before this work, their dynamism had not been systematically characterised, and so, the first priority was to validate what was a speculative anecdotal

observation using a phylogenetic approach. It transpires that, even more so than previously thought, T-box genes are gained and lost at an unprecedented rate in the genus, when compared to neighbouring invertebrate lineages or even beyond, with only 15 to 20 in most vertebrates (Sebé-Pedrós et al. 2013; Papaioannou 2014).

But is it possible that the burgeoning of T-box genes has evolved to merely counter the extent of their loss which has occurred simultaneously? Indeed, members of the T-box family that not only remain static with respect to their conservation throughout the animal kingdom, but are also indispensable to members thereof such as *Brachyury* and *Eomes*, are simply absent in *Caenorhabditis*. Thus, it may be that what drove the rapid expansion of T-box genes in a relatively short period of evolutionary time (approximately over the last 80 million years) was the requirement for the transcriptional regulation offered by T-box genes more generally but not provided by the usual contenders (owing their loss) in the *Caenorhabditis* lineage.

As it was found that underpinning their dynamism at the phylogenetic level was their vulnerability to mutate more broadly — with wild isolates accumulating deleterious mutations in a reciprocal manner among recently derived paralogue pairs — it can be settled that the evolvability of the T-box family in the *Caenorhabditis* genus is remarkable. At the population level, the accumulation of mutations in the new paralogues is evidence for their rapid divergence — though not the kind that would likely lead to the gain, or indeed loss, of new functionality but rather, eventually, the kind of copy number variation there is precedent for in the phylogenetic analysis of the family. This prediction is justified by how all the paralogue pairs that accumulate mutations in the way described most likely functionally overlap in their roles in the developing embryo (as gleaned from their expression as per the scRNA-seq data), so, one gene in the pair may

be lost in a population, but this would have no detrimental effect on development as the other would be retained — proper development could be achieved with only one. In other words, the evolutionary loss of one gene in the pair would not be correlated to a gain or loss of functionality at the phenotypic level. It would, however, presumably limit the ability of those populations to faithfully cross to produce fertile offspring, as genetically speaking, the T-box repertoires in the genomes of these populations would be hard to reconcile (in the context of chromosomal pairing during meiosis, for example). But this is just an untestable theory, though quite possibly the destiny of these populations in thousands, if not millions, of years to come.

The exception to this was seen to be the *tbx-35*/*tbx-36* paralogue pair, which accordingly then became the focus of this investigation. With seemingly divergent roles necessary for the viability of the early embryo, it was hard to foresee how either *tbx-35* or *tbx-36* were dispensable to certain wild populations, that was until, despite initial appearances, their supposed overlapping functionality in muscle specification was revealed, for *tbx-36*, upon a change in temperature. Of interest to geneticists and developmental biologists alike, the intrigue of *tbx-36* will be elaborated on in two respects.

Taking the first, with respect to the genetics at play here: — genetic or functional redundancy is defined as the instance in which "two or more genes are performing the same function such that inactivation of one of these genes has little or no effect on the biological phenotype" (Nowak et al. 1997). Overlapping functionality, while lacking a similarly formalised definition, is often considered (and in practice applied) as falling somewhere on the sliding scale between total non-redundancy and complete functional redundancy.

It is no surprise that gene duplication events are a prolific source of redundant genes, with paralogues being more likely to exhibit any degree of functional overlap than evolutionarily unrelated genes for the most obvious of reasons. When inactivated individually, *tbx-35* and *tbx-36* exhibit similar defects in mid-embryogenesis (only observable at 15 °C in the case in the latter), and in the case of *tbx-35*, to a complete phenotypic penetrance, thereby rendering mutations in it inviable. But to a geneticist, the idea of having paralogous genes which are clearly engaged in the same developmental process, yielding comparable biological phenotypes upon their inactivation, and yet do not exhibit genetic redundancy in its classical sense, is counterintuitive to say the least. Though the two are markedly different, it follows they have to be, else the inactivation of either while still in the presence of the other would not produce a phenotype. The proposed explanation for this is that the ancestor of the gene pair was a master regulator of muscle specification, capable of regulating a whole set of genes that *tbx-35* and *tbx-36* now activate between them for proper muscle specification. Ergo, when either *tbx-35* or *tbx-36* are inactivated on their own, because they are each responsible for regulating different genes required for the same process, similar phenotypes become apparent, though in truth they are derived from different, yet associated, molecular processes going awry (such as the failure to express different subsets of muscle specification genes). But irrefutably, *tbx-35* is more involved in muscle specification than its paralogue, and so it may be said that *tbx-35* has remained more faithful to the role of its progenitor than *tbx-36* which has only a remnant of this functionality (i.e., a partially penetrant phenotype). All this is to say that the evolution of the *tbx-35*/*tbx-36* gene pair in *C. elegans* is reminiscent of specialisation as an evolutionary fate of duplicated genes, though it is unfortunate that, owing to the species-specific duplications of this gene in neighbouring members of the *Caenorhabditis* genus, this cannot be conclusively proven. The experiment that would be performed to provide a decisive answer here would be an attempt to rescue to phenotypes following the loss of the *tbx-35*/*tbx-36*

gene pair in *C. elegans*, but as neighbouring lineages have lost the single-copy ancestor or rather duplicated it independently, this is not feasible.

We will now turn to the developmental significance of the temperature-dependency of *tbx-36*. From the data presented, it can be stated with a degree of confidence that while the role of *tbx-36* in mid-embryogenesis is most probably not temperature-dependent, its role in the very early embryo, as described here, is. The former may be concluded from the lack of *tbx-36* expression volatility at the 24-cell stage at all temperatures tested as part of this investigation, though of course phenotypically, it is only possible to access the role of *tbx-36* at this later stage at 15 °C. The latter, though, is patently clear with the early phenotype temperature-dependent to such a degree that at 25 °C, the (albeit weak) pleiotropy of *tbx-36* is totally masked.

The role of temperature in developmental processes is rarely interrogated. It is the very nature of being in a laboratory environment that such environmental variables can be controlled, and as temperature is rarely at the fore of investigations in developmental biology, all temperature-dependent phenomena, should there be any, are inevitably missed. Though temperature-dependency, of the kind described here, must exist in nature; it is axiomatic that adaptation precludes the ability to survive across a wide range of environmental conditions. In any case, as it stands, we have barely got a handle on the extent of the robustness that needs to be instilled in gene regulatory networks, cellular processes, and morphogenesis, to enable them to proceed faithfully irrespective of the temperature at which they sometimes find themselves *having* to occur.

Undeniably, temperature is a major determinant of the reaction rates of enzymes, and it is for this reason that the regulation of processes on which all life hinges is invariably temperature-

dependent — to varying degrees. The most recent rumination on this topic, while not exclusively focussed on the fidelity of embryogenesis, posited that any biological process, be it at the cellular, physiological, or ecological level, is reliant on enzyme catalysis and is thus affected by temperature (Arroyo et al. 2022). In developing their general theory of temperature dependency using chemical kinetics and statistical physics integrated at the systems level, Arroyo and colleagues suggest that most biological processes (unlike the individual enzymes that collectively catalyse them) have evolved to occur sub-optimally, that there is an inevitable trade-off between the maximum value of a dependent quantity and breadth of performance. This was put by the authors, pithily, as the "jack-of-all-temperatures" but "a master of none" model, and is evolutionarily the most stable strategy so long as the components involved in a given process do not function optimally at the same temperature. So, it must be acknowledged that there is a de facto regularity to embryonic development, but this relative invulnerability is underpinned, more generally, by a universal law that remains at least partially empirically undeciphered. The answer is, in its nature, very unlikely to be associated with one specific gene class, but instead rest with the intrinsic physical properties of all biological molecules.

But what light does this have to shed on *tbx-36*? Here, it is proposed that *tbx-36* may well activate the expression of a gene, or genes, with catalytic activity — that such a gene product, or more likely products, are themselves functionally optimal at higher temperatures to instil robustness in early embryogenesis, should it be required to occur at them. These catalytically active gene products would, as has been indirectly shown, be required for the chromosomal events that happen in the early embryo (chromosome decondensation after the first asymmetric division and polar body extrusion). But upon the inactivation of *tbx-36*, these gene products are not expressed and they are then unable to support the curation of early embryogenesis — of its associated

chromosomal events — at higher temperatures. It is part and parcel of this hypothesis, based on the aforementioned model and data presented of early embryogenesis at lower temperatures, that there would be another suite of catalytically active gene products which function optimally at lower temperatures to regulate early embryogenesis, though these are clearly not regulated by *tbx-36*. All this means that early embryogenesis as a *process* is not temperature-dependent in the wild type scenario (as is known and has been shown here), but that the process is cumulatively reliant on a concert of gene products that are, which is permissible so long as they are not all functionally optimal at the same temperature, in the interests of robustness.

With this post hoc justification, arguably the contradiction in the data only grows. How can it be that some isolates dispense with either *tbx-35* or *tbx-36*? The working model for why this might be is shown in Figure 3.10 and is explained as follows. It is not beyond the realms of possibility, and especially in light of the evidence, that in isolates that have accumulated loss-of-function mutations in *tbx-35*, *tbx-36* is capable of assuming its role. The two are closely related specific-specific paralogues, and so are unsurprisingly similar in sequence. The notion here is not unfamiliar; one only has to look to the vertebrate Hox genes for ample proof of phenotypic rescue in the event of paralogue loss (Hunter and Prince 2002; Zhu et al. 2017a; Zhu et al. 2017b). Evidence supporting this part of the model is in the increased penetrance of 2-fold defects in the event of *tbx-36* knockdown in one such isolate, ECA1185, as compared both N2 and LKC34. Conversely, in those isolates for which the loss of *tbx-36* has proven similarly benign, it is likely that its role in muscle specification during mid-embryogenesis (or rather what is left of it), has been assumed by *tbx-35* — an entirely plausible scenario given the minimal involvement of *tbx-36* in muscle specification. The logically stickier issue to resolve, however, is how early embryogenesis proceeds in the absence of *tbx-36*.

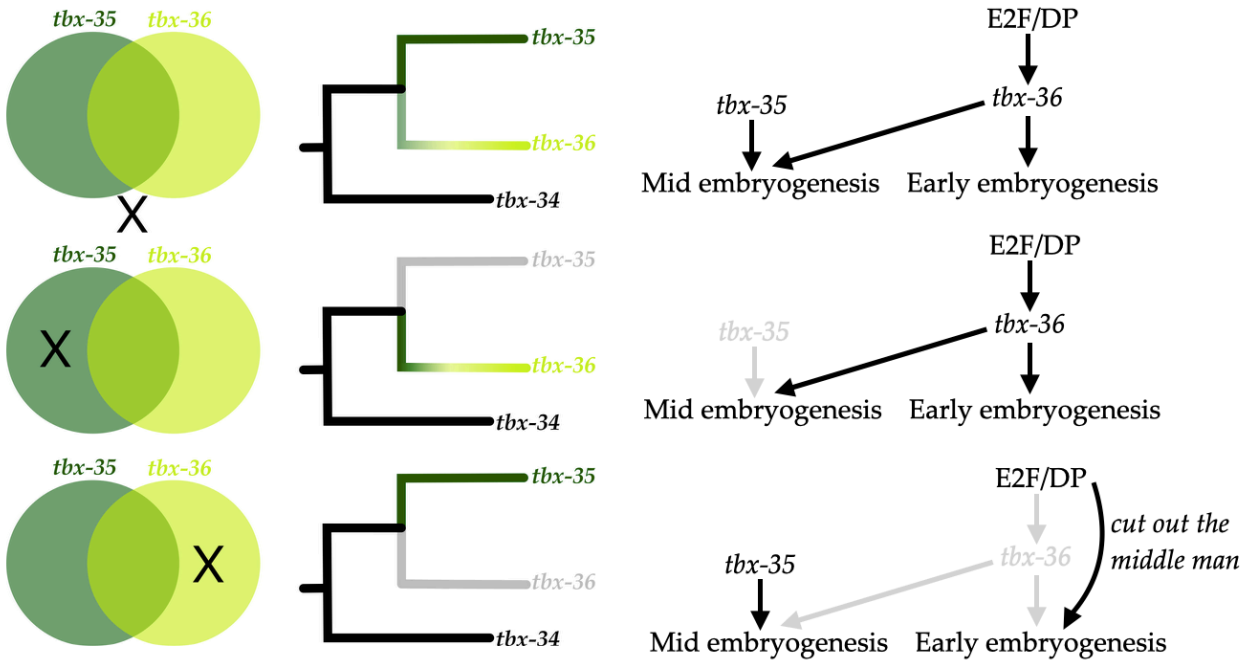


Figure 3.10. A working model for the functional evolution of the *tbx-35/36* gene pair. The three states in which the *tbx-35*/*tbx-36* gene pair have been found to exist are shown in the model, where each paralogue is either retained or accumulates deleterious mutations in one gene but not the other; this is denoted by the position of an 'X' on the Venn diagram. The putative explanations for why this is permitted, unique to both cases, is shown alongside. Where a gene is greyed out, it is marked as lost/loss-of-function; the dark green denotes so-called '*tbx-35*'-like functionality, while the light green denotes so-called '*tbx-36*'-like functionality.

We explored at length how *tbx-36* has neofunctionalised to take on a pivotal role in the early embryo, for which it is reliant on E2F/DP transcriptional regulation. It was found that due to its translocation to an unfamiliar chromosomal environment, *tbx-36* acquired an E2F binding site and in so doing, likely laid the groundwork for its emerging role as a modulator of early chromosomal processes. It is, therefore, almost impossible to foresee how any other T-box gene, not least *tbx-35*, could rescue such a phenotype; no other T-box gene is in the E2F hotspot on chromosome IV, nor is expressed at such an early stage of embryogenesis. And so, it is instead proposed that the genes which are regulated by *tbx-36* are, in isolates that have lost the gene, directly regulated by *efl-1* and *dpl-1*. Presumably, as early embryonic development predates the origin of *tbx-35* and *tbx-36* in *C. elegans* (Baker and Woollard 2019), prior to its derivation by gene duplication, the genes regulated

by *tbx-36* were once under the regulation of E2F/DP anyway (or, of course, another TF altogether). In this model, *tbx-36* is therefore no more than a 'middle man' in the formation of the early embryo — such a one as can be feasibly 'cut out' in isolates, such as ECA1191, that have lost it. All that would have to happen in these populations is the subtle rewiring of a gene regulatory network.

It is impossible to know if *efl-1* and *dpl-1* cause similar temperature-dependent early embryonic effects as in the case of *tbx-36* because the variety, severity, and penetrance of the defects their inactivation gives rise to is so comparably extreme given their status as master regulators of oogenesis and early embryogenesis. Moreover, *tbx-36* is hypothesised to regulate only a tiny proportion of genes under the command of E2F/DP transcriptional regulation. So, to integrate the previous discussion (concerning the nature of the genes regulated by *tbx-36*) with this one, it is unlikely that all the genes regulated by E2F/DP required for early embryogenesis all function optimally at higher temperatures, mostly because they are thought to regulate such a vast array (Chi and Reinke 2006; Chi and Reinke 2009; Petrella et al. 2011).

The neofunctionalisation of *tbx-36* in early embryogenesis is explained by its acquisition of E2F/DP *cis*-regulation, but gaining an on switch is only developmentally tolerated (and thus positively selected for) providing the concomitant gain of an 'off' to counteract it. It is for this reason that following its habitation of an E2F hotspot, and in so doing gaining a transcriptional activator, *tbx-36* evolved a repressive regulatory element to prevent the kind of inviability that is found upon its overexpression. T-box genes found elsewhere in the animal kingdom are known for being comparably dosage-sensitive in their transcriptional activation of genes (Papaioannou 2014). And, while it is of little value to speculate as to the identity of the repressor of *tbx-36*, it may be of use to others to note that repeatedly over the course of evolution, T-box TFs find themselves — both

positively and negatively — regulated by members of the Runx family (Cruz-Guilloty and Pipkin et al. 2009), fellow T-box genes (Stennard and Harvey 2005; Greulich et al. 2011), as well as the now familiar E2F/DP heterodimer (Chang et al. 2016; Nagel and Meyer 2021), and these are sometimes (in the instances of transcriptional repression) in partnership with the co-repressor, Groucho (Farin et al. 2007).

It is to this end that we are able to say that the dynamism observed at the phylogenetic level with respect to the *Caenorhabditis* T-box genes is borne out, and mirrored, in the rates of their functional, and thus regulatory, evolution — of course characterised in detail here with the *tbx-35*/*tbx-36* gene pair but must be just as true for the rest of the family in *C. elegans* as gleaned from their diverse roles and expression patterns from the work of others. What's more, it is an alluring possibility that their functional divergence, coupled with their sequence divergence in wild isolates, may make them agents of speciation among *C. elegans* populations today. Although, it is regrettable that the definitive experiment to form a watertight argument here cannot be performed. Crossing together wild populations with an inverse loss of one of the two members of the *tbx-35*/*tbx-36* paralogue pair (e.g., ECA1185 and ECA1191), and then assessing their ability to produce fertile offspring means very little when contextualised in genomes with a multitude of other comparative variation (Cook et al. 2017). It is though a highly probable scenario based on these data, just not one that can be proven beyond reasonable doubt, that populations like ECA1185 and ECA1191 may one day emerge as distinct species, which may be driven, at least in part, by the divergence of *tbx-35* and *tbx-36*.

CHAPTER 4

Extensive non-redundancy in a recently duplicated developmental gene family

*Chapter 4 is based on work already published in BMC Ecology and Evolution in 2021 — **Baker EA**, Gilbert SPR, Shimeld SM, Woollard A. Extensive non-redundancy in a recently duplicated developmental gene family. BMC Ecol. Evol. 2021 Mar 1;21(1):33.*

Introduction

While the T-box family gave us astonishing insight into the transformative power of gene duplications in the evolution of animal development and by extension their role in speciation, it was undeniably difficult to assess the paralogue dynamics within the family at large owing to their extreme copy number variation and sheer number of members. As such, in order to probe the more fundamental aspects of duplicated gene evolution, it follows we need simpler, smaller, more well-conserved families as models to address such questions. The families chosen throughout the next three chapters of this thesis satisfy this requirement, starting with the Warthog family in this Chapter.

It has been proposed that recently duplicated genes are more likely to be redundant with one another compared to ancient paralogues. The evolutionary logic underpinning this idea is simple, as the assumption is that recently derived paralogous genes are more similar in sequence compared to members of ancient gene families. We set out to test this idea by studying the nematode-specific family of Hedgehog-related genes, the Warthogs. Hedgehog is one of a handful of signal transduction pathways that underpins the development of bilaterian animals. While

having lost a bona fide Hedgehog gene, most nematodes have evolved an expanded repertoire of Hedgehog-related genes, ten of which reside within the Warthog family.

The Warthogs are a family of Hedgehog-related (Hh-r) genes exclusively found in the nematode phylum and are products of many gene duplication events (Bürglin 1996). Unlike their nematode-specific counterpart, the Hedgehog family has diversified little throughout the Bilateria, with most species possessing only one true orthologue. Two rounds of whole genome duplication have given rise to three genes in vertebrates (Sonic Hedgehog (Shh), Indian Hedgehog (Ihh), Desert Hedgehog (Dhh)), and due to an additional round of whole genome duplication, four or five in ray-finned fish (reviewed by Ingham et al. 2011). These vertebrate ohnologues arose approximately 530 million years ago and have taken on distinct, non-redundant, developmental roles. However, one of the teleost-specific ohnologues, *tiggywinkle hedgehog*, is around 350 million years old and appears to be redundant with *shh* in zebrafish retinal development (Stenkamp et al. 2000). This study, however, was limited to gene expression pattern analysis so it remains to be established what the knockout phenotypes would be.

Aside from its conservation in some basal nematode species, including: *Trichuris trichiura*, *Soboliphyme baturini* and *Trichinella zimbabwensis*, most nematodes have lost a Hedgehog gene. They have instead evolved an expanded repertoire of 61 Hh-r genes with partial orthology to the 'Hog' domain, or carboxyl terminus, of Hedgehog proteins. There are no homologues of the 'Hedge' domain, or amino terminus of Hedgehog, in the Hh-r superfamily of genes. The absence of the Hedge domain was surprising upon the initial discovery of Hh-r genes, as fly and mammalian Hedgehog pro-peptides are known to be autocleaved in the endoplasmic reticulum by their enzymatic Hog domain, prior to the release of the Hedge domain for signalling and the Hog

domain for proteasomal degradation (Lee et al. 1994; Porter et al. 1995). In other words, the Warthog family only possess partial orthology to the cleaved and degraded portion of the canonical Hedgehog protein. Nevertheless, the novel amino-terminal domains associated with Hog in nematodes were classified initially as Warthog (WRT) and Groundhog (GRD) (Bürglin 1996), followed by Ground-like (GRL) and Quahog (QUA) (Aspöck, Kagoshima et al. 1999). While all ten Warthogs contain a 'Wart' domain (defined by a consensus sequence of eight cysteine residues), only five family members contain a Hog domain: WRT-1, WRT-4, WRT-6, WRT-7 and WRT-8.

To test the relationship between the age of gene duplicates and the likelihood of functional redundancy in the Warthog family, we set out to investigate their roles in the model nematode *C. elegans* by first characterising their evolutionary history in Nematoda. To systematically elucidate their duplication history, we used a combination of molecular phylogenetic algorithms and then knockout and knockdown approaches in *C. elegans* to assess the functional divergence of paralogous genes.

Results

Widespread gene duplications in the Warthog family

We mined the predicted proteomes of a phylogenetically diverse range of nematodes for the Wart domain and verified the hits individually to ensure they contained a bona fide Wart domain as defined by Bürglin (Bürglin 1996). We exclusively analysed the Warthog repertoires of major parasites and model organisms so as to prevent conclusions about gene family evolution being an artefact of genome quality or the completeness of predicted proteomes (Gilabert et al. 2016). Nematoda is divided into three lineages, namely, Enoplea, Dorylaimia, and Chromadorea,

although orders are commonly organised into five major clades that do not correspond to the divisions of classical taxonomy (Blaxter 1998). The following species were selected for molecular phylogenetic analyses: *Brugia malayi* (Clade III); *Toxocara canis* (Clade III); *Ascaris suum* (Clade III); *Strongyloides ratti* (Clade IV); *Pristionchus pacificus* (Clade V); *Caenorhabditis remanei* (Clade V); *Caenorhabditis brenneri* (Clade V); *Caenorhabditis briggsae* (Clade V) and *Caenorhabditis elegans* (Clade V). Multiple species from Clade I were selected as outgroups (*Trichinella spiralis*, *Trichinella nativa*, *Trichinella murelli*, *Trichinella sp. T6*, *Trichinella sp. T8*, *Trichinella sp. T9*, *Trichinella papuae*, *Trichinella patagoniensis*, *Trichinella nelsoni*, *Trichinella pseudospiralis*, and *Trichuris suis*) as these were the only species in which only a single Wart domain could be detected. *Trichinella zimbabwensis*, *Trichuris trichiura* and *Soboliphyme baturini* were all found to contain at least one Hedgehog homologue, yet no Warthog homologues could be detected. We attempted to use *Trichuris muris* and *Romanomermis culcivorax* as other Clade I/Enoplea representatives in our analyses but no Hedge/Wart/Ground/Qua/Hog/Ground-like domain sequences could be detected in their predicted proteomes. As it was considered very unlikely for a bilaterian animal to have lost all Hedgehog and/or Hedgehog-like genes given their presence in neighbouring lineages, the genomes of *T. muris* and *R. culcivorax* were not deemed to be of sufficient completeness for use in our investigation.

Figure 4.1A summarises the Warthogs present in the nematodes analysed. Mining the genomes of these nematodes for Wart domains revealed multiple hits which had partially lost the consensus sequence (one or more cysteine residues) but otherwise aligned to one of the ten *C. elegans* Wrts. Because they had incompletely lost a typical Wart domain sequence, we classified them as 'degenerate Wrts'. In most cases, degenerate Wrt coding sequences have diverged by more than just their cysteine residues which probably reflects their neofunctionalisation outside of Warthog niches, except for the *wrt-2* orthologues in *C. brenneri* and *C. remanei* which have accumulated a large proportion of repetitive and low complexity DNA.

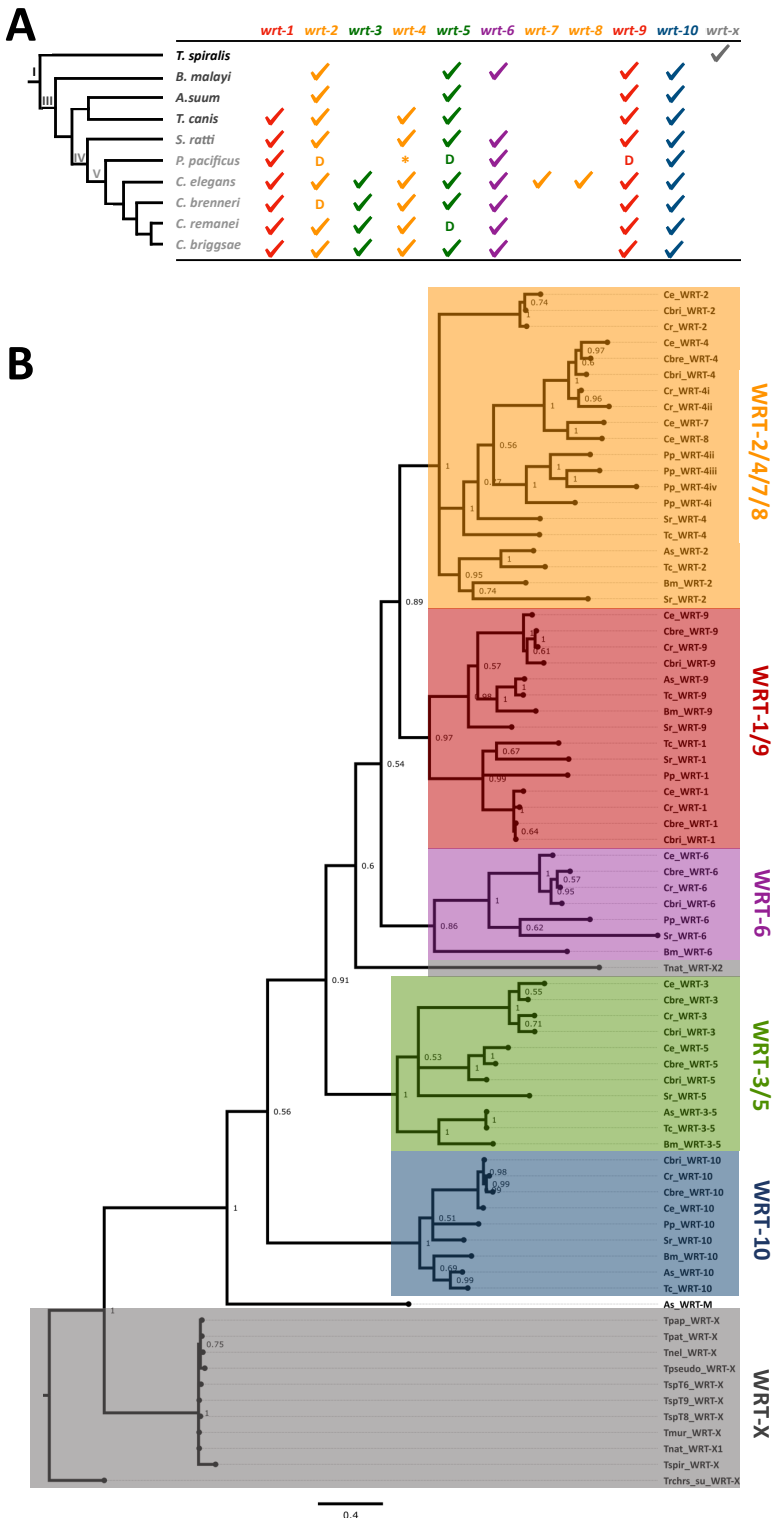


Figure 4.1. Phylogenetic analysis of the Warthog family in *C. elegans* and other nematodes. (A) Cladogram showing relationships between nematodes in this study and a table showing their Warthog orthologues. Coloured ticks indicate that Warthog is present in a respective species. 'D' refers to degenerate Wart domain sequences. '*' refers to the abnormal *wrt-4* complement in *Pristionchus pacificus* which has four bona fide *wrt-4* orthologues and four degenerate *wrt-4* sequences. (B) Phylogram was generated from a multiple sequence alignment of Wart domains

(SD1 in Baker et al. 2021), including *C. elegans* paralogues (stars) and orthologues from selected nematode species. Wart clades are colour coded. Species abbreviations: *Tnat, Trichinella nativa*; Tmur, *Trichinella murelli*; TspT6, *Trichinella sp. T6*; TspT8, *Trichinella sp. T8*; TspT9, *Trichinella sp. T9*; Tpap, *Trichinella papuae*; Tpat, *Trichinella patagoniensis*; Tnel, *Trichinella nelsoni*; Tpseudo, *Trichinella pseudospiralis*; Trchrs_su, *Trichuris suis;* Ts, *Trichinella spiralis*; Bm, *Brugia malayi*; As, *Ascaris suum*; Tc, *Toxocara canis*; Sr, *Strongyloides ratti*; Pp, *Pristionchus pacificus*; Cbre, *Caenorhabditis brenneri*; Cbri, *C. briggsae*; Cr, *C. remanei*; Ce, *C. elegans*. 'As_WRT-M' is our given name to the Warthog in *A. suum* which did not robustly cluster into any of the Wart clades. Node values indicate posterior probabilities for each split. The scale bar indicates average branch length measured in expected substitutions per site.

Two independent phylogenetic analyses were run on the Wart domain alignment. The output of the Bayesian analysis is shown in Figure 4.1B (the Maximum Likelihood IQ-TREE analysis can be found in Appendix VI). Wart domain sequences from other nematodes were named because of their similarity to *C. elegans* sequences (such that the ten *C. elegans* Warthogs remained the basis of this investigation). Since there are more loci in other nematodes than previously named, we propose an updated Warthog nomenclature based on the Wart domain (see Appendix VII).

Both phylogenetic analyses resolved five distinct Wrt clades: WRT-2/4/7/8 (containing WRT-2, WRT-4, WRT-7, WRT-8); WRT-3/5 (containing WRT-3, WRT-5); WRT-1/9 (containing WRT-1, WRT-9); WRT-6 (containing WRT-6 only); WRT-10 (containing WRT-10 only). The single Wart domains in Clade I nematodes root both phylogenetic trees and are taken as the extant representative of the ancestral Wart domain. 'As_WRT-M' is our name given to the Warthog in *A. suum* which did not robustly cluster into any of the Wrt clades.

Unusually, *P. pacificus* contained four *bona fide* 'WRT-4' orthologues (Pp_WRT-4i, Pp_WRT-4ii, Pp_WRT-4iii, Pp_WRT-4iv) and three degenerate 'WRT-4' sequences. Only Pp_WRT-4iii possesses a Hog/Hint domain, while all other paralogues do not, which may suggest only part of the locus is prone to duplicate. An alternative explanation may be inaccurate protein prediction models (Gilabert et al. 2016). The atypical *wrt-4* complement in *P. pacificus* was found to be species-specific

but is probably symptomatic of the gene's repetitive content. The genome instability conferred by repetitive sequences (Bzymek and Lovett 2001) and their tendency to cause the duplication of adjacent regions means that tandem and inverted repeats provide opportunities for gene duplication by providing regions of homology for unequal crossing over. Throughout this investigation, we noticed an abundance of tandem and inverted repeats in and around *C. elegans* Wrt gene sequences, later mined using RepeatMasker (data not shown). As it is known that repetitive elements are similarly distributed on *C. elegans* autosomes (Surzyki and Belknap 2000), and as all Warthog genes contain introns, we propose that all family members have been derived by unequal crossing over as opposed to retrotransposition.

To further probe into the duplication history of these genes, we performed synteny analysis (see Appendix VIII). The extent of genomic reshuffling even within the *Caenorhabditis* genus meant this strategy was not as useful for characterising gene family evolution compared to its illumination of gene diversification in chordates (Lara-Ramírez, Poncelet, Patthey et al. 2017) wherein synteny is more highly conserved. In all clade V nematodes *wrt-1* and *wrt-10* were ~350 bp apart yet in *S. ratti* and *T. canis* they were on different chromosomes, most likely because of lineage-specific reshuffling. The two *C. elegans* specific Warthogs, *wrt-7* and *wrt-8*, were directly adjacent to one another on chromosome V and their loci map to *wrt-4* in other Rhabditina. Outside of clade V (e.g., *S. ratti*, *A. suum*, *T. canis*, *B. malayi*, *T. spiralis*), many microsyntenic relationships break down.

Functions of the *C. elegans* Wrt genes strongly associate with clades of the Warthog phylogeny

Members of the Wrt-2/4/7/8 clade are involved in the development of LR asymmetry

In order to investigate the possible redundancy relationships among duplicated Wrt genes, we first tested the phenotypes of single knockdowns (by RNAi) and single knockouts (using deletion alleles), and later double and triple mutants. All phenotypes reported in this chapter are confirmed by both knockout and knockdown approaches to increase reliability.

Upon initial investigation, it was noted that the characteristic orientation of the gut and gonad with respect to one another was disrupted in *wrt-2(ok2810)* mutant animals. In wildtype (WT) worms, there is an invariant left-right (LR) asymmetry in the middle body (Figure 4.2A) where in the lefthand plane only intestine is visible in the anterior (Figure 4.2B) and only gonad arm is visible in the posterior (Figure 4.2C). Conversely in the righthand plane, only gonad is seen anteriorly (Figure 4.2D), while only intestine is seen posteriorly (Figure 4.2E). Examples of deviations from the WT presentation in *wrt-2(ok2810)* animals are shown in the lefthand plane in both the anterior (Figure 4.2F, H) and the posterior (Figure 4.2G, I); only gut or gonad should be observed respectively, yet both are seen (to variable extents) in the same plane. No other obvious gonad morphology defects were observed in these animals, for example aberrant turns or projectiles (lateral guidance defects) normally associated with dorsoventral (DV) or anteroposterior (AP) axis misguidance. Thus *wrt-2* appears to be involved in specifically regulating LR asymmetry in the middle body of the adult worm.

In order to test whether knockdown of other Wrt genes produces a similar phenotype we performed RNAi knockdown of each family member and recorded the penetrance of defects in the

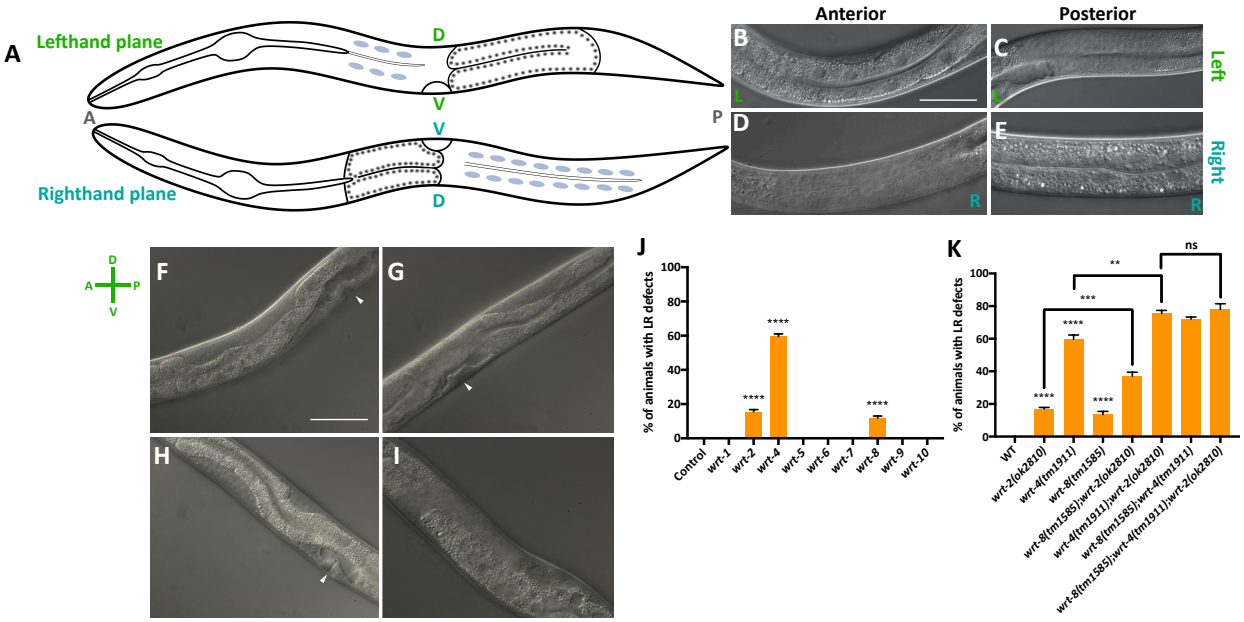


Figure 4.2. Presence of middle body LR asymmetry defects in Wrt family knockdowns and knockouts. (A) Schematic of a wild-type worm in both dorsal and ventral views showing the gut/gonad asymmetry. Panels (B, C, D, E) are wild-type images of the middle body where B and C are taken in the lefthand plane and D and E are taken in the righthand plane. Intestine (B and E) is recognisable for the large nuclei and the gonad (C and D) is most recognisable for being syncytial. Panels F, G, H, I are *wrt-2(ok2810)* animals exhibiting defects in middle body LR asymmetry, where in F and H (lefthand plane) only intestine should be visible yet patches of gonad are observed. In G and I (righthand plane), only gonad should be visible, yet patches of gut are observed. Arrowheads indicate vulvas. Scale bars = 50 μm. (J) % penetrance of LR defects upon knocking down a Wrt family member (x-axis). Empty vector control (L4440) animals displayed no defects in the positioning of their gut and gonads relative to one another (n = 51), nor did *wrt-1* (n = 80); *wrt-5* (n = 71); *wrt-6* (n = 45); *wrt-7* (n = 67); *wrt-9* (n = 85) or *wrt-10* (n = 59) RNAi animals. *wrt-2* (n = 72), *wrt-4* (n = 80) and *wrt-8* (n = 56) RNAi animals did display LR asymmetric defects. *wrt-3* knockdown results in animals with miniaturised or absent gonad arms and/or other disruptions to their middle body anatomy such that LR defects could not be quantified in *wrt-3* defective animals (Additional File 1; Supplementary Figure 6 in Baker et al. 2021). (K) % penetrance of LR defects in wild-type as compared to animals carrying the *wrt-2(ok2810)* allele (n = 65), the *wrt-4(tm1911)* allele (n = 108), the *wrt-8(tm1585)* allele (n = 68) or the following double/triple mutants: *wrt-4(tm1911);wrt-2(ok2810)* (n = 102); *wrt-8(tm1585);wrt-2(ok2810)* (n = 87); *wrt-8(tm1585);wrt-4(tm1911)* (n = 54); *wrt-8(tm1585);wrt-4(tm1911);wrt-2(ok2810)* (n = 103). Black bars show mean + SEM (J, K). Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to Control RNAi with an unpaired *t* test (J) or in the means of Wrt mutants compared to WT with an unpaired *t* test (K).

middle body of the worm compared to EV control RNAi animals (Figure 4.2J). We found that only knocking down *wrt-2*, *wrt-4* or *wrt-8* resulted in LR asymmetric defects with *wrt-4* knockdown

resulting in the highest penetrance of 60% ($P < 0.0001$). Knockdown of *wrt-2* and *wrt-8* gives rise to 16% ($P < 0.0001$) and 12% ($P < 0.0001$) of animals with LR defects, respectively. Thus, all members of the Wrt-2/4/7/8 clade display LR defects upon RNAi knockdown except *wrt-7*. To confirm this, we analysed the phenotypes of *wrt-2(ok2810)*, *wrt-4(tm1911)* and *wrt-8(ok1585)* single mutants, finding concordance with the RNAi data (Figure 4.2K). Next, we tested phenotypic redundancy between Wrt-2 clade members by constructing double and triple mutants and quantifying the penetrance of LR asymmetric defects. We observed the penetrance of defects in the *wrt-2*;*wrt-8* ($P = 0.0003$), *wrt-2*;*wrt-4* ($P = 0.0025$), *wrt-4*;*wrt-8* double mutants to be additive suggesting these pairs of genes do not display redundancy with respect to this phenotype. Moreover, the concomitant inactivation of *wrt-2*, *wrt-4* and *wrt-8* in the triple mutant did not increase the penetrance of LR defects over and above the *wrt-4;wrt-2* double mutant ($P = 0.5478$).

It is worth noting that performing *wrt-7* RNAi-mediated knockdown on *wrt-2(ok2810)*, *wrt-4(tm1911)* and *wrt-8(ok1585)* single mutants and the inverse set of experiments (i.e. *wrt-2, -4* and *-8* RNAi knockdown on *wrt-7(ok3271)* mutant animals) did not reveal a role for *wrt-7* in any obvious biological process. This includes the absence of defects in LR asymmetry in the middle body as there were no phenotypic differences between these and the relevant control animals (see Appendix IX). In addition, *wrt-7* RNAi knockdown in an RNAi-sensitive mutant (*rrf-3(pk1426)*) did not display any abnormal morphologies when compared to control animals (see Appendix IX). Taken together with reports that *wrt-7* is not expressed throughout development (from PolyA+ and Ribozero modENCODE libraries (Gerstein et al. 2010); Hao *et al*. 2006a), we conclude that *wrt-7* is non-functional and has likely pseudogenised. Although the hallmarks of pseudogenisation (e.g., a PTC) are absent in the Bristol N2 strain, many wild isolates of *C. elegans* contain a highly polymorphic copy of *wrt-7* that includes a missing start codon and approximately 50 moderate effect mutations (see Appendix X).

Despite the clear roles of *wrt-2*, *wrt-4* and *wrt-8* in the establishment of LR asymmetry during late larval development in the middle body, we were not able to detect embryonic defects (in either early embryos at the four-to-six cell stage, when LR asymmetry is established in *C. elegans* embryos, or during the intestinal twist at the 1.5-fold stage of mid-embryogenesis) in left-right asymmetry which would have suggested that these genes are global regulators of LR asymmetry (data not shown). Therefore, we infer that these genes are unique in providing a left-right directional signal for the gonad arms as they migrate along the AP and DV axes during larval development (reviewed by Hubbard and Greenstein 2000). No signals were previously implicated in the left-right guidance of gonad morphogenesis because it was considered to be a consequence of AP and DV signalling by molecules such as netrin (Ziel and Sherwood 2010). Notably, *wrt-4*, *wrt-2* and *wrt-8* must not be the only regulators of this aspect of left-right positioning, as no animals were seen with complete reversals of middle body morphology, known as situs inversus, implying other signals are required for this process. Nevertheless, it is striking that members of the Warthog family are involved in the generation of LR asymmetry given the well characterised role of the partially orthologous Shh in the same process during mammalian embryogenesis (Zhang et al. 2001).

Members of the Wrt-3/5 clade are involved in cell fate determination in the developing vulva

Having observed vulval phenotypes in some Wrt family RNAi animals, we crossed in the *ajm-1::gfp* marker (which localises to apical cell membranes (Lynch and Hardin 2010)) in order to visualise and quantify these defects more precisely. RNAi knockdown of each family member revealed that members of the Wrt-3/5 clade are required for vulval fate specification. The hermaphrodite vulva (Figure 4.3A) is a paradigm for organogenesis with a well-elucidated molecular basis underpinned by an inductive RTK-Ras-MAPK signalling cascade and subsequent

lateral Notch signalling between vulval precursors (Kornfeld 1997). Aberrant signalling can cause too many progenitors at the ventral midline to adopt a vulval cell fate giving rise to ectopic non-functional protrusions, or pseudovulvae — a phenotype known as Multivulva (Muv) (Figure 4.3B) (Horvitz and Ferguson 1989).

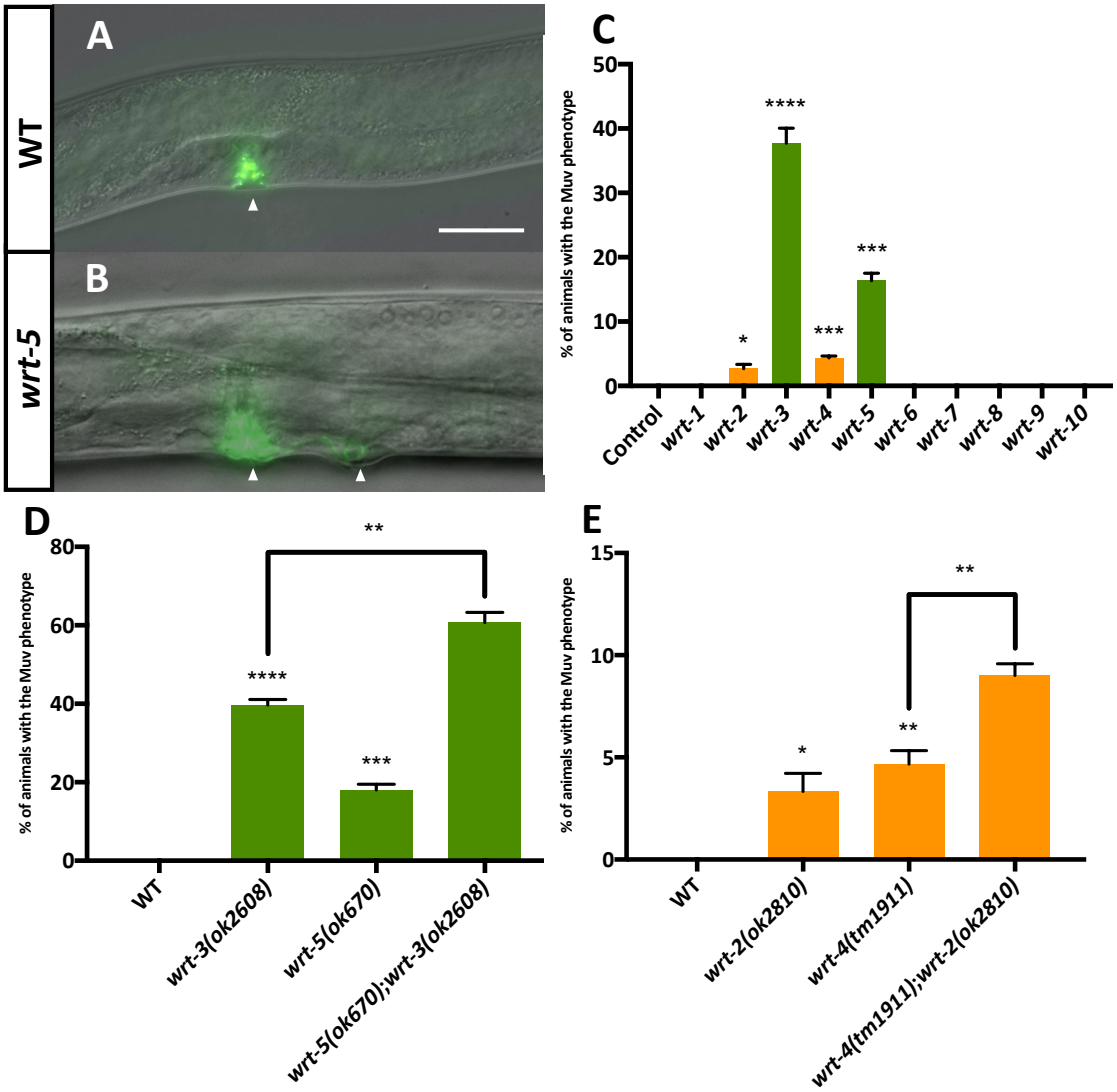


Figure 4.3. Presence of the Multivulva phenotype in Wrt family knockdowns and knockouts. (A) Wild-type L4 animals develop one vulva on the ventral side of the animal, indicated by the single arrowhead. The vulva is made visible using the *ajm-1::gfp* marker, which localises to the vulval cell apical membranes and is used to visualise the vulva using fluorescence optics. Scale bar is 50 µm. (B) *wrt-5(ok670)* L4 animals display a Muv phenotype where more than one ventral site undergoes vulval induction; in this example two developing vulvas are indicated with arrowheads. Scale bar = 50 µm. (C) % penetrance of the Multivulva phenotype upon RNAi knockdown of a Wrt family member (x-axis) in an *ajm-1::gfp* background. Empty vector control (L4440) animals do not display the Muv phenotype (n = 44). 0% penetrance of Muvs was recorded upon knockdown of: *wrt-1* (n =

41); *wrt-6* (n = 40); *wrt-7* (n = 39); *wrt-8* (n = 49); *wrt-9* (n = 43); *wrt-10* (n = 40). The Muv phenotype was recorded upon knockdown of *wrt-2* (n = 45); *wrt-3* (n = 41); *wrt-4* (n = 42) and *wrt-5* (n = 41). (D) % penetrance of Multivulva phenotype in wild-type (n = 40) as compared to animals carrying the *wrt-3(ok2608)* allele (n = 39) or the *wrt-5(ok670)* allele (n = 42) or the double mutant, *wrt-5(ok670);wrt-3(ok2608)* (n = 32), all in an *ajm-1::gfp* background. (E) % penetrance of Multivulva phenotype in wild-type as compared to animals carrying the *wrt-2(ok2810)* allele (n = 44) or the *wrt-4(tm1911)* allele (n = 41) or the double mutant, *wrt-4(tm1911);wrt-2(ok2810)* (n = 45), all in an *ajm-1::gfp* background. Black bars show mean + SEM (C, D, E). Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to Control RNAi with an unpaired *t* test (C) or in the means of Wrt mutants compared to WT with an unpaired *t* test (D, E).

Members of the Warthog family have been implicated in vulval organogenesis previously (Zugasti et al. 2005). Knockdown of *wrt-3* or *wrt-5* resulted in significant Muv defects (40% and 18% penetrance, respectively) whereas none of the other Wrt family members were associated with vulval defects except for the very low penetrance defects (<5%) in *wrt-2* and *wrt-4* knockdowns (but not in *wrt-7* or *wrt-8* knockdowns) (Figure 4.3C). For both gene pairs that exhibited Muv phenotypes in the Wrt-3/5 and Wrt-2/4/7/8 clades, double mutants had additive but not synergistic phenotypes, again suggesting no redundancy (Figure 4.3D, E).

Members of the Wrt-1/9 clade are involved in body size regulation

We also noticed that knockdown of some Wrt family members resulted in shorter worms (Table 4.1). Quantifying this, we observed knockdown of *wrt-1* or *wrt-9* leads to a ~3% decrease in body length in adult worms, whereas none of the other Wrt family members showed this significant decrease. To test for redundancy, we built a *wrt-1(tm1417);wrt-9(ok2732)* double mutant and again found no evidence of redundancy.

Table 4.1. The Role of the Wrt-1/9 clade in Body Size Regulation

RNAi/Genotype	Body size 48h post-L4 (mm)	n number	*P* value[α]
Empty Vector	1.229	35	
wrt-1	1.213	32	<0.0001
wrt-2	1.230	24	ns
wrt-3	1.234	28	ns
wrt-4	1.231	30	ns
wrt-5	1.231	29	ns
wrt-6	1.229	23	ns
wrt-7	1.232	26	ns
wrt-8	1.223	27	0.0229
wrt-9	1.205	33	<0.0001
wrt-10	1.231	22	ns
Wild-type	1.227	30	
wrt-1(tm1417)	1.206	39	<0.0001
wrt-9(ok2732)	1.204	40	<0.0001
wrt-1(tm1417);wrt-9(ok2732)	1.204	48	ns

[α] Unpaired t tests comparing the mean body lengths between Empty Vector Control animals and Wrt gene RNAi animals; wild-type and *wrt-1(tm1417)* and *wrt-9(ok2732)* single mutants; and the *wrt-9(ok2732)* and *wrt-1(tm1417);wrt-9(ok2732)* double mutant. Body lengths of anaesthetised worms were manually measured at high magnification (63 X) in well-fed young adult animals 48 hours post-L4 where all were grown at 20 °C.

Multiple members of the Warthog family are involved in ecdysis

The germline, vulval and body length defects of the Wrt family mutants appear to cluster with particular clades of the phylogeny, however, we observed widespread moulting defects (exemplified in Figure 4.4B, D) upon knockdown of nearly all family members. Moulting is the process by which animals replace their old exoskeleton, or cuticle, with a new one (Lažetić and Fay 2017). The cuticle is a collagenous barrier between the animal and its external environment (Figure 4.4A, C). As ecdysozoans, *C. elegans* like other nematodes undergoes four moults throughout development which mark the start of each larval stage. Bürglin and colleagues (2006) documented the role of *wrt-5* in epidermal development and moulting as well as the cyclical expression pattern (in phase with the moulting cycle) of many Hedgehog-related genes, including the Warthogs. In

light of this, we characterised the presence of moulting defects in Warthog family knockdowns (Figure 4.4E) and found *wrt-1*, *wrt-2*, *wrt-3*, *wrt-4*, *wrt-5* and *wrt-8* all have roles in ecdysis.

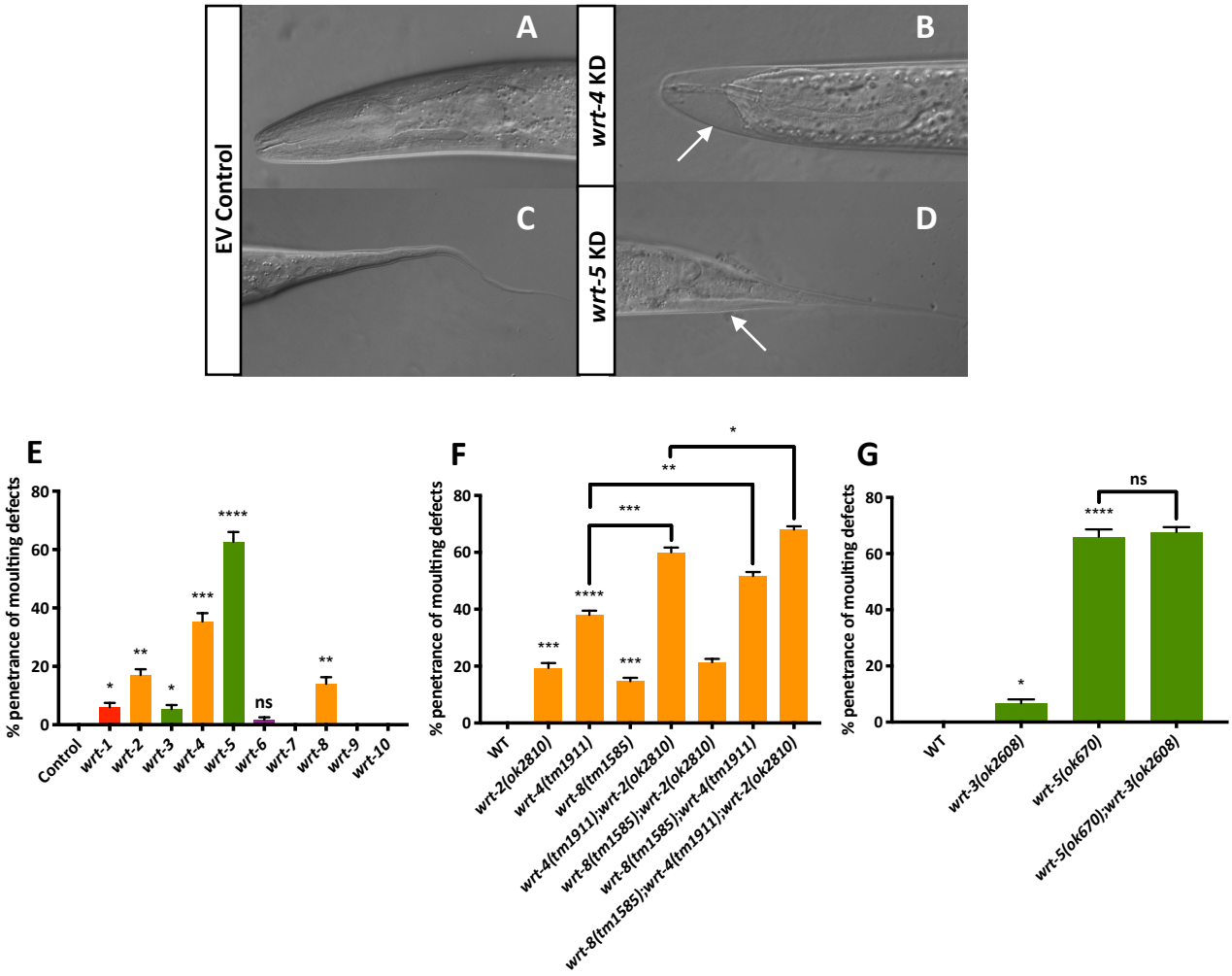


Figure 4.4. Presence of moulting and cuticle defects in Wrt family knockdowns and knockouts. Panels (A, C) are wild-type worms showing head and tail, respectively. (B) *wrt-4* KD animal with an improperly shed cuticle in the head region, referred to as 'head in a bag' phenotype (see arrow). (D) *wrt-5* KD animal with an improperly shed cuticle in the tail region (see arrow). Scale bar = 50μm. (E) % penetrance of moulting defects present upon RNAi knockdown of a Wrt family member (x-axis). Empty vector control (L4440) animals do not display any moulting defects (n = 51). 0% penetrance of moulting defects were recorded upon knockdown of: *wrt-7* (n = 67); *wrt-9* (n = 85); *wrt-10* (n = 59). Moulting defects were recorded upon knockdown of *wrt-1* (n = 80); *wrt-2* (n = 72); *wrt-3* (n = 69); *wrt-4* (n = 80); *wrt-5* (n = 71); *wrt-6* (n = 45); *wrt-8* (n = 56). (F) % penetrance of moulting defects in wild-type (n = 50) as compared to animals carrying the *wrt-3(ok2608)* allele (n = 81) or the *wrt-5(ok670)* allele (n = 71) or the double mutant, *wrt-5(ok670);wrt-3(ok2608)* (n = 67). (G) % penetrance of moulting defects in wild-type as compared to animals carrying the *wrt-2(ok2810)* allele (n = 65), the *wrt-4(tm1911)* allele (n = 108), the *wrt-8(tm1585)* allele (n = 68) or the following double/triple mutants: *wrt-4(tm1911);wrt-2(ok2810)* (n = 102); *wrt-8(tm1585);wrt-2(ok2810)* (n = 87); *wrt-8(tm1585);wrt-4(tm1911)* (n = 54); *wrt-8(tm1585);wrt-4(tm1911);wrt-2(ok2810)* (n = 103). Black

bars show mean + SEM (E, F, G). Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to Control RNAi with an unpaired *t* test (E) or in the means of Wrt mutants compared to WT with an unpaired *t* test (F, G).

While the respective penetrance of moulting defects in the single deletion mutants of *wrt-1(tm1417)*, *wrt-2(ok2810)*, *wrt-3(ok2608)*, *wrt-4(tm1911)*, *wrt-5(ok670)* and *wrt-8(tm1585)* was consistent with the RNAi knockdown data, double and triple mutant analysis showed redundancy is not exhibited between clade members with respect to moulting (Figure 4.4F, G). However, knockout/knockdown combinations of multiple Wrts from different clades, dubbed 'interclade RNAi', revealed that one pair of Wrts from different clades, *wrt-3* and *wrt-9*, are redundant for their role in cuticle biosynthesis, yet not moulting, which was later confirmed by building a *wrt-3(ok2608);wrt-9(ok2732)* double mutant (Figure 4.5). Upon initial imaging of these *wrt-3(ok2608);wrt-9(ok2732)* double mutant animals, it was noted that their cuticles appeared fragile and perforated (Figure 4.5A). Defects in the cuticle integrity of *wrt-3(ok2608);wrt-9(ok2732)* double mutants was assayed using 4',6-diamidino-2-phenylindole (DAPI) uptake (Xiong et al. 2017). Animals in which both of the functions of these genes are abolished have a highly permeable cuticle compared to *wrt-3(ok2608)* and *wrt-9(ok2732)* single mutants, WT, or Wrt mutants that exhibit highly penetrant moulting phenotypes such as *wrt-5(ok670)* (Figure 4.5B, C, D, E, F, G, H).

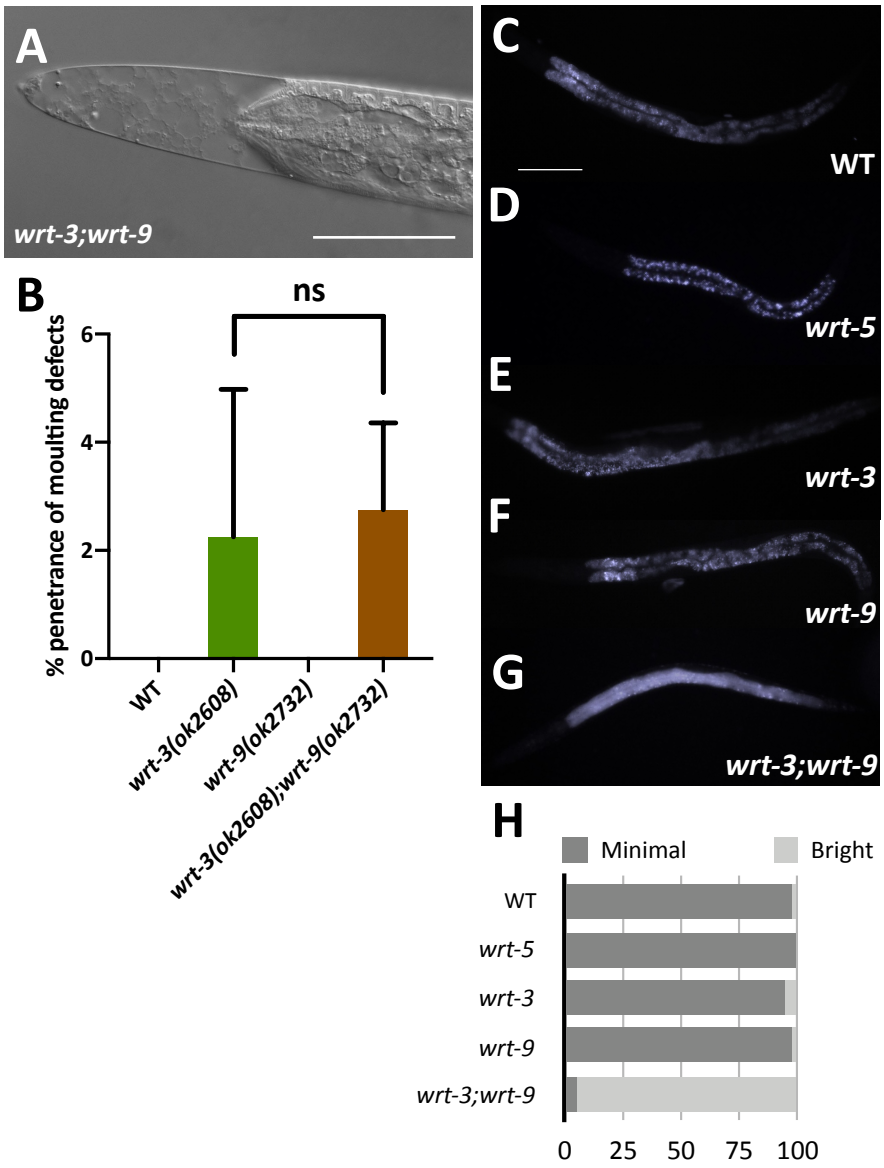


Figure 4.5. Cuticle Integrity of Warthog Mutants. Panel (A) depicts the fragile and perforated cuticles of *wrt-3(ok2608);wrt-9(ok2732)* mutants that was noticed upon the initial construction of these double mutant animals. Panel (B) shows the % penetrance of moulting defects (+SD) in WT (n = 53), *wrt-3(ok2608)* (n = 53), *wrt-9(ok2732)* (n = 57) and *wrt-3(ok2608);wrt-9(ok2732)* (n = 56) animals, respectively. The mean % penetrance of moulting defects present in *wrt-3(ok2608)* single mutant and *wrt-3(ok2608);wrt-9(ok2732)* double mutant animals was compared with an unpaired *t* test and found not to be significant (nsP > 0.05). Panels (C, D, E, F, G) depict worms which have been soaked with DAPI for 15 minutes and imaged using 100msec exposure time. Panel (H) depicts the quantification of DAPI fluorescence using a scoring system established in Xiong et al. 2017 using this DAPI assay where the x-axis is the % of total worms imaged. Panels (C, D, E, F) represent 'Minimal' fluorescence, while Panel (G) represents 'Bright' fluorescence. The fluorescence observed in panels (C, D, E, F) is autofluorescence, rather than DAPI stain. Wild-type (n = 45), 97.44% Minimal; *wrt-5(ok670)* (n = 51), 100% Minimal; *wrt-3(ok2608)* (n = 55), 94.74%

Minimal; *wrt-9(ok2732)* (n = 55), 97.44% Minimal; *wrt-3(ok2608);wrt-9(ok2732)* (n = 59), 95.00% Bright. Scale bars = 50 μm.

It is worth noting that multiple other interclade RNAi combinations were tested over the course of this investigation, yet no additional phenotypes or non-additive effects of phenotypes already recorded were observed (see Appendix XI).

Multiple attempts at knocking down *wrt-6* and *wrt-10* did not result in any apparent phenotypes. Recently generated putative null alleles for *wrt-6* and *wrt-10* using CRISPR/Cas9 gene editing also display no obvious gross morphological phenotype (Sherry et al. 2019), and so their roles in *C. elegans* remain unknown. However, we tested if the *C. elegans* specific substitutions in *wrt-6* and *wrt-10* were driven by positive selection, as indicated by an elevated dN/dS ratio (that is the ratio of non-synonymous to synonymous substitutions), but found the long divergence times were associated with saturation of dS and gave unreliable dN/dS estimation in both cases.

Discussion

The notion that "natural selection merely modified, while redundancy created", has been the fundamental premise to theories of evolution by gene duplication since it was first proposed by Susumu Ohno in his seminal book in 1970 (Ohno 1970). The implication that functional redundancy is simply a transient state of duplicated genes has been widely accepted in the field of evolutionary genetics, but there are instances in which redundancy is maintained between paralogue pairs for over nearly 100 million years of evolution (Tischler, Lehner et al. 2013; Dean et al. 2008). However, the pervasiveness of redundancy in large gene families has been poorly

assessed. It seems intuitive that the functional redundancies in large gene families would occur exclusively between more recent duplicates, while older paralogues would have taken on neofunctionalised, non-redundant roles. To test these ideas, we characterised the duplication history and the roles of the taxon-restricted Warthog family in the nematodes.

Reconstructing the duplication history of the Warthog family

The extensive variation in the Warthog repertoires among nematode species as compared to the static nature of the relatively few Hedgehog genes in the bilaterians is symptomatic of the family's vulnerability to duplication and loss. Due to the generation of high-quality genome assemblies for many species in the nematode phylum in recent years, reconstructing the duplication history of multigene families can now be done in unprecedented phylogenetic detail (Baker and Woollard 2019). By combining phylogenetic, synteny and repeat sequence data, we derived the model for the duplication history of the Warthog genes as shown in Figure 4.6A.

The family have likely derived from a single ancestral gene, *wrt-x*, which is still represented in *Trichinella spiralis*. This ancestral Warthog appears to have duplicated at least twice to yield a Hog-containing (wrt-1/2/4/6/7/8/9) precursor and a Hog-less precursor (wrt-3/5/10) less than 400 mya. These two progenitors presumably then expanded with the radiation of Chromadorea to create a complement of five Warthogs (*wrt-2*, *wrt-5*, *wrt-6*, *wrt-9*, *wrt-10*) which are represented in nearly all the extant Clade III nematodes studied in this investigation, with the exception of the independent loss of *wrt-6* in *A. suum*. Following their generation by tandem duplication, *wrt-2* and *wrt-9* subsequently lost their Hog domains. The Hog-containing progenitor is envisaged to have given rise to *wrt-1* and *wrt-4* in *T. canis* and other lineages (Clade IV and V nematodes), as well as *wrt-7*

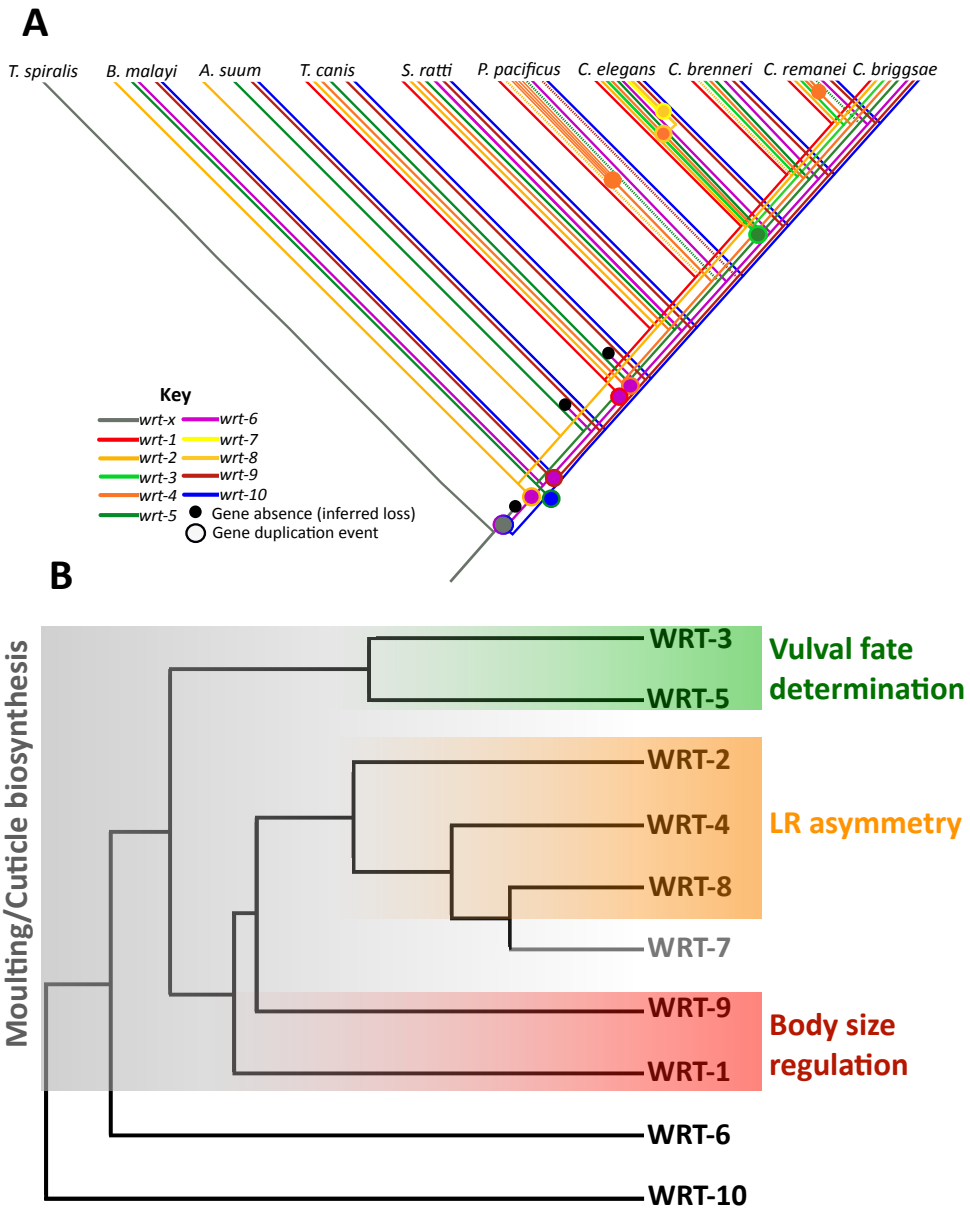


Figure 4.6. Evolution of the Warthog family. (A) Schematic detailing the duplication history of the Warthogs in the nematodes. Degenerate Wrts are represented by a dashed line. The three degenerate '*wrt-4*' paralogues present in *P. pacificus* are not shown for clarity. (B) Cladogram showing the evolutionary relationships between the ten Warthogs and their functions according to the results of this investigation. *wrt-7* (grey) is a pseudogene with no obvious functionality.

and *wrt-8* in *C. elegans* less than 10 mya (Baker and Woollard 2019). The Hog-less wrt-3/5 precursor subsequently duplicated to yield *wrt-3* in the *Caenorhabditis* genus less than 100 mya.

Neofunctionalisation of Warthog family genes reflects cladistic architecture

Because of the genetic tractability of *C. elegans*, evolutionary hypotheses derived from the duplication history of large multigene families can be tested using robust genetic techniques. We sought to test the relationship between the age of gene duplicates and the likelihood of functional redundancy in the Warthog family. Overall, we found hitherto unreported roles for the Warthog family in the generation of middle body LR asymmetry (*wrt-2*, *wrt-4* and *wrt-8*), cell fate specification in the developing vulva (*wrt-3* and *wrt-5*), and body size regulation (*wrt-1* and *wrt-9*). These roles associate strongly with particular clades of the Warthog phylogeny (Figure 4.6B). Thus, we conclude that these clades have neofunctionalised in aspects of post-embryonic development.

Surprisingly, we did not find any instances of complete functional redundancy between family members in these neofunctionalised clades, implying they operate in different pathways. If the Warthogs encode ligands that operate in different signalling pathways, this would explain the additivity of the phenotypes observed in the mutants of even closely related Wrt genes, such as *wrt-2*, *wrt-4* and *wrt-8*; *wrt-3* and *wrt-5*; and *wrt-1* and *wrt-9*. This implies a highly robust network of genes involved in these developmental processes.

wrt-3 is a recently derived Warthog, only found in members of the *Caenorhabditis* genus which arose less than 100 mya. In light of this, the severe and highly penetrant phenotypes that it exhibits are unexpected. Thus, it can be stated that recently duplicated Warthogs are not only non-redundant with one another, but in the instance of *wrt-3*, have also assumed critical developmental roles including in organogenesis. As one of the more recently derived members of the Wrt-2/4/7/8 clade, *wrt-7* appears to have completely pseudogenised, having no obvious functionality or expression pattern throughout development (Hao et al. 2006a).

The roles and redundancies of the Warthog family in ecdysis

Throughout this investigation, we observed many moulting defects in Warthog family mutants. As such, we systematically characterised the role of each member in moulting and found that *wrt-1*, *wrt-2*, *wrt-3*, *wrt-4*, *wrt-5* and *wrt-8* are all involved to some extent in this process. The role of some Warthog family members in moulting and the oscillatory expression patterns of several Warthogs has implicated the family in ecdysis in previous studies (Hao et al. 2006a; Hendriks et al. 2014). As many other members of the Hedgehog-related (Hh-r) (Hao et al. 2006a; Hao et al. 2006b; Hao et al. 2006c) and Patched/Patched-related (Ptc-r) superfamilies are involved in moulting (reviewed by Lažetić and Fay 2017), we propose that ecdysis is the ancestral role of the divergent 'Hedgehog' pathway in Nematoda and that Hh-r and Ptc-r genes were at least ancestrally in the same pathway.

We did not find any instances of redundancy in the Warthog family with respect to moulting, either between those in the same clade or those in different clades. The only instance of functional redundancy observed in this investigation is between Warthogs in different clades, *wrt-3* and *wrt-9*, in cuticle biosynthesis, but not moulting. We propose that these surprising patterns of redundancy are the consequence of paralogue specialisation following gene duplication. It is likely that the ancestor of the Warthog family was a pleiotropic regulator of ecdysis, involved in both shedding the old cuticle and synthesising the new, yet following the generation of the ten members by many tandem gene duplication events, these functions were distributed among paralogues such that *wrt-1*, *wrt-2*, *wrt-4*, *wrt-5* and *wrt-8* all retained moderate roles in moulting, while *wrt-3* and *wrt-9* have independently specialised in cuticle biosynthesis.

Paralogy relationships do not predict redundancy relationships in the Warthog family

The unexpected redundancy relationship between *wrt-3* and *wrt-9* could be explained by their independent specialisation in cuticle biogenesis, giving rise to a rarely described phenomenon of stable redundancy (SR) preserved through unexpectedly long evolutionary timescales. This contrasts with patterns of redundancy often observed between many recently derived paralogous genes, which we term 'Transient-Duplication-Associated-Redundancy' (TDAR). TDAR can be thought of as the evolutionarily unstable short term consequence of duplicated genes, which inevitably exists immediately following a gene duplication event prior to a period of divergence. SR on the other hand, is a possible means by which gene duplications could instil robustness in gene regulatory networks, and thus provides a long term selective advantage which allows it to persist over long evolutionary timescales.

CHAPTER 5

Lessons on asymmetric gene divergence from the Drd family: Pervasive redundancy in spite of radical paralogue diversification

Introduction

The previous two chapters have made it abundantly clear that paralogues can radically alter gene family dynamics when they adopt fates that are not comparable to one another. This as yet unexplored and mechanistically thorny issue (in this thesis and beyond it) may be more helpfully thought of in terms of the extent to which paralogues change over time relative to their ancestor. A state of affairs in which paralogues adopt equal rates of change relative to their ancestor is what one might call 'symmetrical divergence'. With reference to symmetry, it is not suggested that paralogues accumulate similar changes, or even adopt the same evolutionary fates as each other, just that they diverge at a similar rate. When the opposite is true — so-called asymmetric divergence — it can be said that one paralogue accumulates a greater level of mutational change than its counterpart (Holland et al. 2017). Such that, so as to visually represent this in some way, if we were to put a pin in the ancestral state of a gene, following its duplication, symmetric diversification would lead to the equidistant placement of two new pins away from it. In an instance of asymmetric gene diversification, however, it follows that one of these pins would be much further away indeed. This metaphor is intentionally put, as it describes quite literally the outcome of asymmetric gene divergence as it appears in a phylogenetic tree — that is to say a very long branch.

It is at best unhelpful, and at worst downright inaccurate, to refer to members of gene families as being either newer or older relative to one another. One locus, following duplication, is neither the 'parent' nor the 'original'. Both daughter loci are considered the same age as one another. While it may seem overly pedantic to state this so vehemently, it is a fact essential to make explicit in this introduction given that such terminology has the potential to frame the way paralogues are investigated, or interpreted, in any study claiming to characterise the evolution of their function. But this is not to say that one paralogue cannot more closely resemble the ancestral state; *tbx-35* and *tbx-36* are proof of this very idea. Central to dictating how far new duplicates are *able* to or are *permitted* to stray from their ancestor must be factors and mechanisms that are yet to be characterised. How are some paralogues able to innovate so radically and others not? What is the opportunity cost of departing so dramatically from the ancestral state? These are fundamental questions to ask of duplicated genes but are largely unanswerable in studies of big families.

Both the T-box and the Warthog families display multiple instances of asymmetric diversification, but the causes and consequences of any given instance were not wholly understandable per se given the sheer frequency at which they were estimated to have occurred in the families more broadly. Besides, the number of paralogues present in both families leaves open the possibility that more complex dynamics are at play (such as redundancy among more distantly related members as found in the Warthog family). Therefore, the only way we are able to answer the questions posed above more directly is to use, as a model, a much smaller gene family where only one instance of asymmetric paralogue diversification is detected.

The gene family chosen for such a task is one that was identified, named, and characterised, as part of this investigation: the Drd family. Presently, nothing is known about the Drd family in *C.*

elegans. Orthologous to the human fatty acid hydroxylase domain containing protein FAXDC2, *drd-1.1*, *drd-1.2*, and *drd-1.3* have been named during this investigation following their identification by the paralogue mining exercise summarised at the end of Chapter 1. These genes have been named according to the preexisting nomenclature given to what we will be referring to as *drd-1.1* (formerly known as *drd-1* (for Dietary Restriction Downregulated)). After the three genes were identified as being paralogous, it was realised that nothing was known about their roles in the worm. In fact, the only reason *drd-1.1* had been assigned to the 'Drd' class of genes at all is because it was automatically named as such from the RNA-seq experiments performed as part of the modENCODE enterprise (Gerstein et al. 2010; Araya et al. 2014). Beyond this facet of its expression, that is to say it *is* presumably downregulated upon dietary restriction, nothing else is known about *drd-1.1*. Despite not being bona fide members of the Drd class (i.e., not being identified as such from the modENCODE project), it was logical to attribute related names to the two paralogues, and so F49E12.10 became *drd-1.2* and F35C8.5 became *drd-1.3*. Like many small gene families that were identified from the preliminary analysis presented in Chapter 1, the evolution of the Drd family was initially characterised and the hallmark of asymmetric paralogue diversification was readily apparent in it, i.e. one member of the family has a very long branch relative to the other two. This will be fully elaborated on in due course.

There is not a great deal of utility to be gained from introducing the FAXDC2 orthologue present in humans, mostly because so little is known about it, but also because what is known reveals it to be involved in the fairly derived process of megakaryocyte maturation (Jin et al. 2016) — these being hematopoietic cells responsible for the production of blood platelets. Though at a more fundamental mechanistic level, FAXDC2 is known to be regulated via extracellular signal-regulated kinase (ERK) signalling and its downstream effector, RUNX1 (Runt-related TF 1) (Jin et

al. 2016). With ERK signalling (reviewed by Sundaram 2013) and RUNX genes (Kagoshima et al. 2007) being drivers of cell proliferation that are also found in nematodes (demonstrably performing the same role), it is for this reason that this aspect of FAXDC2 activity in humans is potentially valuable to highlight in passing in this introduction. Additionally, the Drd family in nematodes, like FAXDC2 in humans, is predicted to have oxidoreductase activity — enzymes that catalyse oxidoreduction reactions, that is the transference of electrons from one molecule (the oxidant) to another molecule (the reductant). Examples of oxidoreductases are indeed hydroxylases, but also extend to oxygenases and reductases, all of which, collectively, play important roles in both aerobic and anaerobic metabolism.

What the Drd family lack in terms of phenotypic characterisation (at present), they more than make up for in their reduced number of members and thus relative simplicity compared to the families studied in the previous two chapters. This is a considerable advantage for our purposes in this Chapter. It is inherent to taking a reverse evo-devo approach that finding interesting genetic phenomena may lead one to genes or gene families which they know nothing about, or even about which nothing is known. It is of course an advantage, in this kind of study, to probe the evolutionary dynamics of families about which at least basic functionality has already been characterised; but the first priority must be to investigate the evolutionary genetic phenomenon of interest, which in this case is a sole, extreme instance of asymmetric gene divergence. As such, the causes and consequences of asymmetric gene diversification will be thoroughly probed, but not before this sole instance of it in the Drd family alluded to above is fully described.

Results

Characterisation of the Drd family reveals hitherto unknown paralogues in the *Caenorhabditis* genus, one of which appears to be rapidly evolving

On mining exhaustively for paralogues exclusively belonging to the *Caenorhabditis* genus, the Drd family were identified. Found to comprise of: F49E12.9, F49E12.10, and F35C8.5, these genes were subsequently renamed: *drd-1.1*, *drd-1.2*, and *drd-1.3*, respectively. Aforementioned 'quick-and-dirty' phylogenetic analyses were run on sole instances of duplication and triplication in the *Caenorhabditis* genus and from preliminary inspection, it appeared that *drd-1.3* was rapidly evolving — that is to say it had a long branch indicative of asymmetric diversification. Wishing to characterise this in a more phylogenically rigorous way, the evolutionary history of the Drd family was explored so as to ascertain if this was a genuine case of asymmetric paralogue diversification.

Figure 5.1A shows a Bayesian phylogenetic analysis of the Drd family in the *Caenorhabditis* genus, including the single-copy outgroup in *C. japonica.* It is worth taking a moment to point out that a single-copy member of the Drd family is also present in *P. pacificus* (that is to say it has not been lost or independently duplicated therein). But in wishing to comply with the unwritten rule of outgroup selection: *being as closely related as possible to the species used within the gene tree*, the orthologue in *P. pacificus* was sidelined in this phylogenetic characterisation and *C. japonica* was chosen instead. It should be mentioned that the single Drd family member present in *P. pacificus* will be the one that is investigated functionally (in lieu of the orthologue in *C. japonica*), but the reasons for why this is and the results to follow will be returned to later. For now, and based on the Bayesian phylogram shown in Figure 5.1A, it is proposed that the Drd family arose from a single triplication event, and that the long branch leading up to *drd-1.3* **cannot** be underpinned by relaxed selection.

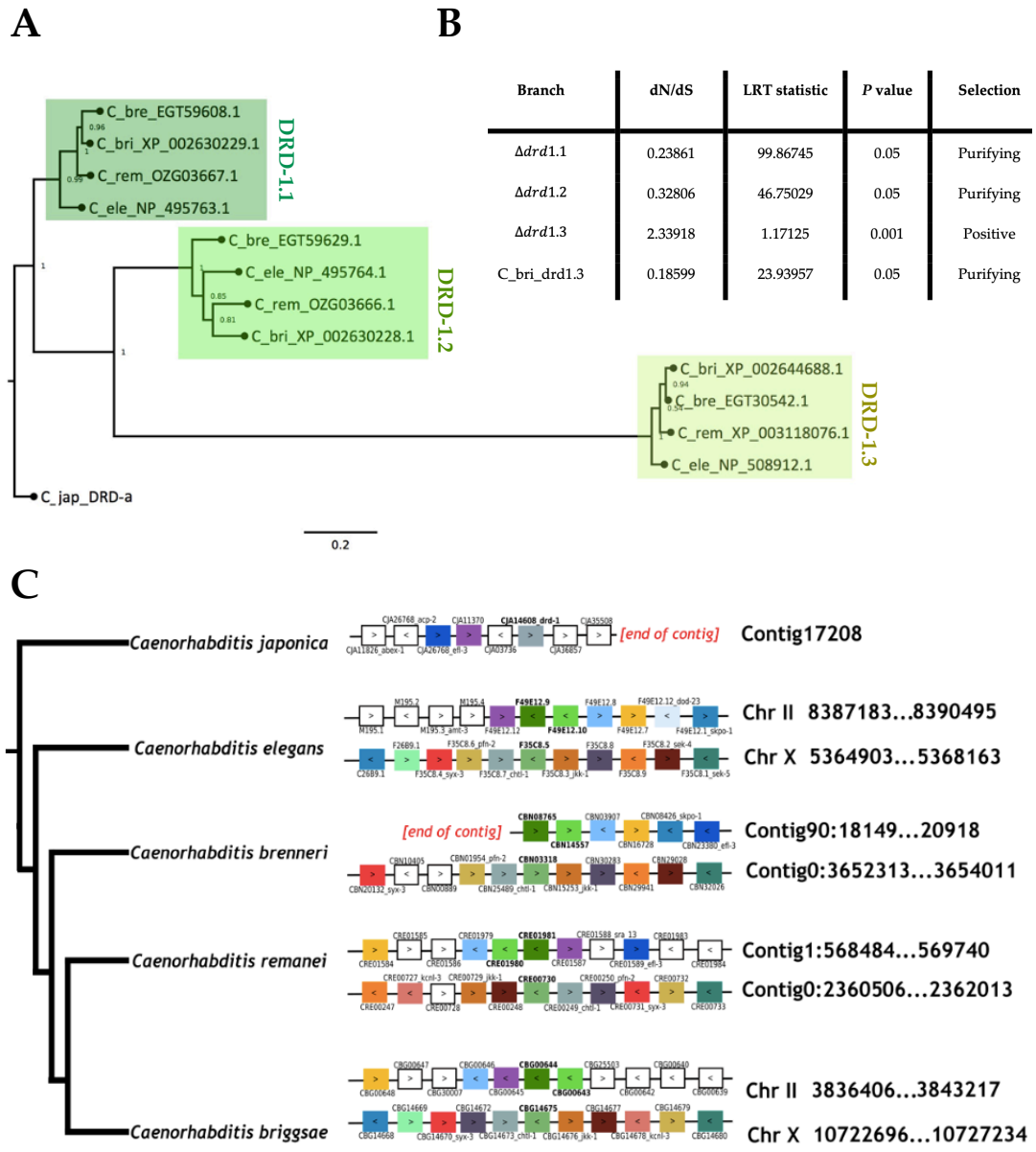


Figure 5.1. Phylogenetic and synteny analysis of the Drd family in the *Caenorhabditis* genus. (A) Bayesian likelihood phylogenetic analysis of the Drd family built from an amino acid sequence alignment of the Drd homologues mined from the predicted proteomes of the *Caenorhabditis* genus. Species abbreviations: Cj, *Caenorhabditis japonica*; Cbre, *C. brenneri*; Cbri, *C. briggsae*; Cr, *C. remanei*; Ce, *C. elegans*. Scale bar is substitutions per site per million years. Node values are posterior probabilities as determined by Bayesian inference. (B) Results of a PAML analysis run on the Drd family in the *Caenorhabditis* genus. Consensus sequences of *drd-1.1*, *drd-1.2*, and *drd-1.3* from each species in which they are present were compiled (to generate *Δdrd-1.1*, *Δdrd-1.2*, and *Δdrd-1.3*) so as to enable the branch leading up to each gene to be tested, rather than any particular paralogue in just one of the species. The final row in the table explicitly tests the kind of selection acting on the *drd-1.3* paralogue in *C. briggsae*. The columns in the table provide the dN/dS ratio of each branch specified as foreground where values in excess of 1 indicate positive selection, and values less than 1 indicate purifying selection. Significance was calculated using the Likelihood Ratio Test (LRT) — the corresponding *P* value is also given indicating the level of significance. The final

column summarises the type of selection in the branch tested. (C) Synteny analysis of the Drd family in the *Caenorhabditis* genus and the outgroup, *C. japonica* (the most basal species in the genus in which a member of the family is present). Synteny was compared manually across species where the tBLASTn of each gene in that region was mapped back to other species. Where genes are assigned a colour, that means they have an orthologue with conserved synteny in another species present in the diagram. Where genes are depicted as white boxes, that means their orthologue is not shown in the microsyntenic region depicted. Genes of interest are shown centrally in the diagram, so *drd-1.1* (F49E12.9, dark green - LG II) *drd-1.2* (F49E12.10, bright green - LG II) and *drd-1.3* (F35C8.5, pale green, - LG X) feature in the middle of all microsyntenic regions.

To take the first of these deductions, the origin of the three genes must have been relatively close to the base of the genus seeing as though the only species in which these paralogues are not detected is the most basal — *C. japonica,* in which only one Drd family member is present. This dates their genesis to approximately 150 to 180 mya (Baker and Woollard 2019). The triplication was most probably in tandem, with *drd-1.1* and *drd-1.2* still remaining adjacent to one another on chromosome II (shown in Figure 5.1C). This implies *drd-1.3* subsequently translocated to the X chromosome more or less straight after the triplication event. We know it cannot have been long after the generation of these three paralogues as no *drd-1.3* orthologues are found alongside *drd-1.1* and *drd-1.2* in any *Caenorhabditis* species.

Next, with regard to the second conclusion stated above, it is considered unlikely that the rapid evolution of *drd-1.3* is due to relaxed selection because while the branch leading up to the DRD-1.3 clade is indeed long, the sub-branches leading to the *drd-1.3* orthologues in the individual species are extremely short. Thus, it is implied that the evolution of the orthologues in these species is slow — that the *drd-1.3* orthologues across *C. elegans*, *C. brenneri*, *C. remanei*, and *C. briggsae* are mostly identical — but the evolution of the gene as it emerged in the genus collectively was fast. If *drd-1.3* were under relaxed selection, that is tending towards pseudogenisation, it is logical to assume that the orthologues in each *Caenorhabditis* species would be rapidly evolving too. Though, this is only an apparently likely scenario based on the phylogram as is stands. No gene tree can

purport to be conclusive proof of any one kind of selection over an other. That is not what they are designed to test. It is therefore necessary to decipher the type of selection directly using methods specifically developed for doing so.

Phylogenetic Analysis by Maximum Likelihood, or PAML for short, is a package of phylogenetic programmes, one of which is capable of detecting the mode of selection acting on a particular sequence in a gene tree (Yang 2007). In essence, the relevant programme implemented in PAML, called CodeML, operates by calculating the ratio of non-synonymous to synonymous substitutions (dN/dS) in any sequence that the user specifies. As with all ratios of this sort, if dN/dS = 1, then the whole coding sequence is considered to be evolving neutrally; when 0 < dN/dS < 1, it is under constraint, and when > 1, the sequence is considered to be under positive selection.

With the aim of detecting the type of selection acting on *drd-1.3* relative to the other two paralogues, dN/dS ratios to test for positive selection were calculated for the Drd family paralogues. However, for reasons outlined above, it would be misleading to specify any one *drd-1.3* orthologue in particular from just one *Caenorhabditis* species; the orthologues themselves are seemingly fixed in each species (and so probably under selective constraint). Not wishing to confound or confuse the result with multiple kinds of selection acting simultaneously, the amino acid sequences of each paralogue were compiled from between the species to yield an 'average' or consensus sequence for each family member. These are denoted: *Δdrd-1.1*, *Δdrd-1.2*, and *Δdrd-1.3* in Figure 5.1B. The results of the PAML assessment (the dN/dS ratio) for each Drd family member will be taken in turn.

First, the selection acting on the *drd-1.1* paralogue at large across all species (*Δdrd-1.1*) is purifying. Second, and likewise, the selection acting on the *drd-1.2* paralogue at large across all species (*Δdrd-1.2*) is purifying. However, in the third row of the table, specifying the long branch leading to all *drd-1.3* orthologues (*Δdrd-1.3*), reveals the paralogue is uniquely and evidently under strong positive selection. And crucially, so as to conclusively resolve that the rapid evolution of *drd-1.3* is not underpinned by its pseudogenisation, the short branch leading solely to the *drd-1.3* orthologue in *C. briggsae* reveals it to be under purifying selection. All values, it goes without saying, are above the threshold for statistical significance. As it is established that *drd-1.3* is rapidly evolving not because it is degenerating in sequence and function, it is now essential to characterise what such strong positive selection means for the role of *drd-1.3* over the course of evolution. Our hypothesis, it would follow, is that it has in some way gained functionality. To establish whether this is indeed true, the whole family needs greater functional characterisation.

A joint enterprise: Overlapping roles for the Drd family in male tail differentiation, gonadogenesis, and the starvation response

Starting from a blank slate in regards to any possible functionality of the Drd family, it was first important to assess the role(s) of individual family members using a reverse genetic approach. The technique deployed for such a task was knockdown by RNAi. By way of very broad summary, the Drd family were all, to varying degrees, implicated in three major developmental and homeostatic processes: male tail development, gonadogenesis, and, as is perhaps expected from their nomenclature alone, dauer entry and maintenance. We shall now explore each of these roles in turn.

Male tail development

Without being in possession of any real prior knowledge concerning the roles of the Drd family, a starting point is to interrogate the only known aspect of their supposed functionality (in the starvation response) by modelling the selfsame RNA-seq dataset from which the nomenclature of *drd-1.1* was derived. Analysis of the temporal expression profiles of the three Drd family members in Figure 5.2A reveals that while they may all be expressed to variable degrees throughout dauer formation and exit — which will be discussed in the fullness of time — the highest level of expression recorded for any member of the family is *drd-1.3* in male larvae. Owing to the exceedingly small proportion of males in ordinary populations of *C. elegans*, it is not generally obvious when male-specific phenotypes arise upon gene knockdown in a wildtype background. Therefore, RNAi knockdown of *drd-1.3*, alongside the other the two family members, was performed in a high incidence of males background, specifically using the *him-8(tm611) IV* allele. When knockdown of *drd-1.3* was first performed by feeding in a *him-8* mutant, it was noticed that the tails of adult male animals appeared deformed and misshapen. Before elucidating the morphological basis of the aberrant male tails in *drd-1.3* knockdowns, it is first necessary to outline the basic development of the male tail in the briefest possible terms.

At the vaguest level of biological organisation, hermaphrodites and males appear the same, yet almost all tissues show some degree of sex-specific specialisation. Outwardly at the gross morphological level, males are slightly shorter and thinner than hermaphrodites, but it is the tail which is the most distinctive feature of the male animal, the development of which is illustrated in Figure 5.2B. In contrast with the tapered tails of hermaphrodites, L4 and adult males possess an elaborate fan structure (see Figure 5.2C), the initial development of which is visible during L3. The cuticular fan embeds 18 sensory rays (9 on each side) creating a sensory organ surrounding the

spicule, used for sperm transfer, with which the male can mate. The rays of the male tail are derived from the posterior seam cell lineages during larval development (Altun and Hall 2009).

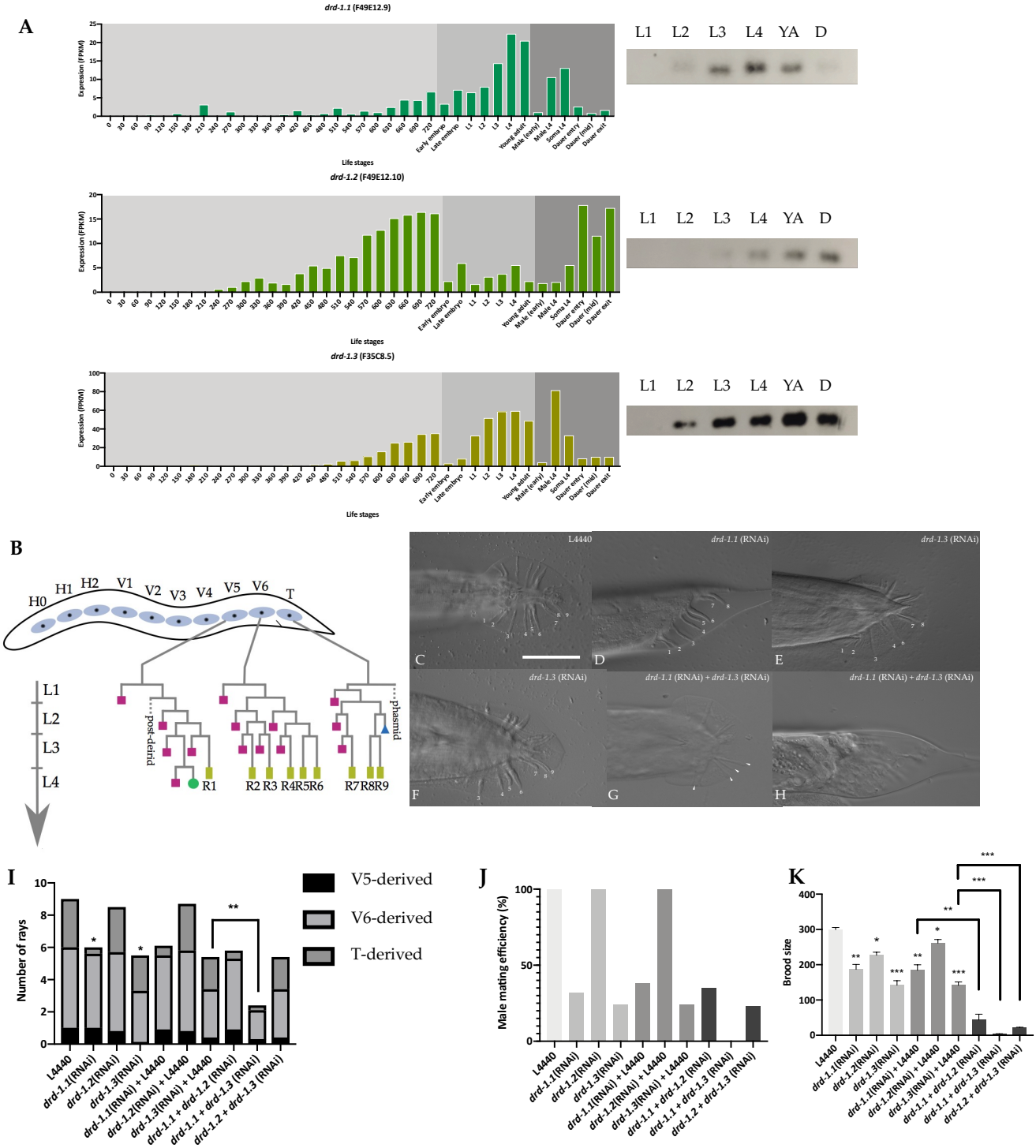


Figure 5.2. Roles of the Drd family in male tail differentiation. (A) RNA-seq analysis where median transcript levels (FPKM) as produced by modENCODE are provided. Expression levels for each life stage are provided including throughout embryogenesis (minutes post-fertilisation) and

at various post-embryonic stages, including dauer entry, mid-dauer, and dauer exit. Note that the Y-axes of each bar graph is scaled differently to enable the changes in the mRNA levels of each gene across the stages to be more easily observed. Beside each bar graph shows the results of an RT-PCR analysis from a single animal (at each stage) which reveals the levels of cDNA (derived from the total transcriptome) in: L1, L2, L3, L4, young adult, and mid-dauer animals. (B) Lineage diagram of the posterior seam cells V5, V6 and T, which produce the nine male-specific sensory rays. Purple squares represent cells that fuse with the hyp7 syncytium; olive green rectangles indicate ray precursor cells. The V and T lineages of the male are identical to those of the hermaphrodite until the end of L2. Divisions are asymmetric and stem cell-like, with the anterior daughter adopting the syncytial fate and the posterior daughter adopting the proliferative fate (Sulston and Horvitz 1977). The characteristic ray sensilla are formed by retraction of the hypodermis surrounding the ray cell groups, leaving finger-like protrusions embedded in the cuticular fan. (C) Shows the appearance of a wildtype adult fan, on treatment with the Empty Vector (EV) control, which shows the nine rays of the male tail numbered as such. (D) Example of a male tail upon *drd-1.1* knockdown shown in a side (lefthand) presentation. Only eight rays are visible, with one from the T-derived lineage being absent. (E) Example of a male tail upon *drd-1.3* knockdown demonstrating that, this time from the dorsal presentation, the absence of rays from multiple lineages (V6 and T), as well as their short, stunted structure. (F) A similarly representative image is also shown of a male tail upon *drd-1.3* knockdown. (G) An example of a male tail upon both *drd-1.1* and *drd-1.3* combinatorial knockdown in a dorsal presentation. Rays appear malformed and missing to such an extent that they cannot be identified (with respect to the lineage from which they were derived) and so numbered meaningfully. (H) Also represents a similar male animal as in the previous. (I) The quantification of the male tail phenotypes in panels (C - H) is shown as a partitioned bar graph where not only are the total number of rays observed but also the relative number derived from each posterior seam cell lineage. Significance levels here are only shown for total numbers of rays in each RNAi treatment for clarity. (J) Results of male mating efficiency assays performed for the Drd family RNAi knockdowns (including relevant dilution controls alongside double RNAi knockdown) where mating behaviour in adult male nematodes was assayed by mating efficiency, i.e., the number of cross progeny sired by males under standard conditions. Being expressed as a % between experimental repeats (to enable 50≥ animals per condition to be tested), error bars are not shown as the results are % values between experimental repeats. (K) Brood sizes of Drd family RNAi knockdowns (including relevant dilution controls alongside double RNAi knockdown)— all progeny scores as part of the brood survived to hatching and beyond L1 such that these can be considered viable broods, i.e., dead eggs and dead L1s were removed from the counts. Broods were scored in triplicate and a mean viable brood was calculated. Black bars show mean + SEM. Black asterisks (****$P \leq 0.0001$, ***$P \leq 0.001$, **$P \leq 0.01$, *$P \leq 0.05$, ns$P > 0.05$) show statistically significant differences in the means compared to RNAi animals where the statistical significance is confirmed by the results of an unpaired t-test.

Collectively, the seam lineage comprises stem cell-like epithelial cells which lie along each side of the worm that undergo reiterative divisions during larval development.

At hatching, seam cells run in lines of ten unfused cells along both lateral sides of the animal and are embedded in the hyp7 syncytium along the body of the worm, starting just posterior to the head and ending anterior to the tail. The T lineage represents the posterior-most end of the seam, located in the tail, where the slightly anterior V5 and V6 lineages, as well as T, produce the sensory rays of the male tail (Sulston and Horvitz 1977).

With this in mind from the outset, the defects in the male tails of Drd family knockdowns were characterised; such that not only were total ray numbers quantified, but where rays were absent, it was noted from which lineage they were absent so as to provide greater insight as to the developmental basis of the roles these genes may play in male tail development. The results of this phenotypic characterisation are shown in Figure 5.2D to I. Surprisingly, *drd-1.3* was not the only family member to be implicated in male tail development; *drd-1.1* also displays *similar*, but not the same, kinds of defects. Yet *drd-1.2*, on the other hand, did not display any male tail defects at all.

Males fed an empty vector RNAi construct have 18 sensory rays distributed evenly across the dorsoventral sides of the adult male animal (numbered 1-9 in Figure 5.2C). Upon knockdown of Drd family members by RNAi, fewer than 9 rays on each side are observed in *drd-1.1* (RNAi) and *drd-1.3* (RNAi) animals. While all rays were at some point seen to be absent or malformed in Drd family knockdowns, *drd-1.1* (RNAi) animals primarily lose rays derived from the seam cell lineage T, rather than V5 and V6 (Figure 5.2D). Meanwhile *drd-1.3* (RNAi) animals, in the main, lose rays derived from the seam cell lineages V5 and V6 rather than T (Figure 5.2E, F). When knocked down together, rays derived from all three posterior seam cell lineages are lost giving rise to greater, more severe defects exemplified in Figure 5.2G and H — and quantified in Figure 5.2I — indicative of partial redundancy between these two paralogous genes. Concordant with the male tail defects

seen in Drd family knockdowns is the inability of these males to successfully mate as assayed by attempting to cross them with females, or *fem-1(hc17) IV* mutants, and observing their capacity (or lack thereof) to produce offspring (method courtesy of J. Hodgkin, pers. comm.). The results of these male mating efficiency assays can be found in Figure 5.2J where it is clearly seen that when abolished together, *drd-1.1* (RNAi) and *drd-1.3* (RNAi) males fail almost entirely to mate at all successfully.

As mentioned above, the RNAi performed to generate the results in Figure 5.2 was delivered by feeding. In the interest of establishing meaningful experimental controls, when the results of a given instance of double knockdown is considered here, the appropriate course of action is, as is shown, to compare it to an instance in which a single Drd family member has been knocked down at half concentration. Reassuringly, in all cases, the efficacy of RNAi knockdown does not appear to be significantly compromised by the dilution of the dsRNA. Though presumably there is a limit to the extent to which dsRNA can be diluted and still yield effective knockdown of sufficient levels of transcript. It is possible to make such a claim because attempts were made to perform knockdown of all three Drd family members concomitantly, and, whether delivered by feeding or injection, triple knockdown was ineffectual with respect to all phenotypes investigated in this Chapter (data not shown as attempts at triple knockdown did not abolish transcripts sufficiently to yield any phenotypes). But whether diluted in double knockdown or concentrated for single knockdown, it is to be acknowledged that, in line with its lack of expression in the same, *drd-1.2* is not required for male-specific aspects of development, including the tail.

With the roles of *drd-1.1* and *drd-1.3* in the male tail now characterised, it is clear that at least some Drd paralogues are implicated in fundamental developmental processes but cannot, as such, be

said to be required to maintain basic fitness because of course males can be dispensed with in the *C. elegans* world. However, it was noticed — upon knockdown of two Drd family members simultaneously in hermaphrodites — that a more fundamental property of organismal fitness was affected by the loss of these genes. Knockdown of **all** family members singly gives rise to a reduction in fitness, that is to say, a lower frequency of surviving progeny in adult hermaphrodites (Figure 5.2K). In a way this is not overly surprising because all three paralogues are expressed at significantly detectable levels in young adults as shown from the analysis of the young adult transcriptome by both RNA-seq and RT-PCR (back in Figure 5.2A). Though this does not implicate the Drd family in fertility per se; there are many cell types and thus kinds of cellular processes occurring at this stage.

Gonadogenesis

To determine the basis of their smaller brood sizes, the hermaphrodite gonads of Drd family single and double knockdowns were imaged. Following the staining of the germline with DAPI to delineate the structure of the gonad more clearly, it became apparent that while the loss of only a single Drd family member has no observable effect, when two members are knocked down together (either *drd-1.1* and *drd-1.2*, or *drd-1.1* and *drd-1.3*, or *drd-1.2* and *drd-1.3*), the hermaphrodite gonad appears to develop improperly.

Wildtype gonad morphology is shown in Figure 5.3A and in direct comparison, the gonad arms of all Drd family double knockdowns appear miniaturised and / or lacking nuclear density (as in Figure 5.3E, F and G). The same cannot be said when only a single member of the family is knocked down (exhibited in Figure 5.3B, C and D). The extent to which the gonads of Drd family double knockdowns appear miniaturised is more apparent when the hermaphrodite gonad is

dissected out of, and stained separately to, the whole animal. In a wildtype scenario (seen in Figure 5.3H), the gonad arm shown is evidently longer than those found in Drd family double

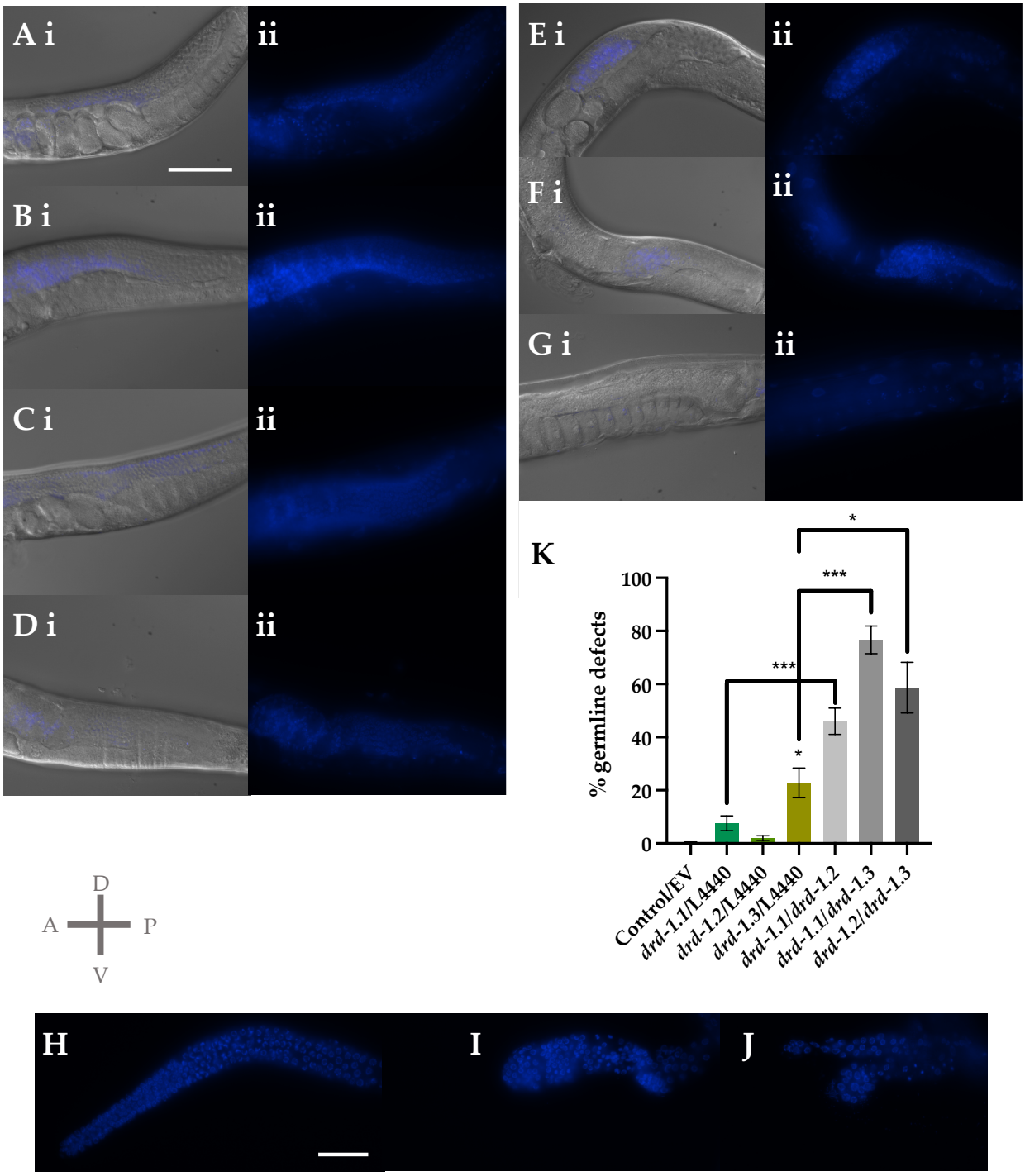


Figure 5.3. Gonad defects in the Drd family. All young adult hermaphrodite gonads pictured in panels (A — G) were DAPI stained in the whole animal (un-dissected) and representative images of the posterior gonad following each (control/single/double) RNAi knockdown were taken and are provided here. All L4 hermaphrodite gonads in panels (H — J) were DAPI stained following dissection following each (control/single/double) RNAi knockdown. (A i and ii) WT/L4440 control DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI shown on its own (ii). (B i and ii) *drd-1.1*/L4440 DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI

shown on its own (ii). (C i and ii) *drd-1.2*/L4440 DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI shown on its own (ii). (D i and ii) *drd-1.3*/L4440 DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI shown on its own (ii). (E i and ii) *drd-1.1*/*drd-1.2* DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI shown on its own (ii). (F i and ii) *drd-1.1*/*drd-1.3* DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI shown on its own (ii). (G i and ii) *drd-1.2*/*drd-1.3* DAPI stained gonad in merged (DIC and DAPI) channels (i) with DAPI shown on its own (ii). (H) Dissected gonad arm (singular) of WT/L4440 control animal stained with DAPI. (I) Dissected gonad arm (singular) of *drd-1.2* + *drd-1.3* (RNAi) animal stained with DAPI. (J) Dissected gonad arms of *drd-1.1* + *drd-1.3* (RNAi) animal stained with DAPI. (K) Showing quantification of germline defects as a percentage penetrance across all conditions: L4440 (Empty Vector) control; dilution controls for *drd-1.1*, *drd-1.2*, and *drd-1.3* where each RNAi feeding construct has been mixed with an equal amount of Empty Vector; *drd-1.1*/*drd-1.2* double RNAi knockdown; *drd-1.1*/*drd-1.3* double RNAi knockdown; and *drd-1.2*/*drd-1.3* double RNAi knockdown. Defects were recorded as a percentage penetrance where any deviation from wildtype gonad morphology was present — invariably here, in the size of the gonad. Quantification in panel K was performed only on dissected young adult gonads such that the full size and scale of the germline was evident, and any defects were clearly observable. Gonads were dissected, fixed and stained as described in Chapter 2. Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to RNAi knockdowns where statistical significance is confirmed by the results of an unpaired t-test. In the instance of quantification in (K), 50≥ animals per treatment were examined. Scale bars = 50 μm.

knockdowns (seen in Figure 5.3I and J) — despite all being dissected, fixed and stained during late L4. These results lead us to assert that the hermaphrodite gonad is a developmental context in which true redundancy is observed among the Drd paralogues. Indeed, this would certainly explain the patterns of redundancy with respect to the diminished broods in the family shown earlier. This too, is the first suggestion of a role for *drd-1.2*. But this newly ascribed role for the whole family still fails to explain the asymmetric divergence of *drd-1.3* given neither the development of the male tail nor gonadogenesis can credibly be said to be aspects of *drd-1.3* functionality that are unique to only it. That being said, the quantification of germline defects shown in Figure 5.3K illustrates that *drd-1.3* is the only member of the family which has a non-redundant role (with its fellow paralogues) in gonadogenesis. By contrast, *drd-1.1* and *drd-1.2* do not display a significant penetrance of germline defects on single knockdown. In the case of all simultaneous knockdowns of two family members together, however, there is redundancy evident

in all possible combinations. This shows that all family members are, to variable extents, involved in gonadogenesis.

Reflecting on these data, it is perhaps only by looking beyond development and reproduction and instead to other processes, i.e., the namesake of the family — the starvation response — may we understand the supposedly radical asymmetric divergence of *drd-1.3*.

Dauer entry and the starvation response

A somewhat different approach was taken in grappling with the role of the Drd family in the starvation response as compared to the reverse genetics deployed immediately in the cases above. This is because, regardless of what the Drd paralogues may or may not be doing that is observable at the physiological level (on the loss of their function), it is expected that at least one member of the family can be defined by its molecular behaviour: on its downregulation upon dietary restriction. It is for this reason that transcriptional reporter constructs were first built for each member of the family — simple enough in conception and construction, merely requiring the promoters of each gene to be 'hooked up' to a fluorophore in a plasmid vector which were then injected as extrachromosomal arrays.

It was seen from the RNA-seq dataset in Figure 5.2A that all family members, apparently, are expressed at some level on entry into, and during the dauer stage, and at dauer exit. But on closer inspection, *drd-1.2* is *unlike* its paralogues in that relative to other life stages (at which dauer can be entered), it actually appears to be upregulated when environmental conditions no longer remain favourable and remains expressed throughout dauer itself. This is in contrast to *drd-1.1* and *drd-1.3* which are seemingly downregulated as they enter into dauer proper, at least as per the temporal

expression profiles provided. But there is a caveat to drawing these overarching conclusions; the expression levels of each family member are not comparable on the face of it. In fact, during dauer, *drd-1.3* is expressed at approximately the same median levels as *drd-1.2*. But when it comes to evaluating the significance of any gene in response to any stressor, it is surely more valuable to consider the relative *change* in the level of a given transcript, rather than any static reading in and of itself. So, with the devil, as it so often is, in the detail, the expression of *drd-1.1*, *drd-1.2* and *drd-1.3* can be more helpfully thought of as decreasing six-fold, increasing three-fold, and decreasing seven-fold from L4 to mid-dauer, respectively. This entirely shapes our understanding of these genes going forward, and at the very least suggests *drd-1.2* is rather inappropriately named.

Expression of transcriptional reporter constructs made using the promoter elements of *drd-1.1*, *drd-1.2* and *drd-1.3* are shown in Figure 5.4. In all images (panels A to G), all family members are expressed, to variable extents, either in L4 animals or young adults on the brink of dauer — that is to say on the heels of starvation, and / or sustainably throughout the dauer stage. The latter though, is only the case for *drd-1.2*. No expression was observed at earlier stages of development, in either well-fed or starved animals.

In Figure 5.4A and B, *drd-1.1p::gfp* is seen in two different L4 animals following their starvation, **prior** to dauer formation. The expression of *drd-1.1p::gfp* is seen non-uniformly throughout the intestine at this stage, and crucially, no expression of *drd-1.1p::gfp* was observed in dauer — this is not shown as it would be impossible to determine whether the lack of *drd-1.1p::gfp* expression in any dauer animal would be due to that animal not expressing the construct at this stage or that it simply did not carry the extrachromosomal array to begin with. In any case, it should be put

plainly — in no uncertain terms — that no dauer animal was ever seen to express the *drd-1.1p::gfp* construct, and nor the *drd-1.3p::gfp* construct for that matter, to which the same justification applies.

Figure 5.4C depicts an L4 carrying the *drd-1.2p::mCherry* construct that is at the same stage to the L4 animals on the brink of dauer in panels A and B (having just been starved); the L4 in panel C is seen to express the construct in posterior intestinal tissue. It is of value to note for the discussion to

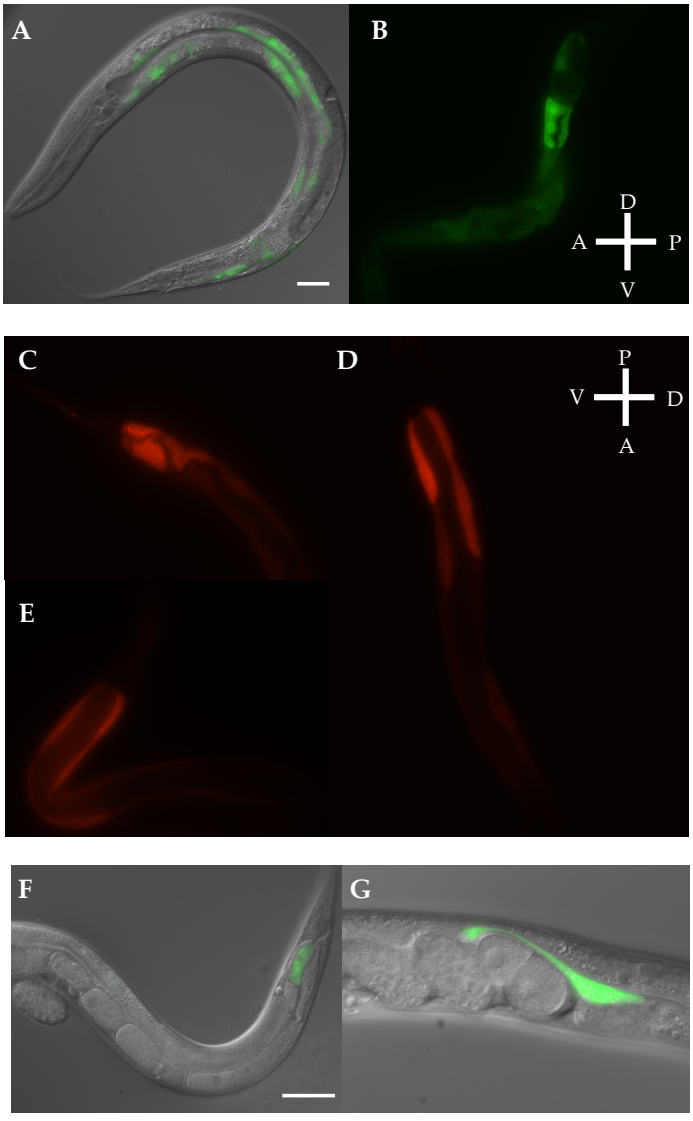


Figure 5.4. Expression patterns of the Drd family in *C. elegans*. (A, B) Expression of the *drd-1.1p::gfp* transcriptional reporter is shown in two different L4 animals on the brink of diverting their developmental programme to dauer having been grown on starvation media since the late L2/L3 stage, and have therefore been dietarily restricted. (C - E) Expression of the *drd-1.2p::mCherry* transcriptional reporter. (C) An L4 on the brink of diverting its developmental programme to dauer after having been grown on starvation media since late L2/L3 (i.e., it has been dietarily

restricted), and the same construct in two established dauers (D, E). (F, G) Expression of the *drd-1.3p::gfp* transcriptional reporter in two different young adult hermaphrodites that have been grown on starvation media since late L2/L3. No expression of *drd-1.3p::gfp* was observed in L4 animals on the brink of dauer. Scale bars = 50 μm. All images (in panels A, B, C, E, F, G) are shown in the regular orientation (with the anteroposterior axis to the left-right and the dorsoventral axis on the up-down axis, respectively, as shown in panel B), except (as described) in panel D.

follow that in no other portions of the intestine was *drd-1.2p::mCherry* observed in animals on the verge of entering dauer. Figure 5.4D and E, on the other hand, show two different animals in an established dauer state, where the expression of *drd-1.2p::mCherry* can be seen in the intestinal tissue, beside a much wider lumen than in non-dauer animals such as those in panels A, B, or C. As an aside for the exacting reader, it should be noted that precisely the same expression pattern was observed for *drd-1.2* when this reporter construct was made using GFP instead of mCherry (data not shown). In other words, it is not the case that mCherry is somehow more stable in intestinal tissue than GFP is; the transcriptional reporter construct for *drd-1.2* was only built using mCherry to facilitate direct comparison with the expression patterns of *drd-1.1* and *drd-1.3* in the same live animal. The results of which are to follow shortly.

And, thirdly and finally, in Figure 5.4F and G, *drd-1.3p::gfp* can be observed only in a posterior stretch of the intestine in young adults having just been starved. No expression of the same was observed at earlier stages on starvation, that is at L4, or during dauer itself, as explicitly stated in the above. For the avoidance of doubt, neither *drd-1.1p::gfp, drd-1.2p::mCherry* nor *drd-1.3p::gfp* were observable in the intestines of well-fed L4 larvae.

For completeness, though in the absence of a fully worked explanation, neither *drd-1.1* nor *drd-1.3* were detectable in the male tail during its development. This could be for any number of speculative reasons, though the most likely of which is that the *cis*-regulatory element responsible

for the expression of *drd-1.1* and *drd-1.3* in the male tail specifically is absent in the transcriptional reporter constructs generated as part of this investigation. It is not beyond the realms of possibility that there is a male tail-specific enhancer element that lies in an intron, perhaps, of these paralogues which is obviously not present in these promoter-only transcriptional reporters. If this were to be the case, only further dissection of their *cis*-regulatory apparatus would give us an answer. While the same reasoning might apply to the absence of expression of these constructs in the gonad, this is much more likely to be due to the general silencing of transgene arrays expressed in the germline (Kelly et al. 1998; Reuben and Lin 2002).

To summarise the evidence so far, while all Drd family paralogues are expressed in response to starvation, it would seem that *drd-1.1* and *drd-1.3* are expressed as an initial response to dietary restriction, but not during the dauer state itself. Inferentially, this implies both *drd-1.1* and *drd-1.3* are implicated in the decision to enter the dauer state; and so logically, these may well belong to a group of genes known as 'dauer decision-makers' — genes which regulate the developmental commitment to enter dauer (Karp 2018). In stark contrast to *drd-1.1* and *drd-1.3* is *drd-1.2* which, while expressed in the posterior intestine of animals in response to starvation, is expressed — in fact more strongly — throughout the intestines of animals in the dauer state itself. This sets *drd-1.2* apart from its two paralogues. And so, to opine on the small insight we had into these genes prior to beginning the results section of this Chapter, it is suggested that *drd-1.1* may have been pinned to the 'dietary restriction downregulated' gene class not because it is nothing to do with the dauer state or that high levels of *drd-1.1* expression actively prohibits it; quite the contrary. In actuality, as one might intuitively expect from a dauer decision-maker, *drd-1.1* is upregulated in response to starvation but downregulated once the dauer state has been achieved, when it is no longer required. This would, therefore, class *drd-1.1* as a gene which is technically (and literally)

downregulated upon 'dietary restriction', but it is, as we have explained, more complicated than that. So with respect to the current nomenclature, it may be said that belonging to the dietary restriction downregulated gene class is possibly misleading (in a functional sense) with respect to the role of *drd-1.1*. Turning to the role and naming of *drd-1.3*, much the same (as has been said for *drd-1.1*) can seemingly be said for it. Meanwhile, *drd-1.2* appears to be required for the same commitment to dauer, but also for the process instigated beyond it — to maintain, not merely initially commit to, the state itself. Or that is the hypothesis to test at least.

To functionally determine the roles of, and possible redundancy relationships between, members of the Drd family in dauer entry and the starvation response, a reverse genetic approach was taken. Unlike the previous instances in which RNAi knockdown was deployed to investigate the functions of the Drd family, two key differences are integral to mention about the precise strategy adopted here. First, RNAi was delivered via microinjection, as opposed to feeding, in order to facilitate Drd knockdown whilst still following a protocol to ensure animals could be starved in a controlled manner. And second, the RNAi was performed in both a wildtype background and in animals carrying the aforementioned Drd family transcriptional reporter constructs so as to illuminate the expression dynamics between the paralogues (given this very much appears to be the pivotal aspect of their divergence in the starvation response).

The results of the functional characterisation of the Drd family in the starvation response are shown in Figure 5.5. Before moving on to any quantification of the dauer phenotype, it was first important to ascertain how the paralogue dynamics changed upon the knockdown of a family member so as to enable us to establish if any functional compensation occurs when one of the genes is compromised. Of course, because these are transcriptional (as opposed to translational)

reporters, it is possible to perform the knockdown of native transcripts and observe any compensatory responses in the transcriptional regulation of that gene, as well as in any related

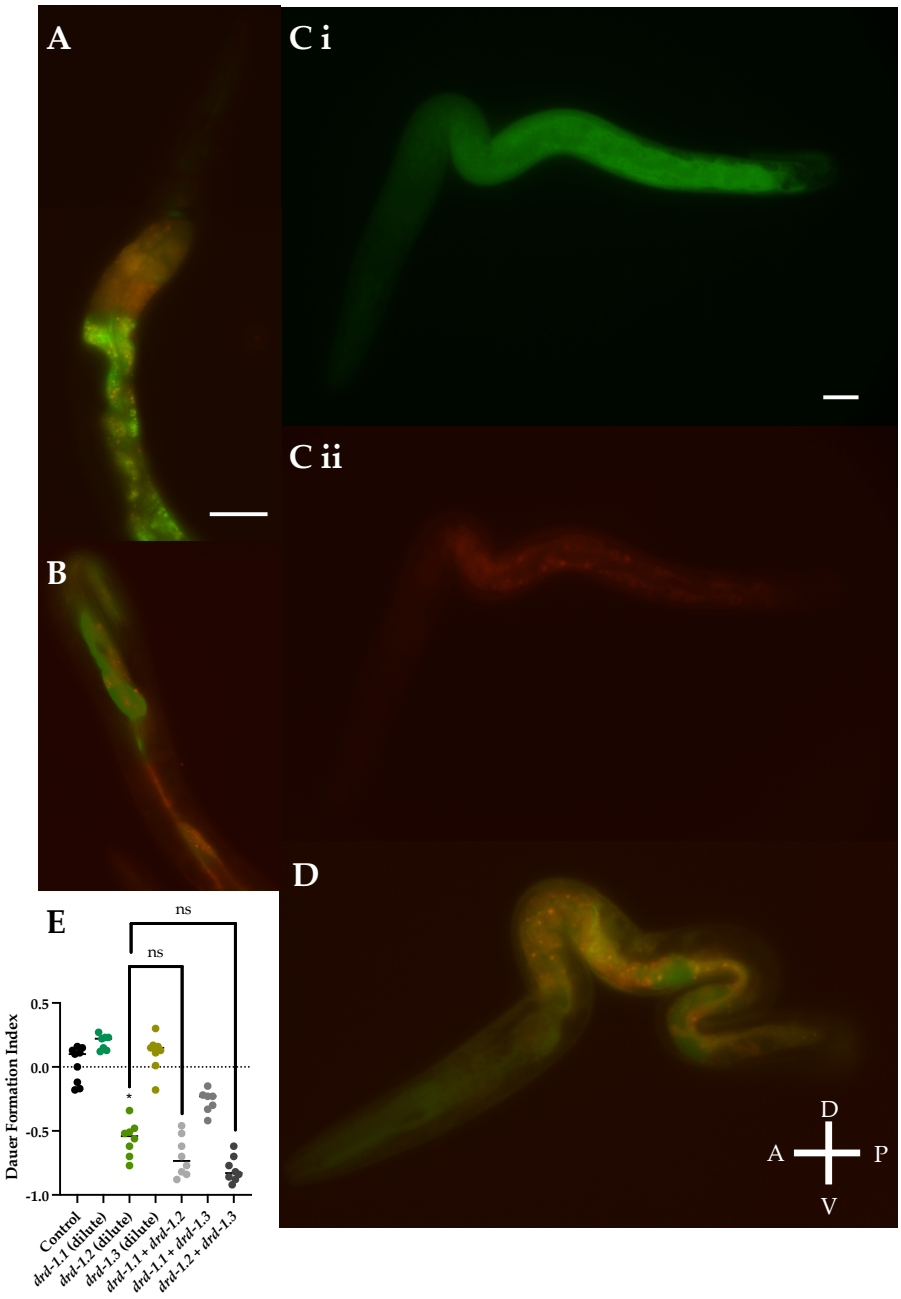


Figure 5.5. Role of the Drd family in the starvation response in *C. elegans*. (A, B) Representative image showing the expression of the *drd-1.1p::gfp* and *drd-1.2::mCherry* transcriptional reporters in two different L4 animals on *drd-1.2* RNAi knockdown and on the brink of diverting their developmental programme to dauer having been grown on starvation media following injection (of *drd-1.2* dsRNA) in the parents. (C, D) Representative image showing the expression of the *drd-1.2::mCherry* and *drd-1.3::gfp* transcriptional reporters in two different L4 animals on *drd-1.2* RNAi knockdown and on the brink of diverting their developmental programme to dauer having as they have been grown on starvation media following injection (of *drd-1.2* dsRNA) in their

parents. (E) Dauer formation assays for Drd family knockdown animals (delivered by injection so dilution of single knockdowns is shown as a relevant control for double knockdown) where n ≥ 50 across all assays. For quantification here, knockdown was performed in a wildtype background. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants with a Welch's t test. Scale bars = 50 μm.

genes, by observing changes in the expression of their reporters. With this in mind, two strains were constructed, one carrying the *drd-1.1p::gfp* reporter as well as the *drd-1.2p::mCherry* reporter, and the other carrying the *drd-1.2p::mCherry* reporter and the *drd-1.3p::gfp* reporter.

RNAi knockdown was performed for each family member on both of these strains, but changes were only observed in the expression of the transcriptional reporters upon the knockdown of *drd-1.2,* so these are the results which are shown. Knockdown of the other two paralogues, it is important to acknowledge, showed no observable difference in the reporters as compared to animals treated with Control RNAi — the reasons for which we shall come on to.

Figure 5.5A and C show the expression of *drd-1.1p::gfp* and *drd-1.2p::mCherry* in two different L4 animals on the brink of diverting their developmental programmes to dauer having been starved, where only the animal in panel C has undergone *drd-1.2* RNAi knockdown. Similarly, Figure 5.5B and D show the expression of *drd-1.2p::mCherry* and *drd-1.3p::gfp* in two different L4 animals on the brink of diverting their developmental programmes to dauer having been starved where only the animal in panel D has undergone *drd-1.2* RNAi knockdown. Comparing both instances with the native expression patterns of these reporters shown in panels A and B (Control conditions) as well as in Figure 5.4 (i.e., all animals with wildtype levels of *drd-1.2* in the transcriptome), reveals clear regulatory responses at the transcriptional level among the whole family upon *drd-1.2* knockdown.

Taking the regulatory response of *drd-1.1* first, Figure 5.5A shows the expression of *drd-1.1p::gfp* (as well as *drd-1.2p::mCherry*) in the absence of *drd-1.2* knockdown, i.e., Control conditions. As compared to Figure 5.5Ci, it is evident that following *drd-1.2* knockdown, there is an upregulation of *drd-1.1* — both in terms of its ubiquity throughout the length of the intestine (here seen throughout the entirety of it, rather than in distinct domains), and also its far greater levels, though admittedly, from such assessments, it is hard to draw reliable conclusions between individual animals due to the variable copy number of the transcriptional reporters in all cases. In Figure 5.5Cii, the regulatory response of *drd-1.2p::mCherry* is shown following knockdown of *drd-1.2* in the same animal. It is evident that the level of *drd-1.2p::mCherry* also appears elevated, in terms of its presence throughout the intestine, though this is entirely in accordance with the expectations one might have of a live organism detecting and responding to insufficient levels of a gene. And so, it would seem that in the absence of *drd-1.2*, the expression domains of *drd-1.1* and *drd-1.2* lose their distinction from one another, essentially compensating to such an extent that they overlap across the landscape of intestinal tissue.

Next, when examining the effect of *drd-1.2* RNAi knockdown on starved L4 animals carrying both *drd-1.2p::mCherry* and *drd-1.3p::gfp*, we see a remarkably similar story being told. Comparing the starved L4 animals in panels B and D, it is seen again that, just as was observed above in the transcriptional response of *drd-1.1*, **the distinct, non-overlapping, expression domains of the two genes are largely lost** when levels of *drd-1.2* are abolished by RNAi knockdown. In the case of *drd-1.3p::gfp* especially, this is rather surprising. Bearing in mind that L4 is a point in development where we previously noted, on starvation, there was no *drd-1.3p::gfp* expression (seeing expression only in the posterior of starved young adult worms), this suggests *drd-1.3* — like *drd-1.1* but to an even greater extent — has retained capabilities in the starvation response beyond what are

ordinarily seen in wildtype circumstances. So, considered altogether, these results suggest that in response to diminished levels of *drd-1.2*, both *drd-1.1* and *drd-1.3* offer a compensatory response in their upregulation.

This would, though, very much suppose that *drd-1.1* and *drd-1.3* have retained functionality beyond what is ordinarily seen from their expression — that they *can* have a role in the starvation response over and above what is seen when wildtype levels of *drd-1.2* are present. But the latter only assesses their expression on starvation, and not their function as such. Accordingly, it is worth asking whether this untapped potential in the starvation response, with respect to their expression at least, is mirrored in any overlapping functionality between the *function* of these paralogues in the starvation response.

In order to investigate the possibility of any redundancy relationships, or shared roles, between members of the Drd family in dauer entry and maintenance, dauer formation assays were performed. Just as how dauer was induced above (the injection of dsRNA(s) followed by the placement of animals on starvation media), the same method was employed here. In order to quantitate their ability to form dauers, the number of dauers and non-dauers were counted and then the Dauer Formation Index (DFI) (Fielenbach and Antebi 2008) was calculated using the following formula: *DFI = No. of Dauers — No. of Non-dauers / No. of Dauers + No. of Non-dauers*. A more positive index value indicates a greater tendency to form dauers while a more negative index value denotes a preference for entering the reproductive developmental trajectory.

The results of the dauer formation assays performed on Drd family knockdowns in a wildtype background are shown in Figure 5.5E. It is seen that for the most part, the only member of the

family to have a role in dauer formation is *drd-1.2*, although there is a reduced DFI on simultaneous knockdown of *drd-1.1* and *drd-1.3* (as compared to when they are knocked down individually), though this is not statistically significant. No change in dauer formation is observed when either of the two other Drd family members are knocked down concomitantly with *drd-1.2*, suggesting that, in fact, there are no redundancy relationships between the family members in this respect, and that *drd-1.2* is very much alone in its importance to dauer entry and maintenance. This is obviously contrary to the expression changes described previously which pointed towards a role for both *drd-1.1* and even *drd-1.3* in the starvation response upon the loss of their paralogue. It is implied then, that the regulatory regions of *drd-1.1* and *drd-1.3* have retained a kind of **memory** for an erstwhile role in dauer formation that their coding sequences evidently cannot support — that this is a function they have lost — because they are, as demonstrated, unable to compensate for the loss of *drd-1.2* functionally speaking.

But this argument is predicated on the assumption that the starvation response is the ancestral role of the family, else the concept of 'memory' applied above would be totally inaccurate. And so, we are left with an obvious question to answer with respect to understanding the evolution of the Drd family: which aspects of Drd family functionality are new and which are demonstrably ancient? Was the role of the Drd ancestor in the starvation response, gonadogenesis, male tail differentiation, or some combination of the three? To ascertain this in a fashion, the expression of the representative of the single-copy ancestor, the lone Drd orthologue in *P. pacificus* (dubbed *Ppdrd* here), was characterised. Attempts were made to knock down *Ppdrd* by RNAi (by feeding and microinjection), but they were unsuccessful, despite this method of gene knockdown being successful in the hands of others — though obviously against different genes — in *P. pacificus* (Cinkornpumin and Hong 2011). So as an alternative methodology, and just as for the Drd

paralogues in *C. elegans*, a transcriptional reporter was constructed for *Ppdrd* and it was injected into both *P. pacificus* and *C. elegans*.

As an aside, it was mentioned previously that there is this same single-copy Drd orthologue present in *C. japonica* (which was selected as an outgroup in the phylogenetic analysis), but it was not used for expression pattern determination because *C. japonica*, unlike *P. pacificus*, is gonochoristic. This means that the generation of transgenic animals is especially difficult. Timing the injection process with subsequent recovery and then mating proved (elsewhere in this thesis in Chapter 6 to follow) to be nigh on impossible. But with *P. pacificus* being hermaphroditic and offering many of the same advantages as *C. elegans* in terms of experimental tractability, it was chosen in lieu of *C. japonica* as the organism in which to investigate the role of the Drd family ancestor. The expression patterns that were formed in the transgenic lines of both *P. pacificus* and *C. elegans* are shown in Figure 5.6.

In Figure 5.6, it is shown that the expression of the *Ppdrdp::gfp* somewhat resembles the collective intestinal expression of the paralogues in *C. elegans* in that expression of *Ppdrdp::gfp* is observed ubiquitously throughout the intestines of starved L4 animals (yet to enter dauer) in both *P. pacificus* (A) and *C. elegans* (B). In neither *P. pacificus* nor *C. elegans* was *Ppdrdp::gfp* seen in well-fed animals or in starved animals at earlier developmental stages (i.e., before L4). In fact, the only distinction between the two expression patterns in the intestines of starved L4 animals of each species is that *Ppdrdp::gfp* appears to be more cytoplasmic in its native species as compared the more nuclear expression of the same in *C. elegans*. But on reflection, when *drd-1.2* was knocked down, cytoplasmic, as well as nuclear expression of all three paralogues was observed, this was in spite of the presence of the same NLS being present and intact in all three constructs just as it is in

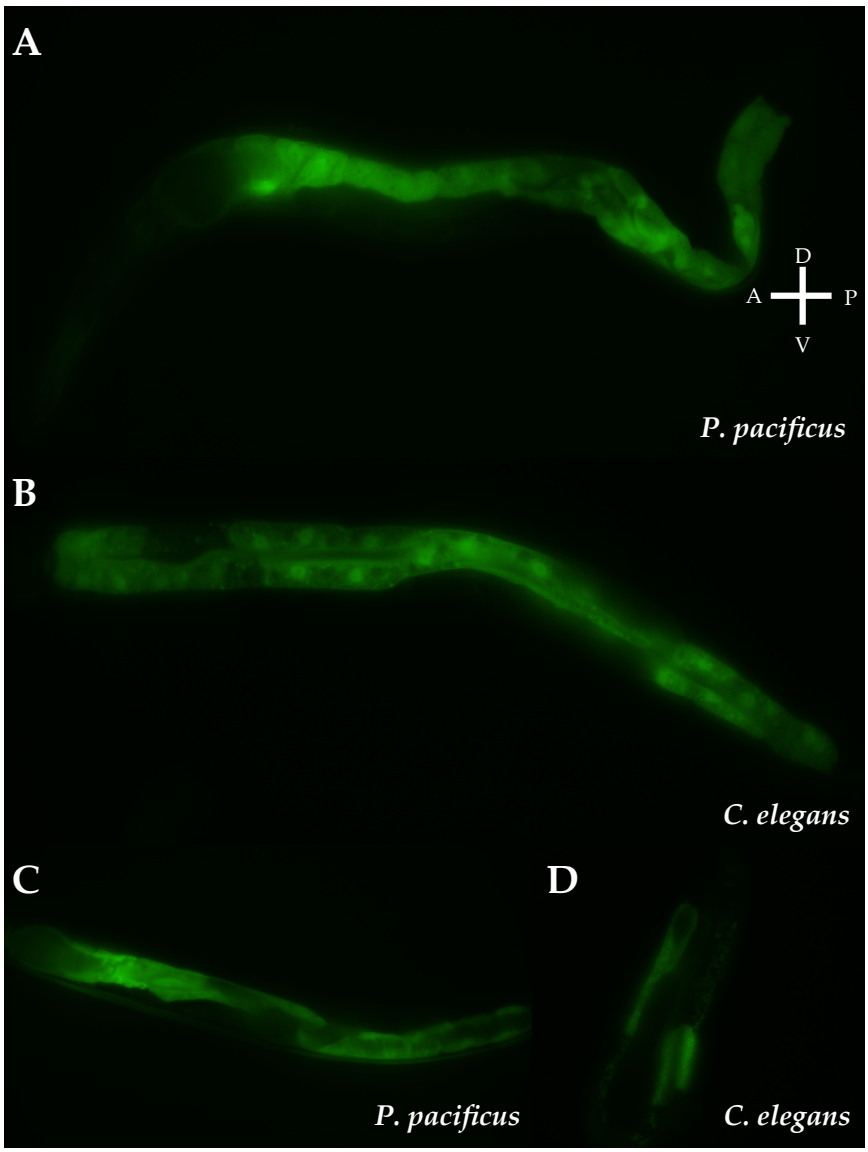


Figure 5.6. The role of the ancestral Drd family member: dauer entry and maintenance. (A) Expression of the *Ppdrdp::gfp* transcriptional reporter in *P. pacificus* at the L4 stage, having been grown on starvation media since late L2/L3. This animal was defined as pre-dauer (about to enter dauer) on the basis of the time it had been starved for (grown on starvation media since late L2/L3) as, just like for *C. elegans*, there are no morphological features that distinguish *P. pacificus* animals yet to divert their developmental programmes dauer from well-fed animals at the same developmental stage (e.g., L4). (B) Expression of the *Ppdrdp::gfp* transcriptional reporter in *C. elegans* at the L4 stage, having been grown on starvation media since late L2/L3. (C) Expression of the *Ppdrdp::gfp* transcriptional reporter in *P. pacificus* in a dauer animal, having been grown on starvation media since late L2/L3. Dauer morphology in *P. pacificus* was found to be remarkably similar to that seen in *C. elegans*, for example, a thin body shape, larger intestinal lumen, an absence of reproduction. (D) Expression of the *Ppdrdp::gfp* transcriptional reporter in *C. elegans* in a dauer animal, having been grown on starvation media since late L2/L3.

Ppdrdp::gfp. Cellular localisation aside, however, the physiological expression pattern of *Ppdrdp::gfp* in starved L4 animals between both species is identical. In other words, the expression of the single-copy 'ancestor' is not concentrated in anterior or posterior intestinal tissue, rather, it is observed throughout the entire length of it. Crucially, in **no** other organs or tissues was *Ppdrdp::gfp* observed, for instance, in the male tail or the hermaphrodite gonad. Also of note is that *Ppdrdp::gfp* was only visible in L4 animals responding to starvation, never in young adults. Figure 5.6C and D show the expression of *Ppdrdp::gfp* in an established dauer state, in *P. pacificus* (C) and *C. elegans* (D), respectively. Just as we observed for *drd-1.2*, strong expression of *Ppdrdp::gfp* is observed beside a wide lumen of a canonical dauer animal.

With this being the case for the expression of this construct in both *P. pacificus* and *C. elegans*, it is suggested that the ancestral role of the family was solely in the starvation response — in dauer entry **and** maintenance, although this is not possible to unequivocally determine from its expression pattern alone. The implication from studying *Ppdrd* is that the roles of the Drd paralogues in male tail differentiation and gonadogenesis are derived functions, evolving only within the last 180 millions years. Furthermore, as the expression of all the Drd family paralogues in *C. elegans* appears restricted to certain domains in the intestine prior to dauer entry, it is suggested that the ancestor has a more generalist role in the starvation response as compared the Drd paralogues in *C. elegans*. Ancestral functionality was therefore, on this basis, distributed between the paralogues, though of course the primary burden in this respect was put on *drd-1.2*. The distribution of function here was manifestly driven by regulatory evolution, though these data arguably imply there is more to be said about their *cis*-regulatory coevolution given their roles in the worm, as told from reverse genetics, that are not matched with an appropriate expression pattern. In any event, in losing most of the functionality associated with the starvation response,

drd-1.1 and *drd-1.3* were freed to perform these other roles — namely in male tail differentiation and gonadogenesis.

This is contrary to expectation. As we saw from the phylogenetic and syntenic characterisation of the Drd family, it is *drd-1.1* and *drd-1.2* that have retained a close association on chromosome II with *drd-1.3* being rehoused on the X chromosome. From their chromosomal affiliations alone, we would expect that if any two paralogues were to share roles, it would be *drd-1.1* and *drd-1.2*, with *drd-1.3* being in a different chromosomal, and therefore regulatory, environment. If this argument sounds familiar, that is because it is. As we saw in Chapter 3 from investigations into the neofunctionalisation of *tbx-36*, translocation to a different chromosome *can be* the catalyst for the acquisition of roles that were not played by an ancestor. But here, in the Drd family, the rehoming of *drd-1.3* on the X chromosome appears not to have had the same transformational effect. That is to say that simply translocating to a different chromosomal environment is not sufficient to drive the gain of new biological functions where paralogues are concerned.

Discussion

The unexpected patterns of divergence in the Drd family yields insight into the evolution of the ancestor

We set out to probe an instance of asymmetric gene diversification — its causes and consequences, and found such an example of one in *drd-1.3* of the Drd family in the *Caenorhabditis* genus. In defining the roles and expression of the family in an array of disparate post-embryonic processes, we have shed light on where the functional niches of *drd-1.1*, *drd-1.2* and *drd-1.3* overlap, and where they are distinct. We have exposed redundancies among all paralogues in the family. In so doing, we have come to understand what it means to — at the genotypic and phenotypic level —

asymmetrically diverge from an ancestral gene, and it is not an insight one would glean on a common sense view alone. Asymmetric paralogue diversification, it has to be said, is nuanced.

This investigation brings us to the conclusion, based on evidence, that *drd-1.3* appears to be rapidly evolving because it has actually **lost**, as well as gained, functionality with respect to the role of the ancestor solely in the starvation response. This does not mean, however, that *drd-1.3* is under relaxed selection. Indeed, we have directly tested the mode of selection that *drd-1.3* is under and found it to be positive. Though, on the face of it, this finding is hard to square with the apparent loss of function of *drd-1.3*, especially when taken with the apparent absence of any functional gain which could reasonably be deduced as new and solely attributable to it. This is seemingly difficult to reconcile because positive selection is intuitively thought of as a process by which functionality is gained — a variant exists within a population that did not exist before, and it happens to confer a fitness advantage, that is, some new functionality which is beneficial (if not essential) for surviving in a given environment; thus resulting in an evolutionary preference for that variant thereby taking a population in a selective sweep. So it is counterintuitive, by this same logic, to think of a loss of function as being positively selected for. But this is a purely naive reading of what is a complex evolutionary genetic process. In spite of the rhetoric surrounding their roles in evolution, duplicated genes are not limitless in their potential. New genes may gain new functions, but this may well be at the *expense* of their old ones. Molecularly speaking, for enzymes especially, this is axiomatic. Being able to catalyse multiple different reactions, with the same active site, defies the notion of specificity at the heart of enzyme catalysis. And such losses are perfectly permissible, even advantageous, providing there are fellow paralogues that hold onto those roles, should those roles require holding on to.

As we have uncovered, *drd-1.3* has pursued roles in male tail development and gonadogenesis, at the cost of its investment in the starvation response. But in the previous two chapters, we have seen how genes are more than capable of maintaining pleiotropy following paralogue generation. Though clearly, this is not the case with respect to the Drd family. We have witnessed firsthand the trade off between members maintaining functionality in the starvation response **as well as** in male tail development and, to a lesser extent, gonadogenesis. It is by this logic we are able to say these are mutually exclusive roles as performed by these genes. It is now pertinent to ask why this is so.

Before delving straight into an answer to this question, it is first wise to evaluate where *drd-1.1* and *drd-1.2* sit functionally in relation to their paralogue. In contrast to *drd-1.3*, *drd-1.1* and *drd-1.2* have retained functionality with respect to the dauer response; the former, however, appearing to be a lot less involved than the latter. It is not clear why the expression of *drd-1.1* is so strong in response to starvation yet appears to have no statistically significant role in starvation when probed functionally. It may well be that a knockout would tell a different tale in relation to the significance of *drd-1.1* in this respect. But in any case, when reflecting on the phenotypic and expression data together, *drd-1.1* appears to straddle a fate somewhere between *drd-1.2* and *drd-1.3* to maintain something of a generalist role — involved in male tail differentiation (to a lesser extent than *drd-1.3*), gonadogenesis (to a lesser extent than *drd-1.3*), and dauer decision entry (not maintenance). Owing to its lack of specialisation in a global sense, *drd-1.1* has no highly penetrant phenotypes of its own, and in this way offers a robust crutch should *drd-1.2* and / or *drd-1.3* — which have heavily invested in the starvation response and male tail development, respectively — fail in any way. Perhaps in this way, the true cost of being a generalist is a lack of importance. *drd-1.1* is really a master of nothing.

As *Ppdrd* — the extant representative of the single-copy ancestor — does not have roles in male tail differentiation, nor gonadogenesis (based on its expression, that is), it is possible to say that the roles of *drd-1.1* and *drd-1.3*, and even to a lesser degree *drd-1.2* (in gonadogenesis), are in fact instances of neofunctionalisation and a *gain* of pleiotropy for the whole family. This implies that following their genesis by triplication, the Drd paralogues acquired two additional roles (in male tail specification and gonadogenesis) that subsequently, very quickly, got doled out among the three, with *drd-1.3* and *drd-1.1* — in that order — taking most of the burden in these respects. In particular, the two sought to resolve their requirement to specify the male tail collectively by specialising in the specification of different ray lineages therein.

As logically abrasive as the concept might seem, a gain of multi-functionality among duplicated genes is not unfamiliar in this thesis thus far; it was a prevalent phenomenon in the Warthog family just as it is present here. It is an issue, therefore, that requires addressing directly.

It seems antithetical to, rather than undergo a division of labour, instead acquire lesser, accessorial, roles in alternative newfound processes. But arguably, this is how duplicated genes are able to contribute in new ways to organismal evolution while still imparting robustness on any process in which they are involved. After all, innovation and robustness are hailed as *the* two advantages of any gene duplication, but on initial reading can seem in conflict with one another; this is just one reason why gene duplication is not always a clearcut path to unbridled novelty. But here at least, it is true that the varying extents of the involvement of Drd family members in male tail development and gonadogenesis have only been possible to acquire anew given their varying degrees of functional loss in the starvation response.

In summary, *drd-1.1* has been able to partially innovate in male tail development and gonadogenesis but has minor involvement in the starvation response; *drd-1.2* is heavily invested in the formation and maintenance of dauer with no discernible role in male tail development, despite somewhat being involved in gonadogenesis; and finally, *drd-1.3* has cut ties with ancestral functionality pretty much altogether, instead focussing on the processes of male tail development and gonadogenesis.

Postulating the mechanisms and process of Drd family evolution

Of course, while we have not directly probed the mechanisms of divergence in the Drd family, it is evident that their *cis*-regulatory divergence has played a critical role in their functional evolution. It was seen how both *drd-1.1* and *drd-1.3* appeared to abut, at the physiological level within intestinal tissue, the expression domain of *drd-1.2* in starved animals about to enter dauer. There has not been, nor will there be, a clearer case of subfunctionalisation in this thesis. It was seen how the expression of the ancestral gene dominated the intestine on starvation, but that not even *drd-1.2* showed such a broad expression domain throughout the length of the intestine. Cumulatively, when considering the expression of all three paralogues on starvation, we should be satisfied that together they make up the expression of the single-copy ancestor prior to dauer entry. And while this is readily observable, and therefore deducible, in the starvation response, there is no denying that expression divergence alone cannot explain the functional evolution of Drd paralogues. Far harder to get at are the changes which are not quite so physiologically apparent, even visual. But this is precisely the kind of change that relates to the asymmetric divergence of *drd-1.3*. It has to be, else the initial test that was run to determine the positive selection acting on *drd-1.3* — using only the amino acid sequence — would not seem to be discordant with the expression data that implies mere subfunctionalisation. Though the two are undoubtedly inextricably linked.

The modularity of *cis*-regulation, whereby one aspect of expression can be tweaked without affecting other aspects, is key. If we accept that *drd-1.3* has lost, in large part, ancestral functionality, the selective constraints it was under to perform a role in the starvation response became relaxed. Quite simply, losing expression domains meant *drd-1.3* was free to accumulate coding sequence change that would otherwise not be open to it. For enzymes, this would, as stated, presumably have an effect of cataclysmic proportions if mutations resulted in changes in the active site. The same can be said too, historically, of *drd-1.1*; though it is true that *drd-1.1* subsequently degenerated to become less integral to the same processes as *drd-1.3* is required for. Although, this argument is put with a chronology of which we cannot be certain for lack of concrete evidence. But being guided by Occam's razor, given the empirical insights we have generated, this must be the most likely scenario.

In acquiring coding sequence change that permitted a role in male tail development, *drd-1.1* and *drd-1.3* were, between them, able to subsequently specialise. For what it is worth, as mentioned in passing previously, the single human orthologue of the Drd family, FAXDC2, is known to rely on regulation by RUNX transcription factors; the same family of transcription factors that are present in the worm and are required for male tail development (Ji et al. 2004; Nimmo et al. 2005). Given that incidences of molecular convergence over the course of evolution are considerably higher than previously thought (Storz 2016), it is certainly a valid hypothesis to test if *drd-1.3* and *drd-1.1* functionally cooperate with RNT-1 to specify the male tail. On this same vein, there is tentative evidence that *rnt-1* is involved in the starvation response in *C. elegans*, though this has not been properly characterised; it is simply known that *rnt-1* mutants do not survive starvation (Baugh et al. 2009; A. Woollard, pers. comm.). Though alluringly, it certainly adds weight to the possibility

that RUNX orchestrates the transcriptional response of the Drd paralogues, possibly using co-factors, unique to each Drd paralogue, to drive the transcription of each family member specifically in the male tail (*drd-1.3*) and following starvation (*drd-1.2*).

Speculation aside, it can be concluded that the regulators (whatever they may be) of *drd-1.3* and *drd-1.1* are different to those of *drd-1.2,* which presumably has regulation more redolent of *Ppdrd* and so too the ancestral gene. But whatever the precise nature of the radical departure from the family *drd-1.3* took, we are able to say it was only possible because *drd-1.2* was left to act by proxy for the ancestor, if you will, in dauer formation and maintenance.

CHAPTER 6

Lessons on domain rearrangements following duplication from the Myrf Family: A metazoan synapomorphy with *Caenorhabditis* exceptions

Introduction

Little is known about evolution at the subgene level. Although the possibility of changes to domain architecture following gene duplication have long been appreciated in theory (Ohno 1970), the instances in which they have been identified (Wu et al. 2011) have never been functionally characterised. Here, we will analyse the genetic consequences of domain evolution following gene duplication in the Myrf family. As far as has been ascertained, all major animal phyla have a single-copy orthologue of the Myrf gene, except some members of Nematoda — those in the *Caenorhabditis* genus. In vertebrates, the MYRF (*my*elin gene *r*egulatory *f*actor) transcription factor is known, and was named, for its role in myelination and is accordingly strongly expressed in oligodendrocytes (Bujalka et al. 2013). Curiously, despite lacking myelinated neurones, *C. elegans* has two Myrf family paralogues, *myrf-1* and *myrf-2,* involved, incidentally, in nervous system development (Meng et al. 2017); this we will come back to later.

Regardless of the additional paralogue in *C. elegans*, it is apparent that there is some intrinsic property to the Myrf family that results in its reluctance to exhibit any kind of phyletic dynamism — unencumbered, and unaided, by gene gain or loss. It is for this reason that the Myrf family can be thought of as the very antithesis of the T-box family (as they evolve in the *Caenorhabditis* genus, at least) explored in Chapter 3. Quite why some gene families are prone to rampant duplication

and others not remains a mystery. This question could be answered in one of two ways. Taken proximately, some genes are less prone to being duplicated, or lost, should they sit in relatively static regions of the genome (Nesta et al. 2021). And so inversely, any gene which lies in a mutational hotspot is more prone to duplication, as it is to other kinds of mutation (Nesta et al. 2021). One obvious factor which directly contributes to the propensity of a gene to duplicate is the frequency of repeats in a region (Bzymek and Lovett 2001), these increasing the likelihood of various molecular processes going awry including DNA replication and recombination — the outcomes being replication slippage and unequal crossing over with respect to each. As repeats are risk factors for duplication events, it follows that the density and identity of genes which lie in regions rich with repeats are themselves traits subject to selection (Chattopadhyay and Weissman et al. 2009; Horton et al. 2021; Nesta et al. 2021; Cano et al. 2023). So the initial question runs much deeper, and should be posed in another way. Rather, why are some genes ultimately selected to lie outside of regions susceptible to duplication events, and inversely, why are others selected to lie within them? This is a question of the evolution of evolvability. The answer to which is harder to get at, but no doubt fundamentally lies with understanding the evolutionary potential some duplicated genes have over others.

Domain rearrangements, (these being structural changes such as the complete or partial loss of an entire protein domain) along with regulatory evolution, seem, intuitively, as though they could be powerful forces that shape the evolution of gene function following duplication. Work by others, and even this thesis thus far, has concentrated on the latter, but it is hard to ignore the potential importance that great changes to domain architecture could have on duplicated gene evolution. The use of the word ‘potential’ there was deliberate — not because domain rearrangements would likely have little or no effect on gene function, but because it is really not known how commonly

they occur among paralogues, or what role they play in the fates they adopt. As such, operating from little by way of a starting point, this Chapter seeks to probe exactly this. We will exploit the domain level conservation (and all round lack of dynamism) in the Myrf family to discover why certain genes remain relatively immutable over evolutionary time with regards not only to their domain structure in the face of duplication but also their duplication at large. This will be achieved by examining the outcome in the rare instance in which deviation from a highly conserved domain structure is observed in a pair of very recently derived Myrf paralogues which, while present in some members of the *Caenorhabditis* genus, are not found in *C. elegans*.

To investigate what the consequences are of straying from the status quo when it comes to domain architecture, it follows we need to set out the expected, that is to say conserved, domain structure of the Myrf family, which in turn relates to their evolutionary origin (Schwarzer et al. 2007; Li and Park et al. 2013; Wu and Zhen et al. 2021).

It is thought that the Myrf family first arose from a kind of gene fusion event mediated by bacteriophage infection. Endosialidases are a unique feature of the end-tail-spikes of many bacteriophages and encoded within them is their own intramolecular chaperone (Li and Park et al. 2013). This chaperone domain enables protein trimerisation and catalyses cleavage at a serine-lysine dyad (Wu and Zhen et al. 2021). It is thought that in a eukaryotic ancestor, one such intramolecular endosialidase chaperone became fused with a NDT80/PhoG-like DNA-binding domain to create the ancestral MYRF (Schwarzer et al. 2007; Li and Park et al. 2013). The result is a biochemical signature conserved among Myrf family members to this day, and that is the autocleavage of a pro-peptide into an N-terminal portion containing the DNA-binding domain, and a C-terminal portion possessing one transmembrane domain (TM), a coiled coil domain

(designated C1), and a functionally unknown MYRF conserved domain 2 (C2) (Schwarzer et al. 2007). Biochemical characterisation of MYRF-1 in *C. elegans* revealed it is cleaved into N-terminal (1-482aa) and C-terminal (483-931aa) fragments, as is to be expected from any MYRF pro-peptide. And so, as might be expected, upon the mutagenesis of the serine in the cleavage site, *myrf-1* loses its role in nervous system development which, as yet, is the only function attributed to *myrf-1*, or *myrf-2* for that matter (Meng et al. 2017).

But *myrf-1* and *myrf-2*, as characterised in *C. elegans*, were shown to exhibit almost complete functional redundancy for their role in synaptic refinement — that is the critical step in nervous system maturation requiring a carefully timed reorganisation of neuronal connections, also known as synaptic rewiring (Meng et al. 2017). Their role in this was exceptionally well-characterised by Meng and colleagues (2017), and so will not be mechanistically probed further in this work, not least because it is outside the scope of it. Rather for our purposes, it is much more pressing to establish why, in knowledge of their strong overlapping functionality as it is known so far, two Myrf paralogues should exist at all. Though as has already been spelled out, MYRF is ubiquitous throughout the animal kingdom and even features outside of it in fungi and protists (Schwarzer et al. 2007; Senoo et al. 2013). It is therefore likely that any roles it has in the nervous system, in *C. elegans* and vertebrates, are more recently acquired aspects of its functionality compared to its other role(s) which must be much more fundamental.

It has been known for some years that generating a MYRF knockout in mice results in embryonic lethality due to the failure to specify major cell types, of which oligodendrocytes are only one (of many) (Emery et al. 2009). Furthermore, the role of MYRF in pre-stalk cell differentiation in *Dictyostelium discoideum* has been similarly well established (Senoo et al. 2013). Taken together, all

this implies that MYRF is a single point of failure for cell fate determination in many taxa, probably having repeatedly evolved to be so given all the gene regulatory networks in which it has been implicated which are by no means conserved across these vast evolutionary distances, just like the cell types that it specifies. It is of value, therefore, to probe the function of seldom generated Myrf paralogues not simply from the perspective of an evolutionary geneticist, but also from that of a developmental biologist.

Results

myrf-1 and *myrf-2* are well-conserved paralogues that arose at the base of the *Caenorhabditis* genus

As has become standard practice before plunging into the functional evolution of a gene family in this thesis, it is germane to first characterise the evolutionary history of the Myrf family using phylogenetics before formulating hypotheses based on the gene duplication events that feature within it. Figure 6.1A shows the maximum likelihood phylogenetic analysis of the Myrf family in the *Caenorhabditis* genus, including the single-copy outgroup in *P. pacificus*. The family falls robustly into two clades, MYRF-1 and MYRF-2, each containing the orthologues in *C. japonica*, *C. elegans*, *C. brenneri*, *C. remanei*, and *C. briggsae* indicating that this well-conserved duplication happened at the base of the genus, approximately 180 — 200 mya (Baker and Woollard 2019). The relationships between all orthologues on the branches are supported by bootstrap values that, for the most part, are assuringly high (>75). This serves as a reliable indicator that the *myrf-1* and *myrf-2* orthologues between species are probably very similar in sequence. The eagle-eyed reader will have noticed the find of an additional *myrf-2* paralogue in *C. brenneri* and *C. remanei* (though

its absence in *C. briggsae* and all other species). At present, we will do no more than to acknowledge the existence of *myrf-2.1* and *myrf-2.2* in these two species, except to say that these hitherto uncharacterised genes are a surprising find given the absence of MYRF paralogues in other taxa. We will return to these additional paralogues in due course.

Figure 6.1B depicts the microsyntenic regions in which these orthologues (and paralogues) are found, showing that there is no conserved microsynteny with the single-copy outgroup as it is found today in *P. pacificus*. Although, it can be said of the genus here more broadly that synteny appears to be well conserved with the local gene order surrounding both Myrf loci remaining stagnant over their approximately 180 million years of evolution. With regards to the newly found paralogues, it is with certainty we are able to say that *myrf-2.1* and *myrf-2.2,* due to their close chromosomal association, arose from a tandem duplication in the ancestor that gave rise to *C. brenneri*, *C. remanei*, and *C. briggsae* — after the divergence of the lineage which gave rise to *C. elegans*. Critically, the analogous region in the *C. briggsae* genome — where we would expect to find the additional *myrf-2* paralogue should it have retained it — is a gene desert, implying that many more orthologous genes have been lost from this region, not only a member of the Myrf family.

Based on their sequence integrity and their level of conservation between species, it is unlikely that either *myrf-1* or *myrf-2* have degenerated to pseudogene status. Though this leaves us with the rather open-ended question of what role these paralogues play in development. It is no exaggeration to say that the single MYRF gene in all animals in which it has been functionally characterised plays roles so critical to development that its loss renders animals inviable; but can the same be true of both *myrf-1* **and** *myrf-2*? This could be put, summarily, as whether the Myrf

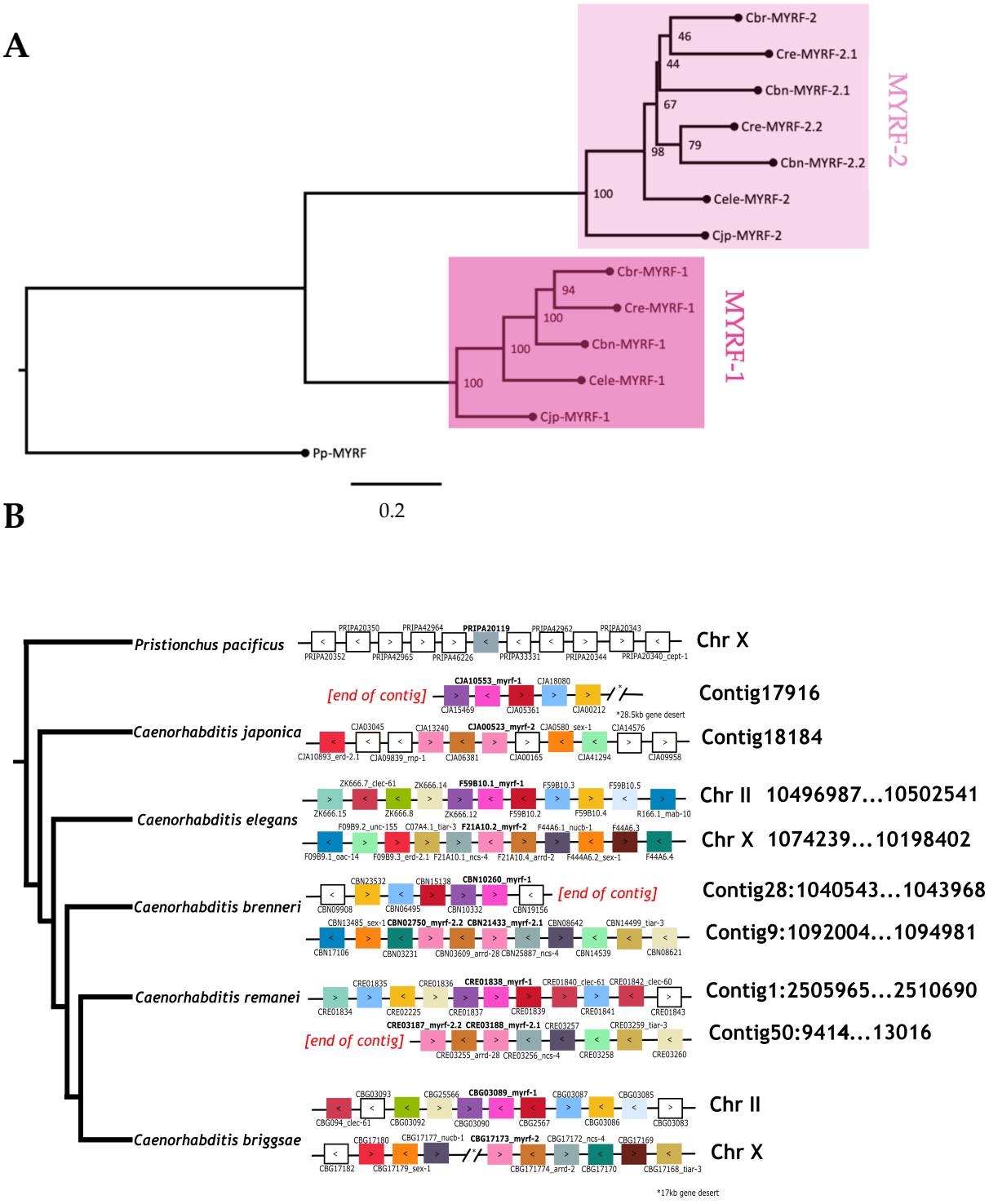


Figure 6.1. Phylogenetic and synteny analysis of the Myrf family in the *Caenorhabditis* genus. (A) Maximum likelihood phylogenetic analysis of the Myrf family built from an amino acid multiple sequence alignment of Myrf family proteins (mined from the predicted proteome). Species abbreviations: Pp, *Pristionchus pacificus*; Cj, *Caenorhabditis japonica*; Cbre, *C. brenneri*; Cbri, *C. briggsae*; Cr, *C. remanei*; Ce, *C. elegans*. Node values are bootstrap values as calculated from 10,000 iterations of the SH-like approximate likelihood ratio test. Scale bar is substitutions per site per million years. (B) Synteny analysis of the Myrf family in the *Caenorhabditis* genus and the outgroup, *P. pacificus*. Synteny was compared manually across species where the tBLASTn of each gene in that region was mapped back to other species. Where genes are assigned a colour, that means they have an orthologue with conserved synteny in another species present in the diagram. Where

genes are depicted as white boxes, that means their orthologue is not shown in the microsyntenic region depicted. Gene deserts (*) (i.e., swathes of chromosome not populated by genes) are abbreviated on the diagram with their size as indicated just below the *. Genes of interest are shown centrally in the diagram, so *myrf-1* (F59B10.1, bright pink - LG II) and *myrf-2* (F21A10.2, baby pink - LG X) feature in the middle of all microsyntenic regions.

family is twice as important to *C. elegans* and its close relatives as compared other animal lineages, in which only one family member is found. Based on their conservation in an entire genus for such a long length of evolutionary time, it seems unlikely *myrf-1* and *myrf-2* would display complete functional redundancy based on our current understanding of duplicated gene evolution, but the only work on these genes to date (in Meng et al. 2017) suggests this in fact might be the case. To probe the functional equivalence, and importance, of these two paralogues in synaptic rewiring and beyond, we are required to use a genetic approach.

Understanding the role of *myrf-1* and *myrf-2* in synaptic refinement

CRISPR/Cas9 gene editing was used to generate putative null alleles for both *myrf-1* and *myrf-2* in *C. elegans* (Meng et al. 2017), and these same alleles were used in this investigation (Figure 6.2). Mutations in both genes are small out-of-frame insertions occurring in the regions encoding the DNA-binding domains (so the N-terminal portions) of *myrf-1* and *myrf-2*; the resulting frameshift, therefore, purportedly compromises not only the TF activity of both paralogues but also the cleavage site, as well as the C-terminal portion containing three domains. Following their generation by Meng and colleagues (2017), strains carrying these alleles were outcrossed twice in this investigation to ensure any off-target effects arising from the gene editing process were not retained for phenotypic characterisation. The primers which enabled allele-specific detection for genotyping are shown. These were used in the generation of the *myrf-1(ybq6);myrf-2(ybq42)* double mutant.

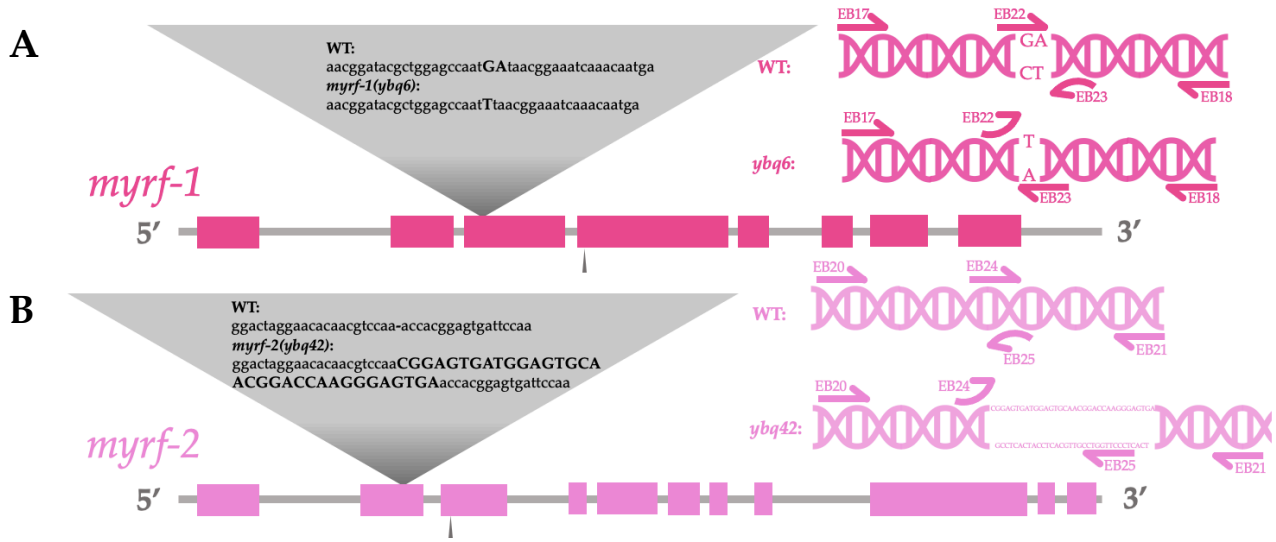


Figure 6.2. Putative null mutations in *myrf-1* and *myrf-2*. In the schematics in panels (A and B), pink boxes indicated exons and grey lines indicate non-coding DNA (e.g. introns). Expanded grey triangles from each gene allow for the magnification of the section of the gene relating to the Myrf mutations used in this investigation. (A) Schematic detailing the *myrf-1(ybq6)* allele generated by CRISPR/Cas9 as part of Meng et al. 2017. The single bp insertion in exon 3 results in a frameshift results in over half the transcript (most of the DNA-binding domain, and all the remaining domains, including the peptidase domain). (B) Schematic detailing the *myrf-2(ybq42)* allele generated by CRISPR/Cas9 as part of Meng et al. 2017. The 34 bp insertion in exon 2 results in a frameshift results in over half the transcript (the DNA-binding domain, and all the remaining domains, including the peptidase domain). The thin, grey arrowheads indicate the location of the cleavage site in both genes. The primers designed and used in this investigation to genotype the alleles (enabling WT and mutant detection) using the tetra-ARMS approach (allele-specific PCR) are included.

While the impact of these mutations cannot be underestimated given the enormity of the consequences that are to be expected from any frameshift which occurs at the start of a gene encoding a large, autocleaved protein, the mutation in *myrf-1* at least, cannot definitively be classed as null yet. Before its annotation as *myrf-1*, the gene sequence F59B10.1 was formerly known as *pqn-47* (meaning very little other than 'Prion-like-(Q/N-rich)-domain-bearing protein', hence its subsequent renaming), in which a knockout was generated by random mutagenesis that was, apparently, homozygous lethal. The knockout in this case was a large deletion spanning the autocleavage site as well as roughly 450 bp either side of it. In the hands of this investigation, this necessarily balanced strain did indeed display homozygous lethality (data not shown), thus calling into question the efficacy of the 'knockout' allele(s) generated by CRISPR/Cas9 detailed above.

However, there is a major caveat to interpreting the two independent *myrf-1* alleles in this way. This is that the allele generated by chemical mutagenesis was not outcrossed prior to its formal registration as homozygous lethal, as can happen due to the often overwhelming volume of mutant alleles generated from a screen. As such, it may be the case that another mutation, possibly not even linked to *myrf-1*, has been induced in this strain that renders it homozygous lethal. We will, accordingly, proceed with characterising the phenotypes of *myrf-1(ybq6)* on the basis that the gene editing process was designed to target *myrf-1* and *myrf-1* only, and that should any off-target effects have occurred, they were presumably purged in the process of outcrossing. As the same cannot be said of the alternative *myrf-1* allele, it will not be relied upon or referred to further in this work. No other alleles have historically been generated for *myrf-2*. Moreover, RNAi constructs were made in this investigation for both genes, and the results of those knockdown experiments — which broadly recapitulate the phenotypes of the knockouts reported in this Chapter — can be found in Appendix XII. RNAi knockdown of neither *myrf-1* nor *myrf-2* alone resulted in lethality at any stage.

It was first a requirement to repeat the results demonstrating functional redundancy between *myrf-1* and *myrf-2* in synaptic rewiring because this has huge implications for any conclusions we may be able to draw about their potential functional diversification. As such, the Myrf mutants were crossed into a *rab-3p::3XFLAG::wrmScarlet::unc-54* marker enabling the visualisation of synaptic vesicle formation, among other neurological processes and anatomies, owing to its pan-neuronal nuclear RFP expression.

It was said by Meng and colleagues (2017) that *myrf-1* and *myrf-2* were not only redundant for their role in synaptic vesicle formation, but that they did not discriminate or give preference with

respect to which neurones they were required for in the process. This is why it was important to use a pan-neuronal marker, such as *rab-3p::wrmScarlet*. Nevertheless, for manageability and consistency, the role of *myrf-1* and *myrf-2* in synaptic rewiring was re-characterised in this investigation using only the dorsal nerve cord, and the number of fluorescent puncta along it were manually counted. The recapitulation of the neuronal phenotypes of these paralogues in both a qualitative (A) and quantitative (B) sense is shown in Figure 6.3.

As compared to wildtype, the *myrf-2(ybq42)* single mutant does not display significant defects in synaptic rewiring, yet *myrf-1(ybq6)* animals do display a significant decrease in the number of puncta along the dorsal nerve cord. While a subtle difference was found in synaptic rewiring in *myrf-1* knockouts by Meng and colleagues (2017), it was not so dramatic as to be statistically significant. Of course, for our purposes, these contradictory results are not immaterial because the phenotypes of *myrf-1* alone have important implications for assessing the functional overlap between these paralogues. It is for this reason we will take a moment to contemplate the basis for the difference in significance here.

It is impossible to rule out a purely stochastic and therefore unknowable basis for this discrepancy, but for the avoidance of doubt, it cannot be as simple as a statistical error due to the application of an unpaired *t*-test in both cases, where an excess of 50 animals were analysed manually in both investigations. This could mean there is a genuine biological reason for the difference observed. And here, this reason may well be related to the presentation of synaptic rewiring defects over the course of a short period of post-embryonic development.

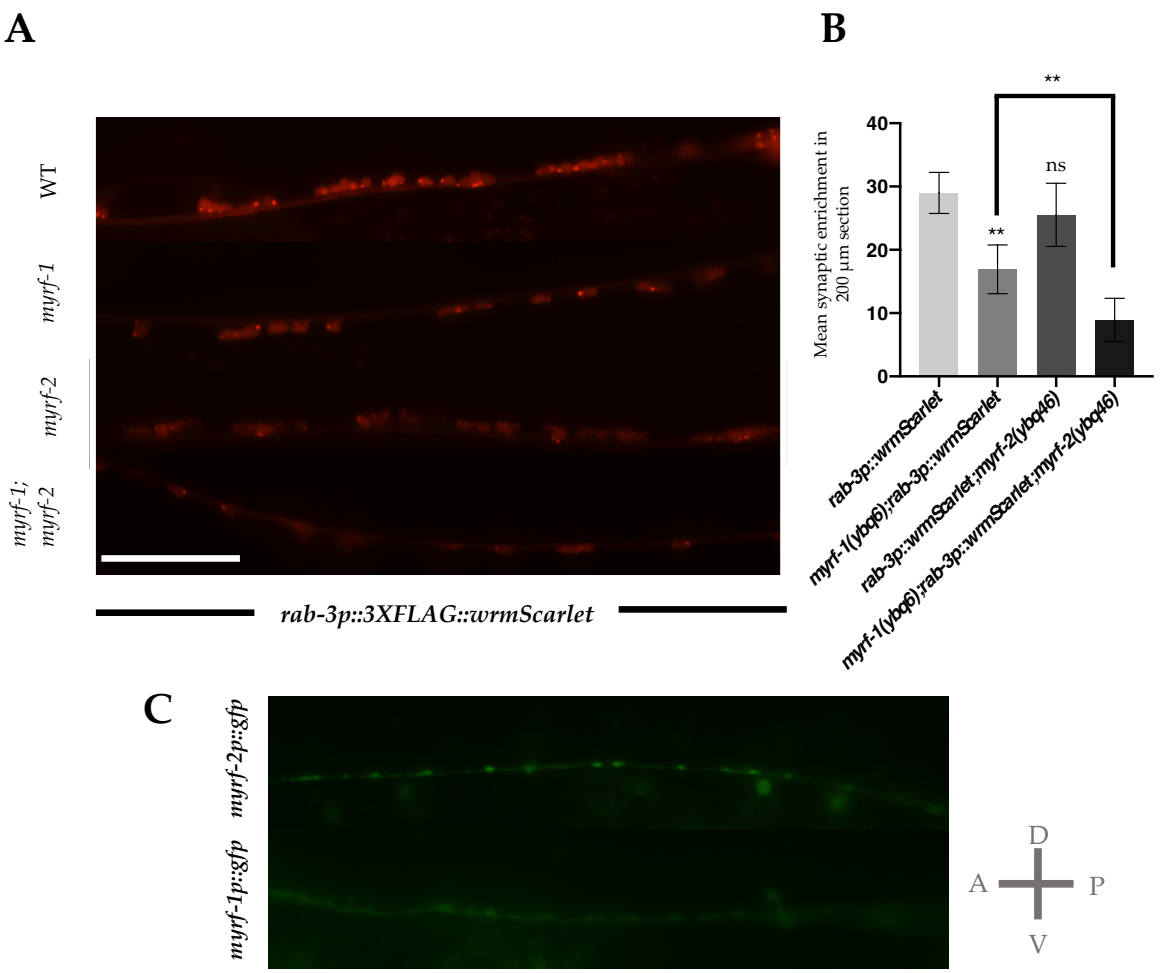


Figure 6.3. Re-characterising the redundant roles of *myrf-1* and *myrf-2* in synaptic rewiring. (A) Synaptic enrichment captured in representative images taken of four genotypes: WT, *myrf-1(ybq6)*, *myrf-2(ybq42)*, and *myrf-1(ybq6);myrf-2(ybq42)* — all in a *rab-3p::3XFLAG::wrmScarlet* background where synaptic vesicles are red fluorescent puncta. Images show a 200 µm stretch of dorsal nerve cord of an animal at the mid-L2 stage (4 hours post-L1). (B) Quantification of mean synaptic enrichment in a 200 µm stretch of dorsal nerve cord at the mid-L2 stage (in a *rab-3p::3XFLAG::wrmScarlet* background): WT (n = 42), *myrf-1(ybq6)* (n = 54), *myrf-2(ybq42)* (n = 51), and *myrf-1(ybq6);myrf-2(ybq42)* (n = 61). Red puncta (synaptic vesicles) were manually counted in all animals at the mid-L2 stage across all genotypes. Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants with an unpaired *t* test. (C) 200 µm stretch of dorsal nerve cord of a mid-L2 worm transformed with *myrf-1p::gfp* and *myrf-2p::gfp* transcriptional reporters: expression of *myrf-1p::gfp* and *myrf-2p::gfp* transcriptional reporters can be seen in the dorsal nerve cord at the mid-L2 stage. Green puncta indicate synaptic vesicles (equivalent to the red puncta in panel (A)). Scale bar = 50 µm.

As per standards set by others in the field (Cuentas-Condori and Miller 2020), quantification of dorsal synapse number here was between late L2 to the start of the L3 stage (operating necessarily within an approximate time window due to the growing asynchrony of worm populations as they

progress throughout their lifecycle). It was subsequently realised that Meng and colleagues (2017) quantified the same phenotype, to quote the only methodological detail provided on the matter, "at L2", and this may not be trivial. On semantic interrogation, it goes without saying that "L2" is not a finite point in time. Rather, it is an 8-hour period encompassing a series of developmental processes, of which synaptic rewiring is but one. Commencing in mid-L1, synaptic refinement takes place over six hours where dorsal motor neurons are not fully 'refined' until late L2 (Cuentas-Condori and Miller 2020). The method of scoring performed by Meng and colleagues (2017) implies that at least some larvae may have been scored for synaptic defects when these structures were not yet fully remodelled (i.e., at early L2). So together, these facts provide us with a possible explanation for the not-quite-significant reduction in dorsal puncta in *myrf-1* mutants observed by Meng and colleagues (2017) which seemingly contradicts the result presented here.

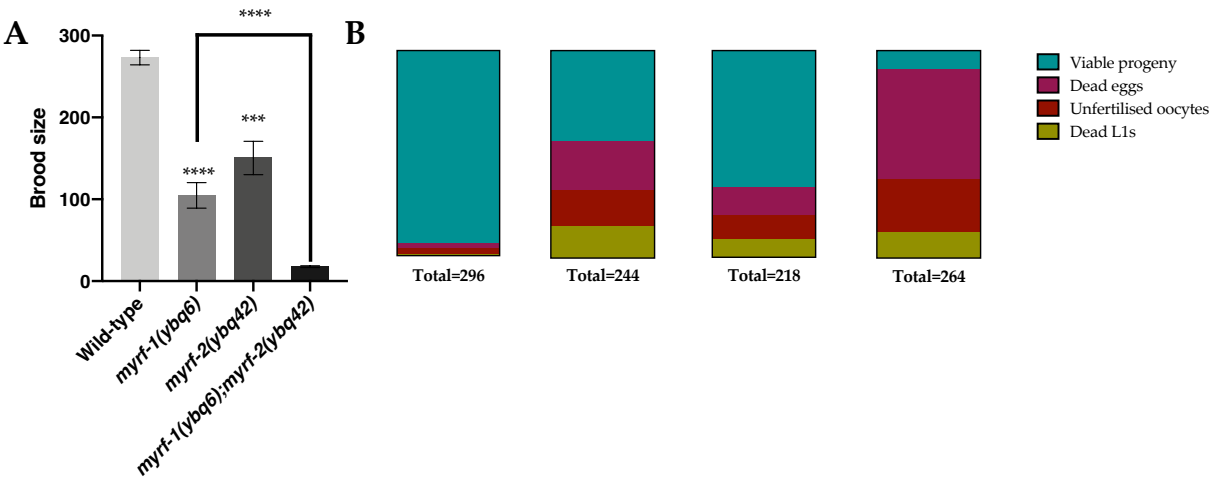
Satisfied by the validity of the results in Figure 6.3A and B (being able to comfortably suggest that *myrf-1* **does** have a role in synaptic rewiring on its own), it is now possible to interrogate defects in the same process in the *myrf-1;myrf-2* double mutant. There are not only a notable lack of dorsal synapses in the double mutant, but their morphology appears defective — being smaller with less defined puncta. The same was not true of the synaptic refinement phenotype observed in *myrf-1* single mutants, which was solely quantitative in nature. This implies that the molecular basis of the two paralogues acting together to regulate synaptic refinement is distinct from that which underpins the relatively small role *myrf-1* plays on its own. Furthermore, upon generating transcriptional reporter constructs for both paralogues (*myrf-1p::gfp* and *myrf-2p::gfp* shown in Figure 6.3C), it was observed that both are, to an approximation, equally expressed along the dorsal nerve cord. All this evidence when taken together points towards at least *some level* of

redundancy between *myrf-1* and *myrf-2* in synaptic refinement, but perhaps not to the extent previously claimed.

myrf-1 and *myrf-2* are pleiotropic paralogues equally critical for basic reproductive and developmental processes

Given the extent to which the Myrf family are utilised in development by other taxa, it is likely that synaptic rewiring barely scratches the surface of their importance to *C. elegans* and its close relatives. In fact, this was all too obvious upon the initial construction of the double mutant, because they appeared, for want of a better term, incredibly sickly. That is to say their movement was uncoordinated and minimal; their body size was reduced; adults appeared more transparent, and the brood size was, even on superficial inspection, but a fraction of the single mutants from which they were built. It was from these initial observations that the basic fitness properties of the Myrf mutants in *C. elegans* went on to be characterised.

Figure 6.4A illustrates the extent to which the broods of *myrf-1;myrf-2* double mutant animals are diminished in comparison to the singles. Unlike with the penetrance of synaptic rewiring defects described above, both *myrf-1* and *myrf-2* appear to play a discrete role in some aspect of basic fitness, though exactly what this is at this stage is of course not possible to decipher from quantifying brood size alone. To dissect the role that *myrf-1* and *myrf-2* play (both alone and together) in development and/or reproduction, the brood sizes were interrogated further by looking not merely at the numbers of viable progeny, but the *kinds* of inviable progeny that were being produced. Figure 6.4B details the relative proportions of inviable progeny types that are laid by these mutants. It is possible for worms to lay: unfertilised oocytes, dead eggs, as well as



C

	WT v *myrf-1*	WT v *myrf-2*	*myrf-1* v *myrf-1;myrf-2*
Dead eggs	*	*	*
Unfertilised oocytes	ns	ns	*
Dead L1s	*	ns	ns

Figure 6.4. Elucidating the morphological basis of the reduction in reproductive fitness in the Myrf family. (A) Mean brood sizes of four genotypes were quantified: WT, *myrf-1(ybq6)*, *myrf-2(ybq42)*, and *myrf-1(ybq6);myrf-2(ybq42)* — all progeny scored were part of the brood that survived to hatching and beyond L1 such that these can be considered viable broods, i.e., dead eggs and dead L1s were removed from the counts. Broods were scored in triplicate and an average was taken for each. All broods were scored at 20 °C and all other variables were kept constant (e.g., plentiful supply of food). (B) Shows a break down of viable and inviable progeny types produced by the aforementioned genotypes to ascertain the basis of the reduction in broods in the Myrf mutants. From left to right, the four genotypes are displayed as follows: WT, *myrf-1(ybq6)*, *myrf-2(ybq42)*, and *myrf-1(ybq6);myrf-2(ybq42).* All progeny classes are proportions relative to the total number of progeny laid by that genotype to enable meaningful comparison between genotypes. Aside from 'viable' progeny (in teal), the relative proportions of dead eggs (pink), unfertilised oocytes (red), and dead L1s (olive green) are also shown. (C) Provides the statistical comparisons for panel (B) which is separated out for clarity. Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants where the statistical significance of the reduction in viable progeny is confirmed by the results of an unpaired t-test (A) or Welch's t-tests (B) as the latter are proportions of a whole.

embryos which complete embryogenesis but die on hatching at L1. This covers every scenario except viability which ensures progeny which are laid and complete embryogenesis but are unable to hatch and develop as larvae are not erroneously categorised as viable.

From the visualisation of the relative proportions in Figure 6.4B and the associated statistics in C, it is evident that as compared to each single mutant, the *myrf-1;myrf-2* double mutant lays significantly more dead eggs and unfertilised oocytes. It is striking, and initially surprising, that on its own, *myrf-1* produces a significant number of dead L1s, though this does not increase when *myrf-2* is knocked out simultaneously. But this is for a very uncomplicated reason, because given the total number of possible offspring that could ever be laid by a worm is just less than 300, that the double mutant produces such a high proportion of unviable progeny as unfertilised oocytes and dead eggs means that in effect, so few remain that could be classed as, or *develop* into, anything else. Biologically speaking, this likely relates to a sort of chronological masking where any role for *myrf-1* and *myrf-2* in L1 viability is effectively masked by the role the two play in fertility and embryogenesis. But regardless of the validity of these two related explanations, it is clear even from the preliminary quantification shown, without a greater understanding of the mechanisms at play here, that *myrf-1* and *myrf-2* are required for processes fundamental to organismal viability. We will return to the relative proportions of dead eggs produced by the *myrf-1* and *myrf-2* single mutants as compared the *myrf-1;myrf-2* double mutant in due course, but for now, we will consider the morphological and developmental basis for the high proportion of unfertilised oocytes laid by these animals.

Logically, the first port of call in attempting to understand the basis for a large of number unfertilised oocytes produced by any mutant is to inspect the gonad. Figure 6.5A depicts the gonads of wildtype (i), *myrf-1(ybq6)* (ii), *myrf-2(ybq42)* (iii), and *myrf-1(ybq6);myrf-2(ybq42)* (iv) animals under DIC optics. It is immediately seen when comparing panels ii, iii, and iv with i that the morphology of the gonad in Myrf family mutants, single or in combination, is not as it should

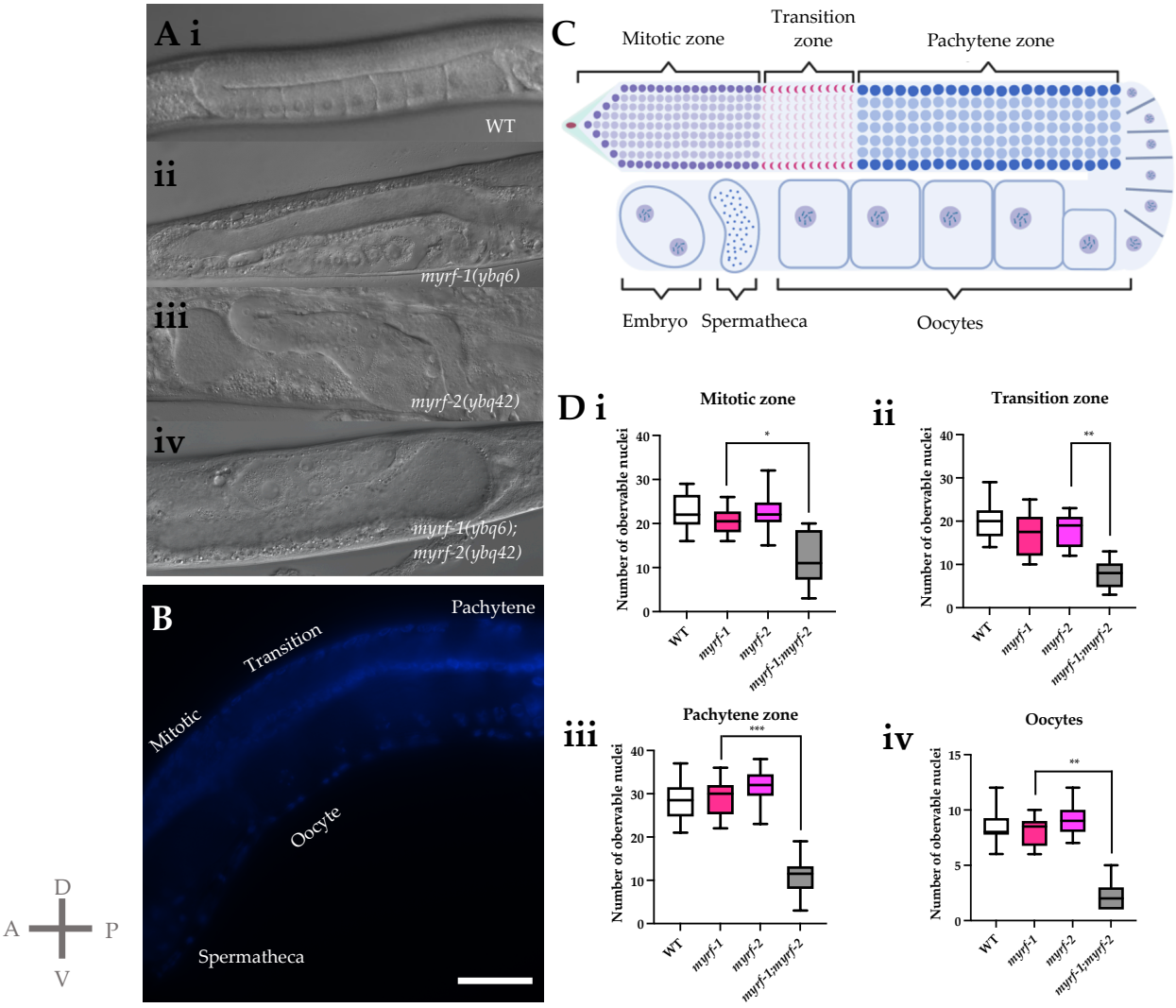


Figure 6.5. Characterising the role of *myrf-1* and *myrf-2* in germline development. (A) Representative DIC images of the anterior gonad arm (in a young adult hermaphrodite) in Myrf family single and double mutants. (i) WT gonad morphology (of the anterior arm); (ii) shows the anterior gonad morphology of a *myrf-1(ybq6)* mutant; (iii) shows the anterior gonad morphology of a *myrf-2(ybq42)* mutant; and shows the anterior gonad morphology of a *myrf-1(ybq6);myrf-2(ybq42)* double mutant. (B) Depicts a representative example of a WT DAPI stained young adult gonad (posterior, in this case) enabling the visualisation, and therefore quantification, of gonadal nuclei across the different mitotic and meiotic zones (shown as labelled), along with the spermatheca. (C) Shows a schematic representation of a gonad whereby the difference mitotic and meiotic regions are labelled, among other features. (D) Shows the quantification of raw numbers of DAPI-stained nuclei in the different regions of the gonad as per the definition in panels (B and C): WT (n = 56), *myrf-1* (n = 61), *myrf-2* (n = 56), and *myrf-1;myrf-2* double mutants (n = 63). The same young adult animals were scored across the four panels for each genotype (such that each genotype has the same n number). (i) First, the mitotic zone. (ii) Second, the transition zone. (iii) Third, the pachytene zone. (iv) Fourth, the number of mature oocytes. Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants with an unpaired *t* test. Scale bar = 50 μm.

be. In the case of all, gonads appeared smaller, and incorrectly migrated. Although, it should be commented on for the sake of the discussion to follow that the gonad arms of the *myrf-1;myrf-2* double mutant are noticeably smaller, and this is the reason for the transparency of the adult animals mentioned previously (because the gonad ordinarily takes up so much volume in the middle body). All animals of each genotype appeared in the way described (these being representative images), and so defective gonad morphology cannot be the reason behind the number of unfertilised oocytes laid by the double mutant — else a similar proportion would be laid by the *myrf-1* and *myrf-2* single mutants, and they are not.

Therefore, the reason has to be more subtle. To investigate this, the gonads were stained with DAPI within the whole animal (as in Figure 6.5B), and the placement and number of nuclei were assessed in each genotype. To illustrate why this was even worth doing — and by extension what underlying tales it can tell about the role of any gene in reproductive fitness — the structure, development, and organisation of the *C. elegans* gonad needs setting out in brief.

The hermaphrodite gonad of *C. elegans* has been dubbed a 'test tube' for cell and developmental biology (Hubbard and Greenstein 2000), and, in reference to its physicality alone (with one of the two gonad arms shown in Figure 6.5C as an example), this is not without obvious reason. In a very literal sense, as is useful to think of it, the gonad is a conveyor belt of nuclei as they pass through the various stages of mitosis and meiosis. To explain this in a way that is required for our purposes, in short, germ cells are derived from a proliferating stem cell population located at the distal end of the gonad (relative to the uterus). While germline stem cells give rise to sperm during larval stages, they subsequently switch to the production of oocytes during adulthood. The distal tip cell of the somatic gonad regulates germline stem cell proliferation and entry into the transition

zone and in this way, entry into prophase of meiosis I. Germ cells are partially enclosed by a plasma membrane, and share a common cytoplasm. Germ cell apoptosis occurs concomitantly with the exit of oocytes from the pachytene stage of meiotic prophase. After pachytene exit, the surviving germ cells grow rapidly and differentiate into individual, fully cellularised oocytes in the proximal germline, and are fertilised as they traverse the spermatheca following their maturation (defined by the shift between diakinesis and metaphase of meiosis I and is accompanied by nuclear envelope breakdown, rearrangement of the cortical cytoskeleton, and meiotic spindle assembly) (Hubbard and Greenstein 2000; Pazdernik and Schedl 2013).

The processes of meiotic maturation, ovulation, and fertilisation are temporally coupled. Acting as a hormone to trigger oocyte meiotic maturation, the major sperm protein galvanises the maturing oocyte to signal its own ovulation thereby setting fertilisation in motion. When the non-motile spermatids enter the spermatheca during the first ovulations, they undergo spermiogenesis to become motile spermatozoa capable of fertilisation, for which several sperm-specific integral membrane proteins are required (Hubbard and Greenstein 2000; Pazdernik and Schedl 2013).

Armed with this knowledge, it is now possible to meaningfully probe the cellular basis for the many unfertilised oocytes produced by the double mutant using the nuclei detection method described above. The quantification of nuclei in each zone of the gonad in wildtype, *myrf-1(ybq6)*, *myrf-2(ybq42)*, and *myrf-1(ybq6);myrf-2(ybq42)* young adults is shown in Figure 6.5D. Images were not included because gonadal nuclei span many many different focal planes (across which they were quantified live) such that it is at best unhelpful, and at worst misleading, to display just one image of each genotype — and especially useless here, given the peculiar morphology of the *myrf-1*, *myrf-2*, and *myrf-1;myrf-2* adult gonads. In the case of all mitotic (i) and meiotic (ii and iii)

zones of the gonad, the *myrf-1;myrf-2* double mutant exhibits a significant reduction in gonadal nuclei, which then relates directly to a diminished number of oocytes (iv). This implies that in the *myrf-1;myrf-2* double mutant, gonad growth is aberrant from its very beginning. The single mutants, however, do not show any signs of a compromised nuclear production capacity in the gonad, despite their unusual morphology. This is the first sign of true redundancy between *myrf-1* and *myrf-2*.

While this certainly goes some way to explain the abysmal brood size of *myrf-1;myrf-2* double mutants, and certainly explains why there are not a hundred more unfertilised oocytes (with germline growth and maintenance clearly being compromised), it does not satisfy the absence of fertilisation. This rests with the sperm. And so, exploiting the ability to detect nuclei with relative ease using DAPI in the whole animal, sperm counts were assessed in each genotype, the results of which are shown in Figure 6.6A. It is seen that *myrf-1;myrf-2* double mutants produce markedly fewer sperm compared to *myrf-1* and *myrf-2* single mutants. Here then, is the second recorded instance of classical redundancy between *myrf-1* and *myrf-2* — the single mutants displaying no phenotype of their own and appear as wildtype, yet the double has a greatly reduced sperm count indeed. Thus, it would seem that with respect to their twofold role in fecundity (germline development and sperm production), *myrf-1* and *myrf-2* are simply surplus to requirements. Although, on the face of it, this creates a glaring contradiction in the brood size data. How can it be that *myrf-1* and *myrf-2* are redundant for their role in reproductive fitness yet the single mutants thereof still yield significantly lower broods? The answer, of course, lies in their role in embryogenesis.

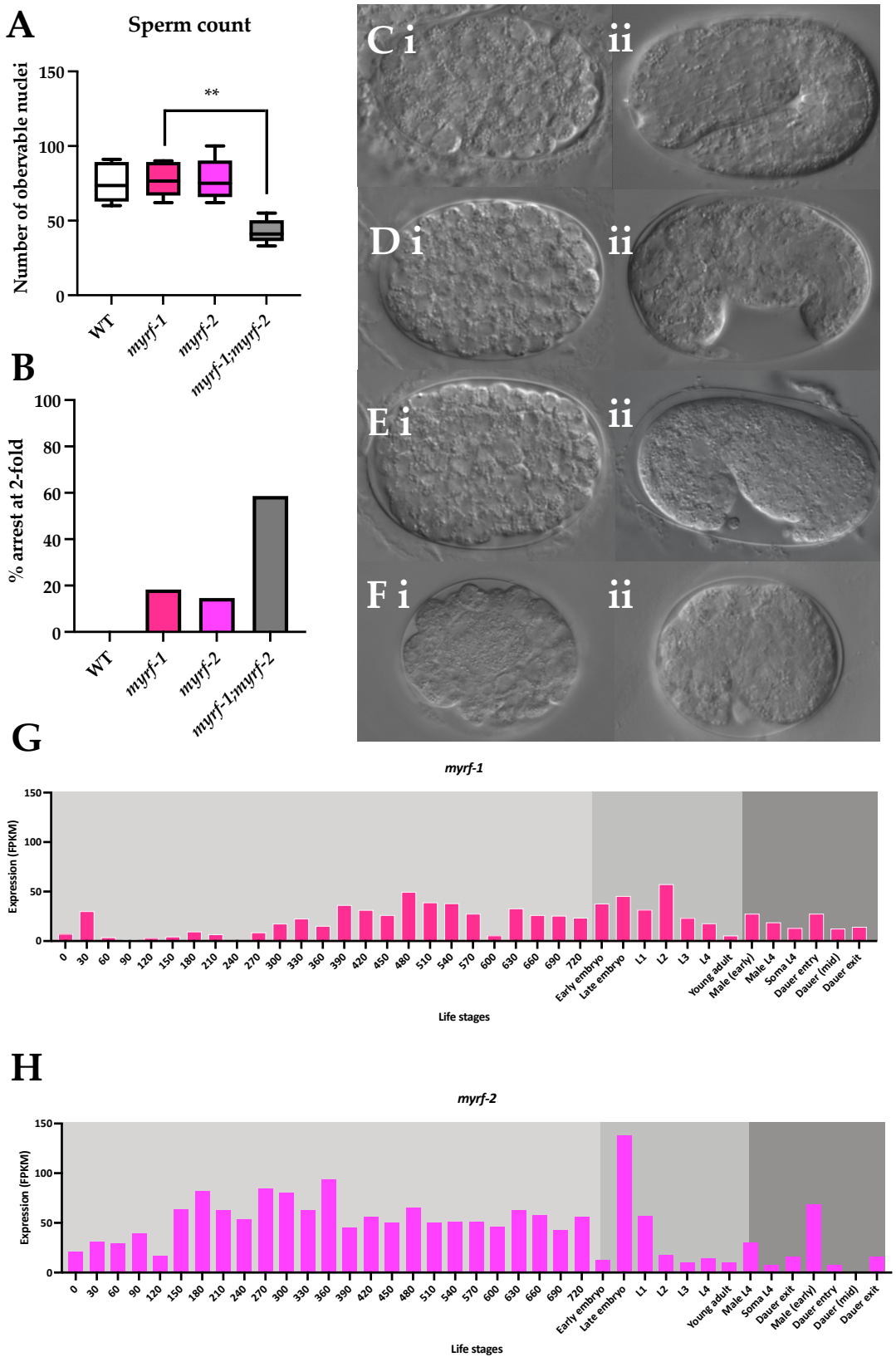


Figure 6.6. Further accounting for the reduction in reproduction fitness in the Myrf family mutants. (A) Sperm counts were quantified for four genotypes: WT, *myrf-1(ybq6)*, *myrf-2(ybq42)*, and *myrf-1(ybq6);myrf-2(ybq42)*. Here, sample sizes (i.e., number of young adult animals) are as

follows: WT (n = 51), *myrf-1(ybq6)* (n = 56), *myrf-2(ybq42)* (n = 52), and *myrf-1(ybq6);myrf-2(ybq42)* (n = 45). Sperm was counted following DNA staining (with DAPI) and the number of observable sperm nuclei in the spermatheca were manually counted. (B) Percentage of embryos that are declared PAT (paralysed arrest at 2-fold) across those four aforementioned genotypes where sample sizes are as follows: WT (n = 54), *myrf-1(ybq6)* (n = 46), *myrf-2(ybq42)* (n = 39), and *myrf-1(ybq6);myrf-2(ybq42)* (n = 60). Embryos were followed from late gastrulation and then throughout elongation. Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants with an unpaired t test. (C, D, E, F) depicts late gastrulation (~64-cell stage) and the 2-fold stage of WT (C i and ii), *myrf-1(ybq6)* (D i and ii), *myrf-2(ybq42),* (E i and ii), and *myrf-1(ybq6);myrf-2(ybq42)* (F i and ii) embryos. (G) Shows the temporal expression of *myrf-1* throughout the worm lifecycle. (H) Shows the temporal expression of *myrf-2* throughout the worm lifecycle. In both (G) and (H), median (mRNA) expression levels are shown (logarithmically in FPKM) for each life stage including minutes post-fertilisation and at various post-embryonic stages — data obtained from the modENCODE library (Gerstein et al. 2010; Araya et al. 2014).

Looking back to the results in Figure 6.4B, it was apparent that each single mutant produced a significant number of dead eggs, as did the double mutant relative to them. The only way to ascertain the role of a gene in embryogenesis is to the trace the embryonic cell lineage of a knockdown or knockout embryo to find the stage at which development goes wrong, just as we did in Chapter 3 for *tbx-35* and *tbx-36*. Though, unlike for *tbx-35* and *tbx-36,* it was readily apparent that *myrf-1* and *myrf-2* were not involved in early embryogenesis, as the eggs which were laid (having reached at least the 30-cell stage just prior to gastrulation), all appeared as wildtype. In following them through gastrulation and morphogenesis, it was noted upon initial inspection that many *myrf-1* and *myrf-2* single mutant embryos failed to complete the 2-fold stage of embryogenesis properly, and in fact became paralysed in the process, ceasing to display signs of life. A comparison of the '2-fold stage' reached by wildtype (Figure 6.6 Ci and ii), *myrf-1(ybq6)* (Di and ii), *myrf-2(ybq42)* (Ei and ii), and *myrf-1(ybq6);myrf-2(ybq42)* (Fi and ii) embryos is pictured in Figure 6.6, with the associated quantification (given as percentage penetrance) provided in Figure 6.6B. The double mutant will be dealt with, but first we are required to explain the presentational abnormality of *myrf-1* and *myrf-2* single mutant embryos.

It is seen that unlike all wildtype embryos that achieve a typical 2-fold, or as it is often referred, the 'plum' stage (reached at around 500 minutes post-fertilisation), *myrf-1* and *myrf-2* embryos appear deformed, adopting an unfamiliar 'croissant-like' form, rather than a canonical plum shape. In order to explain the mechanism underlying the embryonic paralysis, we are required to acknowledge that it is at this stage contractile muscle fibres form and so the muscle itself begins to contract, or as it could be more accurately described, twitch —because it is in fact spontaneous muscle activity due to the incomplete development of synaptic connections from the nervous system at the 2-fold stage. Genes that are essential for this process, identified from forward genetic screens, were historically designated the gene name 'Pat', referring to their 'Paralysed Arrest at Twofold' (Williams and Waterston 1994). There are ten genes in the Pat class; and it is with mutants thereof that *myrf-1* and *myrf-2* embryos share uncanny resemblance.

Genes in the Pat class are not paralogous, and, as far as is understood about their functionality, do not display transcription factor activity. Instead, they are predominantly voltage gated ion channels (for example, *pat-4* and *pat-5*), calmodulin (*pat-10*), or associated integrins (*pat-3*) required for the navigation of extracellular space (Williams and Waterston 1994; Qadota et al. 2012). It is therefore utterly plausible that *myrf-1* and *myrf-2* are required to regulate the expression of Pat class genes, but this has not been tested here. It should be noted, though, that on cursory inspection of the promoter regions of all ten Pat class genes lies the motif 5'-CTGGNAC-3' — this is the conserved DNA-binding motif of all known (and characterised) MYRF TF proteins (Huang et al. 2021); but of course being so short, the presence of this sequence in any Pat gene promoter could mean very little. And while they may or may not regulate Pat class genes already known about, *myrf-1* and *myrf-2* must have unique target genes because the two give rise to distinct phenotypes on their own, though the other side of that particular coin is that these data suggest

myrf-1 and *myrf-2* must also share *at least some* targets too, as is implied by the increased phenotypic penetrance of 2-fold defects in the double mutant. All this is to say that *myrf-1* and *myrf-2* have overlapping functionality in their regulation of the 2-fold stage of embryogenesis: the double mutant displaying a penetrance of 2-fold paralysis defects just slightly above that of the singles.

There is evidently some redundancy in embryogenesis between *myrf-1* and *myrf-2*, proven not just from how the penetrance of 2-fold paralysis defects is above and beyond that of the single mutants, but because defects in *myrf-1;myrf-2* double mutant embryos are evidently more severe (as shown in Figure 6.6Fi and ii). However, this may simply be caused by needing to complete morphogenesis in what is a manifestly smaller embryonic space, with these embryos clearly being smaller and more spherical than those produced by the single mutants and wildtype adults. It is, therefore, impossible to suggest with any degree of certainty if the *myrf-1;myrf-2* double mutant displays a higher penetrance of 2-fold paralysis defects because a) the two share a subset of the same target genes (i.e., proper redundancy) or b) that the misshapen embryos — likely produced this way due to the smaller, misshapen gonads from which they were derived — compromise the fidelity of morphogenesis, as it is required to occur in a smaller embryonic space. To establish the facts one way or the other, we would need to assess the target genes of *myrf-1* and *myrf-2*, which has not been performed as part of this work.

What is clear from this work so far is that *myrf-1* and *myrf-2* are as vital for organismal viability in *C. elegans* as the Myrf family is known for being elsewhere in the animal kingdom. We have seen how these roles stretch from germline development to sperm production to events in late embryogenesis. And, while the expression constructs generated previously (exhibited in Figure

6.3) are not visible in the germline to corroborate this due to their inevitable silencing, as most transgene arrays of this kind are as discussed elsewhere in this thesis, the RNA-seq data shown in Figure 6.6G echoes the newly described roles of *myrf-1* and *myrf-2*. Both *myrf-1* and *myrf-2* are expressed throughout development, but peak, notably, during mid-late embryogenesis and again at L1/L2. The former presumably pertaining to their role at the 2-fold stage and the latter to their role in synaptic rewiring. The expression levels of *myrf-2* are higher across all life stages, which is perhaps surprising given that *myrf-1* is the one in the pair that has discernible roles of its own. But with transcript levels taken in such a static way — there being no way of telling the extent to which *myrf-1* and *myrf-2* mRNA is localised or degraded, what their half-life is or if the MYRF-2 protein is subsequently degraded — not a great deal can be read into this. Once more, given the lack of cellular resolution that is offered by this solely temporal, whole-animal, dataset, it would not be prudent to conclude anything about their role in the germline from expression levels at the adult stage — the worms, by this point, contain so many cells that there is an inherent dilution effect in the assessment of the expression level of any one gene, in any one tissue type. However, the expression of the Myrf family in the germline specifically will be returned to later.

Do many Myrfs make light work? The provably superfluous duplication of *myrf-2* in *C. brenneri* and *C. remanei*

For now, we will revisit the duplication history of the Myrf family in the *Caenorhabditis* genus to establish if the other duplication identified — that produced *myrf-2.1* and *myrf-2.2* in *C. remanei* and *C. brenneri* — has similarly shaped the evolution of nematode development. Reflecting on the fact that *myrf-1* and *myrf-2* are well conserved, equally important, pleiotropic genes that appear to share a great portion of their functionality with one another (despite showing signs of

diversification in synaptic rewiring and late embryogenesis), it is a wonder what the *need* for the additional paralogue in these two lesser known *Caenorhabditis* species is. The additional duplication of *myrf-2*, persisting in *C. remanei* and *C. brenneri* still today, thus poses intrigue.

In an initial attempt to rule out that one of either *myrf-2.1* or *myrf-2.2* is not a pseudogene, both their amino acid sequences were aligned with other family members from the rest of the genus and something altogether more striking than evidence of pseudogenisation was found. Figure 6.7A depicts this alignment where all the Myrf homologues are shown in the genus and their domains are mapped beneath detailing their relative conservation. It is seen that while MYRF-2.1 possesses a TM, C1, and C2 domain, it does not have a DNA-binding domain. Inversely, MYRF-2.2, while in possession of a DNA-binding domain, has neither a TM, C1, nor C2 domain. Indeed, this is mirrored in their protein lengths, as MYRF-2.1 is 501 and 567 amino acids long in *C. brenneri* and *C. remanei*, respectively; and MYRF-2.2 is 524 and 451 amino acids long in *C. brenneri* and *C. remanei,* respectively. Not forgetting that MYRF-1, and likely MYRF-2, autocleaves into two N-terminal (1-482aa) and C-terminal (483-931aa) fragments, this implies that *C. brenneri* and *C. remanei* have resolved this most unique aspect of Myrf family biochemistry — that is, cleaving a very large protein into two distinct products with disparate functions — at the level of the gene, rather than at the level of the protein. In other words, two distinct Myrf products *have* to be made, and rather than yielding one long pro-peptide that gets processed into two functional units as occurs in the rest of the animal kingdom, instead, *C. brenneri* and *C. remanei* have capitalised on a duplication to simply transcribe two different genes. One gene giving rise to the would-be N-terminal fragment (*myrf-2.2*), and the other to the would-be C-terminal fragment (*myrf-2.1*).

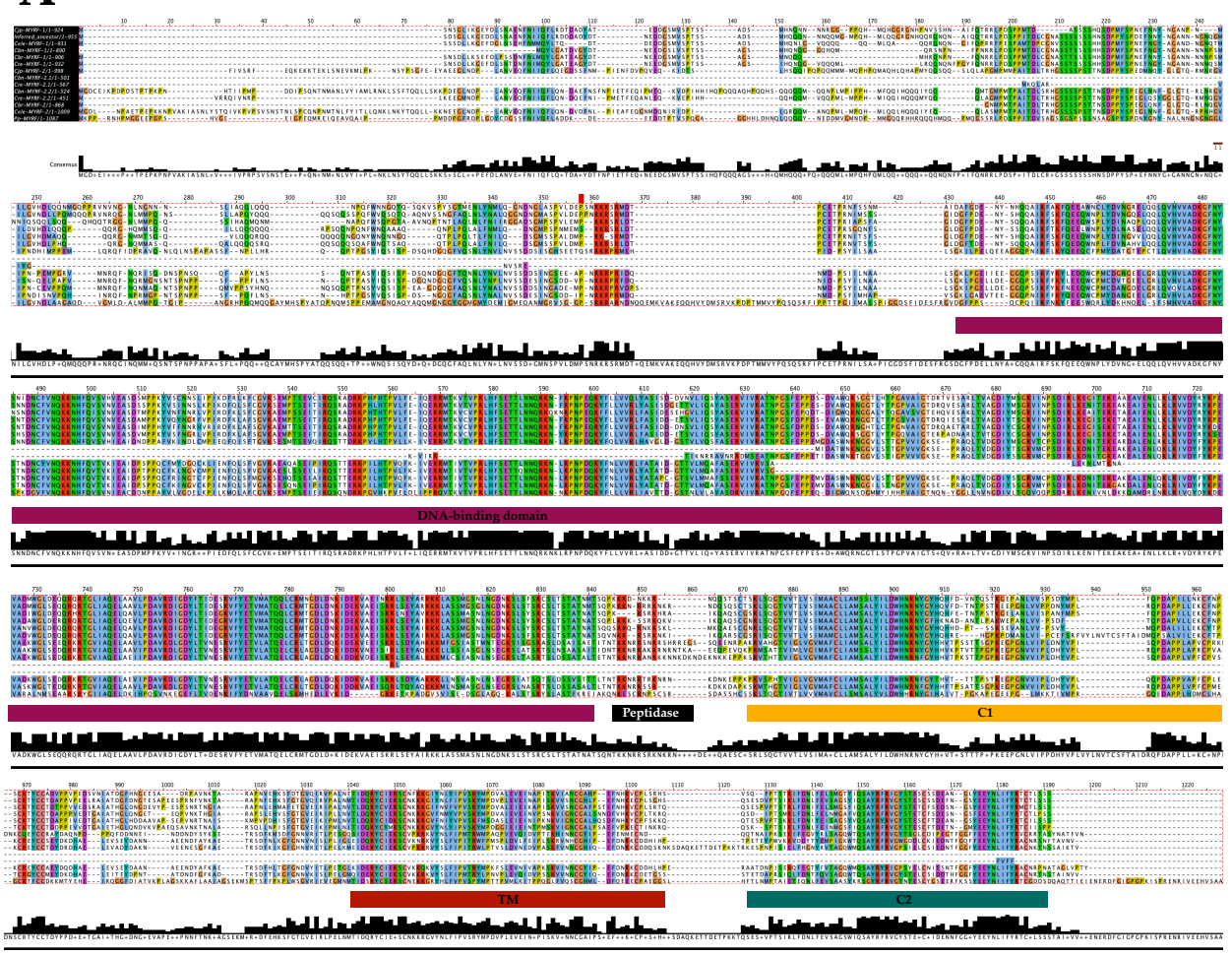


Figure 6.7. Domain architecture evolution in the Myrf family in the *Caenorhabditis* genus. (A) Multiple sequence alignment of the Myrf family visualised in Jalview using MAFFT v. 7.505 using

the E-INS-i iterative refinement strategy. Species abbreviations: Pp, *Pristionchus pacificus*; Cj, *Caenorhabditis* japonica; Ce, *C. elegans*; Cbn, *C. brenneri*; Cre, *C. remanei*; and Cbr, *C. briggsae*. Black bars running along the bottom of the alignment summarise conservation (amino acid consensus shown) where domains mentioned are annotated just above: DNA-binding domain; peptidase domain; C1 domain; TM domain; and a C2 domain. The sequence of the inferred ancestral MYRF, as determined by FastML (fed an alignment of the MYRF proteins, a species tree of the relevant nematode species, and a gene tree of the Myrf family, then processed using FastML default parameters), is shown. (B) Summary schematic of the domain level conservation (in % sequence similarity) between the Myrf homologues in the *Caenorhabditis* genus where the absence of the DNA-binding domain in *myrf-2.2* (only present in *C. remanei* and *C. brenneri*) is clearly shown, as is the absence of the C-terminal domains (C1 domain; TM domain; and a C2 domain) in *myrf-2.1* (only present in *C. remanei* and *C. brenneri*).

Or this is the hypothesis, at least. But first, prior to testing exactly that, Figure 6.7B schematises what has been described above and compares the sequence similarity of each canonical domain from the Myrf homologues to the single ancestral MYRF which presumably once existed at the birth of the genus (approximately 200 mya). Normally it would not be possible, with any degree of reliability, to estimate the ancestral sequence of a gene family; but given their slow rate of evolution within the genus (as indicated by the phylogram shown previously, with only 0.2 substitutions occurring per site per million years), it was indeed possible for the *Caenorhabditis* Myrfs using FastML — an algorithm for maximum likelihood ancestral sequence reconstruction (Moshe and Pupko 2019). It is noteworthy how the DNA-binding domains of MYRF-2.2 in *C. brenneri* and *C. remanei* are similar to those in the ancestor, and the same can be said for the three C-terminal domains found only in MYRF-2.1. This implies, as hypothesised above, MYRF-2.1 and MYRF-2.2 resemble the regular products of MYRF autocleavage in other taxa, and so it does seem that *myrf-2.1* and *myrf-2.2* are two genes doing the job of one. Though computational methods alone cannot settle this convincingly. Extensive sequence similarity does not preclude the possibility that the two have distinct regulatory dynamics. It is with this in mind that the expression of *myrf-2.1* and *myrf-2.2* was assessed by constructing translational reporters. The purpose they serve is twofold. First, and most obviously, to work out the expression pattern of

myrf-2.1 and *myrf-2.2 in vivo* and the extent of their overlap. And second, by generating full-length translational reporters, they are available for use as putative rescue constructs for the Myrf family mutants in *C. elegans*, thus deducing their evolutionary dynamics with respect to the single-copy *myrf-2* from which they were derived.

The C-terminal translational reporter constructs built for *myrf-2.1* and *myrf-2.2* use *mCherry* and *gfp*, respectively. Both were derived from *C. brenneri*, as opposed to *C. remanei*, for reasons of consistency though *C. remanei* would have been equally valid to use here. Great efforts were made to generate transgenic lines with MYRF-2.1::mCherry and MYRF-2.2::GFP in *C. brenneri* and *C. remanei,* though this was repeatedly unsuccessful, most likely because both of these species are gonochoristic and so perfectly timing the injection process with subsequent recovery and then mating was too finicky and experimentally complex. It should be noted that no investigation that has attempted the same procedure has ever been able to generate stable transgenic lines using either species. But as these constructs are still capable of being expressed in *C. elegans*, that is where they were injected instead.

Figure 6.8 depicts the expression patterns of *myrf-2.1::mCherry* and *myrf-2.2::gfp* in *C. elegans*. Bearing in mind that both constructs are invariably overexpressed, it is valuable to note that no stable lines were ever made in a wildtype background, nor in *myrf-1(ybq6)* or *myrf-2(ybq42)* single mutants. This strongly suggests that when overexpressed, Myrf family genes are **lethal** meaning their transcriptional regulation of target genes is dosage-sensitive. Concordant with this justification is how multiple independent lines could only be constructed in a double mutant background, thereby generating *myrf-1;myrf-2*;ouEx900 [*myrf-2.1::mCherry*] and *myrf-1;myrf-2*; ouEx901 [*myrf-2.2::gfp*], which could then be crossed together. Figure 6.8Ai and ii demonstrates the

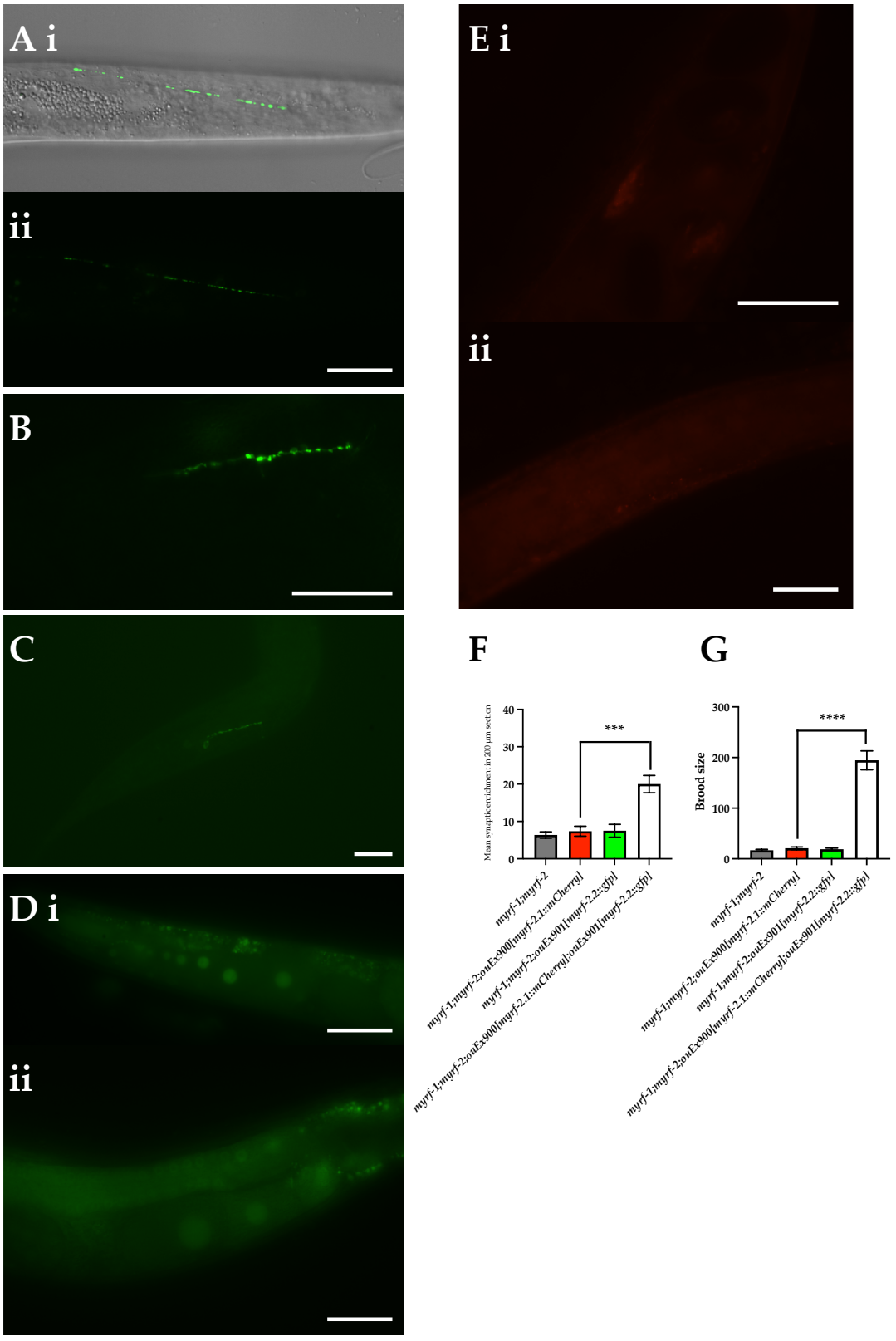


Figure 6.8. Rescuing *myrf-1;myrf-2* phenotypes in *C. elegans* with new paralogues from *C. brenneri*. (A) (i) Merged DIC/GFP image showing a section of dorsal nerve cord in a *myrf-1(ybq6); myrf-2(ybq42)* an animal at the mid-L2 stage carrying both ouEx900[myrf-2.1::mCherry] and ouEx901[myrf-2.2::gfp]; (ii) shows the same with the DIC overlay removed. (B) Higher power image of the restored synaptic rewiring ability in animals carrying both ouEx900[myrf-2.1::mCherry] and ouEx901[myrf-2.2::gfp]; panel (C) depicts the same only at higher

magnification and resolution. (D) (i) Shows the anterior gonad arm of a *myrf-1(ybq6);myrf-2(ybq42)* double mutant animal at the young adult stage rescued with ouEx900[myrf-2.1::mCherry] and ouEx901[myrf-2.2::gfp] — owing to these effectively being rescue constructs, fluorescence (of ouEx901[myrf-2.2::gfp]) is observed in the germline which is rare for most constructs expressed in the gonad (due to silencing). (ii) Depicts the posterior gonad arm (of a different animal to that in (i)) carrying the same transgene arrays derived from *C. brenneri*. (E) (i) Shows the head region of a *myrf-1(ybq6);myrf-2(ybq42)* animal at the mid-L2 stage carrying both ouEx900[myrf-2.1::mCherry] and ouEx901[myrf-2.2::gfp] where expression of ouEx900[myrf-2.1::mCherry] can be observed either side of the pharynx. (ii) Shows a section of the dorsal nerve cord of the same animal in (i) where low level expression of ouEx901[myrf-2.1::mCherry] is observed in synaptic vesicles. (F) Quantifies the synaptic enrichment in a 200 μm stretch in *myrf-1(ybq6);myrf-2(ybq42)* double mutant animals at the mid-L2 stage carrying one, or both, of ouEx900[myrf-2.1::mCherry] and ouEx901[myrf-2.2::gfp]. (G) Quantifies the brood sizes of *myrf-1(ybq6);myrf-2(ybq42)* animals carrying one, or both, of ouEx900[myrf-2.1::mCherry] and ouEx901[myrf-2.2::gfp]. Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants with an unpaired t test. Scale bars = 50 μm.

expression of *myrf-2.2::gfp* along the dorsal nerve cord at late L2, where Figure 6.8B provides a higher magnification example of the same (in a different animal), showing that synaptic rewiring looks, to an approximation, restored in these animals. Although, owing to the variability of its (over)expression, Figure 6.8C shows that the presence of *myrf-2.2::gfp* is not necessarily constant or uniform along the dorsal nerve cord; though where it is evident, the dorsal nerve cord appears as wildtype when compared with the *myrf-1p::gfp*, *myrf-2p::gfp* transcriptional reporters, or even *rab-3p::wrmScarlet* L2 animals, shown back in Figure 6.3.

But how does the expression of the C-terminal domain portion, *myrf-2.1::mCherry*, compare? While less visible, predictable, and uniform than its TF counterpart, MYRF-2.1::mCherry is still associated, albeit in an incredibly localised irregular fashion, with the both the dorsal and ventral nerve cords as shown in Figure 6.8Ei, and very subtly in puncta along the dorsal nerve cord as in Figure 6.8Eii. It is also observed in the head beside the pharynx. The reason for this is not especially clear, and will be discussed in the next section as it may pertain to the functions, and

therefore coevolution of, *myrf-2.1* and *myrf-2.2*. In any event, *myrf-2.1::mCherry* is still required to rescue the double mutant phenotype in synaptic refinement in *C. elegans*, with *myrf-2.2::gfp* alone being **insufficient** to do so. This is quantified in Figure 6.8F. Reconciling this with the number of puncta along the dorsal nerve cord in a wildtype back in Figure 6.3, it is plain that rescue is not 100%, though this is to be expected from the inherent variability of a rescue experiment performed using extrachromosomal arrays. For quantification, the *rab-3p::wrmScarlet* background was crossed in because the expression of *myrf-2.1::mCherry* was deemed to be sufficiently different, and brighter, such that puncta labelled with *rab-3p::wrmScarlet* could be visibly distinguished from those fluorescing red with *myrf-2.1::mCherry.* Alas, the expression of *myrf-2.2::gfp* was not deemed consistent enough along the nerve cords for the quantification.

A considerable advantage to injecting both *myrf-2.1::mCherry* and *myrf-2.2::gfp* as, what are in effect rescue constructs, means that they are visible in the germline unlike the *myrf-1p::gfp* and *myrf-2p::gfp* transcriptional reporters discussed previously. The germline expression of *myrf-2.2::gfp* is shown, in two different animals in both the anterior (i) and posterior (ii) gonad arms, in Figure 6.8D. Expression of this construct (but not *myrf-2.1::mCherry*) is visible in the mitotic and meiotic zones of the gonad, as well as in the oocytes, which reconciles with the overtly redundant roles of *myrf-1* and *myrf-2* in the germline quantified above. This implies that *myrf-2.1* and *myrf-2.2* have quite literally divided, not the roles per se, but the physical attributes of, the *myrf-2* gene between two new genes, rendering them collectively with the same functional potential as the single MYRF-2 protein in other *Caenorhabditis* species. This explains their lack of deviation from the expression and roles of *myrf-1* and *myrf-2* in *C. elegans*, and why *myrf-2.1::mCherry* and *myrf-2.2::gfp* **together, though not individually,** are capable of rescuing — almost entirely — the phenotypes of

the *myrf-1;myrf-2* double mutant in *C. elegans*, including the low brood size as shown in Figure 6.8G.

Discussion

Evaluating the evolution and function of *myrf-1* and *myrf-2*

In stark contrast to the T-box family presented in Chapter 3, the Myrf family is undynamic, with members rarely gained and lost and their functions seldom diversifying to become anything less than essential to the organisms in which they are found. But to all taxa except the *Caenorhabditis* genus (in which additional paralogues are present) making reference to the 'Myrf family' is rather grandiose given it actually only constitutes a single gene. It was from this initial standpoint of exceptionalism that the evolution and function of the Myrf family were probed in the genus. Just as it is only possible to understand the significance of a gene upon its experimentally induced absence — hence the modus operandi of geneticists the world over being to 'break' a gene in some way and then observe the consequences — it is only possible to understand the significance of a pattern or trend in evolution when it too is broken. The duplications of the MYRF gene in the *Caenorhabditis* genus are the break of the trend, and in this way, an experiment which natural selection has already performed for us.

To establish the significance of this idiosyncratic pattern of MYRF evolution, the functions of *myrf-1* and *myrf-2* were characterised in *C. elegans*. Overall, it was shown that both genes display remarkable functional overlap, with true redundancy observed in their roles in germline growth and maintenance and sperm production. However, *myrf-1* and *myrf-2* were found to differentiate in other aspects, including late embryogenesis and synaptic rewiring — with respect to both, the genes individually had roles to play, yet the penetrance of their phenotypes increased upon

generating the double mutant. This strongly suggests that the single Myrf family ancestor (the sequence of which was inferred from this investigation) to *some extent* performed all of these roles, and upon the generation of its paralogue, its many vital roles were divided between two genes such that *myrf-1* and *myrf-2* together are as indispensable, if not more so, for reproductive and developmental processes as the single MYRF gene is in other taxa. And so, it might be suggested that maintaining overlapping functionality in this way has conferred one of two advantages. The first possible advantage being that duplication here implicitly offers belt as well as braces for a multitude of functions critical to organismal viability. Second, that *myrf-1* and *myrf-2* have undergone a necessary division of their once combined labour (in the now extinct ancestor), because together they perform more roles than any single Myrf gene could. This second proposition is predicated on the assumption that the single MYRF found in other taxa is not as pleiotropic, that is to say as vital to development, as *myrf-1* and *myrf-2* are in *C. elegans*. But is this true? And if it is, does that mean two MYRF genes really are better than one?

Vertebrates provide the best example of MYRF pleiotropy to rival the paralogues in *C. elegans*, but there is no doubt that this is because they are comparatively so well studied. A smattering of the roles MYRF has in vertebrates include: myelination (Bujalka et al. 2013), the development of retinal pigment epithelium (Cross et al. 2020), and most recently, cardiac organogenesis and gonadogenesis (of both testis and ovary) (Calonga-Solís et al. 2022; Collin et al. 2022). Being such a pivotal player in so many spatiotemporally distinct aspects of development seems an insult to the idea of gene duplications being necessary to give rise to new functionality, but on closer inspection at the molecular level, the developmental pleiotropy of MYRF is not what it first seems.

From work thus far, it is known that in the roles it is required for listed above, the murine MYRF orthologue interacts directly with Sox10 and/or Cited (Hornig and Fröbb et al. 2013; Calonga-Solís et al. 2022; Collin et al. 2022). Genes which, to be clear, are not found in the *Caenorhabditis* genus. And while it is not uncommon for key nodes of a gene regulatory network to be reused and rehashed in others, MYRF certainly seems to provide an extreme example of this. In order to make conclusive comment, one way or the other, about the functional potential of the two MYRF paralogues in *C. elegans* relative to their single-copy counterpart in vertebrates, the target genes of *myrf-1* and *myrf-2*, and their regulators, would need to be thoroughly characterised in all the roles in which they have been implicated as part of this investigation. One has to be especially mindful not to assume that simply by virtue of being paralogues implicated in the same developmental processes that they are, by extension, part of the same gene regulatory networks with the same interactors therein. While this may be a safe deduction to make in those instances of bona fide redundancy, any other kind of genetic interaction (additivity, overlapping functionality, etc.), this cannot be really true for. In other words, the molecular mechanisms by which they operate must have, at least partially, diverged. This matters because as these kinds of relationships exist between *myrf-1* and *myrf-2*, this says something about their collective potential compared to that of any single gene. Because it is self-evident that if just one gene could perform all the functions of both *myrf-1* and *myrf-2*, this immensely pleiotropic — though admittedly mostly redundant — paralogue pair would be less well-conserved throughout the genus. So while the Myrf family may not be twice, or even close to twice, as important to the development of *C. elegans* as the single MYRF gene is in other taxa, it is still possible to say that they collectively hold more genetic potential than any single MYRF orthologue.

Implicit in this idea of the genetic potential of paralogues then is the acquisition of new molecular interactions, rather than developmental roles per se. With the Myrfs displaying a high degree of pleiotropic competency on the whole, it would at first seem intuitive for more paralogues than there are to have been generated over the course of evolution (to divide functions among paralogues as labour is divided among workers). But the argument for why this is not so is reminiscent of our discussion about why genes that display moonlighting behaviour are similarly not prone to duplication in Chapter 1. It follows that in order to be retained by natural selection, there needs to be a selective advantage to any given gene duplication, else pseudogenisation would be inevitable, or its removal from the genome altogether (Pérez Jurado et al. 1998). With the single MYRF capable of performing so many functions, it follows that these developmental roles in which MYRF is implicated must not compromise one another, that is to say they are independent, whether in space and/or time. In addition, being underpinned by the same couple of molecular interactions, which are so chameleonic in different biological contexts, means that no advantage would be conferred should MYRF duplicate in vertebrates and divide its developmental roles between paralogues. In fact, it would, if anything, increase the mutational vulnerability of any function of MYRF by using more genes to do the job of one. Or as it can be thought of for the sake of this argument, a needless increase in mutational space. The single-copy MYRF in mice gets by just fine performing all these roles because, in essence, they are underpinned by the same few molecular interactions. So actually more than just being unnecessary, sequence diversification of hypothetical MYRF paralogues, as per our understanding of the way in which the gene repeatedly calls upon the same molecular interactions to perform a huge variety of functions, would actually be selected against.

The evolution of *myrf-2.1* and *myrf-2.2*: one Myrf too many

It is commonly held that mutations are random, and in a sense this is true, but this is only with respect to the affect they have on the function of the gene. Mutations cannot be said to be random with respect to the frequency at which they occur across the genome; this is largely undisputed (PWH Holland, pers. comm.). That is to say that different genes, regions of chromosomes, or even entire chromosomes, have very distinct probabilities of accumulating, and then fixing, mutations (see Monroe et al. 2022 for the most recent characterisation of this). There is unlikely to be one overarching reason for these propensities. While the immutability of the Myrf family has already been established with respect to duplication and sequence divergence, the emergence of *myrf-2.1* and *myrf-2.2,* and their radical co-evolution to follow, appears to be the great exception to this rule.

During the course of this investigation, we found and characterised *myrf-2.1* and *myrf-2.2* and their unusual reciprocal domain loss. It is, quite plainly, subfunctionalisation operating at the domain level — each gene complementarily degenerating to lose what the other must retain. But this evolutionary terminology does not go so far as to explain the biochemical nuances going on here. Both *myrf-2.1* and *myrf-2.2* are duplications conserved, in the state they are, because in reality they are — when in partnership — no different to *myrf-2* in other *Caenorhabditis* species. But when reconciling this with our understanding about how they operate as proteins, things start to become unclear.

MYRF proteins, ordinarily, have to localise to the endoplasmic reticulum membrane in order to be cleaved into active N-terminal fragments which then translocate into the nucleus to drive the transcription of target genes. It is thought that the only purpose of the C-termini of MYRF proteins is in the post-translational processing of the N-terminal TF element by facilitating its localisation to

the endoplasmic reticulum membrane in the first instance (Bujalka et al. 2013; Meng et al. 2017). It is thus puzzling why *myrf-2.2* **and** *myrf-2.1* (necessarily in that order), are both retained by *C. brenneri* and *C. remanei* considering MYRF-2.2 encodes a TF without the need for post-translational cleavage of an initial product. What purpose does *myrf-2.1* serve? For the avoidance of doubt, there is no peptidase domain, or even a serine-lysine dyad, in either MYRF-2.1 or MYRF-2.2. But of course it cannot be so simple, else there would be no C-terminal domain in any MYRF protein if its only function was to be confined to the cellular dustbin once it were cleaved off to yield a DNA-binding domain. Indeed, this work irrefutably proves a function for the C-terminus beyond this, because otherwise MYRF-2.1 would not be required to rescue the phenotypes of the *myrf-1;myrf-2* double mutant in *C. elegans*, and it is. Put another way, MYRF-2.2 alone is insufficient to rescue the double mutant phenotypes, implying that the N-termini and C-termini of MYRF proteins interact beyond mere autocleavage. The unexpected localisation of MYRF-2.1::mCherry perhaps holds promise for understanding this — sitting either side of the terminal bulb of the pharynx roughly around what is known as the excretory gland cell (that is the terminus of the nematode excretory system) on the ventral side. Though, it cannot be ruled out that the localisation of MYRF-2.1::mCherry in the head region is not an experimental artefact given it is now appreciated that mCherry contains a fluorescent short protein isoform that interferes with the function of its reporter gene (Fages-Lartaud et al. 2022). But in any event, this goes to show that where genes deviate from an expected scenario they are able to provide insights that conservation, and strict adherence to the rules, cannot.

But paradoxically, in one critical sense, the retention of both *myrf-2.1* and *myrf-2.2* is evidence in support of an earlier argument: that the Myrf family does not utilise gene duplication as an evolutionary tool. If the only other instance, besides the duplication that gave rise to *myrf-1* and

myrf-2, is the generation of an additional paralogue which has only lost functionality, it stands to reason that it is not for lack of opportunity that the Myrf family do not duplicate, it is that they do not make use of duplicates should they arise. In this way, *myrf-2.1* and *myrf-2.2* are the exception that prove the rule.

CHAPTER 7

General Discussion

Synopsis

This thesis began with a premise. This premise was that the existing framework of duplicated gene evolution is unsatisfactory in that its three constituent arms fail to adequately capture and explain the complexity of the real ways in which duplicated genes have already been *shown* to evolve. Examples of paralogue dynamics that proved contrary to the classical framework were described in Chapter 1, however all these examples were extracted from studies that did not set out with the intention of reforming, or even elaborating on, our current perception of the ways in which duplicated genes evolve. It was quite simply the case that the authors of these works stumbled upon more complex paralogue dynamics than current theory could adequately explain. In other words, the current framework proved to be retrospectively outdated. We went on to describe why this was so. Because the theory is exactly that, a theory, and was not built upon empirical foundations. And so, with the aim of rectifying an empirically dispossessed field of evolutionary biology, this thesis set out to probe duplicated gene dynamics directly, to push forward our understanding of a) the roles of duplicated genes in major evolutionary events by establishing the kinds of functions new genes have been recruited to perform, and b) the processes underlying the unusual evolutionary trajectory of certain gene classes, considered stalwarts of the animal kingdom, that have just so happened to diversify in the *Caenorhabditis* lineage of nematodes.

The *Caenorhabditis* genus was chosen as the model taxon in which to solve these problems owing to the enormous, arguably unmatched, genetic tractability of the *C. elegans* paradigm. It was exhibited

throughout this work how, for the most part, computational and experimental procedure could be extended to neighbouring species making it feasible to assess the evolution **and** function of genes over evolutionary time. Indeed, the array of methods deployed in this thesis as a testament to this were described in Chapter 2.

Chapter 3 was where, not only did we delve into our first aim of understanding the ways in which duplicated genes have the power to shape evolutionary processes, but we also began learning — mechanistically — how they do so. The evolutionary path taken by the T-box family in the *Caenorhabditis* genus was shown to be nothing short of extraordinary. We saw how, against a backdrop of historic rampant duplication, reciprocal paralogue loss among extant populations shaped the embryonic development of those populations, though with consequences that likely won't be felt for millions of years. Nevertheless, the reciprocal nature of deleterious mutation accumulation among wild populations in the *tbx-35*/*tbx-36* gene pair, it was shown, was met with newly acquired regulatory innovation in *tbx-36*. The combination of genetic events unfurling in real time in the T-box family enabled us to catch a rare glimpse inside the molecular engine of speciation.

But given the size and the extreme dynamism of the T-box family, Chapter 4 turned instead to the Warthogs to dissect out — in a slightly more manageable model — the redundancy relationships in what is still a developmentally important multigene family. It was found that belying overlapping functionality among paralogues can be the adoption of complex combinations of classical fates concomitantly. And because of this, stronger redundancy relationships can exist among more distantly related members of gene families. In the Warthog family, this phenomenon was underpinned by members' various degrees of association with ancestral functionality. Chapter

5 probed this in more detail using the example of the smaller, simpler Drd family, finding how great degrees of functional loss could be, counterintuitively, selected for provided such loss enabled at least some investment in other functionality. This exposed the fallibility of the idea that generating new genes is always an evolutionary opportunity. Sometimes new genes are simply surplus to requirements, but are still, intriguingly, retained. No where was this more evident than in the Myrf family in Chapter 6. In the case of the duplications in the Myrf family, paralogues were found either to be redundant or to have undergone radical domain loss. With paralogues complementarily losing half of the protein-coding gene to generate no more, though crucially no less, genetic potential than their unduplicated counterpart, the Myrf family gave us insight into the ways in which domain level re-landscaping can dictate the evolutionary opportunities available to duplicated genes.

Collectively, the work in this thesis has illuminated multiple aspects of duplicated gene evolution. It is now the purpose of this Chapter to bring it together.

The Road Less Travelled? Exploring the Nuanced Evolutionary Consequences of Duplicated Genes

Functional redundancy has been a recurring theme in this work, though it is a thorn in the side of classical ideas in duplicated gene evolution. According to the classical framework, redundancy is synonymous with relaxed selection such that redundant genes, for the most part, degenerate to become pseudogenes where a select few go on to partition ancestral regulatory roles and subfunctionalise or, rarer still, some take on new functions and neofunctionalise. But the observation that paralogues exhibit overlapping functionality has been recorded by many authors for many years, this is not new (Tischler et al. 2006). What's more, it was even found to be

surprisingly common (Tischler et al. 2006). What has been uncovered here — what makes this work insightful — is the exploration of *why* it is such a common occurrence, and why it is maintained for unexpectedly long periods of evolutionary time.

It was seen how, hand in hand with the retention of ancestral functionality, paralogous genes gained new roles, commonly reducing their investment in any one biological process to become more and more pleiotropic as gene families expanded. A bit of loss here. A bit of gain there. In the main, no paralogue, across any of the four families, became less multifunctional than their ancestor, with the obvious exceptions of *wrt-7*, *myrf-2.1* and *myrf-2.2* aside which all lost functionality more generally. Distributing ancestral functionality, while acquiring additional (and possibly even unrelated) roles, is proposed here as a kind of bet hedging. This encapsulates the ultimate trade off between the two advantages of generating duplicated genes: innovation and robustness. Paralogues may gain new roles at the expense of their old ones, but this is actually advantageous — in a way that is twofold — providing they do not become totally relieved of ancestral functionality. In our bet hedging scenario, evolutionary opportunities are still made available, novelty is still attainable, (in embryogenesis, organogenesis, stress responses, etc.), but importantly the mutational burden on any one gene is relinquished when it comes to still *having* to perform — a bit of — the role of the ancestor. This then, facilitates innovation with respect to overall function, but robustness with respect to overall mutational risk to the processes in which these genes are involved.

This explanation is directly at odds with an idea that was being developed in an era when subfunctionalisation had its firmest grip on the field of evolutionary genetics, and it is probably because of that that it is reminiscent of it. Escape from adaptive conflict (EAC) (Hittinger and

Carroll 2007; Conant and Wolfe 2008) proposes that a single-copy gene is selected to perform a novel function while still having to maintain its ancestral function. Critically, this gene is constrained from optimising either the novel or ancestral function because of detrimental pleiotropic effects on the other function, or so the theory goes. Therefore, in EAC, gene duplication offers the perfect solution. Following a gene duplication event, one copy is free to improve the novel function, whereas the other is selected to improve the ancestral function. This builds on the idea of subfunctionalisation in its division of ancestral roles (or expression), but develops it in that it is predicated on a genetic conflict that gene duplication purports to resolve. In EAC, a gene is a scarce resource with two competing uses. But if anything, we have only seen instances of the inverse: genes disinvesting — partially — in ancestral roles to become more multifunctional, more pleiotropic, though on the whole less important, less vital, to the processes in which they are implicated.

Though there is no doubt that EAC is a mode of duplicated gene evolution that some paralogues adopt de facto. It is easily foreseen how gene products could have two competing functions that are underpinned by different, opposing spatiotemporal and structural requirements. That being said, the only genetically characterised example of EAC is between paralogues in the anthocyanin biosynthetic pathway in morning glories (*Ipomoea*) (Des Marais and Rauscher 2010). In short, paralogues required for anthocyanin biosynthesis each specialise in particular enzymatic roles. All roles which a single ancestor, in its brief existence, was once believed to perform (Des Marais and Rauscher 2010). As we saw in the Drd family, enzymes are a special case in which juggling multiple roles is not always molecularly feasible. Specificity and pleiotropy are, for enzymes, opposing goals. But while *drd-1.3* was found to disinvest in the starvation response, it concomitantly gained multifunctionality in **both** gonadogenesis and male tail development. One

can only assume that these latter two roles are underpinned by the same catalytic reaction else they too would be in conflict with one another and thus mutually exclusive. And so, the evolution of *drd-1.3* is considerably nuanced. It is true that *drd-1.3* escaped from one adaptive conflict, but equally, gained pleiotropy in another sense.

Altogether more simple, though still tethered to this idea of molecular constraint, is the theory that paralogues retain overlapping functionality — to a greater or lesser extent — quite simply because they have to in order to retain any functionality at all. In Kuzmin et al. 2022, paralogues were only able to diversify their roles as far as the structural constraints of the proteins they encoded enabled them to (Kuzmin et al. 2020; Kuzmin et al. 2022). To formalise this, the authors coined the term 'functional and structural entanglement', which describes the degree to which gene function is constrained by the intrinsic physical forces acting on its protein product. By this logic, non-essential paralogues fall into two basic classes — those that continue to play overlapping roles, and those that have largely diverged in their functions. The former class of paralogues necessarily overlap in function because they cannot, by definition, diversify too much else they risk becoming non-functional. The authors of this work use membrane proteins as the archetypal example of this class where the hydrophobicity of ion channels and receptors must be prioritised to the extent that structural (and so too, functional) divergence seldom occurs in their paralogues because, probabilistically, it would compromise their ability to function at all. With regard to paralogues which fall into the second class, the authors propose TFs as genes which are more able to diversify as DNA-binding domains naturally come in more functionally permissible flavours.

But there is one major problem with this view of paralogue evolution: not all membrane proteins remain redundant and not all TFs diversify their roles. Take *myrf-1* and *myrf-2* as just one example.

Take *tbx-37* and *tbx-38* as another. In fact, we explored how most transcription factors, statistically speaking, retain overlapping functionality because they hypofunctionalise and are locked into a co-regulatory relationship with their paralogue(s). There is no doubt though, that entanglement as an inherent property can adequately describe the evolution of some TFs. It may well underpin the evolution of *tbx-35* and *tbx-36*.

And presumably, duplicated genes can also fail to diversify because to do so would be selected against for any other number of reasons. Perhaps increased dosage is simply advantage enough for some, or as seems appropriate to reach for by way of explanation in the case of the Myrf family, "if it ain't broke, don't fix it", is a maxim that may well apply to others. Even duplicated genes are not limitless in their potential. Limitations by another name, are just adaptations.

If all of this sounds infuriatingly contradictory or lacking a single, unifying theorem, that would be intentional. Because there is no single grand explanation, no one framework that can explain all of duplicated gene evolution. Paralogous genes may be thought of as limitless only in their capacity to adopt any 'fate' which, as we have seen, may really be multiple fates combined. All of this too, sits in conflict with an association to the past. Duplicated genes are in a tug-of-war optimising the two benefits duplication brings in any given circumstance: innovation and robustness. It is, always, going to be out with some old, and in with some new.

But as with any process in evolution, the divergence of paralogous genes has a chronology — admittedly this one takes many forms. All paralogues indeed follow their own evolutionary trajectories, though all are defined by a stronger or weaker association with one another that

changes over time. Such associations may be underpinned by unique combinations of regulatory and coding sequence divergence, or a lack thereof.

To visualise this in some helpful way — to illustrate the point I seek to make — the extent to which paralogues diverge over time from their ancestor can be shown diagrammatically as a spectrum of functional similarity, or redundancy, between duplicated genes (Figure 7.1). At one end, paralogues exhibit total redundancy; at the other, paralogues have diversified such that they exhibit no functional overlap which may be for a multitude of reasons. While the various fates described and discussed in this thesis may be mapped onto this spectrum (just like the genes investigated in it), it is true, of course, that an endless number of other fates are also possible to assume following duplication. It is no coincidence that there has never been any one canonical example set to cite of 'neofunctionalised genes', 'subfunctionalised genes', or 'pseudogenised genes'. It is an intellectually impossible exercise, given all the empirical data presented in this thesis and beyond it, to think what each of these would even look like in reality. Diversity in paralogous genes, it should come as no surprise, is as extensive as the diversity of life itself.

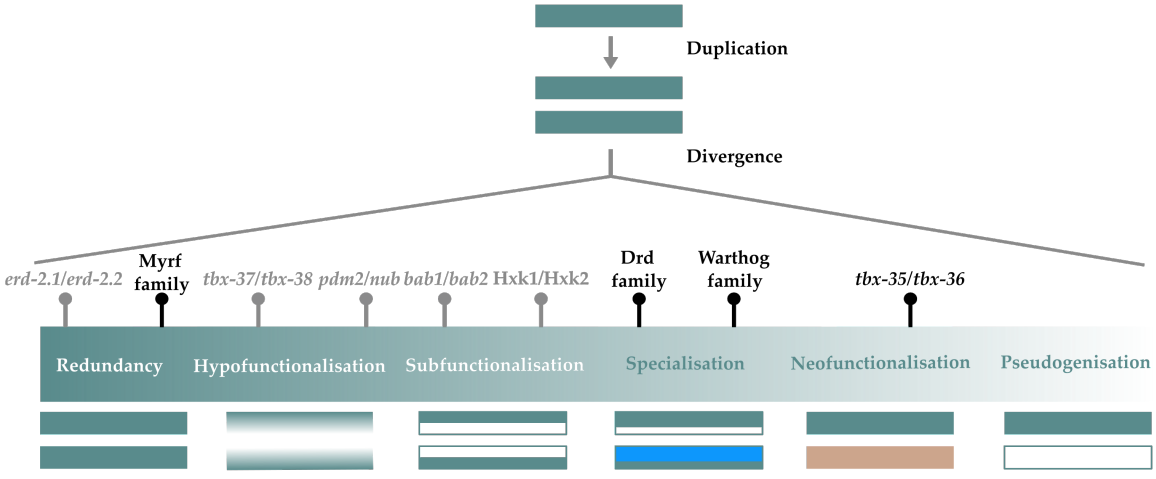


Figure 7.1. Endless fates most bountiful. The spectrum of possible duplicated gene fates is depicted according to the functional similarity between paralogues where complete functional redundancy (far left) represents that greatest degree of functional similarity that can be exhibited

and pseudogenisation (far right) the least. Hypothetical paralogous scenarios are represented by rectangles where the colour therein denotes function. Teal depicts the ancestral gene function; white depicts the loss of ancestral function; sky blue depicts an elaborated function as in the case of specialisation; and pink depicts new functionality as in the case of neofunctionalisation. Throughout the middle of the spectrum there exist a multitude of potential paralogous gene fates with all possible degrees of functional overlap between them. The experimentally determined fates of paralogues discussed and discovered over the course of this thesis are mapped onto the spectrum. These include: *tbx-37/38*; *erd-2.1/erd-2.2*; Hxk1/Hxk2; *bab1/bab2*; *pdm2/nub; tbx-35/36* (Chapter 3); the Warthog family (Chapter 4); the Drd family (Chapter 5); and the Myrf family (Chapter 6).

Concluding remarks

There are endless ways for duplicated genes to evolve; indeed, as many ways as there are paralogous genes. It is, rather unsatisfactorily, impossible to give any more than a flavour of all the evolutionary fates of duplicated genes even in a whole thesis, but this is the point. Our ability to study genetic complexity is now enabling us to catch a glimpse of the staggering array of paralogue dynamics and tease apart the many nuances of the fates duplicated genes adopt. While themes emerge from studies such as this, these fates should not be thought of as strict evolutionary identities that duplicated genes assume in discrete classes. No two sets of paralogues share the same context, be this their unique combinations of age, origin, regulatory demands, and interacting partners. And so, it follows that no two sets of paralogues could ever be expected to share the same fate. It is with this in mind that we as evolutionary geneticists would do well to remember the words of the American poet Robert Frost in his poem 'The Road Not Taken'. Rather than restricting our understanding of paralogue evolution to just three arterial fates, it is incumbent on us to embrace, and seek to understand, the idiosyncratic paths trodden by duplicated genes that meet their own adaptive ends.

Appendices

Appendix I: OMA script for paralogue mining

Pairwise Comparisons Between *Pristionchus pacificus* and *Caenorhabditis* Species

Introduction
We want to use the OMA pairwise comparisons between *Pristionchus pacificus* and the *Caenorhabditis* species to identify genes that are present in multiple copies in some *Caenorhabditis* species but not in (some) others or *Pristionchus pacificus*.
We downloaded pairwise comparisons between *Pristionchus pacificus* and:

1. *C. brenneri*
2. *C. briggsae*
3. *C. japonica*
4. *C. remanei*
5. *C. elegans*

Process
First, we load the data:

```
# Load all 5 tables
pripa_bre_df <- read.table("/Users/Emily/Desktop/DPhil/2019-first-
rotation/analysis/2019-01-10_PRIPA_comparisons/PRIPA_CAEBE.txt",
                           header = FALSE,
                           sep    = "\t")

pripa_bri_df <- read.table("/Users/Emily/Desktop/DPhil/2019-first-
rotation/analysis/2019-01-10_PRIPA_comparisons/PRIPA_CAEBR.txt",
                           header = FALSE,
                           sep    = "\t")

pripa_ele_df <- read.table("/Users/Emily/Desktop/DPhil/2019-first-
rotation/analysis/2019-01-10_PRIPA_comparisons/PRIPA_CAEEL.txt",
                           header = FALSE,
                           sep    = "\t")

pripa_jap_df <- read.table("/Users/Emily/Desktop/DPhil/2019-first-
rotation/analysis/2019-01-10_PRIPA_comparisons/PRIPA_CAEJA.txt",
                           header = FALSE,
                           sep    = "\t")

pripa_rem_df <- read.table("/Users/Emily/Desktop/DPhil/2019-first-
rotation/analysis/2019-01-10_PRIPA_comparisons/PRIPA_CAERE.txt",
                           header = FALSE,
                           sep    = "\t")

# Change column name for all input data frames
table_colnames       <- c("pripa", "caen_genus", "type", "id")
colnames(pripa_bre_df) <- table_colnames
colnames(pripa_bri_df) <- table_colnames
colnames(pripa_ele_df) <- table_colnames
colnames(pripa_jap_df) <- table_colnames
```

```
colnames(pripa_rem_df) <- table_colnames
```

Then, we subset to the genes that have an expansion in members of the *Caenorhabditis* genus but not in *Pristionchus pacificus*.

```
pripa_vs_bre_expansion <- subset(pripa_bre_df, type == "1:m")
pripa_vs_bri_expansion <- subset(pripa_bri_df, type == "1:m")
pripa_vs_ele_expansion <- subset(pripa_ele_df, type == "1:m")
pripa_vs_jap_expansion <- subset(pripa_jap_df, type == "1:m")
pripa_vs_rem_expansion <- subset(pripa_rem_df, type == "1:m")

pripa_vs_bre_expansion_gene <- as.character(pripa_vs_bre_expansion$pripa)
pripa_vs_bri_expansion_gene <- as.character(pripa_vs_bri_expansion$pripa)
pripa_vs_ele_expansion_gene <- as.character(pripa_vs_ele_expansion$pripa)
pripa_vs_jap_expansion_gene <- as.character(pripa_vs_jap_expansion$pripa)
pripa_vs_rem_expansion_gene <- as.character(pripa_vs_rem_expansion$pripa)

# Comparing radiations in C. briggsae and C. brenneri (i.e. '1:m'
relative to P. pacificus in both species)
bre_bri_expansions_in_common <- pripa_vs_bre_expansion_gene %in%
pripa_vs_bri_expansion_gene
bre_bri_genes_in_common <-
pripa_vs_bre_expansion_gene[bre_bri_expansions_in_common == TRUE]


# Comparing radiations in C. briggsae and C. elegans (i.e. '1:m' relative
to P. pacificus in both species)
ele_bri_expansions_in_common <- pripa_vs_ele_expansion_gene %in%
pripa_vs_bri_expansion_gene
ele_bri_genes_in_common <-
pripa_vs_ele_expansion_gene[ele_bri_expansions_in_common == TRUE]


# Comparing radiations in C. briggsae and C. japonica (i.e. '1:m'
relative to P. pacificus in both species)
jap_bri_expansions_in_common <- pripa_vs_jap_expansion_gene %in%
pripa_vs_bri_expansion_gene
jap_bri_genes_in_common <-
pripa_vs_jap_expansion_gene[jap_bri_expansions_in_common == TRUE]


# Comparing radiations in C. briggsae and C. remanei (i.e. '1:m' relative
to P. pacificus in both species)
rem_bri_expansions_in_common <- pripa_vs_rem_expansion_gene %in%
pripa_vs_bri_expansion_gene
rem_bri_genes_in_common <-
pripa_vs_rem_expansion_gene[rem_bri_expansions_in_common == TRUE]
```

The code detailed above isolates genes which are present in multiple copies in any two *Caenorhabditis* species but are single copy in *Pristionchus pacificus*. Now let's check to see if any of them are the same:

```
pripa_ele_df[pripa_ele_df$pripa == "PRIPA08330", ]
##           pripa caen_genus type     id
## 3594 PRIPA08330 CAEEL12360  1:m     NA
```

```
## 3595 PRIPA08330 CAEEL12361  1:m 282274
pripa_bre_df[pripa_bre_df$pripa == "PRIPA08330", ]
## [1] pripa      caen_genus type       id
## <0 rows> (or 0-length row.names)
pripa_bri_df[pripa_bri_df$pripa == "PRIPA08330", ]
##           pripa caen_genus type id
## 2560 PRIPA08330 CAEBR09559  1:1 NA
pripa_jap_df[pripa_jap_df$pripa == "PRIPA08330", ]
## [1] pripa      caen_genus type       id
## <0 rows> (or 0-length row.names)
pripa_rem_df[pripa_rem_df$pripa == "PRIPA08330", ]
##           pripa caen_genus type id
## 3405 PRIPA08330 CAERE21043  1:1 NA
```

The above gene is actually present in the *C. elegans* genome three times yet on my database, it only appears twice. This highlights the main issue with OMA in that it classifies orthology rather mysteriously and sometimes incorrectly. Presumably asymmetric divergence (variable rates of paralogue evolution) is a cause of this problem. Let's try another:

```
pripa_ele_df[pripa_ele_df$pripa == "PRIPA01577", ]
##          pripa caen_genus type     id
## 640 PRIPA01577 CAEEL11122  1:m 514934
## 641 PRIPA01577 CAEEL18480  1:m     NA
pripa_bre_df[pripa_bre_df$pripa == "PRIPA01577", ]
##          pripa caen_genus type     id
## 456 PRIPA01577 CAEBE05963  1:m     NA
## 457 PRIPA01577 CAEBE16328  1:m 514934
pripa_bri_df[pripa_bri_df$pripa == "PRIPA01577", ]
##          pripa caen_genus type     id
## 436 PRIPA01577 CAEBR09942  1:m 514934
## 437 PRIPA01577 CAEBR18660  1:m     NA
pripa_jap_df[pripa_jap_df$pripa == "PRIPA01577", ]
##          pripa caen_genus type     id
## 390 PRIPA01577 CAEJA02988  1:m 514934
## 391 PRIPA01577 CAEJA09401  1:m     NA
pripa_rem_df[pripa_rem_df$pripa == "PRIPA01577", ]
##          pripa caen_genus type     id
## 530 PRIPA01577 CAERE00918  1:m     NA
## 531 PRIPA01577 CAERE04352  1:m 514934
## 532 PRIPA01577 CAERE13847  1:m     NA
```

Paralogy and orthology are predicted correctly in this example.

Appendix II: Strain list

Strain list	Allele / mutation / isolate / background	Extrachromosomal array(s)/ integrants in bold
N2	Bristol wild-type	
PB2801	*C. brenneri*	
LP162	N2	*cp13[nmy-2::gfp + LoxP] I*
CB1489	*him-8(e1489) IV*	
CB4088	*him-5(e1490) V*	
pAW1000	*unc-119 (ed3) III; him-8 (e1489) IV*	*ouEx642 [pAW638+unc-119^{+}]*
pAW1001	*unc-119(ed3) III*	*unc-119*$^{+}$ *tbx-35::gfp*
pAW1002	*unc-119(ed3) III*	*unc-119*$^{+}$ *tbx-36::gfp (construct 1)*
pAW1003	N2	*[nmy-2::gfp + LoxP] I [unc-54p::GFP::H2B] IV*
ECA1191	Hawaiian wild-type	
ECA1185	Hawaiian wild-type	
LKC34	Madagascan wild-type	
pAW1004	ECA1191	*myo-2::mCherry tbx-35::gfp (ECA1191)*
pAW1005	ECA1185	*myo-2::mCherry tbx-35::gfp (ECA1185)*
pAW1006	ECA1191	*myo-2::mCherry tbx-36::gfp (ECA1191)*
pAW1007	ECA1185	*myo-2::mCherry tbx-36::gfp (ECA1185)*
pAW1008	LKC34	*myo-2::mCherry tbx-36::gfp (LKC34)*
pAW1009	*wrt-1(tm1417) II* (outcrossed x 3)	
pAW1011	*wrt-2(ok2810) X* (outcrossed x 4)	
pAW1012	*wrt-3(ok2608) II* (outcrossed x 3)	
pAW1013	*wrt-4(tm1911) X* (outcrossed x 3)	
pAW1014	*wrt-5(ok670) IV* (outcrossed x 3)	
pAW1015	*wrt-7(ok3210) V* (outcrossed x 3)	

Strain list	Allele / mutation / isolate / background	Extrachromosomal array(s)/ integrants in bold
pAW1016	*wrt-8(1585) V* (outcrossed x 3)	
pAW1017	*wrt-9(ok2732) X* (outcrossed x 4)	
pAW60	*wIs78 [ajm-1::gfp + unc-119*$^{+}$*] IV; him-5(e1490) V*	***unc-119***$^{+}$ ***ajm-1::gfp***
pAW718	*rrf-3(pk1426)II; wIs51 [scm::gfp + unc- 119*$^{+}$*] V*	*unc-119*$^{+}$
pAW1019	*unc-119(ed3) III*	*unc-119*$^{+}$ *tbx-36::gfp (construct 2: no 213 bp repressor)*
pAW1020	*unc-119(ed3) III*	*unc-119*$^{+}$ ΔE2F *tbx-36::gfp (no 28 bp E2F site or 213 bp repressor)*
pAW1021	*unc-119(ed3) III*	*unc-119*$^{+}$ *tbx-36::gfp (construct 2: no 213 bp repressor)*
pAW1022	*unc-119 (ed3) III; him-8 (e1489) IV*	*unc-119*$^{+}$ *drd-1.1p::gfp*
pAW1023	*unc-119 (ed3) III; him-8 (e1489) IV*	*unc-119*$^{+}$ *drd-1.2p::gfp*
pAW1024	*unc-119 (ed3) III; him-8 (e1489) IV*	*unc-119*$^{+}$ *drd-1.3p::gfp*
pAW1025	*unc-119 (ed3) III; him-8 (e1489) IV*	*unc-119*$^{+}$ *Ppdrdp::gfp*
PS312	*P. pacificus*	
pAW1026	*P. pacificus*, PS312	*myo-2::mCherry Ppdrdp::gfp*
BA17	*fem-1(hc17) IV*	
WB1141	N2	***wbmIs66 [rab-3p::3XFLAG::dpy-10 crRNA::rab-3 3′UTR] IV***
pAW1027	*unc-119(ed3) III*	*unc-119*$^{+}$ *drd-1.1p::gfp*
pAW1028	*unc-119(ed3) III*	*unc-119*$^{+}$ *drd-1.2p::gfp*

Strain list	Allele / mutation / isolate / background	Extrachromosomal array(s)/ integrants in bold
pAW1029	*unc-119(ed3) III*	*unc-119*$^{+}$ *drd-1.3p::gfp*
pAW1030	*unc-119(ed3) III*	*unc-119*$^{+}$ *Ppdrdp::gfp*
pAW1031	*myrf-1(ybq6)* (outcrossed x 4)	
pAW1032	*myrf-2(ybq42)* (outcrossed x 4)	
pAW1033	*unc-119(ed3) III*	*unc-119*$^{+}$ *myrf-1p::gfp*
pAW1034	*unc-119(ed3) III*	*unc-119*$^{+}$ *myrf-2p::gfp*
pAW1035	*myrf-1(ybq6) II; unc-119(ed3) III; myrf-2(ybq42) X*	
pAW1036	*myrf-1(ybq6) II; unc-119(ed3) III; myrf-2(ybq42) X*	*unc-119*$^{+}$ *ouEx900[myrf-2.1::mCherry]*
pAW1037	*myrf-1(ybq6) II; unc-119(ed3) III; myrf-2(ybq42) X*	*unc-119*$^{+}$ *ouEx901[myrf-2.2::gfp]*
pAW1037	*myrf-1(ybq6) II; unc-119(ed3) III; myrf-2(ybq42) X*	*unc-119*$^{+}$ *ouEx900[myrf-2.1::mCherry]* *ouEx901[myrf-2.2::gfp]*

Appendix III: Maximum Likelihood IQ-Tree of the T-box family

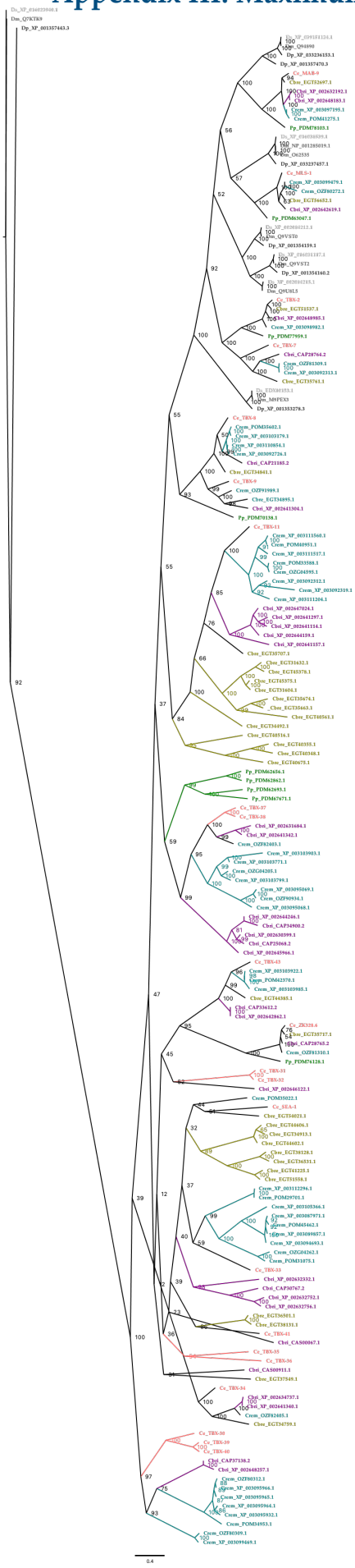


Figure 8.1. T-box family phylogenetic analysis. Maximum likelihood phylogram of all T-box domain (predicted) proteins from the genomes of: *C. elegans* (Ce, pink), *C. brenneri* (Cbre, olive green), *C. remanei* (Crem, teal), *C. briggsae* (Cbri, purple) with a *P. pacificus* (Pp, bright green) outgroup, and further outgroups from the *Drosophila* genus: *D. melanogaster* (Dm), *D. pseudoobscura* (Dp), and *D. simulans* (Ds), in various shades of grey. Scale bar is substitutions per site per million years. Node labels are bootstrap support values missing from the same tree (circularised for clarity) in Figure 3.1.

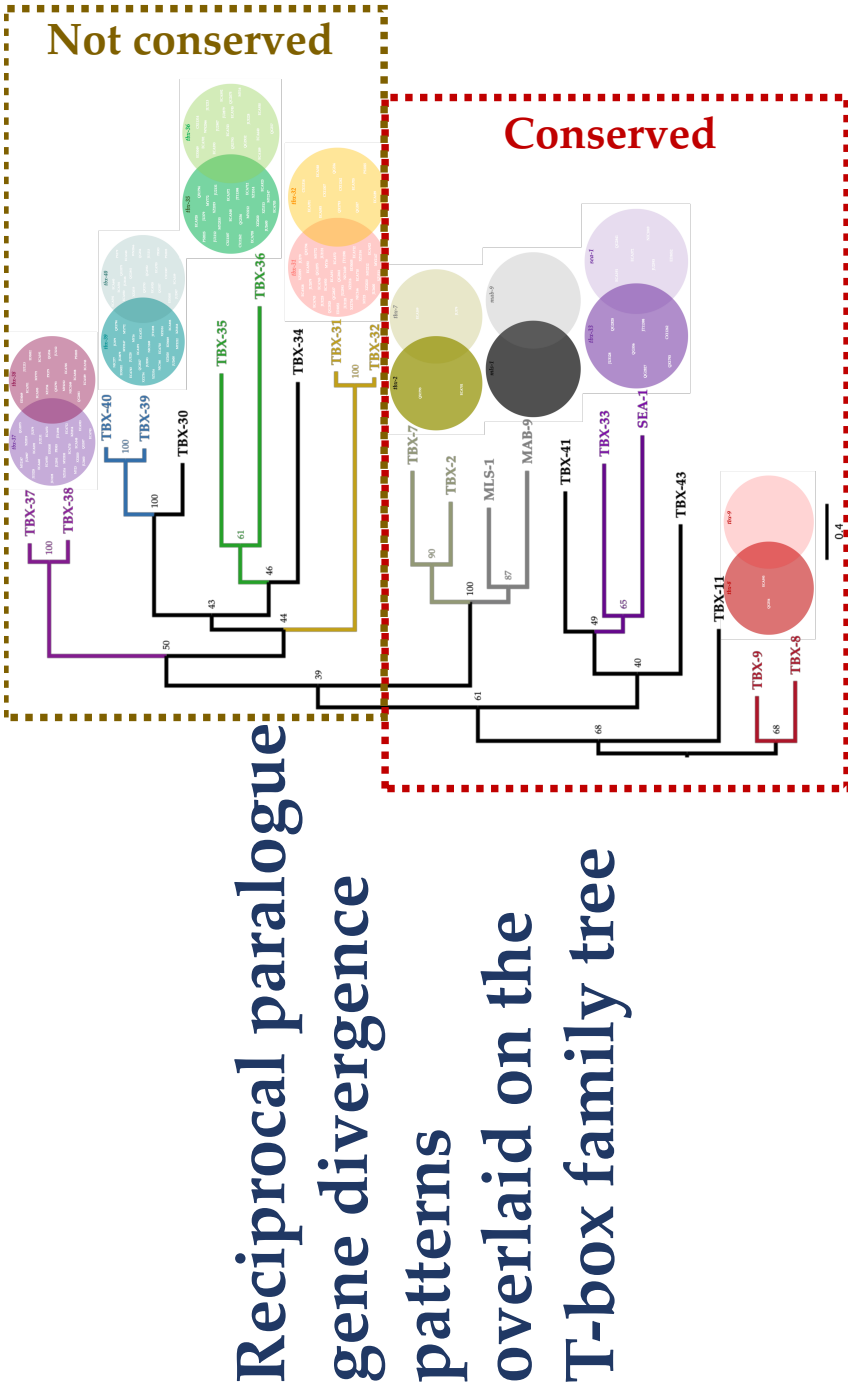
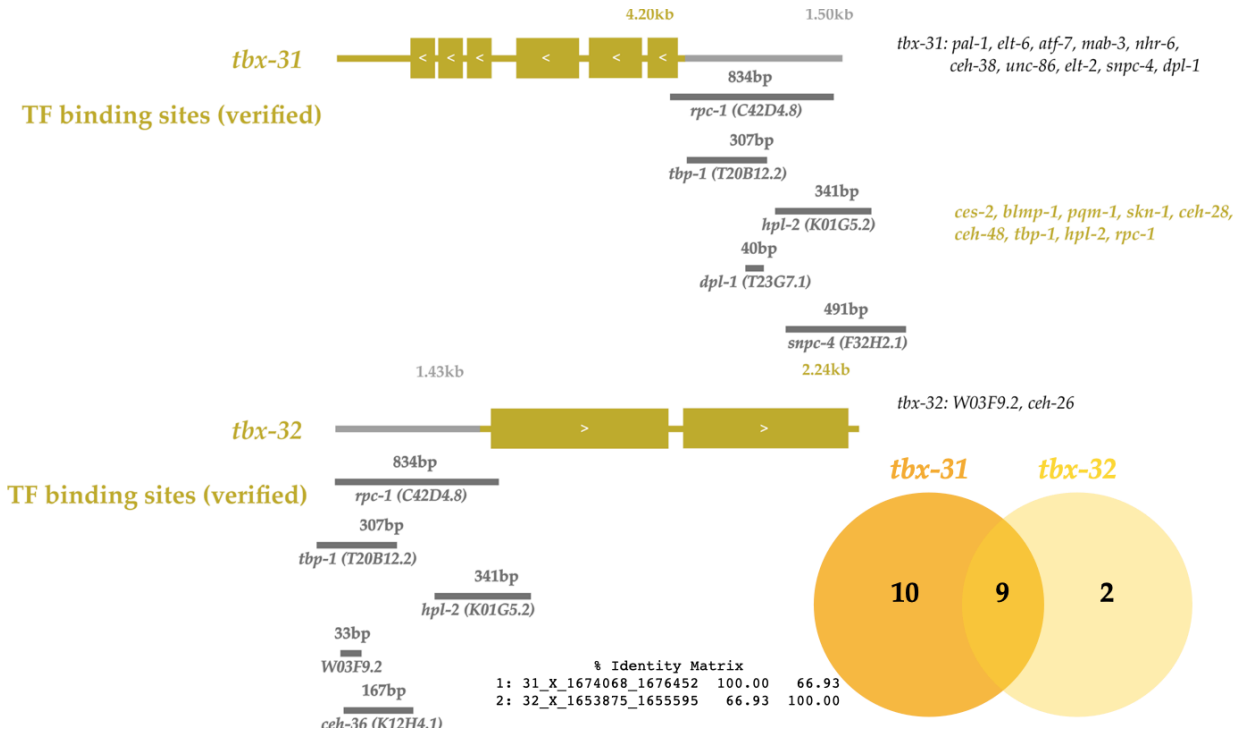
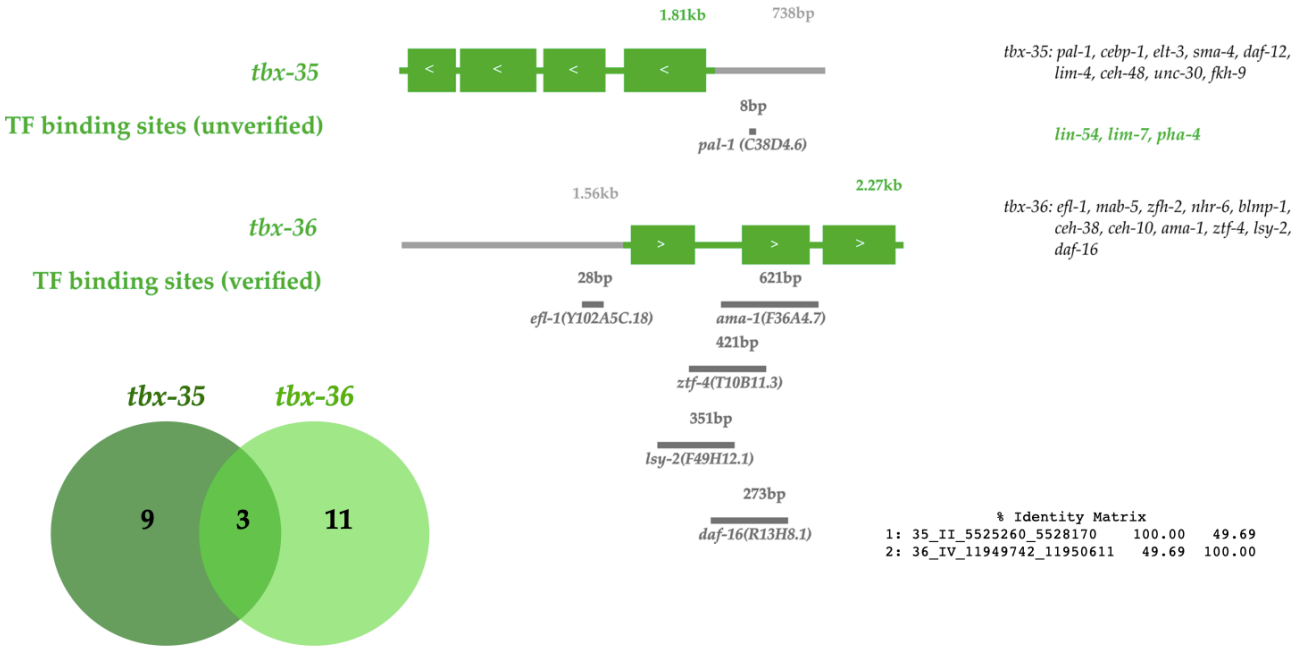


Figure 8.2. Divergence in the T-box repertoire of *C. elegans*: HIGH and MODERATE effect mutation accumulation among wild populations. Gene tree of *C. elegans* T-box domains beside which are Venn diagrams of the respective T-box paralogue pair showing instances of High and Moderate effect mutation accumulation in the pairs among wild isolates. High effect mutations are defined as: premature stop codons; missing start codons; frameshifts (both out-of-frame insertions and deletions); in-frame insertions and deletions; and substitutions which result in a change of

amino acid to one of a different class. Moderate effect mutations include, but are not limited to: any other missense mutations (non-synonymous); and small in-frame deletions. A total of 304 wild isolates were analysed in this experiment — the full library which had a fully sequenced genome which was publicly made available on CeNDR. All isolates which are not featured in this extended supplementary version may be taken to contain intact wildtype coding sequences of both paralogues. This is the majority of the 304 isolates known and characterised. They are not included in the diagram outside of the Venn diagram for clarity. Orange box highlights the *C. elegans* -specific T-box genes, and the pink box covers the more widely conserved T-box genes (among nematodes and other animal groups).

Appendix V: T-box transcription factor binding profile comparisons





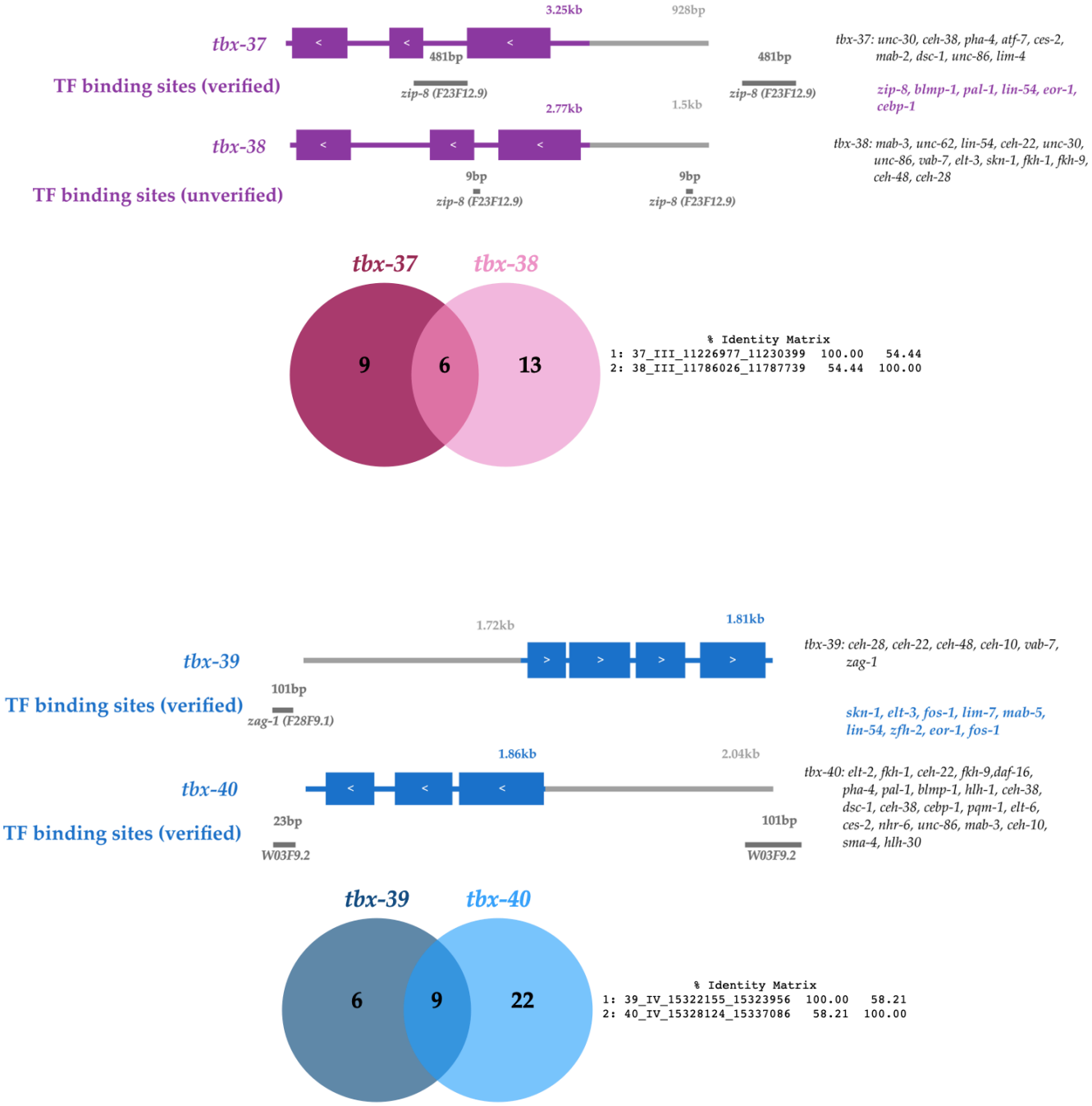


Figure 8.3. T-box transcription factor binding profile comparisons. TF binding site predictions are mapped onto the *C. elegans* specific T-box genes. Binding sites are predicted using JASPAR and, where possible, were verified using modENCODE ChIP-seq datasets — the precise methodology for which can be found in Chapter 2. Summary Venn diagrams of the number of TF binding sits shared between the paralogue pairs are shown below the gene schematics. % sequence similarity is also provided between the non-coding regions (promoters and introns and UTRs) of each of the paralogue pairs, provided in a comparative Identity Matrix for each paralogue pair. From the top: *tbx-31*/*tbx-32* (orange); *tbx-35*/*tbx-36* (green); *tbx-37*/*tbx-38* (purple); and *tbx-39*/*tbx-40* (blue).

Appendix VI: Maximum Likelihood IQ-Tree of the Warthog family

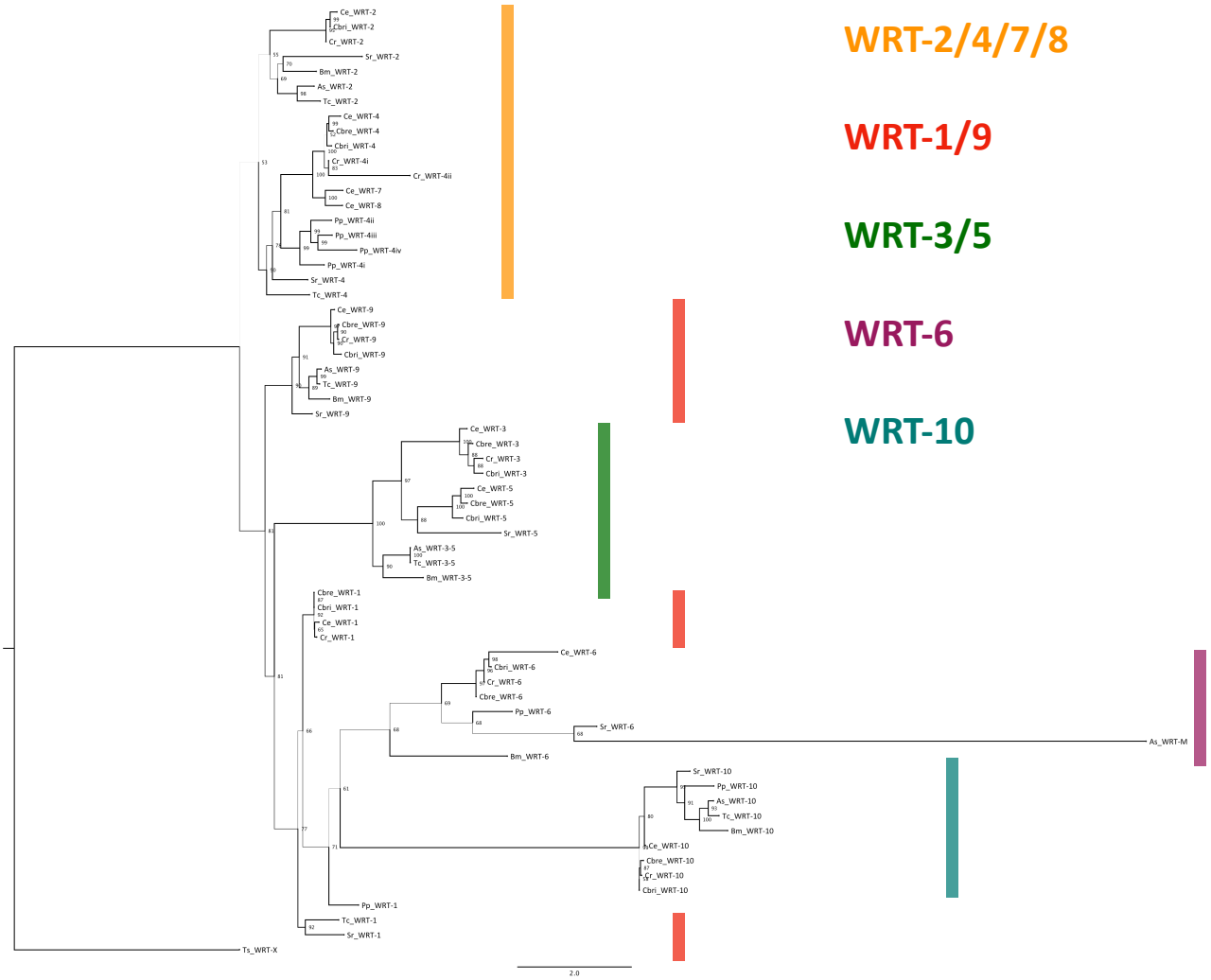


Figure 8.4. Maximum likelihood phylogenetic analysis of the Warthog family in the nematode phylum. IQ-TREE maximum likelihood molecular phylogenetic analysis of the Wart domain sequences mined from selected nematode genomes (*Caenorhabditis briggsae* (Cbri), *Caenorhabditis remanei* (Cr), *Caenorhabditis brenneri* (Cbre), *Caenorhabditis elegans* (Ce), *Pristionchus pacificus* (Pp), *Strongyloides ratti* (Sr), *Toxocara canis* (Tc), *Ascaris suum* (As), *Brugia malayi* (Bm), *Trichinella spiralis* (Ts)). The node labels are ultrafast bootstrap support values. The tree was generated in FigTree.

Appendix VII: Updated Warthog nomenclature

	Warthog identity	Accession number
Trichinella spiralis	***wrt-a***	**KRY35028.1**
Pristionchus pacificus	***wrt-2* -like***	**PDM64744.1**
	wrt-4.1	**PDM79116.1**
	wrt-4.2	**PDM65807.1**
	wrt-4.3	**PDM72491.1**
	wrt-4.4	**PDM69003.1**
	wrt-4* -like	**PDM60764.1**
	wrt-4* -like	**PDM75978.1**
	wrt-4* -like	**PDM80143.1**
	wrt-4* -like	**PDM65807.1**
	wrt-5* -like	**KKA76232.1**
	wrt-9* -like	**PDM64744.1**
Caenorhabditis brenneri	***wrt-2* -like***	**EGT29941.1**
Caenorhabditis remanei	***wrt-5* -like***	**XP_003096339.1**
	wrt-2* -like	**OZG06368.1**

Table 8.1. Table lists the accession numbers of the additional (degenerate) Wart domain containing sequences mined from the nematode species in this investigation. Where legitimate Wart domain containing genes are included, these have been named by our investigation according to our findings, i.e. the previously unannotated four *wrt-4* paralogues in *P. pacificus* and the *wrt-a* orthologue in *Trichinella spiralis* which was previously unannotated in the genome and is thought to be the extant representative of the ancestral Warthog family member.

Appendix VIII: Warthog microsynteny analysis

Species	5' Neighbour	Warthog orthologue	3' Neighbour	Comment
C. elegans	ZK1290.5.1	*wrt-1*	*wrt-10*	
	pccb-1	*wrt-2*	*sec-3*	
	F38E11.6	*wrt-3*	*cng-3*	
	srxa-9	*wrt-4*	*bus-17*	
	pcn-1	*wrt-5*	*ilys-6*	
	sax-3	*wrt-6*	C12D12.3	
	srw-113	*wrt-7*	*clec-228*	Loci (adjacent to one another) map to *Cbri_wrt-4* (*C. briggsae*)
	clec-229	*wrt-8*	C29F3.3	
	H02F09.2	*wrt-9*	H11E01.3	
	wrt-1	*wrt-10*	ZK1290.13	
C. brenneri	Cbr12985	*wrt-1*	*wrt-10*	
	Cbn-*pccb-1*	*wrt-2*	Cbn-*sec-3*	*Degenerate (multiple cysteines missing)*
	Cbn25685	*wrt-3*	Cbn31929	
	Cbn-*srxa-9*	*wrt-4*	Cbn-*bus-17*	
	Cbn-*pcn-1*	*wrt-5*	Cbn08017	
	Cbn-*sax-3*	*wrt-6*	Cbn16646	
	Cbn22677	*wrt-9*	-	
	wrt-1	*wrt-10*	Cbr13535	
C. remanei	Cre26393	*wrt-1*	Cre26108	
	Cre-*pccb-1*	*wrt-2*	Cre-*sec-3*	
	Cre03559	*wrt-3*	Cre-*cng-3*	
	Cre-*srxa-9*	*wrt-4*	Cre-*bus-17*	
	Cre-*pcn-1*	*wrt-5*	-	*Degenerate (multiple cysteines missing)*
	Cre-*sax-3*	*wrt-6*	Cre00890	
	Cre24213	*wrt-9*	Cre24215	
	wrt-1	*wrt-10*	Cre26394	

Species	5' Neighbour	Warthog orthologue	3' Neighbour	Comment
C. briggsae	Cbg12985	*wrt-1*	*wrt-10*	
	Cbg-*pccb-1*	*wrt-2*	Cbg-*sec-3*	
	Cbg21670	*wrt-3*	Cbg25106	
	Cbg-*srxa-9*	*wrt-4*	Cbg-*bus-17*	
	Cbr-*pcn-1*	*wrt-5*	Cbg00237	
	Cbr-*sax-3*	*wrt-6*	Cbg14223	
	Cbg27396	*wrt-9*	Cbg16422	
	wrt-1	*wrt-10*	Cbg25541	
P. pacificus	PPA18368	*wrt-1*	*wrt-10*	
	Pp-*pccb-1*	*wrt-2*	PPA23764	
	PPA17986	*wrt-4iii*	PPA17802	
	PPA06413	*wrt-3-5*	PPA0611	*Degenerate (multiple cysteines missing)*
	PPA31587	*wrt-6*	PPA17221	
	PPA37048	*wrt-9*	PPA00251	*Degenerate (multiple cysteines missing)*
	wrt-1	*wrt-10*	PPA18371	
S. ratti	SRAE_X000259200	*wrt-1*	SRAE_X000259400	
	SRAE_1000112500	*wrt-2*	SRAE_1000112800	
	SRAE_1000168901	*wrt-4*	SRAE_1000169201	
	SRAE_X000039300	*wrt-3-5*	SRAE_X000039500	
	SRAE_X000100400	*wrt-6*	SRAE_X000100600	
	SRAE_2000396500	*wrt-9*	SRAE_2000396700	
	SRAE_X000005500	*wrt-10*	SRAE_X000005700	
A. suum	ASU_02743	*wrt-2*	ASU_02745	
	ASU_01525	*wrt-3-5*	ASU_01527	
	ASU_03946	*wrt-9*	ASU_03948	
	ASU_09186	*wrt-10*	ASU_09188	
T. canis	-	*wrt-2*	-	gene desert

Species	5' Neighbour	Warthog orthologue	3' Neighbour	Comment
	Tcan_04003.1	*wrt-4*	Tcan_04021.1	
	-	*wrt-3-5*	-	gene desert
	Tcan_18250.1	*wrt-9*	Tcan_18243.1	
	Tcan_14303.1	*wrt-10*	Tcan_14311.1	
B. malayi	Bm8326	*wrt-2*	Bm18002	
	-	*wrt-3-5*	-	gene desert
	Bm11541	*wrt-6*	Bm18141	
	-	*wrt-10*	-	gene desert
T. spiralis	T12_9139	*wrt-X*	T12_9141	

Table 8.2. Table details the genes directly adjacent to each Warthog family orthologue in a particular nematode species. Purple font indicates that the adjacent gene is not orthologous to the adjacent gene beside the orthologous Warthog in *C. elegans*. Green font indicates that the adjacent gene is orthologous to the adjacent gene beside the orthologous Warthog in *C. elegans*. '-' indicates that there was a 'gene desert' adjacent to the Warthog locus or that the contig finished and so syntenic relationships could not be established. 'Gene deserts' are arbitrarily defined here as regions on a chromosome that do not feature any open reading frames for over 1.5 kb. All genes are represented with their WormBase/NCBI accession numbers.

Appendix IX: Absence of defects in *wrt-7* animals

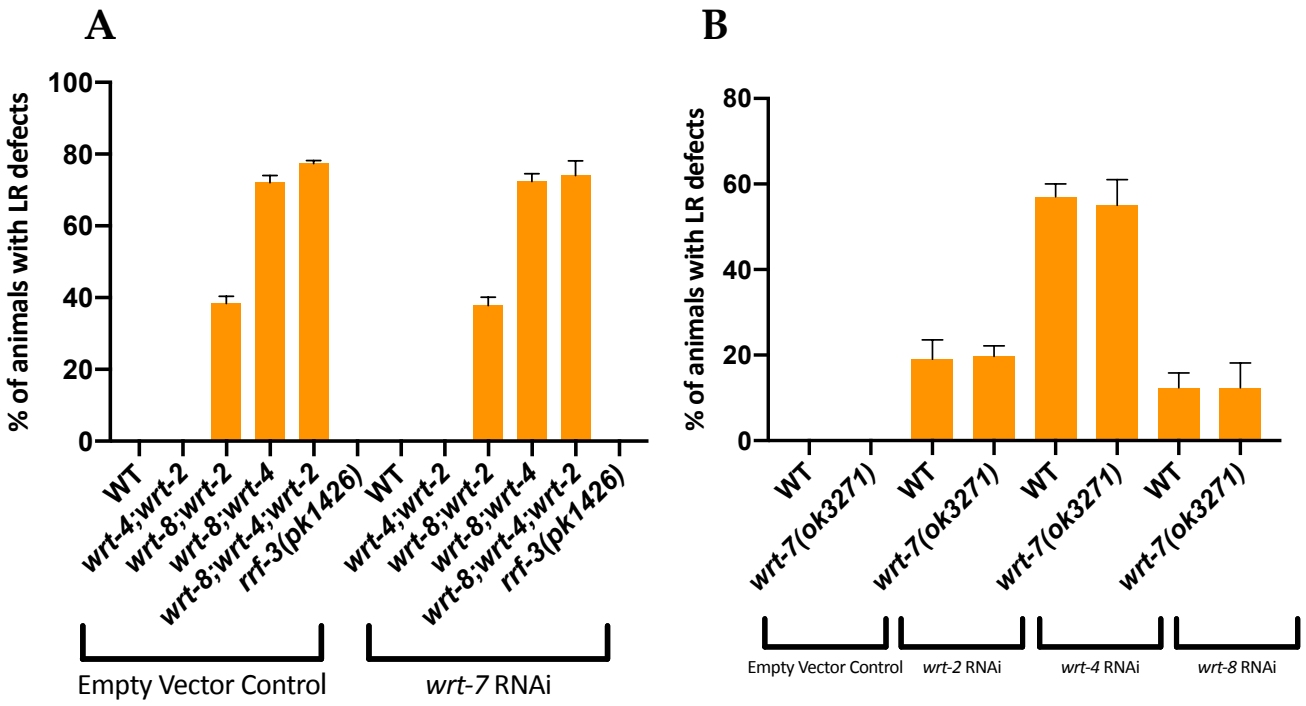


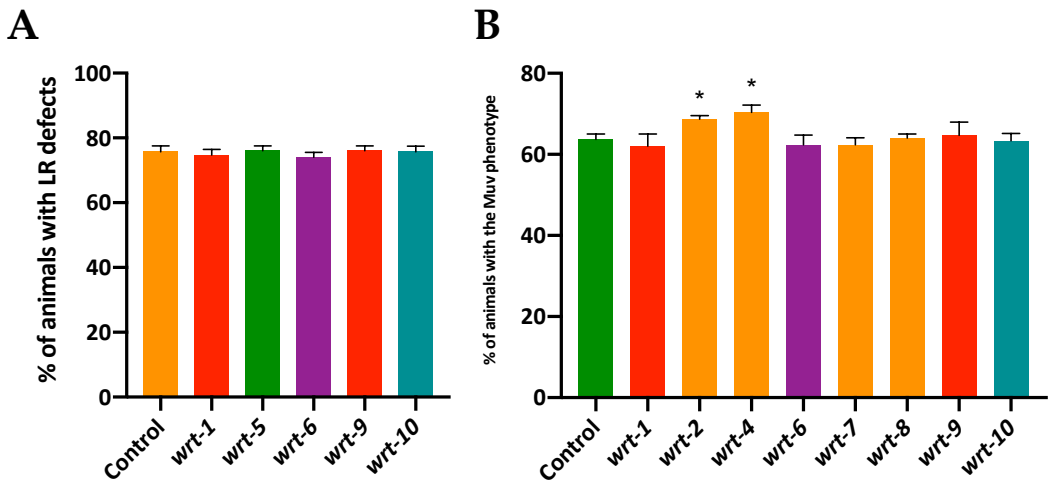
Figure 8.5. Absence of defects in *wrt-7* knockdown or knockout animals. (A) *wrt-7* RNAi knockdown performed on different genotypes. From the left, empty vector control RNAi performed on wildtype, *wrt-4(tm1911);wrt-2(ok2810)*, *wrt-8(ok1585);wrt-2(ok2810)*, *wrt-8(ok1585);wrt-4(tm1911)*, *wrt-8(ok1585);wrt-4(tm1911);wrt-2(ok2810)* and *rrf-3(pk1426)* animals; and *wrt-7* RNAi performed on wildtype, *wrt-4(tm1911);wrt-2(ok2810)*, *wrt-8(ok1585);wrt-2(ok2810)*, *wrt-8(ok1585);wrt-4(tm1911)*, *wrt-8(ok1585);wrt-4(tm1911);wrt-2(ok2810)* and *rrf-3(pk1426)* animals. **(B)** Wrt-2/4/8 knockdowns on WT and *wrt-7(ok3271)* animals. From the left, empty vector control RNAi performed on wildtype and *wrt-7(ok3271)* mutants; *wrt-2* RNAi performed on wildtype and *wrt-7(ok3271)* mutants; *wrt-4* RNAi performed on wildtype and *wrt-7(ok3271)* mutants; and *wrt-8* RNAi performed on wildtype and *wrt-7(ok3271)* mutants. Black bars show mean + SEM. All comparisons are not significant.

Appendix X: Moderate effect mutations in *wrt-7* among wild populations

CHROM	POS	REF	ALT	FILTER	AF	allele	effect	impact	exon_intron_rank	nt_change	aa_change	protein_position	distance_to_feature
V	15334821	A	C	PASS	0.018	C	missense_variant	MODERATE	10/10	c.1407T>G	p.Ile469Met	1407/1446	469/481
V	15334852	C	T	PASS	0.002	T	missense_variant	MODERATE	10/10	c.1376G>A	p.Arg459Lys	1376/1446	459/481
V	15334864	T	A	PASS	0.018	A	missense_variant	MODERATE	10/10	c.1364A>T	p.Tyr455Phe	1364/1446	455/481
V	15334867	C	T	PASS	0.018	T	missense_variant	MODERATE	10/10	c.1361G>A	p.Ser454Asn	1361/1446	454/481
V	15334883	T	C	PASS	0.015	C	missense_variant	MODERATE	10/10	c.1345A>G	p.Thr449Ala	1345/1446	449/481
V	15334884	G	C	PASS	0.006	C	missense_variant	MODERATE	10/10	c.1344C>G	p.Ile448Met	1344/1446	448/481
V	15335248	T	A	PASS	0.018	A	missense_variant	MODERATE	9/10	c.1217A>T	p.Tyr406Phe	1217/1446	406/481
V	15335267	C	T	PASS	0.018	T	missense_variant	MODERATE	9/10	c.1198G>A	p.Val400Ile	1198/1446	400/481
V	15335269	C	T	PASS	0.012	T	missense_variant	MODERATE	9/10	c.1196G>A	p.Arg399Lys	1196/1446	399/481
V	15335276	T	C	PASS	0.073	C	missense_variant	MODERATE	9/10	c.1189A>G	p.Ile397Val	1189/1446	397/481
V	15335285	T	C	PASS	0.009	C	missense_variant	MODERATE	9/10	c.1180A>G	p.Ile394Val	1180/1446	394/481
V	15335290	A	G	PASS	0.003	G	missense_variant	MODERATE	9/10	c.1175T>C	p.Val392Ala	1175/1446	392/481
V	15335305	G	A	PASS	0.006	A	missense_variant	MODERATE	9/10	c.1160C>T	p.Thr387Ile	1160/1446	387/481
V	15335306	T	C	PASS	0.006	C	missense_variant	MODERATE	9/10	c.1159A>G	p.Thr387Ala	1159/1446	387/481
V	15335308	A	G	PASS	0.006	G	missense_variant	MODERATE	9/10	c.1157T>C	p.Leu386Ser	1157/1446	386/481
V	15335360	A	T,G	PASS	0.012	T	missense_variant	MODERATE	8/10	c.1147T>A	p.Ser383Thr	1147/1446	383/481
V	15335384	C	T	PASS	0.024	T	missense_variant	MODERATE	8/10	c.1123G>A	p.Gly375Arg	1123/1446	375/481
V	15335386	G	A	PASS	0.024	A	missense_variant	MODERATE	8/10	c.1121C>T	p.Ala374Val	1121/1446	374/481
V	15335390	T	G	PASS	0.024	G	missense_variant	MODERATE	8/10	c.1117A>C	p.Asn373His	1117/1446	373/481
V	15335411	A	T	PASS	0.003	T	missense_variant	MODERATE	8/10	c.1096T>A	p.Ser366Thr	1096/1446	366/481
V	15335425	T	G	PASS	0.021	G	missense_variant	MODERATE	8/10	c.1082A>C	p.Lys361Thr	1082/1446	361/481
V	15335440	T	C	PASS	0.009	C	missense_variant	MODERATE	8/10	c.1067A>G	p.Asn356Ser	1067/1446	356/481
V	15335459	T	C	PASS	0.021	C	missense_variant	MODERATE	8/10	c.1048A>G	p.Thr350Ala	1048/1446	350/481
V	15335695	T	A	PASS	0.003	A	missense_variant	MODERATE	7/10	c.977A>T	p.Glu326Val	977/1446	326/481
V	15335747	C	T	PASS	0.003	T	missense_variant	MODERATE	7/10	c.925G>A	p.Asp309Asn	925/1446	309/481
V	15335752	C	T	PASS	0.003	T	missense_variant	MODERATE	7/10	c.920G>A	p.Gly307Glu	920/1446	307/481
V	15335757	T	G	PASS	0.003	G	missense_variant	MODERATE	7/10	c.915A>C	p.Glu305Asp	915/1446	305/481
V	15335765	G	T	PASS	0.021	T	missense_variant	MODERATE	7/10	c.907C>A	p.Leu303Met	907/1446	303/481
V	15335809	A	G	PASS	0.003	G	missense_variant	MODERATE	7/10	c.863T>C	p.Val288Ala	863/1446	288/481
V	15335810	C	T	PASS	0.003	T	missense_variant	MODERATE	7/10	c.862G>A	p.Val288Ile	862/1446	288/481
V	15335812	G	C	PASS	0.021	C	missense_variant	MODERATE	7/10	c.860C>G	p.Ala287Gly	860/1446	287/481
V	15335843	T	A	PASS	0.021	A	missense_variant	MODERATE	7/10	c.829A>T	p.Asn277Tyr	829/1446	277/481
V	15336037	A	G	PASS	0.021	G	missense_variant	MODERATE	5/10	c.731T>C	p.Val244Ala	731/1446	244/481
V	15336098	T	C	PASS	0.003	C	missense_variant	MODERATE	5/10	c.670A>G	p.Lys224Glu	670/1446	224/481
V	15336219	G	A	PASS	0.018	A	missense_variant	MODERATE	4/10	c.598C>T	p.Leu200Phe	598/1446	200/481
V	15336234	C	T	PASS	0.018	T	missense_variant	MODERATE	4/10	c.583G>A	p.Val195Ile	583/1446	195/481
V	15336309	C	T	PASS	0.003	T	missense_variant	MODERATE	4/10	c.508G>A	p.Gly170Ser	508/1446	170/481
V	15336314	T	A	PASS	0.003	A	missense_variant	MODERATE	4/10	c.503A>T	p.Lys168Met	503/1446	168/481
V	15336320	A	T	PASS	0.003	T	missense_variant	MODERATE	4/10	c.497T>A	p.Val166Asp	497/1446	166/481
V	15336513	A	T	PASS	0.018	T	missense_variant	MODERATE	3/10	c.381T>A	p.Asp127Glu	381/1446	127/481
V	15336547	A	C	PASS	0.021	C	missense_variant	MODERATE	3/10	c.347T>G	p.Ile116Ser	347/1446	116/481
V	15336959	T	C	PASS	0.003	C	missense_variant	MODERATE	2/10	c.298A>G	p.Asn100Asp	298/1446	100/481
V	15337125	T	A	PASS	0.003	A	missense_variant	MODERATE	2/10	c.132A>T	p.Leu44Phe	132/1446	44/481
V	15337130	C	T	PASS	0.015	T	missense_variant	MODERATE	2/10	c.127G>A	p.Val43Ile	127/1446	43/481
V	15337145	G	A	PASS	0.021	A	missense_variant	MODERATE	2/10	c.112C>T	p.Pro38Ser	112/1446	38/481
V	15337263	G	C	PASS	0.021	C	missense_variant	MODERATE	1/10	c.43C>G	p.Leu15Val	43/1446	15/481
V	15337275	C	A	PASS	0.018	A	missense_variant	MODERATE	1/10	c.31G>T	p.Ala11Ser	31/1446	11/481
V	15337304	AT	A	PASS	0.018	A	frameshift_variant&start_lost	HIGH	1/10	c.1delA	p.Met1fs	1/1446	1/481

Table 8.3. Table details the highly polymorphic *wrt-7* alleles (naturally occurring variations) found in wild isolates of *C. elegans* in the open reading frame of *wrt-7*. The precise mutational change (and corresponding amino acid change) is listed according to the reference (Bristol N2) and the corresponding change and the effect on the predicted protein is given in the eighth and ninth columns, respectively.

Appendix XI: Interclade RNAi of the Warthog family



C

RNAi	Body size 48 hours post L4 (mm)	n number	*P* value [α]
Empty Vector Control	1.201	27	
wrt-2	1.198	23	ns
wrt-3	1.218	24	0.0220
wrt-4	1.999	25	ns
wrt-5	1.209	24	ns
wrt-6	1.203	19	ns
wrt-7	1.198	22	ns
wrt-8	1.200	24	ns
wrt-10	1.192	20	ns

[α] Unpaired t tests comparing the mean body lengths between Empty Vector Control and Wrt gene RNAi on *wrt-1(tm1417);wrt-9(ok2732)* double mutant animals

Figure 8.6. Interclade RNAi in the Warthog reveals no additional functional redundancy. (A) % penetrance of LR defects recorded on specific Wrt family RNAi knockdown on *wrt-8(tm1585);wrt-4(tm1911);wrt-2(ok2810)* triple mutants (Empty Vector Control, n = 20; *wrt-1*, n = 23; *wrt-5*, n = 20; *wrt-6*, n = 19; *wrt-9*, n = 25; *wrt-10*, n = 20). *wrt-3* knockdown was not performed due to the severe middle body morphological defects it gives rise to (Additional File 1; Supplementary Figure 6 in Baker et al. 2021). (B) % penetrance of Muv defects recorded on specific Wrt family RNAi knockdown on *wrt-5(ok670);wrt-3(ok2608)* double mutants. (Empty Vector Control, n = 19; *wrt-1*, n = 19; *wrt-2*, n = 21; *wrt-4*, n = 24; *wrt-6*, n = 22; *wrt-7*, n = 26; *wrt-8*, n = 20; *wrt-9*, n = 29; *wrt-10*, n = 20) (C) Body size (48 h post L4) of *wrt-3(ok2608);wrt-9(ok2732)* double mutants upon knockdown of specific Wrt family members. Black bars show mean + SEM. Black asterisks show *P ≤ 0.05, all other comparisons are not statistically significant, i.e. P > 0.05.

Appendix XII: RNAi knockdown of *myrf-1* and *myrf-2*

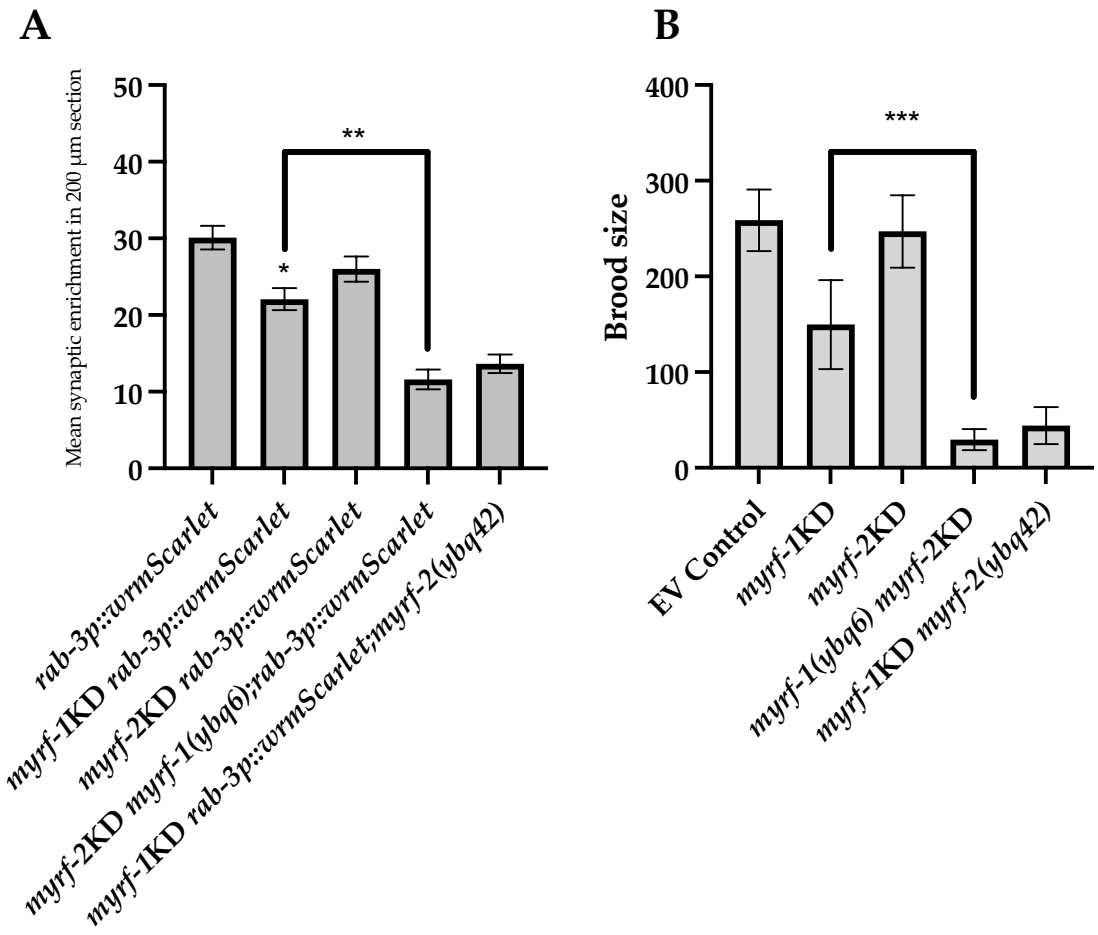


Figure 8.7. *myrf-1* and *myrf-2* RNAi knockdown recapitulates knockout phenotypes. The RNAi knockdown of either *myrf-1* or *myrf-2* was performed on an N2 wildtype background or *myrf-1(ybq6)* or *myrf-2(ybq42)* single mutants. RNAi knockdown was performed by feeding and no double knockdown was attempted. (A) Quantifies the synaptic enrichment in a 200 µm stretch in *myrf-1* and *myrf-2* single knockdowns either in a wildtype background, or in a *myrf-1(ybq6)* mutant background (for *myrf-2* RNAi), or a *myrf-2(ybq42)* mutant background (for *myrf-1* RNAi). As elsewhere in Chapter 6, mean synaptic enrichment was quantified in animals at the mid-L2 stage. (B) Quantifies the mean brood sizes following *myrf-1* and *myrf-2* single RNAi knockdown either in a wildtype background, or in a *myrf-1(ybq6)* mutant background (for *myrf-2* RNAi), or a *myrf-2(ybq42)* mutant background (for *myrf-1* RNAi). All progeny scored were part of the brood that survived to hatching and beyond L1 such that these can be considered viable broods, i.e., dead eggs and dead L1s were removed from the counts. Broods were scored in triplicate and an average was taken for each. All broods were scored at 20 °C and all other variables were kept constant (e.g., plentiful supply of food). Black bars show mean + SEM. Black asterisks (****P ≤ 0.0001, ***P ≤ 0.001, **P ≤ 0.01, *P ≤ 0.05, nsP > 0.05) show statistically significant differences in the means compared to mutants with an unpaired t test.

REFERENCES

- A. Hao, L. Aspock, G. Burglin, T.R. The hedgehog-related gene wrt-5 is essential for hypodermal development in Caenorhabditis elegans. Dev. Biol. 2006. 290:323–336.
- Alcorn MR, Callander DC, Lo´pez-Santos A, Torres Cleuren YN, Birsoy B, Joshi PM, Santure AW, Rothman JH. 2016 Heterotaxy in Caenorhabditis: widespread natural variation in left – right arrangement of the major organs. Phil. Trans. R. Soc. B. 371, 20150404. (doi:10.1098/rstb.2015.0404)
- B. Hao, L., Johnsen, R., Lauter, G., Baillie, D. & Burglin, T. R. Comprehensive analysis of gene expression patterns of hedgehog-related genes. BMC Genomics. 2006. 7:280.
- Brenner, S. The genetics of Caenorhabditis elegans. Genetics. 1974. 77:71–94
- C. Hao, L. K., Mukherjee, S. Liegeois, D. Baillie, M. Labouesse, T.R. Burglin. The hedgehog-related gene qua-1 is required for molting in Caenorhabditis elegans Dev. Dyn., 235 (2006), pp. 1469-1488.
- F. Marletaz, P. N. Firbas, I. Maeso, J. J. Tena, O. Bogdanovic, M. Perry, C. D. R. Wyatt, E. de la Calle-Mustienes, S. Bertrand, D. Burguera, R. D. Acemel, S. J. van Heeringen, S. Naranjo, C. Herrera-Ubeda, K. Skvortsova, S. Jimenez-Gancedo, D. Aldea, Y. Marquez, L. Buono, I. Kozmikova, J. Permanyer, A. Louis, B. Albuixech-Crespo, Y. Le Petillon, A. Leon, L. Subirana, P. J. Balwierz, P. E. Duckett, E. Farahani, J.-M. Aury, S. Mangenot, P. Wincker, R. Albalat, È. Benito-Gutiérrez, C. Cañestro, F. Castro, S. D'Aniello, D. E. K. Ferrier, S. Huang, V. Laudet, G. A. B. Marais, P. Pontarotti, M. Schubert, H. Seitz, I. Somorjai, T. Takahashi, O. Mirabeau, A. Xu, J.-K. Yu, P. Carninci, J. R. Martinez-Morales, H. R. Crollius, Z. Kozmik, M. T. Weirauch, J. Garcia-Fernàndez, R. Lister, B. Lenhard, P. W. H. Holland, H. Escriva, J. L. Gómez-Skarmeta, M. Irimia, Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature. 2018. 564, 64–70.
- Ferguson, E.L. Horvitz, H.R. The multivulva phenotype of certain Caenorhabditis elegans mutants results from defects in two functionally redundant pathways. Genetics. 1989. 123(1):109–121.
- Hendriks, G.J., Gaidatzis, D., Aeschimann, F. & Grosshans, H. Extensive oscillatory gene expression during C. elegans larval development. Mol. Cell. 2014. 53:380–392.
- Hermann, G. J. Leung, B. Priess, J. R. Left-right asymmetry in C. elegans intestine organogenesis involves a LIN-12/Notch signaling pathway. Development. 2000. 127:3429 -3440.
- Hubbard, E.J., and Greenstein, D. The Caenorhabditis elegans gonad: a test tube for cell and developmental biology. Dev. Dyn. 2000. DC1.218:2–22
- Ingham, P. W. Nakano, Y. Seger, C. Mechanisms and functions of Hedgehog signalling across the metazoa. Nat. Rev. Genet. 2011. 12: 393-406
- Johnson, B. R. Taxonomically Restricted Genes Are Fundamental to Biology and Evolution. Front. Gen. 2018. 9:407-412
- Kamath, R.S., and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. Methods. 30:313–321.

- Lara-Ramıŕez, R. Poncelet, G. Patthey, C. Shimeld, S. M. The structure, splicing, synteny and expression of lamprey COE genes and the evolution of the COE gene family in chordates. Dev. Genes Evol. 2017. 227:319-338.
- Lee, J.J., Ekker, S.C., von Kessler, D.P., Porter, J.A., Sun, B.I., and Beachy, P.A. 1994. Autoproteolysis in hedgehog protein bio- genesis. Science. 266: 1528–1537.
- Phillips, B. T. Kimble, J. A New Look at TCF and β-Catenin through the Lens of a Divergent C. elegans Wnt Pathway. Dev. Cell. 2009. 17:27-34
- Porter, J.A., von Kessler, D.P., Ekker, S.C., Young, K.E., Lee, J.J. Moses, K., and Beachy, P.A. 1995. The product of hedgehog autoproteolytic cleavage active in local and long-range signalling. Nature. 374: 363–366.
- Prabh, N. Rödelsperger, C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs?. BMC Bioinformatics. 2016. 17(1):1–13. doi:10.1186/s12859-016-1102-x.
- Smit AFA."RepeatMasker."URL: http://www.repeatmasker.org (1996-2005)
- Sulston, J. E. and Hodgkin, J. (1988). The nematode Caenorhabditis elegans. Methods. W. B. Wood, Cold Spring Harbor Laboratory
- Timmons, L., Court, D.L., and Fire, A. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. Gene. 263:103–112.
- Yu, J.-K., Holland, L. Z. and Holland, N. D. An amphioxus nodal gene (AmphiNodal) with early symmetrical expression in the organizer and mesoderm and later asymmetrical expression associated with left–right axis formation. Evol. Dev. 2002. 4:418-425.
- Zhang X.M., Ramalho-Santos M., McMahon A.P.(2001) Smoothened mutants reveal redundant roles for Shh and Ihh signaling including regulation of L/R asymmetry by the mouse node. Cell 105:781–792.
- Abrams EW, Cheng YL, Andrew DJ.. 2013. Drosophila KDEL receptor function in the embryonic salivary gland and epidermis. PLoS One. 8:e77618.
- Altenhoff A. “The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces” Nucleic Acids Research, 2018, 46 (D1): D477-D485
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402. 1.
- Altun, Z. F. and Hall, D. H. (2009). Introduction. WormAtlas.
- Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D, Niu W, Boyle AP, Xie D, Ma L, Murray JI, Reinke V, Waterston RH, Snyder M. Regulatory analysis of the C. elegans genome with spatiotemporal resolution. Nature. 2014 Aug 28;512(7515):400-5
- Arroyo JI, Díez B, Kempes CP, West GB, Marquet PA. A general theory for temperature dependence in biology. Proc Natl Acad Sci U S A. 2022 Jul 26;119(30):e2119872119. doi: 10.1073/pnas.2119872119.

- Ascencio D, Diss G, Gagnon-Arsenault I, Dubé AK, DeLuna A, Landry CR. Expression attenuation as a mechanism of robustness against gene duplication. Proc Natl Acad Sci U S A. 2021 Feb 9;118(6):e2014345118.
- Aspöck, G. Kagoshima, H. Niklaus, G. Bürglin, T.R. Caenorhabditis elegans has scores of hedgehog-related genes: sequence and expression analysis. Genome Res. 1999. 9:909–923.
- Baker EA, Gilbert SPR, Shimeld SM, Woollard A. Extensive non-redundancy in a recently duplicated developmental gene family. BMC Ecol Evol. 2021 Mar 1;21(1):33. doi: 10.1186/s12862-020-01735-z.
- Baker EA, Woollard A; The road less travelled? Exploring the nuanced evolutionary consequences of duplicated genes. Essays Biochem 8 December 2022; 66 (6): 737–744.
- Baker EA, Woollard A. How Weird is The Worm? Evolution of the Developmental Gene Toolkit in Caenorhabditis elegans. J Dev Biol. 2019 Sep 28;7(4):19.
- Baugh, L. R. and Sternberg, P. W. (2006). "DAF-16/FOXO Regulates Transcription of cki- 1/Cip/Kip and Repression of lin-4 during C. elegans L1 Arrest." Current Biology 16(8): 780-785.
- Baugh LR, Demodena J, Sternberg PW. RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science. 2009 Apr 3;324(5923):92-4. doi: 10.1126/science.1169628.
- Bienkowska, D. and C. R. Cowan (2012). "Centrosomes can initiate a polarity axis from any position within one-cell C. elegans embryos." Curr Biol 22(7): 583-589.
- Birchler JA, Veitia RA. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci USA. 2012. 109:14746–14753
- Birchler JA, Veitia RA. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci USA. 2012. 109:14746–14753
- Birchler JA, Veitia RA. One hundred years of gene balance: how stoichiometric issues affect gene expression, genome evolution, and quantitative traits. Cytogenet Genome Res. 2021. 161: 529–550
- Blaxter, M. L. De Ley, P. Garey, J.R. Liu, L.X. Scheldeman, P. Vierstraete, A. Vanfleteren, J. R. Mackey, L.Y. Dorris, M. Frisse, L.M. Vida, J.T. Thomas, W. A. molecular evolutionary framework for the phylum Nematoda. Nature. 1998. 392 (6671): 71–75.
- Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., Van de Peer, Y., 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7, R43.
- Bourbon HG, Benetah MH, Guillou E, Mojica-Vazquez LH, Baanannou A, Bernat-Fabre S, Loubiere V, Bantignies F, Cavalli G, Boube M. A shared ancient enhancer element differentially regulates the bric-a-brac tandem gene duplicates in the developing Drosophila leg. PLoS Genet. 2022 Mar 16;18(3):e1010083.
- Brabin, C. Appleford P. Woollard, A. (2011) The Caenorhabditis elegans GATA Factor ELT-1 Works through the Cell Proliferation Regulator BRO-1 and the Fusogen EFF-1 to Maintain the Seam Stem-Like Fate." PLoS Genetics. 7(8): e1002200
- Brenner, S. The genetics of Caenorhabditis elegans. Genetics. 1974. 77:71-94.

- Broitman-Maduro G, Lin KT, Hung WW, Maduro MF. Specification of the C. elegans MS blastomere by the T-box factor TBX-35. Development. 2006 Aug;133(16):3097-106. doi: 10.1242/dev.02475.
- Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V., Robinson-Rechavi, M., 2006. Gene loss and evolutionary rates following whole- genome duplication in teleost fishes. Mol. Biol. Evol. 23, 1808–1816.
- Bujalka H, Koenning M, Jackson S, Perreau VM, Pope B, Hay CM, Mitew S, Hill AF, Lu QR, Wegner M, Srinivasan R, Svaren J, Willingham M, Barres BA, Emery B. MYRF is a membrane-associated transcription factor that autoproteolytically cleaves to directly activate myelin genes. PLoS Biol. 2013;11(8):e1001625. doi: 10.1371/journal.pbio.1001625.
- Bürglin T.R. Warthog and Groundhog, novel families related to Hedgehog. Curr. Biol. 1996. 6:1047–1050
- Byerly, L., Cassada, R. C. and Russell, R.L. (1976). "The life cycle of the nematode Caenorhabditis elegans: I. Wild-type growth and reproduction." Developmental Biology 51(1): 23-33.
- Bzymek, M. Lovett, S. T. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc. Natl Acad. Sci. USA. 2001. 98:8319–8325.
- C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. (1998);282:2012–2018
- Calonga-Solís V, Fabbri-Scallet H, Ott F, Al-Sharkawi M, Künstner A, Wünsch L, Hiort O, Busch H, Werner R. MYRF: A New Regulator of Cardiac and Early Gonadal Development-Insights from Single Cell RNA Sequencing Analysis. J Clin Med. 2022 Aug 18;11(16):4858. doi: 10.3390/jcm11164858.
- Cano AV, Gitschlag BL, Rozhoňová H, Stoltzfus A, McCandlish DM, Payne JL. Mutation bias and the predictability of evolution. Philos Trans R Soc Lond B Biol Sci. 2023 May 22;378(1877):20220055. doi: 10.1098/rstb.2022.0055.
- Cassada, R. C. and Russell, R. L. (1975). "The dauer larva, a post-embryonic developmental variant of the nematode Caenorhabditis elegans." Developmental Biology 46(2): 326- 342.
- Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, Mathelier A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022 Jan 7;50(D1):D165-D173.
- Chang F, Xing P, Song F, Du X, Wang G, Chen K, Yang J. The role of T-box genes in the tumorigenesis and progression of cancer. Oncol Lett. 2016 Dec;12(6):4305-4311. doi: 10.3892/ol.2016.5296.
- Charest et al., Combinatorial Action of Temporally Segregated TFs 2020, Developmental Cell 55, 483–499

- Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. Proc Natl Acad Sci U S A. 2009 Jul 28;106(30):12412-7. doi: 10.1073/pnas.0906217106.
- Cheeks, R. J., J. C. Canman, W. N. Gabriel, N. Meyer, S. Strome and B. Goldstein (2004). "C. elegans PAR proteins function by mobilizing and stabilizing asymmetrically localized protein complexes." Curr Biol 14(10): 851-862.
- Chi W, Reinke V. Promotion of oogenesis and embryogenesis in the C. elegans gonad by EFL-1/DPL-1 (E2F) does not require LIN-35 (pRB). Development. 2006 Aug;133(16):3147-57. doi: 10.1242/dev.02490. Epub 2006 Jul 19. Erratum in: Development. 2006 Sep;133(17):3495.
- Cinkornpumin JK, Hong RL. RNAi mediated gene knockdown and transgenesis by microinjection in the necromenic Nematode Pristionchus pacificus. J Vis Exp. 2011 Oct 16; (56):e3270. doi: 10.3791/3270.
- Collin J, Hasoon MSR, Zerti D, Hammadi S, Dorgau B, Clarke L, Steel D, Hussain R, Coxhead J, Lisgo S, Queen R, Lako M. Single cell RNA sequencing reveals transcriptional changes of human choroidal and retinal pigment epithelium cells during fetal development, in healthy adult and intermediate age-related macular degeneration. Hum Mol Genet. 2023 Jan 16:ddad007. doi: 10.1093/hmg/ddad007.
- Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938-50. doi: 10.1038/nrg2482.
- Cook DE, Zdraljevic S, Roberts JP, Andersen EC. CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Res. 2017 Jan 4;45(D1):D650-D657
- Copley RR. The Unicellular Ancestry of Groucho-Mediated Repression and the Origins of Metazoan Transcription Factors. Genome Biol Evol. 2016 Jun 27;8(6):1859-67. doi: 10.1093/gbe/evw118.
- Cowan, C. R. and A. A. Hyman (2004). "Centrosomes direct cell polarity independently of microtubule assembly in C. elegans embryos." Nature 431(7004): 92-96.
- Cross SH, Mckie L, Hurd TW, Riley S, Wills J, Barnard AR, Young F, MacLaren RE, Jackson IJ. The nanophthalmos protein TMEM98 inhibits MYRF self-cleavage and is required for eye size specification. PLoS Genet. 2020 Apr 1;16(4):e1008583. doi: 10.1371/journal.pgen.1008583.
- Cruz-Guilloty F, Pipkin ME, Djuretic IM, Levanon D, Lotem J, Lichtenheld MG, Groner Y, Rao A. Runx3 and T-box proteins cooperate to establish the transcriptional program of effector CTLs. J Exp Med. 2009 Jan 16;206(1):51-9. doi: 10.1084/jem.20081242.
- Cuenca, A. A., A. Schetter, D. Aceto, K. Kemphues and G. Seydoux (2003). "Polarization of the C. elegans zygote proceeds via distinct establishment and maintenance phases." Development 130(7): 1255-1265.
- Cuentas-Condori A, Miller Rd DM. Synaptic remodeling, lessons from C. elegans. J Neurogenet. 2020 Sep-Dec;34(3-4):307-322. doi: 10.1080/01677063.2020.1802725.
- de Celis Ibeas J.M. and Bray S.J., Bowl is required downstream of Notch for elaboration of distal limb patterning. Development, 2003. 130(24): p. 5943–52

- Deppe, U., E. Schierenberg, T. Cole, C. Krieg, D. Schmitt, B. Yoder and G. von Ehrenstein (1978). "Cell lineages of the embryo of the nematode Caenorhabditis elegans." Proc Natl Acad Sci U S A 75(1): 376-380.
- Des Marais DL, Rausher MD. Parallel evolution at multiple levels in the origin of hummingbird pollinated flowers in Ipomoea. Evolution. 2010 Jul;64(7):2044-54. doi: 10.1111/j.1558-5646.2010.00972.x.
- Dulai KS, von Dornum M, Mollon JD, Hunt DM. (1999). "The Evolution of Trichromatic Colour Vision by Opsin Gene Duplication in New World and Old World Primates." Genome Res 9: 629-638.
- Emery, B., Agalliu, D., Cahoy, J.D., Watkins, T.A., Dugas, J.C., Mulinyawe, S.B., Ibrahim, A., Ligon, K.L., Rowitch, D.H., and Barres, B.A. (2009). Myelin gene regulatory factor is a critical transcriptional regulator required for CNS myelination. Cell 138, 172–185.
- Espinosa-Cantú A, Ascencio D, Barona-Gómez F, DeLuna A. Gene duplication and the evolution of moonlighting proteins. Front Genet. 2015 Jul 7;6:227. doi: 10.3389/fgene.2015.00227.
- Fages-Lartaud M, Tietze L, Elie F, Lale R, Hohmann-Marriott MF. mCherry contains a fluorescent protein isoform that interferes with its reporter function. Front Bioeng Biotechnol. 2022 Aug 9;10:892138. doi: 10.3389/fbioe.2022.892138.
- Farin HF, Bussen M, Schmidt MK, Singh MK, Schuster-Gossler K, Kispert A. Transcriptional repression by the T-box proteins Tbx18 and Tbx15 depends on Groucho corepressors. J Biol Chem. 2007 Aug 31;282(35):25748-59. doi: 10.1074/jbc.M703724200.
- Fernando D K Tria, William F Martin, Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes, Genome Biology and Evolution, Volume 13, Issue 10, October 2021
- Fielenbach N, Antebi A. C. elegans dauer formation and the molecular basis of plasticity. Genes Dev. 2008 Aug 15;22(16):2149-65. doi: 10.1101/gad.1701508.
- Fierst, J.L. Willis, J.H. Thomas, C.G. Wang, W. Reynolds R.M. Reproductive mode and the evolution of genome size and structure in Caenorhabditis nematodes. PLoS Genet. 2015. 11: e1005323 (erratum: PLoS Genet. 11: e1005497)
- Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998 Feb 19;391(6669):806-11. doi: 10.1038/35888.
- Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999 Apr;151(4):1531-45.
- Gancedo, C., Flores, C. L., and Gancedo, J. M. (2014). Evolution of moonlighting proteins: insight from yeasts. Biochem. Soc. Trans. 42, 1715–1719. doi: 10.1042/BST20140199

- Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dosé AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM 3rd, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Rätsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X; modENCODE Consortium; Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010 Dec 24;330(6012):1775-87. doi: 10.1126/science.1196914.
- Gibson-Brown JJ, Agulnik SI, Silver LM, Niswander L, Papaioannou VE. Involvement of T-box genes Tbx2-Tbx5 in vertebrate limb specification and development. Development. 1998 Jul;125(13):2499-509
- Goldstein B, Macara IG. The PAR proteins: fundamental players in animal cell polarization. Dev Cell. 2007 Nov;13(5):609-622. doi: 10.1016/j.devcel.2007.10.007.
- Goldstein, B. and S. N. Hird (1996). "Specification of the anteroposterior axis in Caenorhabditis elegans." Development 122(5): 1467-1474.
- Good K, Ciosk R, Nance J, Neves A, Hill RJ, Priess JR. The T-box TFs TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in C. elegans embryos. Development. 2004 May;131(9):1967-78
- Gout JF, Lynch M. Maintenance and loss of du- plicated genes by dosage subfunctionalization. Mol Biol Evol. 2015;32:2141–8.
- Gouy M., Guindon S. & Gascuel O. (2010) SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2):221-224.1.
- Greenberg L. and Hatini V., Essential roles for lines in mediating leg and antennal proximodistal patterning and generating a stable Notch signaling interface at segment borders. Dev Biol, 2009. 330(1): p. 93–104.
- Greulich F, Rudat C, Kispert A. Mechanisms of T-box gene function in the developing heart. Cardiovasc Res. 2011 Jul 15;91(2):212-22. doi: 10.1093/cvr/cvr112.
- Grill, S. W., J. Howard, E. Schaffer, E. H. Stelzer and A. A. Hyman (2003). "The distribution of active force generators controls mitotic spindle position." Science 301(5632): 518-521.

- Grill, S. W., P. Gonczy, E. H. Stelzer and A. A. Hyman (2001). "Polarity controls forces governing asymmetric spindle positioning in the Caenorhabditis elegans embryo." Nature 409(6820): 630-633.
- Hamill, D. R., A. F. Severson, J. C. Carter and B. Bowerman (2002). "Centrosome maturation and mitotic spindle assembly in C. elegans require SPD-5, a protein with multiple coiled-coil domains." Dev Cell 3(5): 673-684.
- Hammell CM, Hannon GJ. Inducing RNAi in Caenorhabditis elegans by Injection of dsRNA. Cold Spring Harb Protoc. 2016 Jan 4;2016(1):pdb.prot086306. doi: 10.1101/pdb.prot086306.
- Harrison, PM, Zheng D, Zhang Z, Carriero N, Gerstein M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. 2005. 33:2374-2383.
- Hird, S. N. and J. G. White (1993). "Cortical and cytoplasmic flow polarity in early embryonic cells of Caenorhabditis elegans." J Cell Biol 121(6): 1343-1355.
- Hird, S. N., J. E. Paulsen and S. Strome (1996). "Segregation of germ granules in living Caenorhabditis elegans embryos: cell-type-specific mechanisms for cytoplasmic localisation." Development 122(4): 1303-1312.
- Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. Nature. 2007 Oct 11;449(7163):677-81. doi: 10.1038/nature06151.
- Holland PW, Marlétaz F, Maeso I, Dunwell TL, Paps J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philos Trans R Soc Lond B Biol Sci. 2017 Feb 5;372(1713):20150480. doi: 10.1098/rstb.2015.0480.
- Holland, PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. (2017) “New genes from old: asymmetric divergence of gene duplicates and the evolution of development.” Phil. Trans. R. Soc. B 372: 20150480.
- Hornig J, Fröb F, Vogl MR, Hermans-Borgmeyer I, Tamm ER, Wegner M. The transcription factors Sox10 and Myrf define an essential regulatory network module in differentiating oligodendrocytes. PLoS Genet. 2013 Oct;9(10):e1003907. doi: 10.1371/journal.pgen.1003907.
- Horton JS, Flanagan LM, Jackson RW, Priest NK, Taylor TB. A mutational hotspot that determines highly repeatable evolution can be built and broken by silent genetic changes. Nat Commun. 2021 Oct 19;12(1):6092. doi: 10.1038/s41467-021-26286-9.
- Huang H, Zhou F, Zhou S, Qiu M. MYRF: A Mysterious Membrane-Bound Transcription Factor Involved in Myelin Development and Human Diseases. Neurosci Bull. 2021 Jun;37(6):881-884. doi: 10.1007/s12264-021-00678-9.
- Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754-755.
- Hunter MP, Prince VE. Zebrafish hox paralogue group 2 genes function redundantly as selector genes to pattern the second pharyngeal arch. Dev Biol. 2002 Jul 15;247(2):367-89. doi: 10.1006/dbio.2002.0701.
- Hutter, H. and R. Schnabel (1994). "glp-1 and inductions establishing embryonic axes in C. elegans." Development 120(7): 2051-2064.

- Hutter, H. and R. Schnabel (1995). "Establishment of left-right asymmetry in the Caenorhabditis elegans embryo: a multistep process involving a series of inductive events." Development 121(10): 3417-3424.
- Irimia, M. Rukov, JL. Penny, D. Vinther, J. Garcia-Fernandez, J. Roy, SW. Origin of introns by 'intronization' of exonic sequences. Trends in Genetics. 2008. 24(8):378-81.
- Jeffery CJ. Protein moonlighting: what is it, and why is it important? Philos Trans R Soc Lond B Biol Sci. 2018 Jan 19;373(1738):20160523.
- Ji YJ, Nam S, Jin YH, Cha EJ, Lee KS, Choi KY, Song HO, Lee J, Bae SC, Ahnn J. RNT-1, the C. elegans homologue of mammalian RUNX transcription factors, regulates body size and male tail development. Dev Biol. 2004 Oct 15;274(2):402-12. doi: 10.1016/j.ydbio.2004.07.029.
- Jin Q, Ren Y, Wang M, Suraneni PK, Li D, Crispino JD, Fan J, Huang Z. Novel function of FAXDC2 in megakaryopoiesis. Blood Cancer J. 2016 Sep 30;6(9):e478. doi: 10.1038/bcj.2016.87.
- Johnston WL, Krizus A, Dennis JW. The eggshell is required for meiotic fidelity, polar-body extrusion and polarization of the C. elegans embryo. BMC Biol. 2006 Oct 16;4:35.
- Juan Carlos del Pozo, Elena Ramirez-Parra, Whole genome duplications in plants: an overview from Arabidopsis, Journal of Experimental Botany, Volume 66, Issue 22, December 2015, Pages 6991–7003
- Kagoshima H, Shigesada K, Kohara Y. RUNX regulates stem cell proliferation and differentiation: insights from studies of C. elegans. J Cell Biochem. 2007 Apr 1;100(5):1119-30. doi: 10.1002/jcb.21174.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587-589.
- Kanzaki, N. Tsai, I.J. Tanaka, R. Hunt, V.L. Tsuyama, K. Liu, D. Maeda, Y. Namai, S. Kumagai, R. Tracey, A. et al. Biology and genome of a newly discovered sibling species of Caenorhabditis elegans. Nat Commun. 2018. 9:3216.
- Karp X. Working with dauer larvae. WormBook. 2018 Aug 9;2018:1-19. doi: 10.1895/wormbook.1.180.1.
- Kelly WG, Fire A. Chromatin silencing and the maintenance of a functional germline in Caenorhabditis elegans. Development. 1998 Jul;125(13):2451-6. doi: 10.1242/dev.125.13.2451.
- Kemphues, K. J., J. R. Priess, D. G. Morton and N. S. Cheng (1988). "Identification of genes required for cytoplasmic localization in early C. elegans embryos." Cell 52(3): 311-320. Morton, D. G., J. M. Roos and K. J. Kemphues (1992). "par-4, a gene required for cytoplasmic localization and determination of specific cell types in Caenorhabditis elegans embryogenesis." Genetics 130(4): 771-790.
- Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PW, Chu KH, Hui JH. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. Heredity (Edinb). 2016 Feb;116(2):190-9.
- Kiontke, KC. Félix, M. Ailion, M. Rockman, MV. Braendle, C. Pénigault JB. and Fitch, D (2011) “A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits” BMC Evolutionary Biology 11: 339.

- Kirby, C., M. Kusch and K. Kemphues (1990). "Mutations in the par genes of Caenorhabditis elegans affect cytoplasmic reorganization during the first cell cycle." Dev Biol 142(1): 203-215.
- Kispert, A., Herrmann, B. G., Leptin, M. and Reuter, R.(1994). Homologs of the mouse Brachyury gene are involved in the specification of posterior terminal structures in Drosophila,Tribolium, and Locusta. Genes Dev. 8,2137-2150
- Kondrashov, F.A. and Kondrashov, A.S. Role of selection in fixation of gene duplications. J. Theor. Biol. 2006. 239:141–151
- Kornfeld, K. Vulval development in Caenorhabditis elegans. Trends Genet. 1997. 13:55-61
- Kostas SA, Fire A. The T-box factor MLS-1 acts as a molecular switch during specification of nonstriated muscle in C. elegans. Genes Dev. 2002 Jan 15;16(2):257-69. doi: 10.1101/gad.923102.
- Koszul, R. et al. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. 2004. EMBO J. 23, 234–243
- Kusch, T. and Reuter, R. (1999). Functions for Drosophila brachyenteron and forkhead in mesoderm specification and cell signalling. Development 126,3991-4003.
- Kuzmin E, Taylor JS, Boone C. Retention of duplicated genes in evolution. Trends Genet. 2022 Jan;38(1):59-72. doi: 10.1016/j.tig.2021.06.016.
- Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, Tresenrider A, Pryszlak M, Hu MC, Varriano B, Costanzo M, Knop M, Moses A, Myers CL, Andrews BJ, Boone C. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science. 2020 Jun 26;368(6498):eaaz5667.
- Labbe, J. C., E. K. McCarthy and B. Goldstein (2004). "The forces that position a mitotic spindle asymmetrically are tethered until after the time of spindle assembly." J Cell Biol 167(2): 245-256.
- Lazetic, V. Fay, D.S. Molting in C. elegans. Worm. 2017. 6: p. e1330246
- Lemons D, McGinnis W. (2006) “Genomic evolution of Hox gene clusters." Science 313: 1918-1922.
- Lewin TD, Royall AH, Holland PWH. Dynamic Molecular Evolution of Mammalian Homeobox Genes: Duplication, Loss, Divergence and Gene Conversion Sculpt PRD Class Repertoires. J Mol Evol. 2021 Jul;89(6):396-414.
- Lewis, EB. Pseudoallelism and gene evolution. Cold Spring Harb Symp Quant Biol. 1951. 16:159–74.
- Li Z, Park Y, Marcotte EM. A Bacteriophage tailspike domain promotes self-cleavage of a human membrane-bound transcription factor, the myelin regulatory factor MYRF. PLoS biology. 2013;11:e1001624
- Loker R, Mann RS. Divergent expression of paralogous genes by modification of shared enhancer activity through a promoter-proximal silencer. Curr Biol. 2022 Aug 22;32(16):3545-3555.e4
- Lynch, M. Conery, J. S. The evolutionary fate and consequences of duplicate genes. Science. 2000. 5494:1151–2254.

- Maduro, M. and D. Pilgrim (1995). "Identification and cloning of unc-119, a gene expressed in the Caenorhabditis elegans nervous system." Genetics 141(3): 977-988.1.
- Mathews EA, Stroud D, Mullen GP, Gavriilidis G, Duerr JS, Rand JB, Hodgkin J. Allele-specific suppression in Caenorhabditis elegans reveals details of EMS mutagenesis and a possible moonlighting interaction between the vesicular acetylcholine transporter and ERD2 receptors. Genetics. 2021 Aug 9;218(4)
- Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc Biol Sci. 2014 Jan 22;281(1778):20132881. doi: 10.1098/rspb.2013.2881.
- McClintock, B. The Relation of Homozygous Deficiencies to Mutations and Allelic Series in Maize. Genetics. 1944. 29:478-502.
- Mello, C. and A. Fire (1995). "DNA transformation." Methods in cell biology 48: 451-482.
- Meng J, Ma X, Tao H, Jin X, Witvliet D, Mitchell J, Zhu M, Dong MQ, Zhen M, Jin Y, Qi YB. Myrf ER-Bound Transcription Factors Drive C. elegans Synaptic Plasticity via Cleavage-Dependent Nuclear Translocation. Dev Cell. 2017 Apr 24;41(2):180-194.e7. doi: 10.1016/j.devcel.2017.03.022.
- Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M, Kliebenstein D, Weng ML, Imbert E, Ågren J, Rutter MT, Fenster CB, Weigel D. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022 Feb;602(7895):101-105. doi: 10.1038/s41586-021-04269-6.
- Moore RC, Purugganan MD. The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A. 2003 Dec 23;100(26):15682-7.
- Moshe, A., and Pupko, T. 2019. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. Bioinformatics. 35(15):2562-2568.
- Munro, E., J. Nance and J. R. Priess (2004). "Cortical flows powered by asymmetrical contraction transport PAR proteins to establish and maintain anterior-posterior polarity in the early C. elegans embryo." Dev Cell 7(3): 413-424.
- Nagel S, Meyer C. Establishment of the TBX-code reveals aberrantly activated T-box gene TBX3 in Hodgkin lymphoma. PLoS One. 2021 Nov 22;16(11):e0259674. doi: 10.1371/journal.pone.0259674.
- Naseeb, S. Ames, RM. Delneri, D. Lovell, SC (2017) “Rapid functional and evolutionary changes follow gene duplication in yeast”. 284. Proceedings of the Royal Society Biological Sciences
- Nesta AV, Tafur D, Beck CR. Hotspots of Human Mutation. Trends Genet. 2021 Aug;37(8):717-729. doi: 10.1016/j.tig.2020.10.003.
- Nimmo R, Antebi A, Woollard A. mab-2 encodes RNT-1, a C. elegans Runx homologue essential for controlling cell proliferation in a stem cell-like developmental lineage. Development. 2005 Nov;132(22):5043-54. doi: 10.1242/dev.02102.

- Nowak MA, Boerlijst MC, Cooke J, Smith JM. Evolution of genetic redundancy. Nature. 1997 Jul 10;388(6638):167-71. doi: 10.1038/40618.
- O'Connell, K. F., K. N. Maxwell and J. G. White (2000). "The spd-2 gene is required for polarization of the anteroposterior axis and formation of the sperm asters in the Caenorhabditis elegans zygote." Dev Biol 222(1): 55-70.
- Ohno, S. Evolution by gene duplication. Springer-Verlag. 1970. ISBN 0-04-575015-7.
- Olsen EN. (2006). “Gene Regulatory Networks in the Evolution and Development of the Heart.” Science 313: 1922-1927.
- Osche, G. Systematik und Phylogenie der Gattung Rhabditis (Nematoda). Zool. Jb. (Abt. 1). 1952. 81: 90–280.
- Papaioannou VE. The T-box gene family: emerging roles in development, stem cells and cancer. Development. 2014 Oct;141(20):3819-33
- Pazdernik N, Schedl T. Introduction to germ cell development in Caenorhabditis elegans. Adv Exp Med Biol. 2013;757:1-16. doi: 10.1007/978-1-4614-4015-4_1.
- Pérez Jurado LA, Wang YK, Peoples R, Coloma A, Cruces J, Francke U. A duplicated gene in the breakpoint regions of the 7q11.23 Williams-Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK. Hum Mol Genet. 1998 Mar;7(3):325-34. doi: 10.1093/hmg/7.3.325.
- Petrella LN, Wang W, Spike CA, Rechtsteiner A, Reinke V, Strome S. synMuv B proteins antagonize germline fate in the intestine and ensure C. elegans survival. Development. 2011 Mar;138(6):1069-79. doi: 10.1242/dev.059501.
- Plageman TF Jr, Yutzey KE. T-box genes and heart development: putting the "T" in heart. Dev Dyn. 2005 Jan;232(1):11-20
- Pocock R, Ahringer J, Mitsch M, Maxwell S, Woollard A. A regulatory network of T-box genes and the even-skipped homologue vab-7 controls patterning and morphogenesis in C. elegans. Development. 2004 May;131(10):2373-85.
- Priess, J. R. and J. N. Thomson (1987). "Cellular interactions in early C. elegans embryos." Cell 48(2): 241-250.
- Qadota H, Moerman DG, Benian GM. A molecular mechanism for the requirement of PAT-4 (integrin-linked kinase (ILK)) for the localization of UNC-112 (Kindlin) to integrin adhesion sites. J Biol Chem. 2012 Aug 17;287(34):28537-51. doi: 10.1074/jbc.M112.354852.
- Qian W, Liao BY, Chang AY, Zhang J. Maintenance of duplicate genes and their functional redundancy by reduce expression. Trends Genet. 2010;26:425–30.
- Qiao, X., Li, Q., Yin, H. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol 20, 38 (2019).
- Redmond AK, Gundappa MK, Macqueen DJ, McLysaght A. Extensive Lineage-specific Rediploidisation Masks Shared Whole Genome Duplication in the Sturgeon-paddlefish Ancestor. bioRxiv 2022.05.16.492067; doi: https://doi.org/10.1101/2022.05.16.492067

- Reuben M, Lin R. Germline X chromosomes exhibit contrasting patterns of histone H3 methylation in Caenorhabditis elegans. Dev Biol. 2002 May 1;245(1):71-82. doi: 10.1006/dbio.2002.0634.
- Rodriguez, A., De La Cera, T., Herrero, P., and Moreno, F. (2001). The hexokinase 2 protein regulates the expression of the GLK1, HXK1 and HXK2 genes of Saccharomyces cerevisiae. Biochem. J. 355, 625–631.
- Rody, H.V.S., Baute, G.J., Rieseberg, L.H. et al. Both mechanism and age of duplications contribute to biased gene retention patterns in plants. BMC Genomics 18, 46 (2017).
- Rose L, Gönczy P. Polarity establishment, asymmetric division and segregation of fate determinants in early C. elegans embryos. WormBook. 2014 Dec 30:1-43. doi: 10.1895/wormbook.1.30.2.
- Royall AH, Frankenberg S, Pask AJ, Holland PWH. Of eyes and embryos: subfunctionalization of the CRX homeobox gene in mammalian evolution. Proc Biol Sci. 2019 Jul 24;286(1907):20190830 Rose L, Gönczy P. Polarity establishment, asymmetric division and segregation of fate determinants in early C. elegans embryos. WormBook. 2014 Dec 30:1-43. doi: 10.1895/wormbook.1.30.2.
- Sadler, P. L. and D. C. Shakes (2000). "Anucleate Caenorhabditis elegans sperm can crawl, fertilize oocytes and direct anterior-posterior polarization of the 1-cell embryo." Development 127(2): 355-366.
- Schlientz AJ, Bowerman B. C. elegans CLASP/CLS-2 negatively regulates membrane ingression throughout the oocyte cortex and is required for polar body extrusion. PLoS Genet. 2020 Oct 7;16(10):e1008751.
- Schwager, E.E., Sharma, P.P., Clarke, T. et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol 15, 62 (2017).
- Schwarzer D, Stummeyer K, Gerardy-Schahn R, Muhlenhoff M. Characterization of a novel intramolecular chaperone domain conserved in endosialidases and other bacteriophage tail spike and fiber proteins. The Journal of biological chemistry. 2007;282:2821–31.
- Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, Adamski M, Adamska M, Hughes TR, Gómez-Skarmeta JL, Ruiz-Trillo I. Early evolution of the T-box transcription factor family. Proc Natl Acad Sci U S A. 2013 Oct 1;110(40):16050-5
- Semenza JC, Hardwick KG, Dean N, Pelham HR.. 1990. ERD-2, a yeast gene required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway. Cell. 61:1349–1357.
- Senoo H, Araki T, Fukuzawa M, Williams JG. A new kind of membrane-tethered eukaryotic transcription factor that shares an auto-proteolytic processing mechanism with bacteriophage tail-spike proteins. J Cell Sci. 2013 Nov 15;126(Pt 22):5247-58. doi: 10.1242/jcs.133231.
- Session, A., Uno, Y., Kwon, T. et al. Genome evolution in the allotetraploid frog Xenopus laevis. Nature 538, 336–343 (2016).
- Showell C, Binder O, Conlon FL. T-box genes in early embryogenesis. Dev Dyn. 2004 Jan;229(1):201-18

- Singer, J. B., Harbecke, R., Kusch, T., Reuter, R. and Lengyel,J. A. (1996). Drosophila brachyenteron regulates gene activity and morphogenesis in the gut. Development 122,3707-3718
- Smithies, O. Chromosomal rearrangements and protein structure. Cold Spring Harb. Symp. Quant. Biol. 1964. 29, 309–319
- Smythe, A.B. Holovachov, O. Kocot, K. M. Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. BMC Evol. Biol. 2019. 19:121.
- Spring, J. Vertebrate evolution by interspecific hybridisation--are we polyploid? FEBS Lett. 1997. 400:2-8
- Stennard FA, Harvey RP. T-box transcription factors and their roles in regulatory hierarchies in the developing heart. Development. 2005 Nov;132(22):4897-910. doi: 10.1242/dev.02099.
- Sterken MG, Snoek LB, Kammenga JE, Andersen EC. The laboratory domestication of Caenorhabditis elegans. Trends Genet. 2015 May;31(5):224-31. doi: 10.1016/j.tig.2015.02.009.
- Stevens L, Félix MA, Beltran T, Braendle C, Caurcel C, Fausett S, Fitch D, Frézal L, Gosse C, Kaur T, Kiontke K, Newton MD, Noble LM, Richaud A, Rockman MV, Sudhaus W, Blaxter M. Comparative genomics of 10 new Caenorhabditis species. Evol Lett. 2019 Apr 2;3(2):217-236. doi: 10.1002/evl3.110.
- Storz JF. Causes of molecular convergence and parallelism in protein evolution. Nat Rev Genet. 2016 Apr;17(4):239-50. doi: 10.1038/nrg.2016.11.
- Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol. 1977 Mar;56(1):110-56. doi: 10.1016/0012-1606(77)90158-0.
- Sulston, J. E., E. Schierenberg, J. G. White and J. N. Thomson (1983). "The embryonic cell lineage of the nematode Caenorhabditis elegans." Dev Biol 100(1): 64-119.
- Sulston, J.E. Horvitz, H.R. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev. Biol. 1977. 56:110-56.
- Sundaram MV. Canonical RTK-Ras-ERK signaling and related alternative pathways. WormBook. 2013 Jul 11:1-38. doi: 10.1895/wormbook.1.80.2
- Taylor, J.H. et al. The organization and duplication of chromosomes as revealed by autoradiographic studies using tritium-labeled thymidine. Proc. Natl. Acad. Sci. U. S. A. 1957. 43, 122–128.
- Ting CT, Tsaur SC, Sun S, Browne WE, Chen YC, Patel NH, Wu CI. Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. Proc Natl Acad Sci U S A. 2004 Aug 17;101(33):12232-5.
- Tintori SC, Osborne Nishimura E, Golden P, Lieb JD, Goldstein B. A Transcriptional Lineage of the Early C. elegans Embryo. Dev Cell. 2016 Aug 22;38(4):430-44.
- Tischler J, Lehner B, Chen N, Fraser AG. 2006. Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. Genome Biol. 7:R69

- Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016 Jul 8;44(W1):W232-5.
- Vallejo-Marín M, Buggs RJA, Cooley AM, Puzey JR. Speciation by genome duplication: Repeated origins and genomic composition of the recently formed allopolyploid species Mimulus peregrinus. Evolution. 2015 Jun;69(6):1487-1500.
- Vanin, E., F. Processed pseudogenes: characteristics and evolution. Annual Review of Genetics. 1985. 19: 253–72
- Veitia RA. Gene duplicates: Agents of robustness or fragility? Trends Genet. 2017. 33:377–9.
- Vinckenbosch, N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A. 2006. 103:3220-3225.
- Wagner, A. (2011). “The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems”. Oxford University Press. ISBN-13: 978-0199692606
- Wallenfang, M. R. and G. Seydoux (2000). "Polarization of the anterior-posterior axis of C. elegans is a microtubule-directed process." Nature 408(6808): 89-92.
- Watts, J. L., B. Etemad-Moghadam, S. Guo, L. Boyd, B. W. Draper, C. C. Mello, J. R. Priess and K. J. Kemphues (1996). "par-6, a gene involved in the establishment of asymmetry in early C. elegans embryos, mediates the asymmetric localization of PAR-3." Development 122(10): 3133-3140.
- Williams B.D. , Waterston R.H. Genes critical for muscle development and function in Caenorhabditis elegans identified through lethal mutations. J. Cell Biol. (1994);124:475–490.
- Williams, B. D., B. Schrank, C. Huynh, R. Shownkeen and R. H. Waterston (1992). "A genetic mapping system in Caenorhabditis elegans based on polymorphic sequence-tagged sites." Genetics 131(3): 609-624.
- Wood, W. B. (1991). "Evidence from reversal of handedness in C. elegans embryos for early cell interactions determining cell fates." Nature 349(6309): 536-538.
- Wood, W. B. Introduction to C. elegans Biology. Cold Spring Harbour. 1988. 1-16.
- Woollard A, Hodgkin J. The caenorhabditis elegans fate-determining gene mab-9 encodes a T-box protein required to pattern the posterior hindgut. Genes Dev. 2000 Mar 1;14(5):596-603.
- Wu W, Zhen X, Shi N. DNA-binding domain of myelin-gene regulatory factor: purification, crystallization and X-ray analysis. Corrigendum. Acta crystallographica Section F, Structural biology communications. 2017;73:713.
- Wu YC, Rasmussen MD, Kellis M. Evolution at the subgene level: domain rearrangements in the Drosophila phylogeny. Mol Biol Evol. 2012 Feb;29(2):689-705. doi: 10.1093/molbev/msr222.
- Xia, B. and Zhang, W. et al. The genetic basis of tail-loss evolution in humans and apes. bioRxiv. https://doi.org/10.1101/2021.09.14.460388
- Xiong H, Pears C , Woollard A (2017) An enhanced C. elegans based platform for toxicity assessment. Sci Rep 7, 9839.

- Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug;24(8):1586-91. doi: 10.1093/molbev/msm088.
- Yang, Z. (2007). PAML 4: a program package for phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 24: 1586-1591
- Ye, S., Dhillon, S., Ke, X., Collins, A. R. and Day, I.N. (2001). "An efficient procedure for genotyping single nucleotide polymorphisms." Nucleic Acids Research 29(17): e88.
- Zhang, Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 2003. 13:2541-2558)
- Zhu K (A), Spaink HP, Durston AJ. Hoxc6 loss of function truncates the main body axis in Xenopus. Cell Cycle. 2017 Jun 3;16(11):1136-1138. doi: 10.1080/15384101.2017.1317415.
- Zhu K (B), Spaink HP, Durston AJ. Collinear Hox-Hox interactions are involved in patterning the vertebrate anteroposterior (A-P) axis. PLoS One. 2017 Apr 11;12(4):e0175287. doi: 10.1371/journal.pone.0175287.
- Ziel, J.W., and D.R. Sherwood. Roles for netrin signaling outside of axon guidance: a view from the worm. Dev. Dyn. 2010. 239:1296–1305.