

A characterisation of gene duplicability:
its determining factors and impact on
duplicate evolution

by

Zoe Vance

A thesis submitted to
The University of Dublin
for the degree of
Doctor of Philosophy

Smurfit Institute of Genetics
Trinity College
The University of Dublin



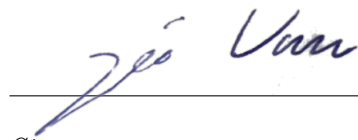
March, 2022

Declaration of Authorship

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018)



Signature

3/3/2022

Date

Acknowledgements

The work contained in this thesis would not have been possible without the help and support of many people.

Firstly, a thank you to Aoife. It's been a long four years, but your support and understanding for all of us, especially during these 'unprecedented times', has always been much appreciated. There's a lot of trust involved in letting me choose my own way as much as I have, so really, a big thanks for that. I firmly believe that helping when needed and knowing when to just leave me to it is the only reason I've stuck it out.

Thank you to all the McLysaght lab members past and present who have helped me along the way. I am so privileged to have worked with such great scientists and, more to the point, fantastic human beings. I can't even begin to explain the positive impact of endless coffee breaks ('group therapy'), scratchcards, cans of monster, lab baby quotes, snowflake making and extended rants. It may not have been a positive impact on productivity, but a positive impact nonetheless. I hope we eventually crack how to recreate even the half of it over a Zoom call. Special mention to Alan: it's been almost five years now, we can only live in hope that I will someday stop needing supervision.

To all my other Smurfit pals, especially the remnants of the Bench E crew, I have missed coffee and lunchtime chats so much over the past 18 months. There are not many people I could honestly say I want to see every single weekday again,

but ye make the cut. It wouldn't have been the same without the drama, the gossip, the dyeing of hair and the Pav pints. We're all almost there and I'm so proud of us all for sticking with it. To Emily especially, I'm so glad I had my best friend right there with me this whole time. Having someone you can text to rant about badly formatted datasets at any time is a truly underutilised PhD coping mechanism.

Mum and Dad, and all the dogs. Thanks for being home base the past few years. I know I can always turn to you for support, advice, or even just a hot meal I didn't have to cook myself. Also thank you for not questioning what I'm doing with my life when I said I was doing four more years of college: I'm still not sure.

Finally, last but not least. Kevin, you're my port in a storm. I wouldn't have picked anyone else to be stuck in lockdown with for the better part of a year and a half while finishing a PhD; how we got through this long sitting within two feet of each other all day, every day I don't even know. You've always been there through all the meltdowns, even if you have had to interrupt the tears to ask what exactly goes in a thesis introduction, and even though you still don't know what a goddamn ohnolog is. Love you to the moon and back.

Contents

1	Introduction	1
1.1	Gene duplication	1
1.1.1	Mechanisms of gene duplication	2
1.1.2	Broad impact of gene duplication	5
1.2	Fates of gene duplicates	9
1.2.1	Routes to non-redundancy	9
1.2.2	The Dosage Balance Hypothesis	13
1.2.3	The relative frequency of duplicate fates	14
1.3	The interaction between WGD and environmental factors	15
1.3.1	Alternative contexts for WGD	18
1.4	The concept of gene duplicability	19
1.4.1	Sequence and structure based biases	20
1.4.2	Regulatory and expression based biases	21
1.4.3	Constraint based biases	22
1.4.4	Functional biases	23
1.4.5	Biased fractionation	24
1.4.6	Outstanding questions on the intrinsic factors impacting gene duplicability	25
1.4.7	Contribution of factors to gene duplicability	25
1.5	Aim	28

2	Materials & Methods	29
2.1	Ortholog and paralog detection	29
2.1.1	How is homology determined?	29
2.1.2	Potential issues with homology assignment	30
2.2	Determination of duplication type	33
2.2.1	Use of existing 2R ohnolog data	33
2.2.2	Definition of ohnologs in bony fish dataset	34
3	Genes permitting different modes of duplication differ across a wide range of features	36
3.1	Introduction	36
3.2	Materials & Methods	38
3.2.1	Data sources and processing	38
3.2.2	Statistical methods	43
3.3	Results	46
3.3.1	Structural and sequence-based features: ohnologs are longer and more structurally complex than SSDs or singletons . .	46
3.3.2	Regulation-based features: ohnologs are more broadly ex- pressed and show more complex regulation	48
3.3.3	Conservation and essentiality related features: ohnologs are slower evolving and less dispensable	52
3.3.4	Duplicate categories show differences in functional enrich- ments	54
3.3.5	Enriched/depleted functions show a clearer complementary pattern between singletons and ohnologs than between SSDs and ohnologs	57
3.3.6	Differences in duplicate age between SSDs and ohnologs do not fully account for observed effects	60

3.3.7	Feature importance rankings show ohnologs and SSDs are primarily distinguished by domain structure, expression level and haploinsufficiency	61
3.4	Discussion	73
4	Evidence from <i>Drosophila</i> supports higher duplicability of faster evolving genes	77
4.1	Introduction	77
4.2	Materials & Methods	80
4.2.1	Inference of orthologies	80
4.2.2	Selection of singleton groups and tree building	83
4.2.3	Dataset filtering and evolutionary rate calculation	83
4.2.4	PAML branch models	87
4.2.5	Quantification of potential confounders	88
4.2.6	Reassessment of duplication status for fast evolving singletons	90
4.2.7	Additional statistical methods	91
4.3	Results	91
4.3.1	Faster evolving genes are more duplicable in the <i>Drosophila</i> lineage	91
4.3.2	Potential confounders do not fully explain the faster evolution rate of duplicable genes	94
4.3.3	Weak evidence for other implications of redundancy-driven rate increase	97
4.3.4	Fast evolving singletons may be explained by limitations of orthology inference	101
4.4	Discussion	103
5	Consistency of gene duplicability varies across the genome . .	108

viii CONTENTS

5.1	Introduction	108
5.2	Materials & Methods	111
5.2.1	Data sources and processing	111
5.2.2	Family size distribution	114
5.2.3	Cross-event overlap definition	114
5.2.4	Modelling	115
5.2.5	Investigation into relationship between reciprocal retention and other gene features	117
5.2.6	GO function analysis	117
5.3	Results	118
5.3.1	Gene family size distribution is consistent with a broad spread of intragenomic duplicability	118
5.3.2	Significant overlaps exist in retention following independent WGD events	118
5.3.3	Gene duplicability can be modelled as a continuous value .	121
5.3.4	Families showing retention across all events are not strongly reciprocally retained	123
5.3.5	Families with consistent duplicability show indications of dosage sensitivity	124
5.3.6	Weak evidence for lineage-specific preferential retention/loss of WGD duplicates	127
5.4	Discussion	137
6	Conclusions	151
	Bibliography	158

List of Tables

3.1	Duplicate classification by dataset	46
3.2	Regression formulae	63
3.3	Change in rank for correlated features considered in isolation . . .	67
3.4	Change in rank for correlated features considered in isolation using centered data	68
3.5	Change in rank for correlated features considered in isolation for age controlled model	68
3.6	Change in rank for correlated features considered in isolation for centered data using age controlled model	73
4.1	Data location for species used	81
4.2	Abbreviations used	82
5.1	Genome assemblies used in this analysis	113
5.2	Source of <i>Ks</i> limits used in WGD placement	116
5.3	Top 20 enriched GO terms for most strongly reciprocally retained families	128
5.4	Top 20 enriched GO terms for overlap families	129
5.5	Functional enrichment and depletion for families showing teleost 3R retention	132
5.6	Functional enrichment and depletion for families showing sturgeon 3R retention	133

5.7	Functional enrichment and depletion for families showing salmonid 4R retention	134
5.8	Functional enrichment and depletion for families showing cyprinid 4R retention	135
5.9	Highest retention functions for teleost 3R	138
5.10	Lowest retention functions for teleost 3R	139
5.11	Highest retention functions for sturgeon 3R	140
5.12	Lowest retention functions for sturgeon 3R	141
5.13	Highest retention functions for salmonid 4R	142
5.14	Lowest retention functions for salmonid 4R	143
5.15	Highest retention functions for cyprinid 4R	144
5.16	Lowest retention functions for cyprinid 4R	145

List of Figures

1.1	A summary of WGD events in vertebrates	4
1.2	NAHR causes small-scale gene duplication	5
1.3	The biases affecting the set of retained duplicates	11
1.4	A summary of post-duplication fates	12
1.5	The dosage balance hypothesis	16
2.1	Asymmetrically diverging paralogs are prone to orthology assignment errors	32
3.1	Exclusion of retroduplicates	40
3.2	Sequence based feature comparisons	49
3.3	Regulation related feature comparisons	50
3.4	Comparison of expression at the tissue level	51
3.5	Constraint related feature comparisons	55
3.6	Overlap of enriched/depleted GO terms between categories	59
3.7	Features regressed on duplicate age and duplicate type	62
3.8	Duplicate type prediction	69
3.9	Duplicate type prediction using scaled data	70
3.10	Duplicate type prediction controlling for age differences	71
3.11	Duplicate type prediction with scaled data, controlling for age differences	72

xii LIST OF FIGURES

4.1	Confirmatory rate comparisons in duplicability assessment clade with more relaxed starting singleton definition	84
4.2	Proxy ancestral rate comparisons with more relaxed starting singleton definition	84
4.3	Filtering performed on initial set of singletons	85
4.4	Proxy ancestral rate comparisons with no dS filter applied	86
4.5	Confirmatory rate comparisons with no dS filter applied	87
4.6	Project strategy	92
4.7	Duplicable genes are faster evolving in a clade with duplications allowed	93
4.8	Duplicable genes are ancestrally faster evolving	93
4.9	Confounding features show not significant difference between singleton and duplicable genes	95
4.10	OLS regression of rate on confounding features	96
4.11	Residual comparisons show significant changes in p-value	97
4.12	LOWESS regression on confounders	98
4.13	Comparison of LOWESS residuals from regressing rate on confounders	99
4.14	Comparison of differing rate models does not support rate acceleration or rate asymmetry post-duplication	99
4.15	Fast-evolving singletons are a distinct group partially explained by HDF	104
5.1	Bony fish species considered in this work	112
5.2	Family size distributions show intragenome duplicability variation	119
5.3	Significant overlaps exist across ohnolog retention sets	120
5.4	Identification of highly overlapping retention	121
5.5	Distribution of reciprocal retention	123

5.6	Comparison of lambda values between overlapping set and whole set	125
5.7	Distribution of GO terms by percentage of families retained . . .	136
5.8	Hypothetical groupings for consistency of duplicability	150
6.1	A proposed model for consistency of duplicability	157

Abbreviations

CAI	Codon adaptation index
CNVs	Copy number variants
DBH	Dosage balance hypothesis
DDC	Duplication-degeneration-complementation
EAC	Escape from adaptive conflict
GMM	Gaussian mixture model
HDF	Homology detection failure
IDRs	Intrinsically disordered regions
LOF	Loss of function
LOWESS	Locally weighted scatterplot smoothing
Mya	Million years ago
NAHR	Non-allelic homologous recombination
OLS	Ordinary least squares
PPIs	Protein-protein interactions
ss4R	Salmonid-specific 4th round

SSD	Small-scale duplication
SSDs	Small-scale duplicates
TPM	Transcripts per million
ts3R	Teleost-specific 3rd round
WGD	Whole-genome duplication

Summary

Duplication, either through whole genome duplication or small-scale duplication, is a critical process in evolution. The duplication of genetic material provides ample opportunity for adaptation and innovation through the creation of new genes. However, not all genes are capable of successful duplication. This property of genes is referred to as gene duplicability and varies considerably across the genome. The work described in this thesis presents a more cohesive picture of duplicability and its contributions to gene duplication outcomes than has previously been available.

Our knowledge of the factors affecting gene duplicability is fragmented, with few studies dedicated to the quantification of this property of genes. As a result, it is challenging to obtain a complete picture of how different gene features impact on gene duplicability, particularly given that our existing knowledge base is heavily biased in the lineages represented. In chapter 3, we present a full examination of the gene features which impact duplicability across mechanisms in the human genome. We find evidence that paralogs derived from whole-genome duplication represent a highly-conserved and complex group both in terms of structure and regulation, in contrast to the relatively unconstrained small-scale duplicates. Having comparisons across a broad range of features within the same organism and duplicate sets allowed us to estimate the relative importance of features, a question which has gone largely unanswered up to now. We find that expression level

and haploinsufficiency are major contributors to determining small-scale duplicability vs. whole-genome duplicability, which should inform future considerations of differences in duplicate evolution between the two groups.

Although there is a large body of literature examining post-duplication evolutionary processes, few have considered any role for the prior duplicability of genes in contributing to observed patterns. In chapter 4 we consider the role of duplicability in creating the observed lower constraint in small-scale duplicates. We compare possible explanations for this lower constraint and find that, while we can not rule out the established model of rate acceleration due to redundancy post-duplication, duplicable genes are faster-evolving even prior to duplication. This work underlines the importance of considering the pre-existing biases in which genes have duplicated in the first place when developing hypotheses of duplicate evolution.

A large outstanding question in the consideration of duplicability is whether or not this property of genes applies consistently across time and contexts. In chapter 5 we provide an initial investigation of this question in the context of whole-genome duplication in the bony fish lineage. We find that broad trends of duplicability appear to be conserved across events when considering duplicates as a group, however the picture is more complex at the level of single gene families. For individual gene families, consistency of duplicability appears to be a variable property itself and could be largely determined by how strong the influence of intrinsic gene features is relative to random or environmental effects.

Gene duplicability is a key consideration in the study of the process of gene duplication. The work here begins to establish how the probability of duplication for a given gene is determined and how such factors impact on our understanding of duplication.

Chapter 1

Introduction

1.1 Gene duplication

Gene duplication is a frequent occurrence in genomes (Crow and Wagner, 2006; Lipinski et al., 2011; Lynch et al., 2008) and has long been known to play an important role in evolution. As first proposed by Susumu Ohno (Ohno, 1970), the creation of new genetic material for selective forces to act upon has vast potential for biological innovation. Duplicate copies of existing functional gene sequence have a particular advantage in this regard as, in contrast to emergence of new coding sequence from previously non-coding sequence, these copies are ready made fully functional genes, with pre-existing functional domains and interactions providing clearer routes to functional impact on the genome. Additionally, new genes created through duplication (paralogs) are generally thought to be redundant at the point of creation, as distinct from orthologs which are related through speciation events. This promotes exploration of new functions in paralogs not possible in the case of orthologs, as reflected in conservation of function across orthologs.

The importance of duplication is clearly illustrated in two seminal examples;

2 INTRODUCTION

the globins and the Hox genes. The globin family, responsible for oxygenation of blood, has expanded and diversified in function within vertebrates through gene duplication. This diversification has allowed for key innovations, such as the development of functionally distinct globins at different stages of development (Storz et al., 2013). The Hox genes, homeobox genes, act as master regulators in development with roles in determination of body plan. Duplication has played an important role in shaping this family, both in the creation of their clustered structure and in subsequent duplications of the entire cluster. These cluster duplications within vertebrates are suspected to relate to important vertebrate-specific developments (Soshnikova et al., 2013; Wagner et al., 2003), again underlining the critical impacts of gene duplication in the course of evolution.

1.1.1 Mechanisms of gene duplication

Gene duplication can be broadly grouped into two categories: whole-genome duplication (WGD) which comprises genome doublings; and small-scale duplication (SSD) which may include both tandem duplications and retroduplication through retrotransposition. This work will focus primarily on WGD and tandem duplication and further references to SSD will refer to tandem duplication.

WGD can be divided into autopolyploidy and allopolyploidy. Autopolyploidy is a genome doubling within the organism usually caused by errors at cell division, while allopolyploidy refers to a hybridisation event. Regardless of the mechanism of doubling, these are large scale events encompassing the entire genome. Such events are present in the history of all major extant lineages, though at differing frequencies. There are a multitude of examples in plants (Meyers and Levin, 2006), with all angiosperms sharing at least one ancestral genome duplication and many plant lineages showing additional lineage-specific events (Li et al., 2015; Liu, 2015; Ming et al., 2015; Vanneste et al., 2015). Duplications have also been found

in single-celled lineages such as yeast and *Paramecium* (Aury et al., 2006; Kellis et al., 2004; Wolfe and Shields, 1997), with the duplication in yeast particularly well studied.

Within metazoans, the majority of studies concerning WGD have been focused in the vertebrates (events summarised in Figure 1.1). The best known of these WGD events are the generally accepted two rounds (2R) of duplication at the base of the vertebrate tree (Dehal and Boore, 2005; Nakatani et al., 2007; Putnam et al., 2008) now proposed to be a single round of duplication in the vertebrate ancestor followed by lineage specific duplication and triplication in the gnathostomes and cyclostomes respectively (Nakatani et al., 2021; Simakov et al., 2020). Also within the vertebrates, there exist a series of events in fish subsequent to the 2R duplications. There has been at least one 3R event within the Acipenseriformes (paddlefish and sturgeons) (Du et al., 2020; Cheng et al., 2021) as well as a teleost-specific 3R (ts3R) (Braasch et al., 2016; Jaillon et al., 2004; Meyer and Peer, 2005) and multiple subsequent events within the teleosts; the salmonid-specific 4R (ss4R) at approximately 80-100 million years ago (Mya) (Berthelot et al., 2014; Lien et al., 2016) and cyprinid-specific 4R at 12-14 Mya (Chen et al., 2019; Xu et al., 2019). Recent genomic characterisation of these younger or more well-preserved events in fish have allowed more insight into the earlier stages in the resolution of genome duplication, discussed further below. Further events are also proposed within the arthropods (Kenny et al., 2016; Li et al., 2018; Nong et al., 2021), although the position of proposed events in chelicerates is not yet fully known and subsequent work has not supported the large number of proposed duplications in hexapods (Nakatani and McLysaght, 2019; Roelofs et al., 2020). The existence of these genomic duplication events across plants and many diverse animal lineages emphasises how ubiquitous these events are across the tree of life.

The second mechanisms of duplication to be discussed here, SSD, involves

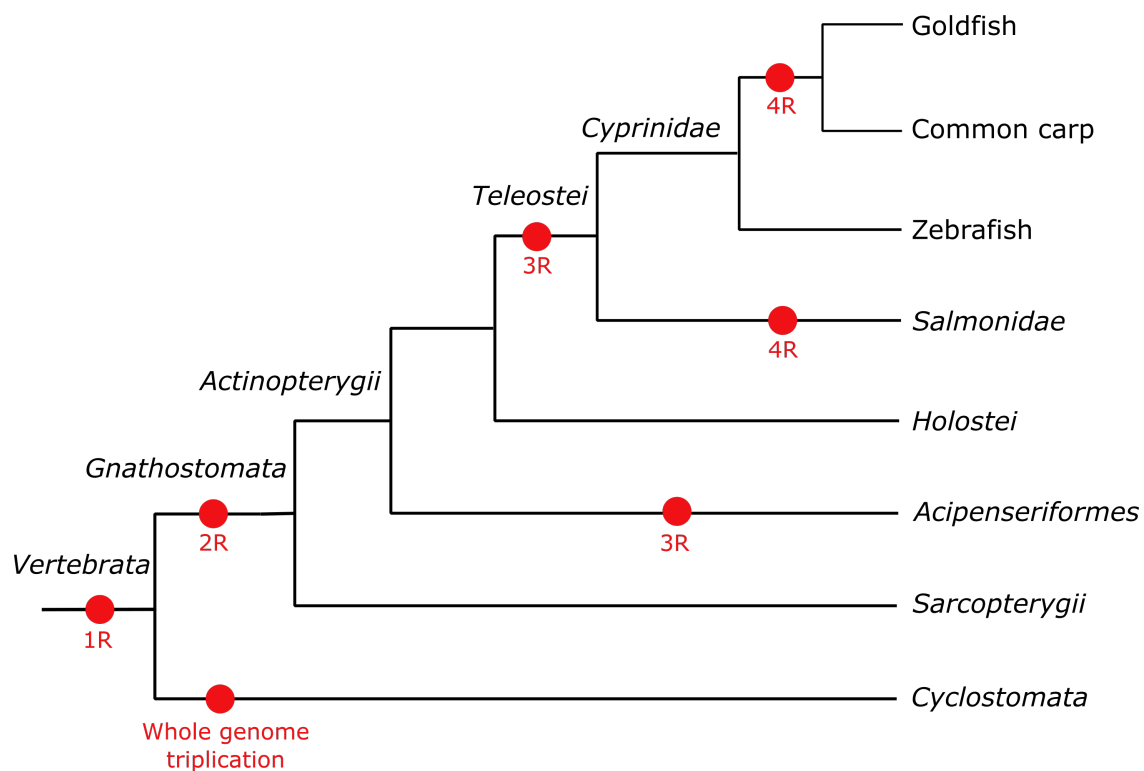


Figure 1.1 | A summary of WGD events in vertebrates. A pruned vertebrate tree showing relevant lineages. Whole genome multiplication events are indicated with red points, branch lengths are not to scale and so event placement is approximate. Resulting ploidy is tetraploid for duplications (1R-4R) and hexaploid for the triplication.

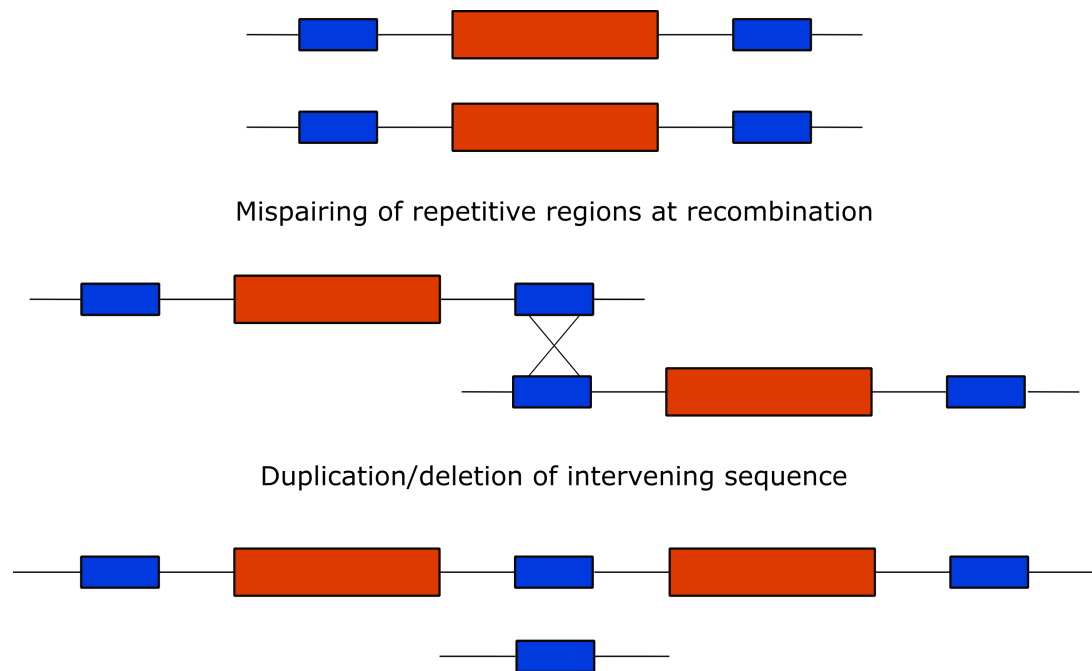


Figure 1.2 | NAHR causes small-scale gene duplication. A hypothetical small-scale duplication scenario. Gene is represented in red with flanking repetitive sequences shown in blue. Repetitive sequences pair incorrectly at recombination creating a gene duplication/deletion.

far fewer genes, with the causal segmental duplications typically comprising segments of up to 200kb (Bailey et al., 2001). In this case the duplication event is typically caused by mispairing at recombination, with non-allelic homologous recombination (NAHR) the major mechanism involved. NAHR occurs when repetitive regions mispair, causing a misalignment at recombination which creates a duplication/deletion of the intervening segment (see Figure 1.2), generating copy number variants (CNVs).

1.1.2 Broad impact of gene duplication

In both cases, the duplication event can impact genomic structure. Tandem duplication creates further repetitive sequence, promoting further instability. Repetitive sequence is associated with 'hotspots' of duplication in the genome which are more likely to create copy-number variants (Dittwald et al., 2013; Mefford and Eich-

6 INTRODUCTION

ler, 2009). However WGD has a more far-ranging impact. Tetraploid organisms eventually restore their diploid state through a process known as rediploidization (Bennett and Johnson, 2003; Gerstein et al., 2006), possibly due to hypothesised issues with tetravalent pairing at meiosis (Stebbins, 1971). This can involve large genomic rearrangements and extensive gene loss (Inoue et al., 2015; Lien et al., 2016; Nakatani et al., 2007; Sémon and Wolfe, 2007), although other studies have suggested a more gradual segmental process that lacks the dramatic genomic alterations (Du et al., 2020). The rediploidization process is completed when the entire process has reverted to disomic, rather than tetrasomic inheritance, a state which has not fully been achieved in some more recent duplications. For example, the salmonid 4R duplication is not fully resolved with parts of the genomes affected remaining tetrasomic to the present day (Lien et al., 2016). Thus, gene duplication impacts not only the gene set present within the genome, but also the structure of the genome itself.

Both types of duplication are also heavily associated with adaptation and innovation. For SSDs, there are examples across diverse categories of their role in niche colonisation. These duplicates are frequently implicated in adaptation to new environmental challenges (Kondrashov, 2012) such as toxic materials (Soo et al., 2011; Yang et al., 2010), exploitation of new resources as seen in dietary examples (Perry et al., 2007) and even in man-made adaptations with some related to domestication (Bickhart et al., 2012). Such a relationship to adaptation is expected within the ideas presented by Ohno (1970), wherein duplication and the ensuing redundancy between copies allows exploration of the evolutionary space with fewer consequences. Colonisation of novel niches is a prime example of how such explorations could be beneficial in the course of evolution.

WGD is generally implied to have even more impact in adaptation with many well-studied duplications associated with important functional innovation

or species radiations as in the vertebrates, angiosperms and salmonids. However, a number of details relating to the exact relationship between WGD and adaptation/species diversification is debated.

Firstly, polyploids are considered to frequently represent evolutionary 'dead ends' (Arrigo and Barker, 2012; Mayrose et al., 2011; Mayrose et al., 2015), which are extremely unlikely to establish themselves (Fowler and Levin, 2016). They are reproductively isolated from the population they occur in, making reproduction difficult and also suffer a host of issues relating to complications at meiosis and genomic instability (Comai, 2005; Marsit et al., 2021; Morgan et al., 2020). On the face of it, we might be surprised that any polyploids have persisted for long enough to give rise to the diverse set of lineages that we observe as paleopolyploids today, however there are advantages to polyploidy that may explain their short-term survival.

Polyploids are hypothesised to be more robust than their diploid counterparts in a number of ways. Their increased genome copy number provides mutational robustness as, in a tetraploid, a single recessive mutation is buffered by three wild-type alleles. Thus, the probability of a recessive mutant phenotype reduces substantially from q^2 in the diploid to q^4 in the tetraploid, where q is the frequency of the mutant allele. This creates greater standing variation in a polyploid population and is supposed to increase evolvability in such genomes (Shi et al., 2015; Soltis et al., 2014), with this likely of particular relevance in allopolyploidy, which brings together previously diverged genes in a hybridisation event. Perhaps because of this, polyploids are observed to show greater stress tolerance than diploids with respect to various environmental stressors (Chao et al., 2013; Ruiz et al., 2016; Yang et al., 2014), including resistance to pathogens (Oswald and Nuismer, 2007). This higher stress tolerance may explain in part how polyploids persist and potentially out-compete diploid individuals.

The role of WGD in adaptation concerns the time period after this initial retention, in the duplication’s long term impact on the evolution of the lineage. Various models exist regarding how a WGD might contribute to, for example, diversification of species within the lineage where it has occurred. While there is circumstantial evidence for WGD contributing to explosive species radiations, for example in the teleost lineage, these are often temporally separated from the WGD event. The radiation time lag model (Schranz et al., 2012) proposes that species radiations can lag behind the occurrence of the WGD by several million years, a hypothesis with mixed support (Clarke et al., 2016; Laurent et al., 2017; Tank et al., 2015), but one that also is in concordance with the LORe (lineage-specific ohnolog resolution) model (Robertson et al., 2017). This model describes how adaptation due to WGD could be delayed for some time due to delayed rediploidization. As long as ohnologs exist in a tetraploid rather than diploid state they do not behave as duplicates, but as alleles. It is only on resolution into a diploid state that the duplicates can begin to diverge and provide material for any potential adaptation. Under the LORe model resolution is delayed until after a speciation event, leading to ohnolog resolution occurring in a lineage-specific manner which is more likely to lead to disparate outcomes than if ohnolog divergence began in the ancestor under common selective pressures and was followed by speciation. This model’s incorporation of delayed rediploidization provides a possible mechanistic basis for the radiation time lag model and therefore a possible refutation for claims that lack of concurrency between WGD and species radiations/ adaptations rules out a role for WGD in promoting large scale adaptation.

Overall, it seems difficult to dismiss that gene and genome duplications have an impact on adaptation. Though in the case of WGD it may be disputed how exactly the duplication event relates to the observed functional novelties that follow, the establishment and long term survival of polyploids in the face of many

challenges in achieving this suggests that they must secure some advantage relative to diploids. This may take the form of increased evolvability or robustness and it seems unlikely such events would not have far-reaching impacts on the evolution of the lineages they occur in.

1.2 Fates of gene duplicates

There are a number of steps that stand between a gene and a successful small-scale duplication (see Figure 1.3, top panel). Creation of a duplicate requires an initial mutation step, which is not equally likely across the genome. As discussed above, certain parts of the genome are considered hotspots for copy number changes, leading to differences in the likelihood of ever duplicating. The subsequent fate of the newly created duplicate then relies on overcoming fixation and retention biases which determine its survival in the genome across a longer timescale. Achieving fixation and long-term retention generally requires that the duplication be in some way beneficial to the organism. This set of biases has less relevance in WGD (see Figure 1.3, lower panel), where no mutation or fixation bias exists. Either the entire genome duplicates, or no genes duplicate. Either the entire WGD event becomes fixed, or no duplicates from the event will have any chance to be retained. It is only the final stage, long-term retention of duplicates where biases across different genes can be observed

1.2.1 Routes to non-redundancy

There are several potential fates for a newly created duplicate, summarised in Figure 1.4. The most likely fate for a new duplicate is nonfunctionalization. A large majority are quickly lost again, with half-lives on the scale of a few million years for mammalian SSDs (Lynch and Conery, 2000). This may be due to

deleterious effects of duplication which drive the purging of the extra copy or a more passive accumulation of disabling mutations in redundant gene copies. Most duplicates are assumed to be redundant at the point of creation and, as such, must in some way overcome this to avoid passive loss.

The first model commonly put forward for duplicate genes to escape redundancy is subfunctionalization, a situation where the sister duplicates share the previous parental function Figure 1.4. This can apply in a variety of ways, either through a literal partitioning of functional domains, sharing of the overall expression dosage or sharing of expression domains (Klüver et al., 2005; Lan and Pritchard, 2016). Such processes can happen without a requirement for adaptive processes, as in the duplication-degeneration-complementation (DDC) model where both copies gradually accumulate mutations until both are required to perform the ancestral function due to complementary loss of function (Force et al., 1999). There are also arguments in favour of an adaptive role in some cases however, such as the escape from adaptive conflict (EAC) model, which invokes a situation where a single gene performs multiple function which are at odds with each other's optimisation, so that duplication provides a solution to the conflict by allowing partitioning of the conflicting functions between the duplicates (Des Marais and Rausher, 2008; Hittinger and Carroll, 2007).

A second model, neofunctionalization, describes how one paralog may escape redundancy by gaining an entirely new function relative to the parent gene. In this case, both copies are retained as one is required to perform the ancestral function and one is required to perform the novel function. One well-known example of this fate is antifreeze glycoproteins in Antarctic notothenioid fish. Neofunctionalization of one of the many trypsinogen paralogs has lead to the creation of these proteins which improve tolerance of freezing temperatures (Chen et al., 1997). An additional example in fish is found in electric fish in the Mormyroid

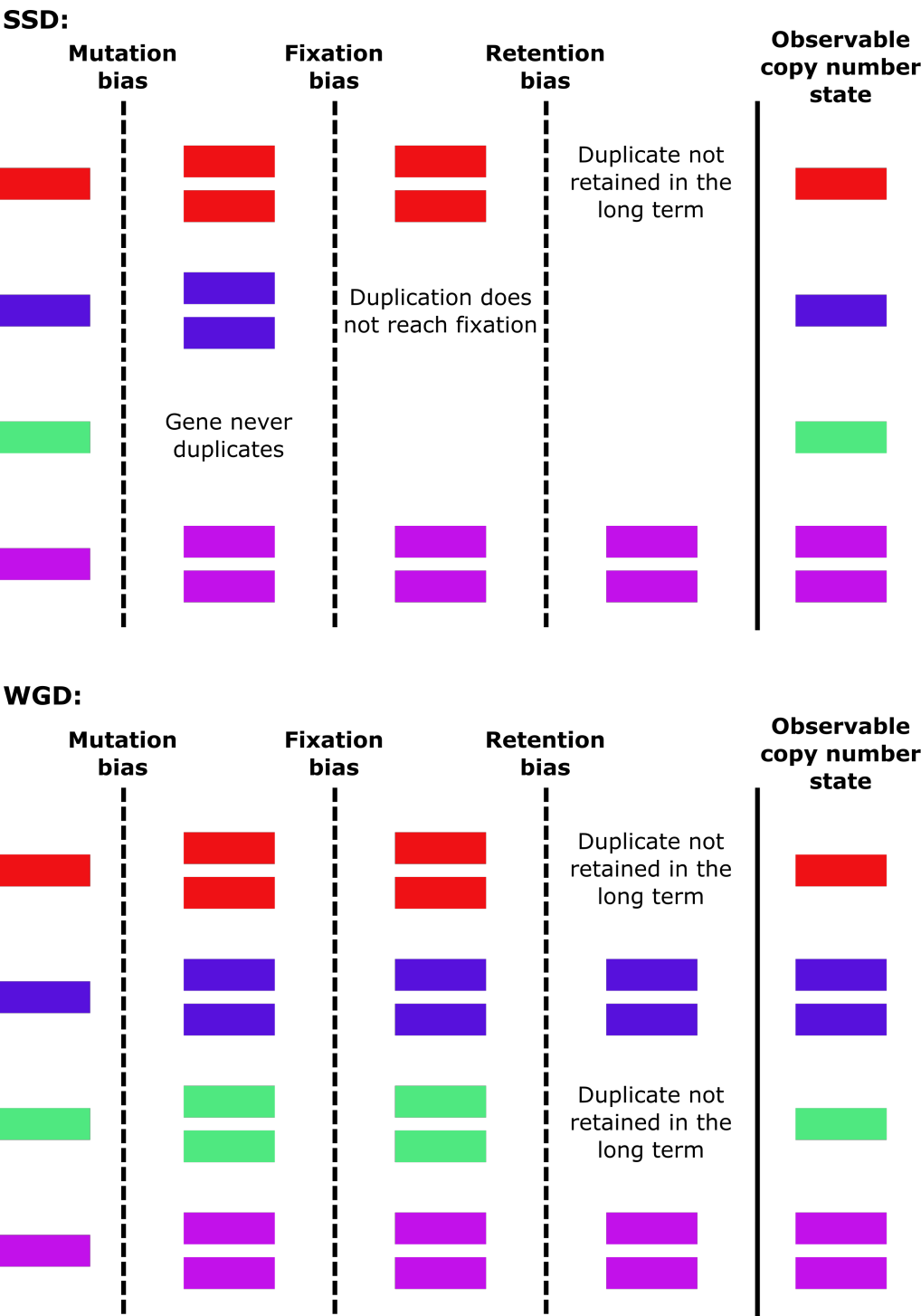


Figure 1.3 | The biases affecting the set of retained duplicates. In the case of the SSD panel, only the purple gene is successfully duplicated in terms of long term retention, with other genes failing at earlier stages. Note that we cannot know if, for example, a green gene duplicate would become fixed or retained. WGD panel shows the scenario where the WGD event occurs and is successfully fixed, any other scenario would result in no long-term retention of WGD duplicates.

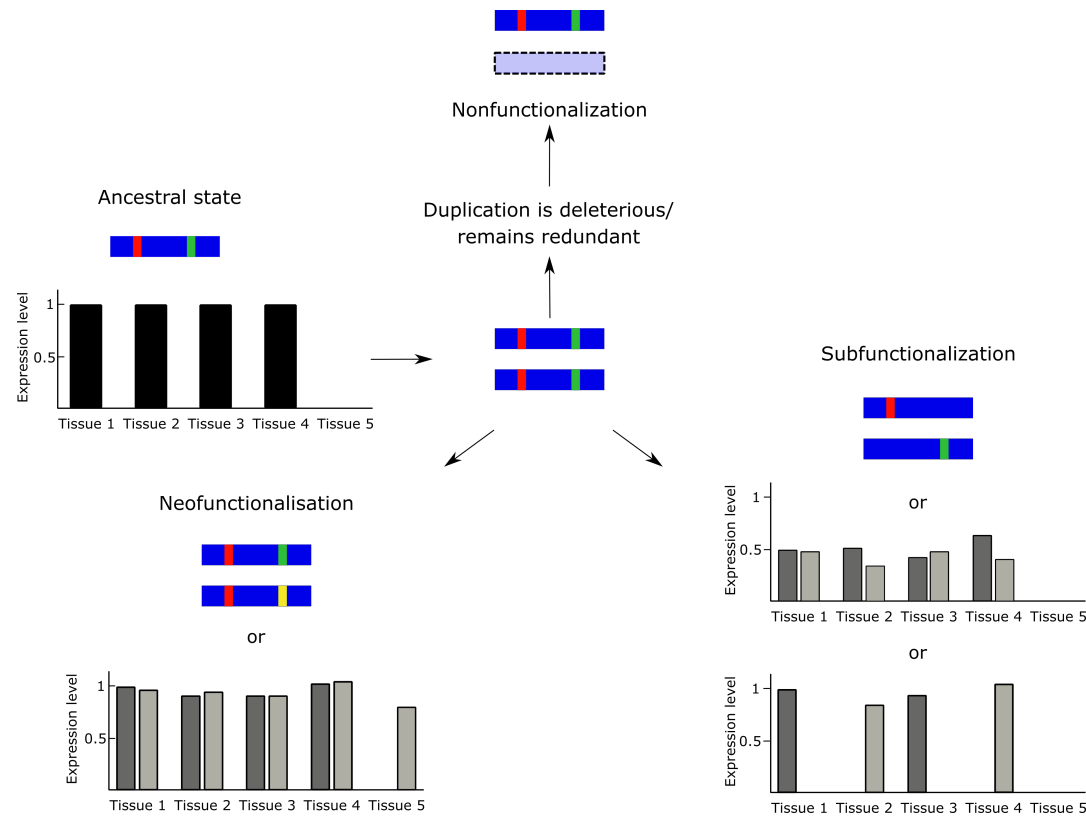


Figure 1.4 | A summary of post-duplication fates. Coloured bands represent different functional domains. Expression levels shown for duplicate pairs (paralogs shown in differing shades of grey) are hypothetical situations under each scenario. The first expression scenario for subfunctionalization refers to dosage subfunctionalization wherein the sum of paralog expression equals the ancestral level, while the second refers to expression domain subfunctionalization wherein one paralog is responsible for maintenance of expression in a given tissue where the ancestral gene was expressed. Expression is expressed relative to the ancestral level such that 0.5 is half the expression of the ancestral gene.

and Gymnotiform lineages. In both cases a duplication of a sodium channel is co-opted to develop a new muscle-derived tissue which acts as an electric organ (Thompson et al., 2014).

1.2.2 The Dosage Balance Hypothesis

In addition to these models of escape from redundancy, it is possible for duplicates to be non-redundant at the point of creation. This can be true in the case of the increased dosage being advantageous, such that the retention of the duplicate copy is favoured by selection, or in the case of dosage balance (Papp et al., 2003; Veitia, 2002). The dosage balance hypothesis (DBH) is an important concept in explaining duplicate fate under different mechanisms. Under the DBH, genes which must maintain their relative stoichiometry, either due to complex assembly or regulatory requirements, are in dosage balance and have their duplicability affected accordingly. Under dosage balance, SSD is deleterious as duplication of a single gene disrupts balance (Figure 1.5). However, loss of a duplicate following WGD is similarly problematic as the initial duplication does not impact on dosage balance but subsequent duplicate loss would. There are multiple reasons why this imbalance may prove to be an issue. The most obvious, as shown by the simplified version of events in Figure 1.5, is that fewer complexes can be formed relative to the production level of subunits, however this can be amplified in the case of higher order complexes (see Figure 1.5, lower panel), particularly in the case of involvement of bridging complex members (Oberdorf and Kortemme, 2009; Veitia and Birchler, 2015), so that the number of completed complexes is even lower than we would expect due to the lower dose of one subunit (Veitia, 2003). In the case of too high a dose of one subunit, the unused protein is not only an energetic waste but may also prove toxic if it is to form aggregates or interfere with the functional complexes in the absence of normal complex assembly (Veitia,

2009).

1.2.3 The relative frequency of duplicate fates

The relative frequency of these varying duplicate fates is not known conclusively. Some studies have identified higher rates of neofunctionalization than subfunctionalization (Chen et al., 2019; DeGiorgio and Assis, 2021), however in the case of De Giorgio and Assis (2020) this is somewhat questionable. The authors use a classifier approach to determine the fate of paralog pairs using various parameters. The highest frequency retention mechanism according to their classifier is 'neofunctionalized child'. 'Child' in this instance seems to refer to the more diverged paralog, with 'parent' referring to the paralog most similar to the ancestral state. This is a misrepresentation of the situation at duplication, with the paralogs indistinguishable from each other at that point it is probably more appropriate to refer to them as 'siblings' than parent and child. It is unclear what impact this artificial designation has on the classifier's assignments. Also, many studies that have attempted to investigate post-duplication evolutionary trends have solely considered expression patterns, with neofunctionalization indicated by a gain of new expression domains as shown in Figure 1.4. This is less complex to quantify than the existence of a new function within the existing expression domain of the parent gene, but this higher frequency of neofunctionalization may not be applicable to that scenario. In order for sequence-based neofunctionalization to occur, it would be required that a beneficial mutation arise (relatively unlikely in and of itself (Charlesworth and Charlesworth, 1998)) and that it be positively selected. This would seem unlikely to be more frequent an outcome than subfunctionalization which can occur through purely neutral accumulation of mutations. That said, duplicate fates are not necessarily mutually exclusive (discussed below) and the likelihood of neofunctionalization might be altered by the influence of

other mechanisms. In this case, however, it is not clear how meaningful it is to quantify relative frequencies of these duplicate fates at all.

While to this point we have presented alternative duplicate fates as distinct possibilities, this is likely not the case over evolutionary time. It has been credibly suggested that processes such as dosage balance and subfunctionalization, which can occur more easily than neofunctionalization as they do not require adaptive processes, could act as initial preservation steps post-duplication (He and Zhang, 2005; Chapman et al., 2006). We cannot even necessarily rule out non-functionalization as having a role in future neofunctionalization. In the case of the electric fish (Thompson et al., 2014), the paralog which was later recruited to the development of an electric organ showed a convergent decrease in expression prior to this neofunctionalization. Thompson et al. (2014) suggest that this decrease in expression was necessary to allow for such alteration to take place. They propose that duplicate expression follows a random walk post-duplication with only the total expression level under selection. In this case, this lead to the eventual loss of importance of one paralog, allowing it to eventually follow a novel path as the loss of its contribution to the collective expression level by then had little effect.

1.3 The interaction between WGD and environmental factors

WGD events have been observed to cluster around times of extreme environmental stress. Many younger plant WGDs are dated around the K-Pg boundary, contemporaneous with the end Cretaceous mass-extinction event (Vanneste et al., 2014), and the teleost event can be placed at a similar time point to the end-Permian extinctions as has also been suggested for older plant WGDs (Van De Peer et al.,

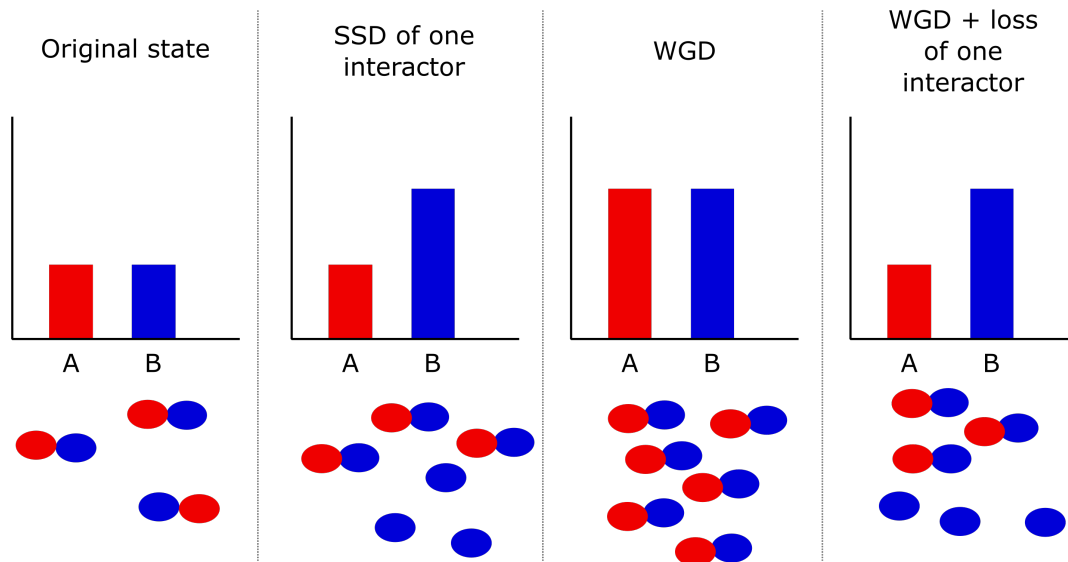
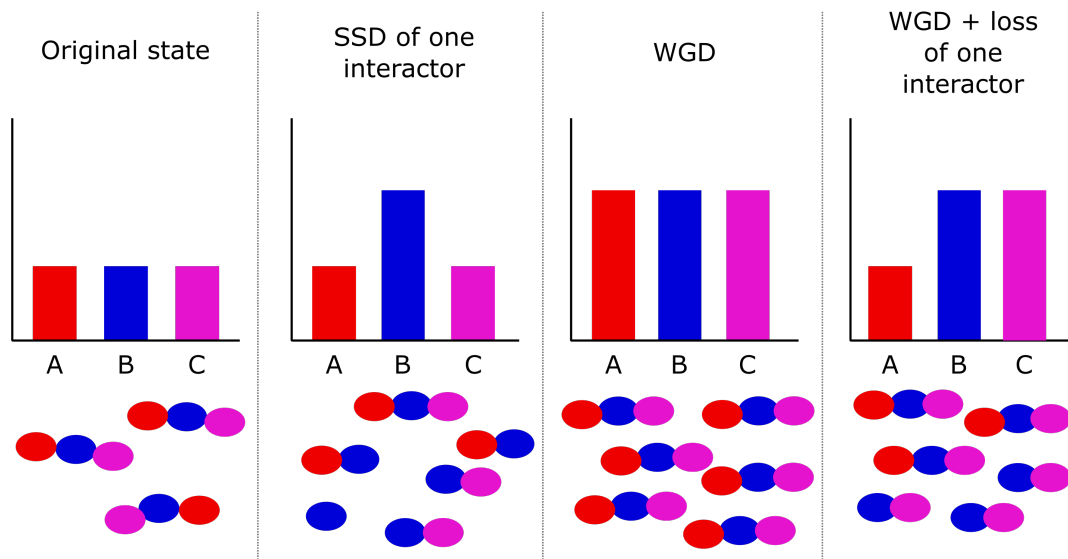
Imbalance:**The impact of bridging subunits:**

Figure 1.5 | The dosage balance hypothesis. The upper panel shows a heterodimer as an illustration of the basic concept of the DBH. Relative quantities of two interacting proteins, A and B, are shown under different duplication and loss scenarios. In the case of both SSD of B and WGD followed by loss of the A duplicate, the B protein is present in excess and represents an energetic waste. The lower panel shows the case where B is a bridging subunit in a complex with A and C, which can not interact without B. In the case of SSD of B fewer complexes are formed, in both SSD and loss of one duplicate subunit post-WGD excess malformed complexes are created.

2017). Although in some cases this relationship with environmental conditions is little more than speculation, possibly amounting to searching for correlation where one does not necessarily exist (Soltis and Burleigh, 2009) or technical artifacts relating to the dating of WGD events (Tiley et al., 2018), the association with many independent plant WGD and the K-Pg boundary does stand up to statistical testing. Vanneste et al. (2014) used Bayesian methods to compare the hypothesis of WGD clustering at this timepoint to a null of random accumulation of WGD and find support for clustering of these events at this time. Additionally, a large number of more recent plant WGDs have been associated with a timepoint in the late Miocene following multiple environmental shifts (Estep et al., 2014).

Although more work would be necessary to state with confidence that WGDs are associated to environmental change or stress, and this may be almost impossible to prove for more ancient events with fewer data points to consider, this would correspond well to the observed trends of increased adaptability and robustness in polyploids and lend support to these claims. It additionally helps to resolve the issue of how polyploids persist at all. While it might not be plausible that the benefits of polyploidy outweigh the disadvantages in the case of a stable environment, it is likely that increased robustness and evolvability would come into their own as advantages in times of increased environmental stressors and instability. It has also been suggested that an increased extinction rate in diploids in conditions where mass extinctions are seen could provide more opportunity for polyploids to establish themselves (Levin and Soltis, 2018).

An alternative explanation for the increase in WGDs in times of environmental turmoil is that stress increases the rate of WGD as a mutation through an influence on meiotic errors. This is true of stress in general (Bomblies and Madlung, 2014; Kreiner et al., 2017; Mable et al., 2011; Ramsey and Schemske, 1998) and particularly for temperature (De Storme and Geelen, 2020; Pécirix et al., 2011).

An increased rate of polyploid creation, even with no corresponding increase in the rate of polyploid survival should lead to an elevation in the number of polyploids that we observe to have persisted. Of course, it is also possible that the increased frequency of observed WGDs during times of stress is a combination of these factors, with increased number of polyploids arising combining with a more favourable environment for their establishment.

1.3.1 Alternative contexts for WGD

Polyploidy can also be seen outside of the context of whole-organism level evolution, in the cellular environment. Somatic polyploidy created through endoreplication is a well-known phenomenon, in particular within hepatocytes; approximately 30% of human liver cells are polyploid (Donne et al., 2020). This type of polyploidy is programmed and essential in many developmental programs (Edgar et al., 2014), in particular in insects (Ren et al., 2020), but any relevance to duplication on the evolutionary level remains to be explored. For example, does programmed polyploidy have an impact on retention of duplicates in genes which function at that point in development by creating a buffer to dosage balance effects through massively increased copy number? These kinds of questions are interesting to consider, and WGD at the sub-organismal level could represent an interesting new environment for the consideration of dosage effects. However, the study of polyploidy in the cellular context and the evolutionary context have been largely separate to this point.

A second example of the relevance of polyploidy in the cellular environment is polyploidy in cancer. WGD events occur in 28.2-37% of cancers in the course of tumorigenesis (Matsumoto et al., 2021) and are associated with poorer prognosis (Dewhurst et al., 2014). In contrast to programmed somatic polyploidy, WGD in cancer has more relevance to the evolutionary significance of WGD discussed

here as cancer progression also represents an evolutionary timeline, albeit one with very different selective pressures. Although polyploid cancer cells do not possess the same concerns as polyploid organisms regarding reproductive isolation or long-term persistence of the organism's lineage, they do mirror the pattern of adaptability and robustness presented above. The poorer prognosis in the case of cancers which have undergone WGD indicates an advantage for these cancer cells over diploid equivalents in terms of proliferation, quite possibly due to increased adaptability to stress as seen in the fact that these cancers are more resistant to drugs used in their treatment (Coward and Harding, 2014). The study of polyploidy in these cellular contexts then is worthwhile in determining the evolutionary significance of WGDs as it reflects a consistent advantage for polyploids under stress in disparate selective contexts.

1.4 The concept of gene duplicability

A key consideration in the study of duplicate evolution is that not all genes are duplicated with the same frequency. Within any genome, there are genes which readily undergo SSD and those which will have SSD-created duplicates quickly purged, while there are also groups which are likely to be retained in duplicate following WGD, those which will be easily lost in the same context and those which do not duplicate under any circumstance. This property of gene duplicability varies considerably depending on the mode of duplication under consideration, as we might expect given the role of dosage balance in promoting duplicate retention and the disparate effects on relative dosage posed by WGD and SSD. In fact, genes which are retained in duplicate post-WGD rarely duplicate by SSD (Makino and McLysaght, 2010) as to duplicate by SSD would pose much the same issue as to lose WGD duplicates (Figure 1.5). This pattern of duplicate retention,

where WGD duplicates are always retained but SSD duplicates never, is referred to as reciprocal retention and is known to be a common retention pattern for dosage-balanced genes (Tasdighian et al., 2017).

While dosage balance is clearly a key contributor to the duplicability of genes, we also know it is not the only feature of a gene which is related to duplicability. While the sets of genes retained in duplicate following WGD and SSD are largely non-overlapping, exceptions to this do exist and so we cannot conclude that strict dosage balance is the only reason for post-WGD retention. Furthermore, dosage balance can not be invoked to explain the single-copy status of many genes. If the reason for lack of SSD in these genes were dosage balance, they would presumably also be retained in duplicate post-WGD. Consequently, other gene features must be considered in the context of duplicability.

1.4.1 Sequence and structure based biases

Beyond dosage, there are a variety of features which show biases within groups of differing duplicability. In terms of features of a gene which may impact on its duplicability, existing work has shown that features as basic as the length of a gene may relate to how likely it is to successfully duplicate, along with other structural features such as intron count (Chapman et al., 2006; Jiang et al., 2013) and intrinsic features of the gene sequence such as GC content or codon optimisation (Banerjee et al., 2017; Bu et al., 2011; Jiang et al., 2013). These features are generally observed to reflect greater complexity and constraint in genes which do not undergo SSD, such as ohnologs and singletons. Genes which do undergo SSD are consistently found to be shorter, with less complex gene structure in terms of intron/exon arrangement and are less constrained in terms of translational efficiency as measured by codon usage bias at the sequence level. These differences in structure and sequence based features may reflect differences

in constraint, such as more deleterious effects of translational errors in genes that are not duplicable by SSD, or mechanistic difficulties in the safe duplication of large, more complex gene structures. Established mechanisms of pathogenicity of CNVs include the disruption of genes and creation of fusion transcripts where breakpoints occur within genes (Holt et al., 2012; Zhang et al., 2009). Longer and more complex genes are more likely to be interrupted by a breakpoint and also more likely to have a functional region disrupted, thus making them more prone to difficulties in duplication through this mechanism.

1.4.2 Regulatory and expression based biases

Regulation and expression are another area where features have been found to give an indication of the duplicability of a gene. Expression differences are reliably found between groups, along with differences in breadth of expression (Jiang et al., 2013; Seoighe and Wolfe, 1999). Regulation of ohnologs is also more complex than that of SSDs (Alvarez-Ponce and Fares, 2012; Amoutzias et al., 2010; Keller and Yi, 2014), with ohnologs also having more transcript isoforms (Jiang et al., 2013). Again, this can be ascribed to differences in constraint. If we assume that ohnologs and singletons are more constrained than SSDs, as potentially indicated by their more stringent regulation, differences in expression and regulation make sense. The broader expression pattern of non-SSDs mean that any alterations in copy number at the genomic level would have far wider-ranging impacts. While ohnologs do show narrower expression domains than singletons, they are generally expressed in tissues with strong constraints such as neural tissue (Roux et al., 2017). There is some indication that the constraints on ohnologs might be tissues specific, for example the higher number of alternative splicing isoforms may reflect a drive to create 'internal paralogs' (Modrek and Lee, 2003) which are able to explore new evolutionary paths in a tissue specific manner. This alternative route

to 'duplication' differs from a straightforward increase of copy number through SSD as it does not have to affect all tissues in which the gene is expressed and so any ill effects in tissues where a duplication would be deleterious can be avoided.

The higher level of expression seen in ohnologs relative to other groups could also be related to constraint, with higher expressed genes generally coming under stronger selective pressure to avoid the risk of protein misfolding (Drummond et al., 2005; Pál et al., 2001), however alternative explanations exist. The duplication of a highly expressed genes through SSD is more costly in the sense of the absolute quantity of protein to be produced. Protein production is one of the most energetically expensive processes in the cell (Tomala and Korona, 2013) both in consumption of translational resources and pressure on the protein homeostasis and post-translational modification machinery (Morrill and Amon, 2019). Although there is not always a direct correspondence between genomic copy number, transcriptional level and protein abundance (Ishikawa et al., 2017; Wang et al., 2020), an increase in copy number of a highly expressed gene is hypothesised to put pressure on the cellular machinery and could impact on the ability of the cell to maintain necessary expression of other genes (Dong et al., 1995), a concept referred to as the 'zero-sum' model of gene expression (Rice and McLysaght, 2017b). This effect is not present in the case of duplication by WGD as the pathways impacted by the increased expression burden have been concurrently duplicated in their entirety, leaving WGD as the only mechanism of duplication available to highly expressed genes and giving the effect that ohnologs as a group are more highly expressed.

1.4.3 Constraint based biases

Perhaps the most critical category of differences between duplicability categories relates to the level of conservation of the gene, as is evidenced by the repeated

invocation of constraint as an explanatory factor for other differences between duplicate groups. Ohnologs are more constrained at the sequence level than SSDs (Chapman et al., 2006), which have repeatedly been found to show faster rates of evolution than other groups (Defoort et al., 2019; He and Zhang, 2006; O’Toole et al., 2018; Qiao et al., 2018). This is true even of post-duplication evolutionary patterns, with ohnologs generally diverging from each other much more slowly than do SSDs in a number of respects (Guan et al., 2007; Hakes et al., 2007; Qiao et al., 2018; Zhu et al., 2013), having been shown to evolve under purifying selection while SSDs evolve under positive selection (Ezoe et al., 2021). The lower level of constraint in SSDs could reflect that the least constrained genes are the safest to duplicate. The ‘dangerous duplicates’ hypothesis (Singh et al., 2012), provides an explanation for why more constrained genes are less likely to duplicate by SSD, and potentially WGD given that some genes do not duplicate by either mechanism. Under this hypothesis, some genes avoid duplication due to strong pressure against dominant negative mutations. Duplication of the sequence creates a larger mutational target for such mutations to occur in either copy and interfere with normal gene function. This idea is also applied to the preservation of ohnologs, with any divergence between the paralogs taking a risk of such an interfering mutation occurring. As such, the level of constraint that applies to the sequence of a given gene is heavily related to its likelihood of duplication.

1.4.4 Functional biases

There are also clear differences in the functional role of genes belonging to varying duplicability groups. Ohnologs typically function in development, in particular neural development but also enrichments are also seen for more generally morphogenesis-related functions. Genes relating to cellular signalling and regulation are also typically over-represented in ohnolog sets, with the enrichment

for transcription factors well documented (Blanc and Wolfe, 2004; Conant et al., 2014; Freeling, 2009; Paterson et al., 2006). In contrast, SSDs usually show enrichment for immune related functions, even across diverse lineages (Conrad and Antonarakis, 2007; Qiao et al., 2018), as well as other functions relating to interaction with the external environment, tallying with their role in niche adaptation as discussed earlier. Finally, singletons show enrichments for basic cellular functions such as biogenesis of cellular components.

1.4.5 Biased fractionation

It would be an incomplete discussion of retention biases without including the issue of subgenome dominance in allopolyploids, also known as biased fractionation (see overview in Bird et al. (2018)). Allopolyploidy involves the bringing together of previously diverged genomes and in many cases it is observed that one genome is dominant over the other, with genes of one genome retained much more frequently than the other in the case of the loss of one copy. This may be due to incompatibilities between the genomes, wherein genes of genome A are better optimised for interactions with other genome A genes and so retention of a given gene in that genome precipitates a pressure for retention of its intragenomic interaction partners (as supported by findings that dividing paralog pairs according to the highest expressed subgenome yields functionally distinct groups (Xu et al., 2019)), although it is also proposed that differences in expression level mask loss of the lower-expressed genome (Freeling et al., 2012). Although this matter is included here as another example of biased retention patterns, we only discuss it briefly as it does not pertain to gene duplicability. Gene duplicability refers to the differences in duplicate retention across the genome while subgenome dominance refers to intergenomic dynamics.

1.4.6 Outstanding questions on the intrinsic factors impacting gene duplicability

In summary, much work has examined the differences defining groups of differing duplicability. However, there remain gaps in our knowledge in this area. Much of the work carried out up to this point has been concentrated in plants and there is not sufficient evidence to confirm that these findings can be extended to other lineages such as vertebrates. More relevant to our understanding of duplication, few studies have attempted to synthesise our knowledge of duplication biases into a cohesive model of how they relate to the duplication process. It is not possible to deduce from the current body of literature the relative importance of the observed differences in duplicability, or to determine which features are important in their own right rather than as passengers due to correlation with other features. For example, the bias towards neural functions/nervous system expression in teleost ohnologs appears to be the driving force in sequence constraint differences between ohnologs and genes not retaining duplicates from the WGD event (Roux et al., 2017). In chapter 3 - 'Genes permitting different modes of duplication differ across a wide range of features', we investigate these questions with the aim of estimating the relative importance of specific gene features in determining gene duplicability.

1.4.7 Contribution of factors to gene duplicability

The study of the observed biases in duplicate retention could tell us much about the process of gene duplication and its role in evolution. Determining the precise reason for the relationship between a given feature and duplicability, be it a direct link to the process of duplication or retention or a confounding relationship, is informative in deciphering the mechanism of how genes duplicate and are either retained, gradually lost or rapidly purged. For example, in chapter 4 -

'Evidence from *Drosophila* supports higher duplicability of faster evolving genes', we investigate the relationship between faster rates of evolution and duplicability. However, it is unlikely that any one feature of a gene is fully deterministic in its retention in duplicate, with the duplicability of a gene likely a function of a number of factors. Furthermore, in all likelihood even fully characterising a given gene will not allow us to predict with certainty whether or not the gene will duplicate within a given lineage at a given point in time. Despite the importance of determining the relative contribution of all relevant factors in utilising biases in gene duplication to better understand the process, little conclusive work exists on the topic. However, we can speculate on possible contributing factors to duplicability which require further study. In addition to the factors contributing to the intrinsic duplicability of a gene, we should also consider extrinsic factors, for example the genomic or environmental context in which a duplication occurs, or the duplication history of the gene itself.

The relationship between duplicate retention and environment is of particular interest given the assumed role of duplication in adaptation and the potential requirement for environmental stressors in providing more accommodating conditions for polyploid establishment. It is unclear if duplicable genes are necessarily duplicable in the absence of the need for adaptation, in the sense that they may not be successfully retained in duplicate if the requirement for innovation within their specific functional area has not arisen. For example, drought resistant genes may be retained in duplicate in the case that this is the prevailing stress at the time of duplication, but would obviously have little cause to do so when flood tolerance is required. While we might expect these effects to be transient (i.e. duplicate copies would be lost again if there were a change in prevailing stress. However, this is dependent on whether the stress itself is transient and, as proposed for dosage balance and subfunctionalisation, these initial requirements may act as an initial

preserving force that allows duplicates the time to escape redundancy. Work in plants contributes to this idea of environmentally driven duplicate retention, with Wu et al. (2020) finding consistent functional patterns across retained genes from independent WGD events occurring within the same timeframe.

The genomic context in which a duplication takes place is also likely to be relevant to the likelihood of duplication. Evidence has shown that the background ploidy in which a duplication occurs has an impact on the phenotypic consequences of the duplication, likely due to differing relative impacts of a single gene gain or loss in different base copy number states (Birchler and Veitia, 2012). There is also evidence of differences in the impacts or resolution of WGD depending on lineage-specific genomic features. For example, *Paramecium* is hypothesised to be more permissive of WGD due to the unusual genome structure in this lineage (McGrath et al., 2014). The macronucleus, which acts as the somatic nucleus from which transcription occurs, is normally highly polyploid. This polyploidy originates from amplification of the inherited micronucleus which is transcriptionally silent. As normal function of this genome requires a programmed amplification of a genome to a suitable ploidy, mechanisms exist in *Paramecium* for sensing genome copy number. These mechanisms might mask the increase in dosage that occurs in WGD by altering the amplification to accommodate it, i.e. a duplicated genome would be sensed as twice as much DNA and be amplified half as much as usual. Initially then, only the 'germline' genome of the organism, which does not participate in normal cellular processes, would show any noticeable change. Such genome-specific quirks impact the overall influence of WGD and should be investigated further for impacts on duplicate retention and evolution

Finally, the process of duplication itself could act to impact the likelihood of any further successful duplications. If the specific route to retention followed by the paralogs after a single duplication is not replicable, for example if the gene

function has been divided through subfunctionalization and can not be further subdivided then the feature of the gene that facilitated its duplicability has been spent. In this way, it may be that duplicability is a consumable property in some cases and this should be taken into account in assessments of duplicate behaviour.

The question of the relative contributions of these factors to duplication outcomes is really a question of how consistent duplicability is as a gene property. If duplicability for a given gene is a constant across time, duplication events and diverse contexts, then the majority contributor to duplicate retention will be factors intrinsic to the gene itself. Deviations from this scenario will indicate an impact from the extrinsic factors just discussed. In chapter 5 - 'Consistency of gene duplicability varies across the genome' we present an initial investigation into the consistency of post-WGD behaviour.

1.5 Aim

This thesis examines the concept of gene duplicability as an intrinsic feature of genes across multiple lineages. Chapter 3 investigates differences between genes duplicable by SSD vs WGD and attempts to assign their relative importance. Chapter 4 delves into differences specifically in evolution rate between single copy and duplicable genes and shows these differences could be related to duplicability rather than a consequence of duplication. Finally, chapter 5 examines the idea of consistency of gene duplicability as a property by considering groups of genes which are consistent in their retention pattern across multiple genome duplication events.

Chapter 2

Materials & Methods

This chapter provides an introduction and detailed background to some of the key methods used within this thesis.

2.1 Ortholog and paralog detection

Determining homology (i.e. evolutionary relatedness) between sets of genes is a key part of any evolutionary analysis. Here, the definition of homology, and particularly paralogy (relatedness through a duplication) vs. orthology (relatedness through a speciation), is a critical element. Without reliable paralogy assignments, attempting to study duplicability for a gene would be largely meaningless.

2.1.1 How is homology determined?

Homology inference tools can be broadly divided into clustering/graph based methods and tree-based methods (Altenhoff et al., 2019). Certain tools also use a combination, using graph-based methods as a basis to then build gene trees, as this is less computationally costly with increasing numbers of genomes to be considered, and there also exist 'meta-method' tools which leverage assignments

from multiple approaches to generate more robust orthology calls.

Graph-based methods such as InParanoid and OrthoMCL use sequence similarity, estimated using BLAST (Camacho et al., 2009) for example, to measure how closely related a pair of sequences are and build a graph of relatedness based on this. A clustering algorithm is then applied to group the sequences in the graph into orthologous groups. In contrast, tree-based methods rely on inferring whether the nodes in a gene tree represent speciation or duplication events through reconciling the gene tree with the species tree. Gene trees differ from species trees through a variety of processes acting at the single-gene level such as duplication, gene loss etc. and reconciliation methods aim to resolve differences between the two by inferring duplication and loss events. The reconciled tree is usually inferred under either parsimony or maximum likelihood.

Of the orthology inference methods used within this thesis, all are at least partially graph-based. Orthofinder (Emms and Kelly, 2019) uses BLAST scores to build a similarity graph which is then clustered using MCL to cluster to generate orthogroups. These groups are then used as input to build gene trees, build a species tree and perform reconciliation in order to give the final set of homologous relationships. OMA (Train et al., 2017) performs Smith-Waterman alignments for all possible pairings in the gene set and uses these scores to select sufficiently similar pairs for further analysis, then applying a clustering algorithm.

2.1.2 Potential issues with homology assignment

Systematic errors exist in the inference of orthology (Natsidis et al., 2021), particularly pertaining to genes with features that predispose them to faster decay of similarity scores. Faster-evolving, shorter sequences quickly fall below significance thresholds. This effect is magnified in more diverged species as orthology also becomes more difficult to assign with increasing phylogenetic distance.

The increase in error with faster-evolving genes is particularly pertinent to the definition of paralogs. Paralogs complicate the picture for orthology inference, it is not enough to know that the sequence is similar to another within the same genome, it must be determined whether the duplicates represent in-paralogs (those originating from duplications within the species set under consideration) or out-paralogs (those whose duplication pre-dates the earliest species divergence). This determines whether the paralogs should be placed within the same orthogroup or form two separate ones.

An additional issue is that paralogs diverge more quickly from each other than do orthologs (Altenhoff et al., 2012), perhaps as a result of the requirement for some duplicates to diverge from each other in function to escape redundancy and loss. In particular, it might be expected that paralogs would sometimes diverge from each other asymmetrically, the pattern expected in some cases of neofunctionalization as one copy gains a novel function and the other preserves the previous one. This poses a particular problem for orthology inference as, in sequence similarity terms, one paralog now appears much more similar to its orthologs than its sister. This leads to false assignments of out-paralog status as shown in Figure 2.1 (Train et al., 2017) and also false assignments as the faster evolving paralog would be determined to be paralogous to the other genes in the orthogroup through an ancestral duplication when, depending on the timing of the duplication, it may be paralogous only to its paralog within the same genome and orthologous to the rest of the group.

Different orthology inference tools take different steps to circumvent these issues. Orthofinder performs normalisation of similarity scores by sequence length and phylogenetic distance prior to determining the pairs to be included in the graph-building and clustering steps, thus avoiding the dependence between similarity and species relatedness and the disadvantaging of shorter sequences. OMA

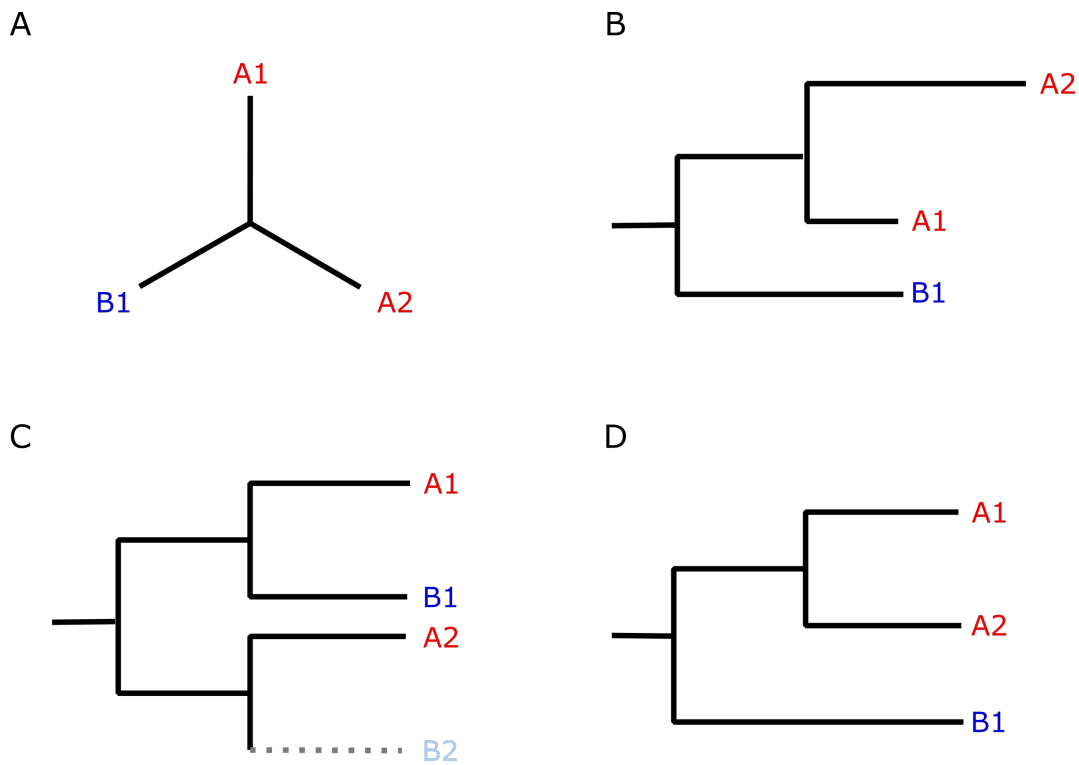


Figure 2.1 | Asymmetrically diverging paralogs are prone to orthology assignment errors. (A) An example of an unrooted gene tree for species A and B containing a duplicate pair in A. (B) The true tree topology, showing a faster rate of evolution in the A2 paralog. (C) and (D) Possible scenarios given the unrooted tree in (A). (C) shows the case where A1 and A2 originated from an ancestral duplication, following by loss of a paralog in B. (D) shows the case where A1 and A2 resulted from a lineage specific duplication in A. Asymmetric evolution in the paralogs causes the orthology assignment to lean towards the scenarios presented in (C) as A1 appears closer to B1 than A2 does. This separates A1 and A2 into separate orthogroups and would give the false conclusion that these genes had not undergone duplication since the A-B divergence. An additional false conclusion would be that A2 and B1 are paralogous rather than orthologous.

applies tolerance limits when determining which pairs are significantly similar to include to avoid loss of paralogous relationships and additionally includes a verification step to remove putative orthologous pairs which are actually paralogous. These pairs arise when, through gene loss, genes which are actually related through a duplication event appear to be orthologs. OMA's verification step includes a 'witness of non-orthology', a third species which has retained both copies and so can be used to determine the correct relationship.

2.2 Determination of duplication type

As we consider differences in duplicability between different modes of duplication in the course of this work, the correct classification of duplicates by their mode of creation is almost as important as the recovery of the pairs themselves. Ohnologs are typically distinguished from duplicates arising from SSD through a combination of syntenic analysis and comparison to outgroups which have not undergone the WGD in question. Segments of the genome which are duplicates of each other through a WGD will typically show some degree of conserved gene order and this can be used to infer whether a pair of paralogs have originated in a WGD event. An alternative method is to use Ks distributions between paralogs to identify peaks where a WGD has taken place, however these are not straightforward to apply to older events such as the vertebrate 2R or teleost 3R due to saturation of Ks over time (Tiley et al., 2018). All ohnologs sets used here are based on conserved synteny (detailed below).

2.2.1 Use of existing 2R ohnolog data

In the case of identifying ohnolog pairs from the vertebrate 2R duplications, we used a number of pre-existing datasets. The human ohnolog dataset from Makino

and McLysaght (2010) was defined through a combination of clustering vertebrate genes which shared the same best hit in a tunicate outgroup, filtering this set and building 'paralogons' - blocks of conserved gene order that remain from the genome duplication despite genomic rearrangements over time - to determine ohnologous relationships between genes.

Singh et al. (2019) (used here for both human and spotted gar ohnologs in Chapters 3 and 5 respectively) use a similar method (initially developed in Singh et al. (2015)), involving identification of candidate ohnolog pairs from syntenic comparisons both with invertebrate outgroups and within vertebrate genomes, filtering these by age and assessing the statistical significance of the syntenic blocks. This yields multiple ohnolog sets of different confidence levels: 'strict', 'intermediate' and 'relaxed'.

Our final set of human ohnologs were defined using syntenic segments within the human genome from Nakatani et al. (2021). These segments had previously been defined as originating from the 2R WGD through reconstruction of the proto-vertebrate and proto-gnathostome genomes using a probabilistic macrosynteny model. We defined paralog pairs in this case as ohnologs if they (a) linked two macrosyntenic segments and (b) existed in a block of at least 3 linking pairs with no more than 8 genes separating any two pairs.

2.2.2 Definition of ohnologs in bony fish dataset

We defined ohnologs within the bony fish dataset using the software i-ADHoRe (Proost et al., 2012). No pre-existing dataset encompasses all the WGD events which were considered and so all were considered together to minimise any variance across events introduced by differences in method of identification. i-ADHoRe functions by initially identifying collinear blocks between an initial two segments of genome. This is achieved through finding significant diagonal clustering when

the gene orders of the two segments are plotted against each other. The segments are then aligned and this combined data compared to a new segment, iteratively building up a 'profile' of collinear blocks from which ohnolog pairs are determined.

Chapter 3

Genes permitting different modes of duplication differ across a wide range of features

3.1 Introduction

The duplicability of a gene, i.e. the likelihood of its duplication and retention, differs vastly depending on the mechanism of duplication. In fact, most genes with retained WGD duplicates rarely successfully duplicate by SSD and vice versa (Makino and McLysaght, 2010), although this is not an absolute rule and does not obviously operate equally in both directions. In addition to the overall opposing patterns in gene retention between ohnologs and SSDs, the longevity of duplicates differs greatly, with many ohnologs being retained in the long term while a vast majority of duplicates that arise through SSD are lost as quickly as they arise (Lynch and Conery, 2003). These striking patterns imply that there is much to be learned regarding the mechanisms of duplicate retention in comparing the two groups. Can we define the type of gene likely to successfully duplicate by each

mechanism and what can this tell us about the evolutionary processes at play?

There are a number of known features in each duplicate group which lend themselves well to explaining the observed differences. Ohnologs are known to be enriched for functions associated with constrained dosage such as development, neural processes, regulation and signalling (Blomme et al., 2006; Brunet et al., 2006) and their evolutionary patterns and response to copy number variation support the hypothesis of dosage balance (Makino and McLysaght, 2010; Rice and McLysaght, 2017a). Under the dosage balance hypothesis (Birchler and Veitia, 2012) the WGD event is not disruptive as it does not cause a change in relative dosage across dosage balanced genes (1:1 becomes 2:2), while a subsequent loss of the duplicate copy in a dosage balanced gene would (2:2 becomes 2:1), creating an immediate preservation constraint. SSD of dosage-balanced genes leads to a similar imbalance and would therefore create purifying selection against such duplicates. Similarly, genes which are able to duplicate through SSD likely do not possess these dosage constraints and therefore have no pressure for retention post-WGD and are eventually lost.

While this is a satisfying explanation for the contrasting retention in these groups, genes do not fall exclusively into one category or the other e.g. some genes retained following WGD also experience SSD. Additional differences in the gene sets may therefore also play a role in the overall duplicability of a gene. A broad range of features have been compared in the existing literature, encompassing basic gene features such as gene length and sequence composition (Banerjee et al., 2017; Chapman et al., 2006; Jiang et al., 2013; Qiao et al., 2018; Zhu et al., 2013), to regulation and conservation related measures (Amoutzias et al., 2010; Guan et al., 2007; Hakes et al., 2007; He and Zhang, 2005; He and Zhang, 2006; Keller and Yi, 2014), with a general pattern emerging that ohnologs are large, complex and highly constrained genes, SSDs are shorter and less constrained,

and singletons, which are not observed as successful duplicates, fall somewhere in between in most cases.

Despite this wealth of known comparisons, there are some issues with using the existing body of work to create an overarching picture of what kind of gene may successfully duplicate by either mechanism. Firstly, current work is carried out across multiple diverse lineages with a bias towards plants and yeast. There is comparatively little work comparing vertebrate ohnologs with SSDs and it is not guaranteed whether conclusions can be extrapolated across species and/or whole genome duplication events. A second issue is that features have often been compared in isolation with a given study examining only a small subset. As studies are not generally consistent in the organism considered or the definition of duplicate groups, it is difficult to quantify any kind of relationship or relative importance across features which have not been considered together within one piece of work.

Here, we compare human duplicate groups across a broad range of gene features. We investigate possible confounders for the observed differences and quantify relationships between features before finally estimating their relative importance in the definition of each duplicate group.

3.2 Materials & Methods

3.2.1 Data sources and processing

Paralog and ohnolog sets

A list of 19,548 human protein coding genes, as well as a list of human paralog pairs with an estimate of duplication timing (last common ancestor) was obtained from Ensembl v99 (Yates et al., 2019). Regions of macrosynteny originating from

the 2R vertebrate genome duplication were obtained from (Nakatani et al., 2021) and used to define ohnolog pairs. These pairs were defined as Ensembl paralog pairs that linked macrosyntenic regions and existed in blocks of at least 3 linking pairs with no more than 8 genes separating any two paralogs on the same segment. This ohnolog set is referred to as '2021 ohnologs'. Additional ohnolog datasets were obtained from Makino and McLysaght (2010 and Singh et al. (2019) for cross referencing to generate the final duplicate category assignments.

Considering each ohnolog set separately, a paralogous pair was designated as originating from WGD if the pair was present in the list of ohnolog pairs, and as SSD if not. Pairs designated as SSD were further examined for evidence of retroduplication, as this mode of duplication is outside the scope of this work and it can not be assumed that these pairs behave similarly to SSDs. Paralog pairs were defined as potentially retroduplicated if one member of the pair has zero introns while the other has 3 or more, or if at least one member of the pair has zero introns and the other has less than 3 *and* there is no conserved microsynteny between the duplicates. Microsynteny was defined as having at least one other paralogous pair linking the surrounding region (within 5 genes either side of the genes being tested), following from methods used in Jun et al. (2009) (Figure 3.1). The additional microsynteny check was used for low intron genes as there is a possibility that the second member of the pair is a tandem duplicate and reached zero introns through intron loss rather than retroduplication or, in the case where both members of a pair have zero introns, that the original parent gene had zero introns prior to duplication. The possibility of intron gain in a retroduplicated pair was not considered as intron gain is rare and losses typically outnumber gains (Roy and Penny, 2007). Genes were assigned their mode of duplication status based on the pairs they are present in. Ohnologs are present only in WGD pairs, SSD genes only in SSD pairs and genes not present in any paralogous

pairs within the vertebrate lineage designated as singletons. Singletons can be further divided into genes with no paralogs at all and genes with no paralogs since the vertebrate divergence (duplication node as given by Ensembl is Chordata or earlier). Genes belonging to any pair classed as resulting from retroduplication are excluded, as are genes with pairs belonging to more than one duplication type. The final classification for each gene was based on 'majority rules' between the three ohnolog datasets.

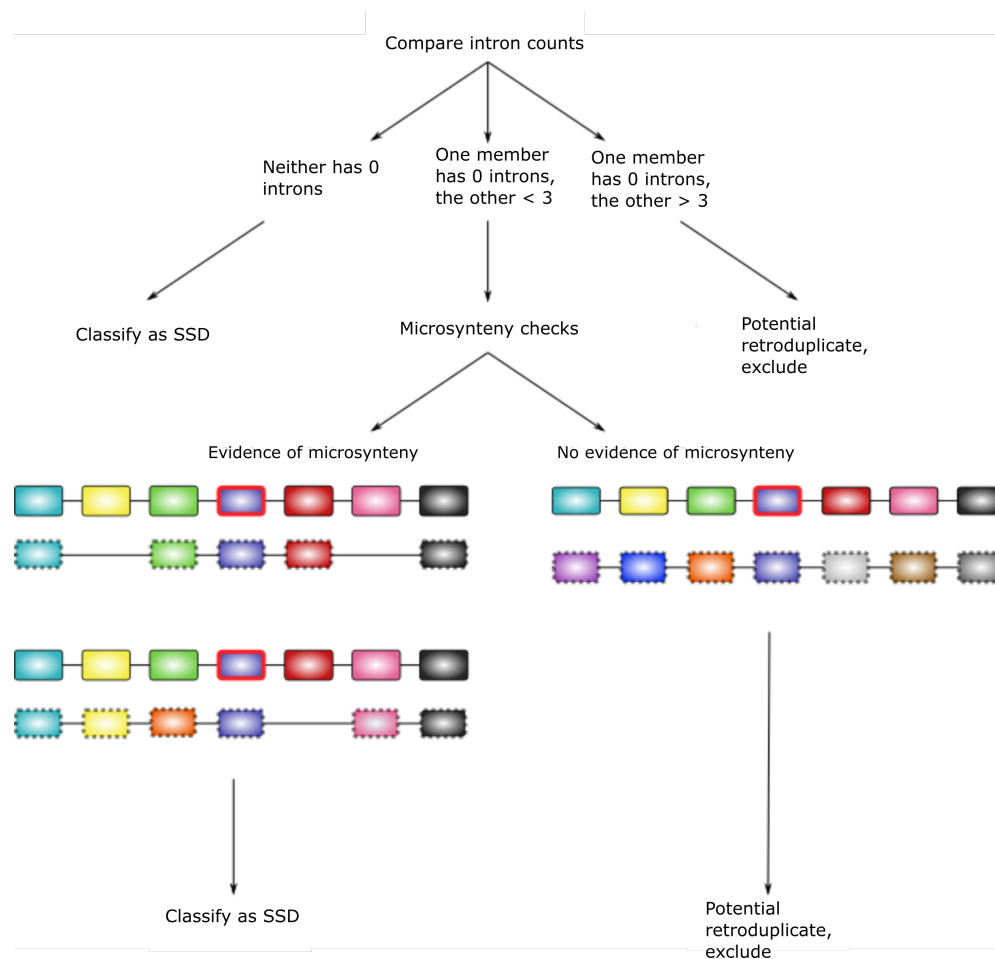


Figure 3.1 | Exclusion of retroduplicates. Gene of interest is highlighted by a red outline with neighbouring genes shown on either side. Paralogs of a given gene are shown in the same colour but with a dashed outline. There is no paralogous relationship between genes shown in different colours.

Measures of gene essentiality

A measure of cellular essentiality, the CRISPR score, was taken from Wang et al. (2015). This score is derived from a proliferation screen defined as the average $\log_2(\text{fold-change})$ in the abundance of all sgRNAs from the library used which target a given gene, i.e. the change in sgRNAs causing disruption to a gene. The greater the decrease in the disruptive sgRNAs, the more essential a gene is, so smaller CRISPR scores indicate greater essentiality. The minimum CRISPR score across the four cell lines in the dataset (i.e. the maximum essentiality) was taken. We use the negative of this score and refer to it as ‘cellular essentiality’.

We also examine a number of population-based measures of gene essentiality. These metrics differ from the CRISPR score as they use modelling to estimate the expected frequency of mutation or loss of function in a given gene and define essentiality based the difference between this expected frequency and that observed. We examine a number of these metrics including pLI (Lek et al., 2016), Phi (Bartha et al., 2015), LoFtool (Fadista et al., 2016), S_{het} (Cassa et al., 2017), RVIS (Petrovski et al., 2013), EvoTol (Rackham et al., 2015) and missense Z score (Samocha et al., 2014).

Of these metrics of essentiality, each has distinct features but there are some commonalities in the statistical methods they employ. RVIS and EvoTol both estimate residual variance after regressing the number of common variants on total number of variants for a given gene, thus defining genes as essential (intolerant of variation) if they have fewer common variants than expected given the total functional variance for the gene. Although they share this framework, these metrics are distinct in the data used with RVIS considering variants using only information from humans, while EvoTol incorporates whether a variant is ‘damaging’ across evolutionary time. Similarly, Phi, pLI and S_{het} all use estimates from probabilistic models (Poisson mixture models in the case of Phi and pLI, Bayesian

in the case of S_{het}) to measure tolerance of mutation, estimating probability of haploinsufficiency, probability of LOF intolerance and the selection coefficient respectively. The missense Z score quantifies the deviation in missense mutation frequency relative to the expectation of a neutral mutation model, making it the only metric included here which focuses fully on sequence changes rather than including mainly truncating variants. Finally, LoFTool is the only metric included which is built on multiple of the others, combining information from both EvoTol and the missense Z score to consider a breadth of functional information.

Expression and expression specificity

Gene expression levels (median values, in TPM) were obtained from GTEx (v8) (GTEx Consortium, 2020). Expression data for differing developmental stages was obtained through Expression Atlas (fetal expression data from FANTOM5, expression at various stages of prenatal brain development from the Human Developmental Biology Resource and expression at various stages of development, project numbers E-MTAB-3358, E-MTAB-4840 and E-MTAB-6814 respectively).

Expression specificity was calculated as the tissue specificity index τ which is given by $\tau = \frac{\sum_{i=1}^N (1-x_i)}{N-1}$ where N is the number of tissues and x_i is the expression value for the i^{th} tissue scaled by the highest expression value for the gene. τ essentially represents the average difference across tissues from the maximal gene expression, scaled relative to the maximal gene expression; a value of 0 indicates a housekeeping gene while a value of 1 indicates a tissue specific gene. For the purposes of this calculation, different development stages were treated as separate tissues. This measure of tissue specificity was determined to be the most robust when benchmarked by Kryuchkova et al. (2016).

Other features

Protein-protein interactions (PPIs) were obtained from the Human Interactome (Luck et al., 2020). CAI (codon adaptation index) and proportion of residues in IDRs (intrinsically disordered regions) were determined from CDS sequence using IUPred (Mészáros et al., 2018) and CAIcal (Puigbò et al., 2008) respectively. All other features used were obtained from Ensembl biomart or Ensembl API (or derived from data available through these sources). This included genomic length, CDS length, evolution rate (dN/dS with macaque orthologs), number of introns, average intron length, intron coverage, regulatory motifs, protein domains, unique protein domains (from domains), % GC content, % GC3 content (determined from CDS sequence from Ensembl). In cases where a feature may differ between different transcripts/protein products of a gene, the value for the longest transcript was used.

3.2.2 Statistical methods

Two-tailed Mann-Whitney U tests were used for all direct comparisons of features between duplicate types and between duplicates and singletons. P-values were Bonferroni corrected for multiple testing when applicable.

Depletion and enrichment of GO terms in each category was determined using gProfiler (Raudvere et al., 2019) using FDR to control for multiple tests with a threshold of 0.05 and using the total gene set as background.

Regression models

Regression models were built to investigate if duplicate differences were explained by differences in the age of the duplicate pairs. Models were built using OLS (ordinary least squares) regression with the statsmodels Python package (Seabold and Perktold, 2010), data transformations were selected from none, log trans-

formation and Box-Cox transformation according to which minimised deviation from normality (lowest Jarque-Bera test statistic) and whether or not to include an interaction term was determined based on whether adding the term yielded a significant improvement in AIC (a decrease of 2 or more units). Final formulas used for each feature are given in Table 3.2. Duplicate age here is intended to measure how long-lived duplicates are and represents the oldest duplication node for a given gene from Ensembl, one age unit corresponds to 50MY of divergence time using time estimates from TimeTree (Kumar et al., 2017).

Random forest

A random forest classifier for determining feature importance was constructed with all SSD and WGD genes with values available for all features considered using the scikit-learn Python package (Pedregosa et al., 2011). This method was selected over, for example, a regression model as it is a better choice in cases where there may be complex interactions between features.

Hyperparameters (i.e. parameters of the model set prior to training) for these classifiers were determined from a randomised search with 10-fold cross validation set to maximise F1 score, with a model trained on the selected gene features. Following this step, the estimated best hyperparameters were the same as defaults except in the case of the number of estimators (39 rather than the default of 100). Classes were weighted (`class_weight='balanced'`) to account for an imbalance in class frequency (3,374 WGD to 1,236 SSD). The dataset was split 80-20 for training and testing.

Feature importances were calculated using the `rfpimp` python package (Parr and Turgutlu, 2020), using permutation importance as the method of calculation. This method was chosen over the default feature importances from `sklearn`, which are based on mean impurity decrease, as this method can be biased by variable

scale or number of categories (Strobl et al., 2007). Permutation importance records the drop in accuracy caused by randomly permuting each feature relative to a base accuracy.

Model training and feature importance calculation was repeated 100 times using the final hyperparameters and feature set with each pass using different randomly selected training and validation datasets in order to estimate how variable the importance rankings were due to randomness in the model.

To control for the effects of duplicate age differences, another model was constructed and importances calculated in the same manner using residuals from regressing each feature on age (equation of the form $feature \sim age$) as input under the assumption these values capture variation in the feature not explained by age variation. Additional models were also constructed using scaled data (Z-scores for each value) in order to check for any effects of differences in scale between features on the importance estimates.

For groups of correlated features, each feature in the group was considered in isolation in order to obtain an importance estimate independent of any compensation from other, correlated features. The same model was constructed but with all but one feature in a given group dropped and the mean importance and accuracy estimates across 100 iterations in the same manner as above. 'Previous rank' were assigned based on the position in the ranked list of mean importance estimates in the model including all features. 'New rank' was assigned based on the position the new mean importance would place the feature under consideration within the previous ranked list. Rank and mean accuracy comparisons are given in Table 3.3 - Table 3.6 for all four datasets.

Table 3.1 | Duplicate classification by dataset.

Category	2021 ohnologs	Makino & McLysaght	Singh et al.
WGD pairs	13,934	7,074	6,799
Non-WGD pairs	124,274	131,134	131,409
<i>Excluded pairs:</i>			
Retroduplicated pairs	22,065	23,177	23,135
Pre-vertebrata duplications	83,668	86,692	87,010
Presumed SSD pairs	18,541	21,265	21,264
WGD genes	3,111	5,814	5,738
SSD genes	3,844	3,180	3,134
Singletons	10,398	8,538	8,697
Excluded genes	2,195	2,016	1,979
<i>Final gene counts:</i>			
WGD	5,327		
SSD	3,352		
Singleton	8,932		

3.3 Results

3.3.1 Structural and sequence-based features: ohnologs are longer and more structurally complex than SSDs or singletons

We classified 19,548 human genes as one of singleton (no duplicate copies detected within vertebrates), SSD paralog, or WGD paralog based on the consensus of duplication annotation across three published datasets (Table 3.1; see Methods). For stringency, we assessed paralog pairs using three ohnolog datasets and made the final assignment based on the majority classification for a given gene across the three. Across all datasets singletons make up the largest class, with SSD and WGD genes more evenly split.

Various sequence composition and gene structure based features have been studied previously in relation to duplicate category differences, mainly in the

context of yeasts and plants. Most agree that WGD genes are longer, in terms of total genomic length, than others (Jiang et al., 2013; Zhu et al., 2013) and that WGD genes possess more structural domains (He and Zhang, 2005; Jiang et al., 2013). For other features the trend is less clear with some reports claiming protein coding sequence length is higher in WGD genes (He and Zhang, 2005; Jiang et al., 2013) but others finding no difference (Qiao et al., 2018). Similarly contradictory results were observed for %GC3 content (G+C content at third codon positions) in two different plant genomes (Jiang et al., 2013; Qiao et al., 2018). Other features have not been examined multiple times and so consistency can not be judged, for example intrinsic disorder (highest in WGD genes, lowest in SSDs - Banerjee et al. (2017)) and average intron length (higher in genes retained post-WGD - Jiang et al. (2013)).

Our analysis confirms the longer length of WGD genes. Notably, length-related features such as genomic length, CDS length and intron count are among the most dramatically different between WGD and SSD genes, along with the number of domains/unique domains, followed by further intron-related features such as % intron coverage (Figure 3.2). Sequence composition features such as %GC3, codon usage and intrinsic disorder also differ between the duplicate types but much more weakly; both sets of duplicable genes seem to differ more from singletons than from each other in these cases.

There are multiple possible explanations for this difference in lengths between the duplicate categories. One possibility is that the mechanics of tandem gene duplication by non-allelic homologous recombination may bias the size of gene likely to undergo duplication by this method, as larger genes with more functional elements may be less likely to successfully duplicate in their entirety. However, the existence of large segmental duplications that have occurred throughout primate evolution (Bailey and Eichler, 2006) suggests that this should not be a significant

factor. The longer length of WGD genes is, however, consistent with a 'zero sum game' mode of duplicability (Rice and McLysaght, 2017b) which implicates the expression costs of duplicating longer genes as a limiting factor.

3.3.2 Regulation-based features: ohnologs are more broadly expressed and show more complex regulation

Previous work on features which impact the regulation of gene expression and interactions gives a clear picture that WGD genes occupy a considerably more complex regulatory context than other groups. This is true on both the RNA level (higher *cis*-regulatory motifs and number of transcription factors binding (He and Zhang, 2005), differences in promoter methylation (Keller and Yi, 2014)) and on the protein level (more post-translational modifications (Amoutzias et al., 2010)). These extra layers of regulation may reflect the network properties of WGD genes; they are more likely to be members of complexes and interact more with the rest of the proteome, occupying densely connected parts of metabolic networks (Conant, 2020; Hakes et al., 2007; Zhu et al., 2013). WGD genes have also been observed to be more highly expressed and paralog groups have been observed occupy distinct expression domains in terms of tissue or developmental stage depending on their mechanism of creation. (Conant, 2020; Jiang et al., 2013; Session et al., 2016).

The most striking difference across categories for our comparisons is that seen for expression specificity, with SSD genes showing much more restricted expression than either of the other two groups (Figure 3.3). Isoform count and maximal expression level also show robust differences across all comparisons, with WGD genes showing highest expression and number of isoforms and SSD genes lowest for both features. Differences in the number of regulatory motifs and protein-protein interactions (PPIs) are comparably small, though WGD genes possess

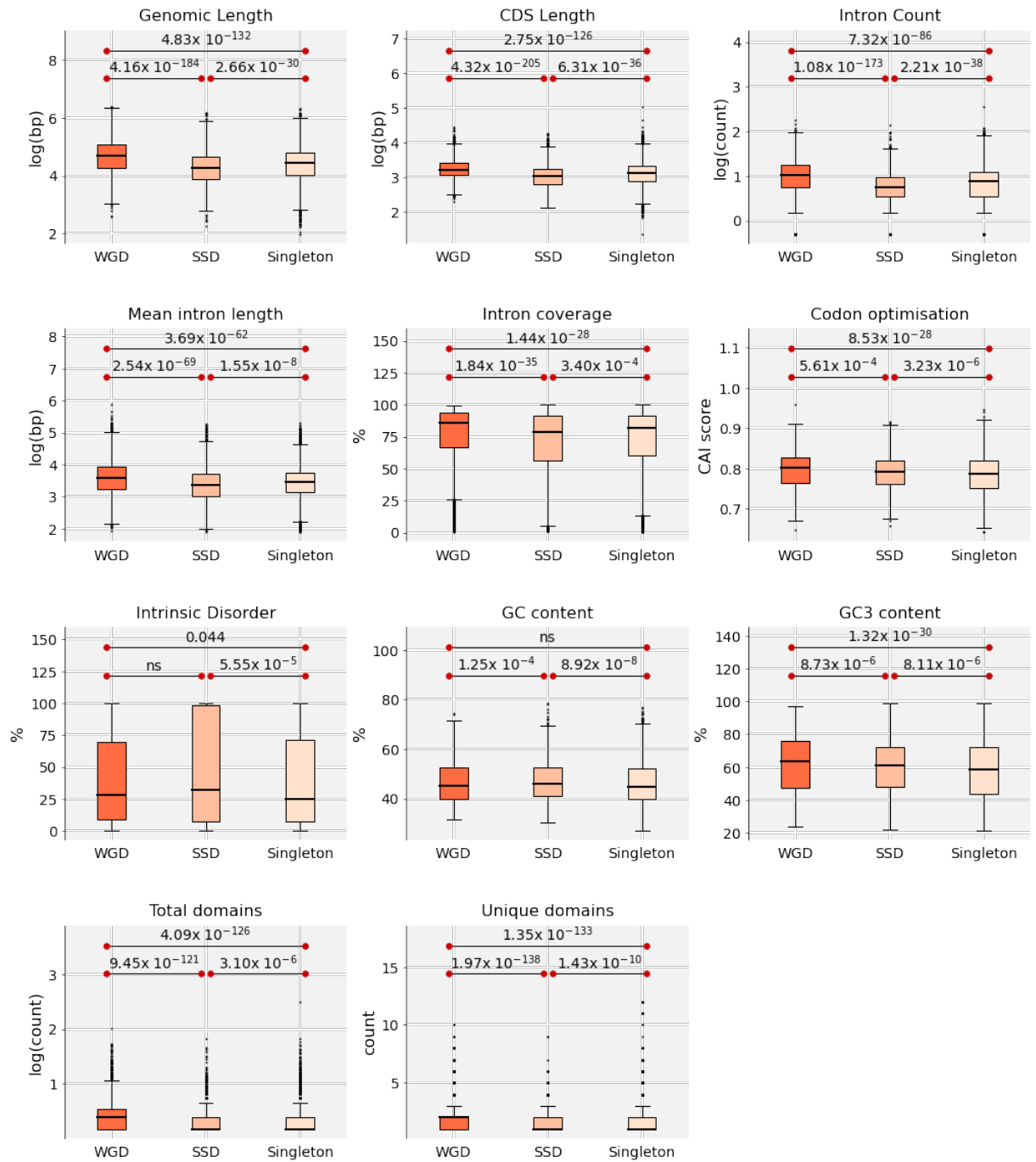


Figure 3.2 | Sequence based feature comparisons. Comparisons for features classed as related to genomic sequence or gene structure. For each feature, different categories of gene were compared using the Mann-Whitney U test. All P -values are Bonferroni-corrected, given above the line connected the groups compared.

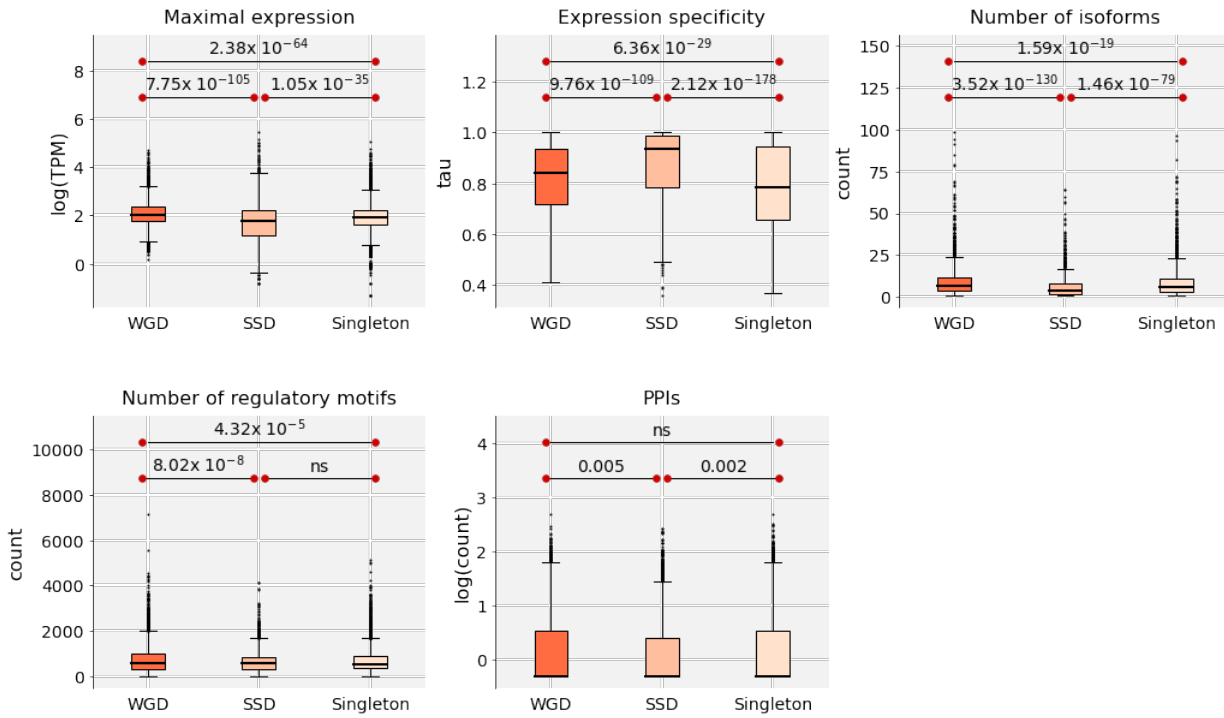


Figure 3.3 | Regulation related feature comparisons. Comparisons for features relating to gene features affecting gene expression or behaviour carried out using the Mann-Whitney U Test. All P -values are Bonferroni-corrected, given above the line connected the groups compared.

significantly more regulatory motifs than the other two groups and SSDs possess fewer PPIs. The weak difference for PPIs is potentially an artefact of the dataset used (see Discussion).

As SSD genes show a strong bias towards narrow expression, we chose to investigate expression level on a per-tissue basis to determine if it is generally true that WGD genes are more highly expressed than SSD or if SSD genes overtake in the tissues where their expression is concentrated. (Figure 3.4). Even with the removal of genes expressed at under 1TPM (transcripts per million) in each tissue, the median expression of WGD genes is always higher than SSD genes, suggesting WGD genes are the more highly expressed group regardless of tissue context.

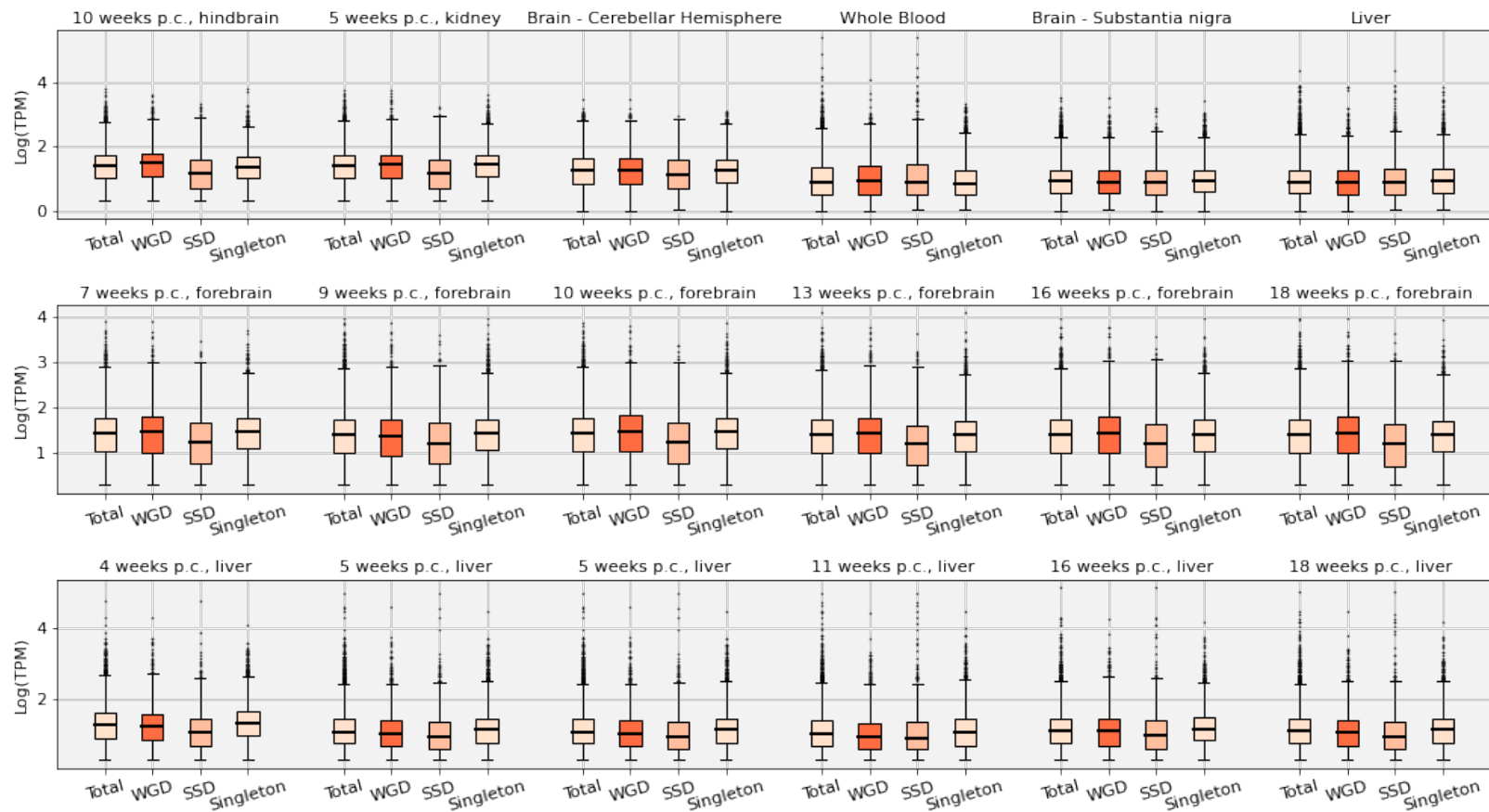


Figure 3.4 | Comparison of expression at the tissue level. The top panel shows representative tissues with the highest and lowest difference in median expression between ohnologs and SSDs. The lower two panels show change over development time for a tissue that shows large differences (forebrain) and one which shows a comparatively small difference (liver). 'Total' refers to the distribution of all genes in the genome.

3.3.3 Conservation and essentiality related features: ohnologs are slower evolving and less dispensable

Conservation and constraint may be the most studied category of features with respect to duplication type differences, though it is usually assessed in the context of paralog pair divergence post-duplication rather than as a broader property of the genes undergoing duplication. That said, the broad trend does show WGD genes to be more constrained in that context: WGD pairs overlap more in function and are better able to complement each other (Guan et al., 2007; Hakes et al., 2007; Zhu et al., 2013); WGD genes evolve more slowly than others (Brunet et al., 2006; Qiao et al., 2018); and also are more likely to have synonymous substitutions than non-synonymous (Chapman et al., 2006). This might at first appear to contradict the general trend showing duplicates as a group to be less essential than singletons (Gu, 2003; He and Zhang, 2006). However, vertebrate WGD genes are more essential than SSD (Makino et al., 2009).

We examined evolutionary constraint in the duplicate groups by comparing evolutionary rate, population genetics based metrics (missense Z score, pLI, Phi, S_{het} , RVIS, EvoTol and LoFtool - see Methods), as well as direct estimation of essentiality in cell lines (Wang et al., 2015). Across comparisons, evolutionary rate, missense Z score and RVIS are among the most distinct between duplication categories (Figure 3.5). These population-based measures represent two of the three measures which include missense mutations in the estimate rather than only nonsense, frameshift or splicing mutations, thus are consistent with the observations regarding evolutionary rates. This suggests that these groups differ in essentiality not only in terms of avoiding loss of function but also in terms of a more general sequence constraint. Also relevant is that the largest differences for other metrics are seen when comparing WGD genes to either SSD genes or singletons, but not when comparing SSD genes to singletons. For example, this

pattern is clear for both pLI and Phi, metrics are used in defining haploinsufficiency. Haploinsufficiency is a well-known form of dosage balance requirement; this perhaps indicates a difference in dosage balance between WGD genes and other groups but not between the groups not retained post-WGD.

Two features show unusual patterns relative to the others: cellular essentiality and EvoTol. Cellular essentiality indicates both groups of duplicable genes are similarly essential, and singletons are the most essential group. As this is the only metric based on cell lines rather than on population allele frequencies, we might infer that this disagreement indicates the functions which cause WGDs to be constrained in a population are not required in the context of a single cell e.g. developmental functions. This would also explain the results in previous work indicating all duplicates are less essential than singletons. Much of this work is carried out in yeast, a context where this result may well be true, while it does not hold in the context of vertebrate ohnologs.

It is less clear why EvoTol indicates all groups are similarly tolerant to mutation. It is the third population-based measure that uses missense mutations, as such we might expect it to behave similarly to the missense Z score and RVIS, however it adds the additional step of classifying variants as 'damaging' or 'tolerated' based on broader evolutionary constraint. It is unclear why this additional information would cause this disagreement with the other metrics, given the overall slow evolutionary rate of WGD genes.

In the case of evolutionary rate we use the ratio of non-synonymous mutation rate to synonymous mutation rate (dN/dS). Although this value is not a 'rate' in strict terms, but rather a ratio of two rates, convention is to refer to it as measuring the rate at which a gene evolves and so we label it 'evolutionary rate' here. More correctly, this metric represents the mode of selective pressure a gene is evolving under, such that values over 1 (excess of non-synonymous changes)

indicate positive selection while values under 1 (depletion of non-synonymous changes) indicate negative selection. Here we interpret lower values of dN/dS as indicating greater evolutionary constraint or 'slower' evolution. In this context we determine that ohnologs and singletons are slower evolving or more constrained than SSDs, although it may be more correct to say that these groups evolve under differing selective conditions, as shown in previous work examining the selection pressures affecting preservation of the two duplicate types (Ezoe et al., 2021).

3.3.4 Duplicate categories show differences in functional enrichments

Different classes of duplicate show specific functional enrichments, with existing work showing that WGD genes are consistently enriched for developmental, neural, signalling and regulatory functions while they show depletions for base cell functions such as translation (Blomme et al., 2006; Brunet et al., 2006; Conant, 2020; Hakes et al., 2007; Kassahn et al., 2009; Makino and McLysaght, 2010; Qiao et al., 2018; Session et al., 2016). SSDs, on the other hand, are usually enriched for functions related to response to external factors, most notably immune response (Hakes et al., 2007; Qiao et al., 2018).

We find that these enrichments and depletions can be replicated in our dataset. Among the most enriched functions for WGD genes are developmental functions (GO:0048731, 'system development', $p = 1.05 \times 10^{-138}$; GO:0009653, 'anatomical structure morphogenesis', $p = 2.90 \times 10^{-129}$), in particular neural development (GO:0007399, 'nervous system development', $p = 2.65 \times 10^{-122}$). The other known WGD functions show similarly strong enrichments (GO:0065007, 'biological regulation', $p = 8.42 \times 10^{-138}$; GO:0023052, 'signaling', $p = 3.04 \times 10^{-118}$; GO:0045202, 'synapse', $p = 1.76 \times 10^{-87}$). Similarly for the known depletions in WGD genes, we find depletions for base cell functions, particularly translation

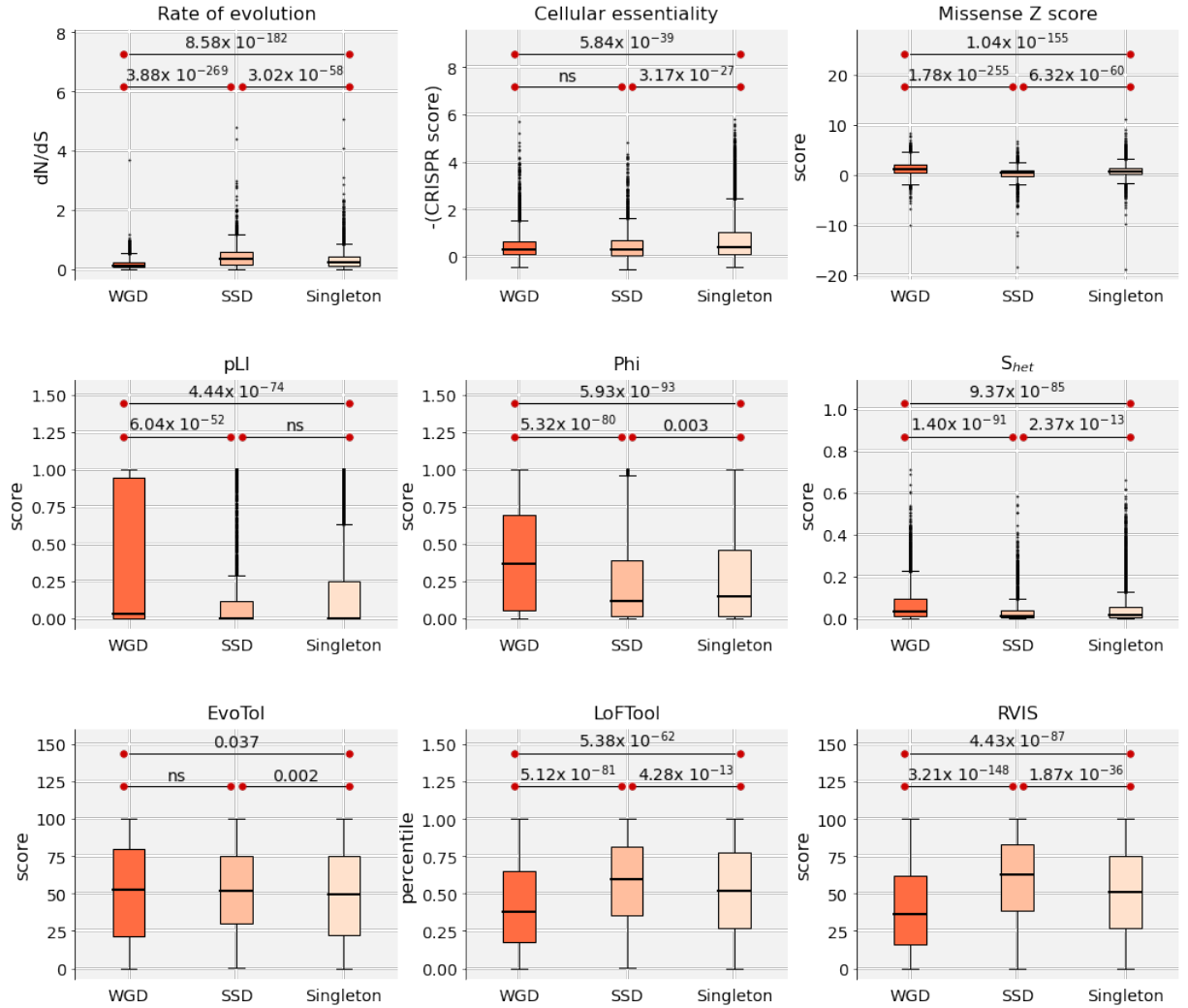


Figure 3.5 | Constraint related feature comparisons. Comparisons for features relating to how constrained or essential a gene is. Rate of evolution given by dn/ds , cellular essentiality by CRISPR score (Wang et al., 2015), missense Z score by the Z-score of missense variation in a gene compared to a null model. pLI = probability of LOF intolerance, Phi = probability of haploinsufficiency, S_{het} = selective effects on heterozygous protein truncating variants, EvoTol = evolutionary tolerance, LOFtool = percentile for LOFtool score (lower percentiles are more intolerant to variation), RVIS = residual variance intolerance score. For each feature, different categories of gene were compared using the Mann-Whitney U test. All P -values are Bonferroni-corrected, given above the line connected the groups compared.

and mitochondrial functions (GO:0042254, 'ribosome biogenesis', $p = 6.17 \times 10^{-24}$; GO:0005759, 'mitochondrial matrix', 7.26×10^{-18} ; GO:0006399, 'tRNA metabolic process', 1.60×10^{-14}).

In the case of SSD genes, we confirm an enrichment for functions related to sensing with significant enrichments for many terms related to receptors and response to stimulus (GO:0048018, 'receptor ligand activity', $p = 7.78 \times 10^{-28}$; GO:0043207, 'response to external biotic stimulus', 7.92×10^{-20}). These enrichments may be partially explained by an enrichment for immune genes among SSDs, as the vast majority of significant enrichments in this group are directly related to immune response (GO:0008009, 'chemokine activity', $p = 2.07 \times 10^{-22}$; GO:0006955, 'immune response', $p = 5.41 \times 10^{-22}$; GO:0098542, 'defense response to other organism', $p = 5.41 \times 10^{-22}$). Similarly to WGD genes, SSD genes show strong depletion of terms related to basic cellular functions, though they do differ slightly in the ranking of terms, with the most depleted terms for SSDs being related to organelles or binding (GO:0043226, 'organelle', $p = 2.01 \times 10^{-72}$; GO:1901363, 'heterocyclic compound binding', $p = 1.64 \times 10^{-63}$; GO:0000166, 'nucleotide binding', $p = 3.03 \times 10^{-52}$).

The most highly enriched/depleted terms for singletons essentially mirror those of the WGD genes with the highest enrichments in translation and mitochondrial functions (GO:0022613, 'ribonucleoprotein complex biogenesis', $p = 6.75 \times 10^{-44}$; GO:0005759, 'mitochondrial matrix', $p = 2.19 \times 10^{-32}$) and depletions in signaling, regulation and development (GO:0023052, 'signaling', 1.25×10^{-93} ; GO:0065007, 'biological regulation', 5.61×10^{-58} ; GO:0048731, 'system development', 3.93×10^{-53}).

3.3.5 Enriched/depleted functions show a clearer complementary pattern between singletons and ohnologs than between SSDs and ohnologs

Given the overlap in terms observed above, we decided to investigate the overall trend in overlapping enriched/depleted terms across the duplication categories (Figure 3.6). These overlaps may be informative to resolve the reasons for observed duplicate retention patterns for each mechanism. For example, we know that most WGD duplicable genes do not duplicate by SSD. If it is also the case that functions enriched in WGD genes are depleted in SSD genes and vice versa then the opposing retention patterns could be explained by some functions contributing to the likelihood of retention under one mechanism, but harming the chance of retention in the other.

What we observe is that WGD genes and singletons seem to show a clearer complementary pattern to their enriched/depleted functions, with 53.3% of terms enriched/depleted in WGD genes showing the opposite trend in singletons, versus only 29.5% for SSDs. For singletons, 71.4% of terms show an opposite enrichment pattern in WGD genes, compared to only 14.6% of terms for SSD. If we consider terms that show the same enrichment pattern (i.e. term is either enriched in both or depleted in both), singletons and SSDs are most concordant groups with 40.2% of SSD terms showing the same enrichment pattern in singletons, compared to 5% for WGDs. However, there is some asymmetry in these patterns as SSDs and singletons do not share any enrichments and this overlap in patterns is driven solely by large overlaps in depleted terms (48.5% of terms depleted in SSDs are also depleted in singletons). Based on the nature of this overlap between the two groups not retained post-WGD, we can infer that these depletions are shaped by preferential retention of genes with these functions amongst WGD genes. When

considered in light of opposite retention patterns of the duplicate types, asymmetry in overlaps between SSD and WGD terms is also informative. While 71.2% of SSD depleted terms are enriched in WGD genes, only 7.5% of terms enriched in SSDs are depleted in WGD genes, suggesting that while terms enriched in WGD genes cannot duplicate by SSD but there is not necessarily the same pressure preventing SSD enriched terms from being retained post-WGD. Overall, there is a noticeable trend of WGD gene functions being shaped by what functions promote retention post-WGD while SSD functions are more shaped by avoiding functions of genes which would be detrimental to duplicate in this manner. This is reflected in both these asymmetrical overlaps in terms but also the extreme differences in the number of terms enriched versus depleted for each group, with WGD genes mainly having considerably more terms enriched than depleted and SSD genes showing the opposite pattern. This could have the interesting implication for duplicability that WGD duplicability and SSD duplicability are not perfectly mutually exclusive. While genes solely duplicable by WGD certainly exist, we do not identify examples of the opposite case: functions enriched in SSDs that prevent duplicate retention following WGD.

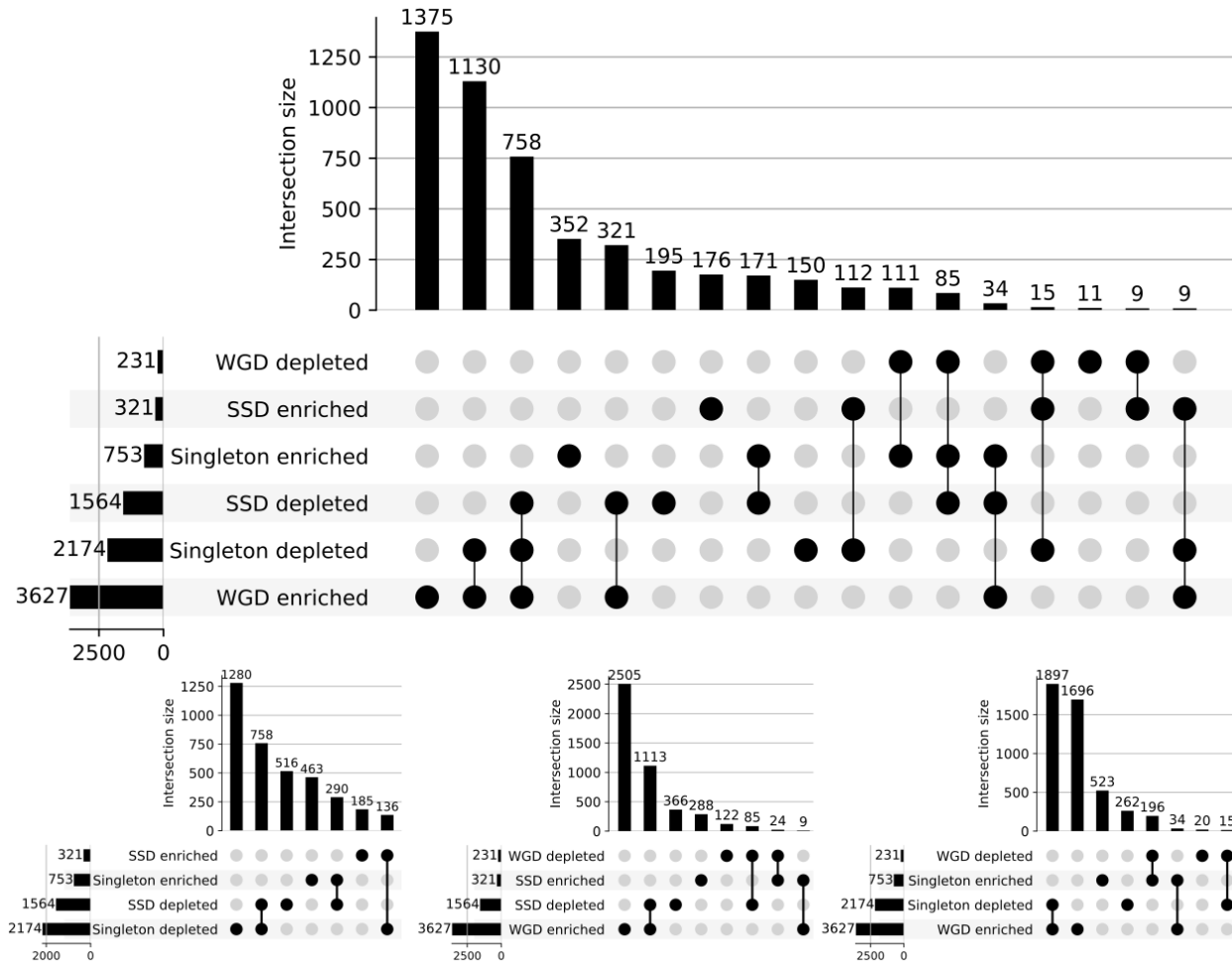


Figure 3.6 | Overlap of enriched/depleted GO terms between categories. Top panel shows intersection size, filled in dots and connections below indicate which groups are overlapping for a given intersection and the side panel shows the total number of GO terms significantly enriched/depleted for each duplicate group. Lower plots are simplified versions of the top plot showing only overlaps between two groups at a time.

3.3.6 Differences in duplicate age between SSDs and ohnologs do not fully account for observed effects

SSD and WGD genes do not differ only in the mechanism they may duplicate by, WGD duplications all date to early in the vertebrate lineage while SSD duplications may have occurred anywhere in the vertebrate tree. WGD duplications are also known to be long-lived while only a small portion of SSDs are retained in the long term. With this knowledge, we must investigate the possibility that the differences we observe are differences between long-lived and short-lived duplicates rather than due to duplication mechanism. To test this possibility, we used multiple regression models with duplication age and duplication type as predictors for each feature. (Figure 3.7). Theoretically, if we still see a significant effect for duplication type in a model controlling for age then we can assume the difference in mechanism exists independently of age.

For the most part, features where duplication type has no significant effect either did not significantly differ in our pairwise comparisons or differed weakly relative to other features (codon adaptation, %GC, PPIs, cellular essentiality, EvoTol).

Better candidates for having any observed difference explained by age are those where a significant interaction between duplicate type and age exists i.e. duplicate types respond differently to increasing age (See Methods and Table 3.2 for features including a significant interaction and how these were determined). In these cases (genomic length, CDS length, mean intron length, evolution rate and pLI) it is possibly that older SSDs approach the same values as WGDs. The main effect for duplication type is less relevant here as it only tells us if the average difference across age is significant; more informative is whether the SSD and WGD values converge with increasing age. For all these features except for pLI we do see a pattern of older SSDs appearing more and more similar to WGDs, however WGDs

remain longer and slower evolving than SSDs of the same age (except possibly in the case of mean intron length). In the case of pLI, SSDs actually appear to diverge more from WGDs the older they are, with this age effect causing the average duplication type effect to be non-significant. There is no clear reason for this behaviour, SSDs are not generally thought to become less essential the older the duplication and adding an interaction term to the models predicting other measures of essentiality does not significantly improve the model so this does not seem to be a general pattern. Another unusual case is intrinsic disorder, which is not significant in our pairwise comparisons but does show a significant effect of duplication type in the regression model. This is likely due to the transformation applied to the data prior to building the regression model.

Overall, while we do see evidence for differences between younger and older SSDs for some features, this does not seem to be a good explanation for the observed differences between WGD genes and SSD genes.

3.3.7 Feature importance rankings show ohnologs and SSDs are primarily distinguished by domain structure, expression level and haploinsufficiency

Our final aim was to integrate the observed feature differences to determine the most important features in distinguishing duplication classes. To this end, we built a random forest classifier to classify the two duplication types using a final set of 18 features which differed between WGD and SSD genes in both the pairwise comparisons and the regression models (Figure 3.8).

We find a high degree of interdependence between features (Figure 3.8A), with length and intron based features in particular heavily interrelated. This makes sense, as more/longer introns are likely to increase the length of a gene,

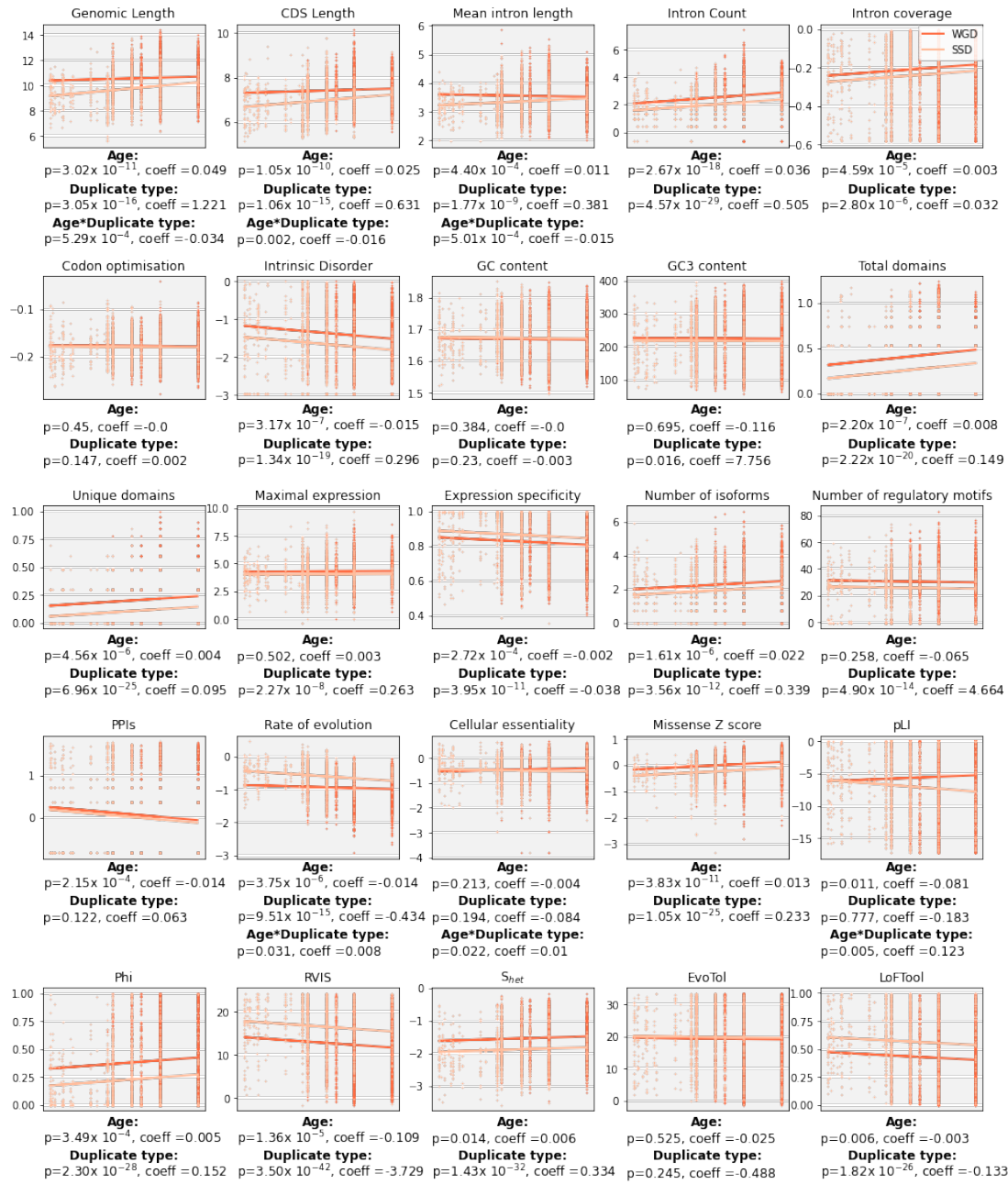


Figure 3.7 | Features regressed on duplicate age (see Methods) and duplicate type. Scatter and regression line for each feature with age increasing left to right. WGD shown in darker colour, SSD in lighter. Co-efficients and associated p values for each predictor are given, including for the interaction term where applicable (see Methods). Units for the features are as in previous figures but transformed according to formulas given in Table 3.2

Table 3.2 | Regression formulae. Transformations for each feature and predictors included in the regression model.

Feature	Regression formula
Genomic Length	$\text{Box-Cox}(\text{Genomic Length}) \sim \text{Age} + \text{Duplication_type} + \text{Age} * \text{Duplication_type}$
CDS Length	$\text{Box-Cox}(\text{CDS Length}) \sim \text{Age} + \text{Duplication_type} + \text{Age} * \text{Duplication_type}$
Rate of evolution	$\text{Log}(\text{Rate of evolution}) \sim \text{Age} + \text{Duplication_type} + \text{Age} * \text{Duplication_type}$
Unique domains	$\text{Log}(\text{Unique domains}) \sim \text{Age} + \text{Duplication_type}$
GC content	$\text{Log}(\text{GC content}) \sim \text{Age} + \text{Duplication_type}$
GC3 content	$\text{Box-Cox}(\text{GC3 content}) \sim \text{Age} + \text{Duplication_type}$
Total domains	$\text{Box-Cox}(\text{Total domains}) \sim \text{Age} + \text{Duplication_type}$
Expression specificity	$\text{Expression specificity} \sim \text{Age} + \text{Duplication_type}$
Number of isoforms	$\text{Box-Cox}(\text{Number of isoforms}) \sim \text{Age} + \text{Duplication_type}$
Intron Count	$\text{Box-Cox}(\text{Intron Count}) \sim \text{Age} + \text{Duplication_type}$
Mean intron length	$\text{Log}(\text{Mean intron length}) \sim \text{Age} + \text{Duplication_type} + \text{Age} * \text{Duplication_type}$
Maximal expression	$\text{Box-Cox}(\text{Maximal expression}) \sim \text{Age} + \text{Duplication_type}$
Cellular essentiality	$\text{Log}(\text{Cellular essentiality}) \sim \text{Age} + \text{Duplication_type} + \text{Age} * \text{Duplication_type}$
Number of regulatory motifs	$\text{Box-Cox}(\text{Number of regulatory motifs}) \sim \text{Age} + \text{Duplication_type}$
PPIs	$\text{Box-Cox}(\text{PPIs}) \sim \text{Age} + \text{Duplication_type}$
Intron coverage	$\text{Box-Cox}(\text{Intron coverage}) \sim \text{Age} + \text{Duplication_type}$
Missense Z score	$\text{Log}(\text{Missense Z score}) \sim \text{Age} + \text{Duplication_type}$
pLI	$\text{Box-Cox}(\text{pLI}) \sim \text{Age} + \text{Duplication_type} + \text{Age} * \text{Duplication_type}$
LoFTool	$\text{LoFTool} \sim \text{Age} + \text{Duplication_type}$
S_{het}	$\text{Log}(S_{het}) \sim \text{Age} + \text{Duplication_type}$
Phi	$\text{Phi} \sim \text{Age} + \text{Duplication_type}$
RVIS	$\text{Box-Cox}(\text{RVIS}) \sim \text{Age} + \text{Duplication_type}$
EvoTol	$\text{Box-Cox}(\text{EvoTol}) \sim \text{Age} + \text{Duplication_type}$
Codon optimisation	$\text{Box-Cox}(\text{Codon optimisation}) \sim \text{Age} + \text{Duplication_type}$
Intrinsic Disorder	$\text{Box-Cox}(\text{Intrinsic Disorder}) \sim \text{Age} + \text{Duplication_type}$

however this may make it difficult to untangle which feature is most informative for classification.

Our classifier performs relatively well, with a mean accuracy of 0.7963 (95% CI: 0.7938, 0.7988) and mean F1 score of 0.8674 (95% CI: 0.8657, 0.8691) across 100 iterations. Looking at each group specifically, again we see satisfactory performance on precision (proportion of classifications which are correct) (mean = 0.669, 95% CI = 0.6623, 0.6749 for SSDs and mean = 0.827, 95% CI = 0.8245, 0.8298 for WGD); Figure 3.8B). However when examining recall (proportion of a class classified as the correct class) we identify more of an imbalance (mean = 0.481, 95% CI = 0.4749, 0.4877) for SSD, mean = 0.912, 95% CI = 0.9101, 0.9143 for WGD), indicating bias towards classifying duplicates as WGD. With regard to relative importance in predicting the duplicate class, we find that evolution rate consistently ranks as the highest importance feature (mean rank across 100 iterations of 1.61, SD of 1.04; Figure 3.8C), followed by missense Z score, which is somewhat more variable (mean rank of 4.85, SD of 2.87). The following 3 features (genomic length, CDS length and intron count) all show a relatively similar mean rank (5.87, 5.65 and 6.73 respectively) relative to the next highest ranked feature (specificity at 10.31) and also show higher variability than the other highly ranked features (SD of 3.62, 3.75 and 3.61). This combined with these features showing high dependence on each other could indicate that they are sharing importance. As they can predict each other, once one is included as a predictor the others appear less important as much of the predictive power they contribute is also provided by the other dependent feature.

An additional observation that can be made from these importance rankings is that many of the highest ranking features are those that have shown a significant age interaction in our regression analysis. As these are cases where duplicate age differences contribute to the other differences observed between SSD and

WGD duplicates, we chose to repeat the classification using residual values from regressing each feature on duplicate age Figure 3.10. These values should represent the variation present in each feature not accounted for by variation in duplicate age and indicate whether our previous rankings were influenced by age differences between duplicates.

We find some minor differences in dependency between features in the age controlled data Figure 3.10A, however the main differences occur in model accuracy and feature rankings. We observe an increase in accuracy on controlling for age relative to the previous model (mean accuracy = 0.8625, 95% CI = 0.8604, 0.8646; mean F1 score = 0.9089, 95% CI = 0.9074, 0.9104) Figure 3.10B. This increase seems to be primarily driven by improvements in SSD classification, with large increases in both precision (mean = 0.794, 95% CI = 0.7880, 0.7996) and recall (mean = 0.658, 95% CI = 0.6518, 0.6640) for this group. We suspect that removing the influence of duplicate age has had this effect due to our previous model misclassifying older SSD duplicates as WGD duplicates. As there are no younger WGD duplicates, this effect was one sided.

In terms of feature rankings, we see a dramatic change on controlling for duplicate age, with the highest ranking features from our previous model (in particular CDS length and intron count) dropping considerably. Our most important features in this case are the number of unique types of domain and maximal expression, with little variation in their position in the rankings. Also outranking our previous set of high-ranking features are the total number of domains (possibly related to the high ranking of unique domains) and Phi, the probability of haploinsufficiency.

Feature importance estimates can be impacted by specific properties of the dataset, most notably differences in scale and correlation between features. Although permutation importance shows a relatively small bias towards larger scale

predictors than other measures (Strobl et al., 2007), we chose to investigate whether normalising features has any meaningful impact on our feature rankings. We find little to no difference between scaled and unscaled data for either our original model or the age controlled case (Figure 3.9, Figure 3.11), suggesting this did not have a meaningful impact on our original analysis.

Our checks for dependence between features (Figure 3.8A, Figure 3.10A) had indicated that there were relationships between the features themselves. While this shouldn't impact model accuracy, it may lead to reduced importance estimates for the correlated features. This issue becomes more prevalent with increasing correlation strength and number of correlated features (Gregorutti et al., 2017), as the impact of removing/permuting a single feature in the group becomes less and less with an increasing number of other correlated features to compensate the lost information. To investigate how this might impact our importance estimates and final feature rankings, we defined two groups of dependent features and repeated our permutation importance calculations including only one feature in the group at a time (Table 3.3, Table 3.5). For example, when considering genomic length we would include only genomic length and drop CDS length, mean intron length, intron count and intron coverage. We take this new importance estimate to represent how informative the feature truly is including information it may share with other correlated features. We then rank the feature among the average importance estimates from the model including all features as if it had this new importance level. Considering these features in isolation seems to indicate that correlation has impacted the importance estimates in our original analysis. In the original model, any of genomic length, CDS length and intron count would be ranked second when considered in isolation while they rank third, fourth and fifth respectively when considered together. Additionally we see very little impact on model accuracy when only considering one feature from the group, suggesting

Table 3.3 | Change in rank for correlated features considered in isolation.

Feature	Previous rank	New rank	Previous accuracy	New accuracy
<i>Group 1</i>				
Genomic length	4	2	0.7935	0.7859
CDS length	3	2	0.7935	0.7864
Mean intron length	8	4	0.7935	0.7847
Intron count	5	2	0.7935	0.7821
Intron coverage	15	6	0.7935	0.7807
<i>Group 2</i>				
Unique domains	11	16	0.7935	0.7924
Domains	16	11	0.7935	0.7929

that only one of these features is required to supply almost the same amount of information to the classifier as the entire group. When we consider the age controlled model, there is comparatively little impact on rankings when features are considered in isolation, possibly due to some of the information shared between features also correlating with age. Similarly to previous analysis, we do not see any real impact on using scaled data (Table 3.4, Table 3.6).

On controlling for the effect of age differences between duplicate classes, the best features for distinguishing duplicate type are the number of unique domain types (and, to a lesser degree, the absolute number of domains), maximal expression and Phi. Both length and constraint, which are among the features most frequently investigated and consistently found to differ between duplicates, seem to have their strong relationship to duplication type mostly explained by a relationship to age rather than mechanism, though features in these groups do still contribute to our model. These results are robust to both correlation within our group of features and differences in scale across features.

Table 3.4 | Change in rank for correlated features considered in isolation using centered data.

Feature	Previous rank	New rank	Previous accuracy	New accuracy
<i>Group 1</i>				
Genomic length	4	2	0.7983	0.7906
CDS length	3	2	0.7983	0.7904
Mean intron length	7	3	0.7983	0.7917
Intron count	5	2	0.7983	0.7882
Intron coverage	18	6	0.7983	0.7850
<i>Group 2</i>				
Unique domains	10	9	0.7983	0.7991
Domains	12	10	0.7983	0.7960

Table 3.5 | Change in rank for correlated features considered in isolation for age controlled model.

Feature	Previous rank	New rank	Previous accuracy	New accuracy
<i>Group 1</i>				
Genomic length	6	6	0.8651	0.8628
CDS length	11	11	0.8651	0.8640
Mean intron length	12	10	0.8651	0.8630
Intron count	10	9	0.8651	0.8664
Intron coverage	18	16	0.8651	0.8624
<i>Group 2</i>				
Unique domains	1	1	0.8651	0.8587
Domains	4	3	0.8651	0.8533

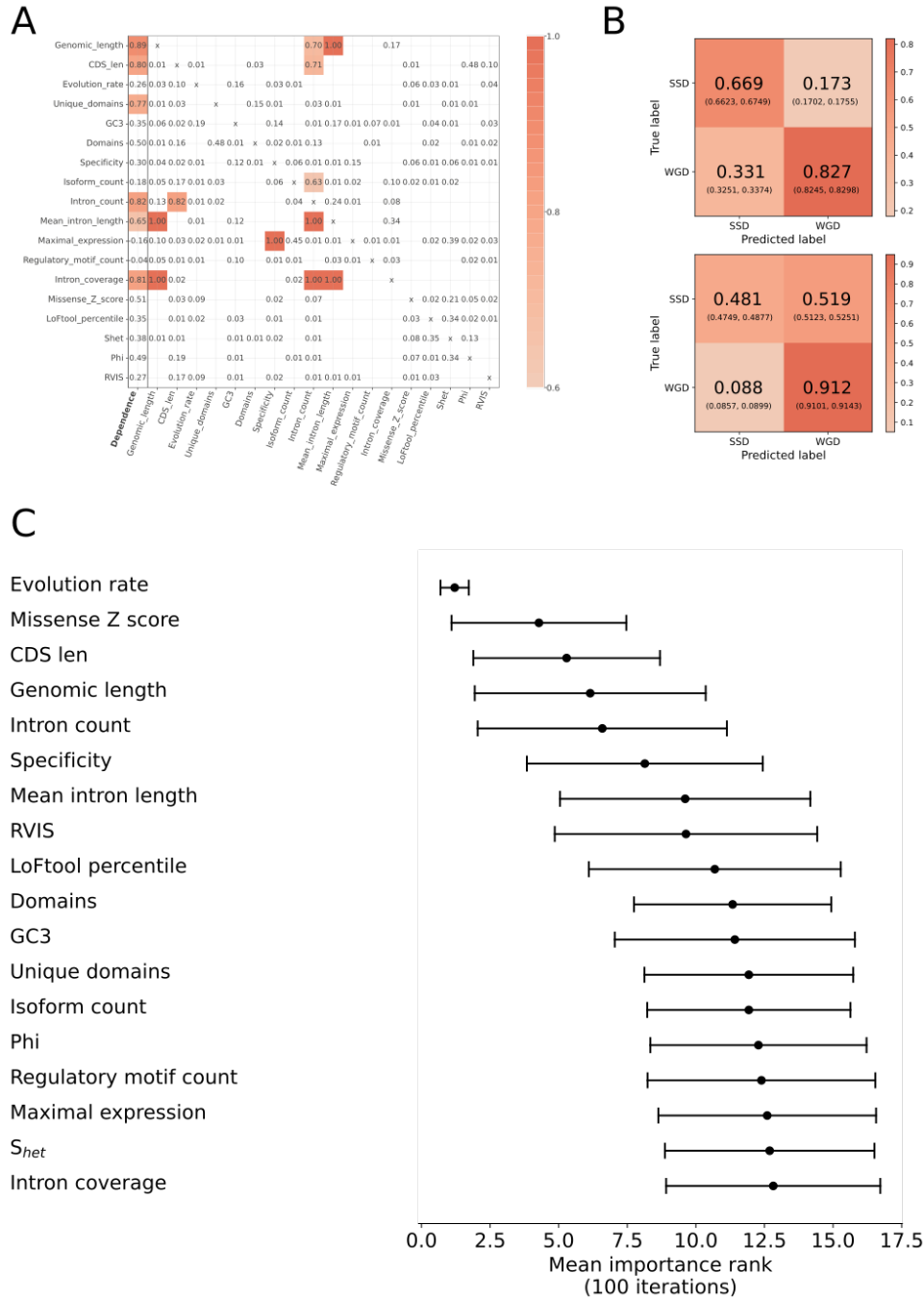


Figure 3.8 | Duplicate type prediction. (A) Feature dependencies, based on a fitted classifier with each feature as the target and all other features as predictors. The 'Dependence' column gives the R^2 value for the model i.e. how well the feature can be predicted by the others. Other values give the importance of each variable in the given predictive model. Values are bounded at 0 and 1, zero values not shown. (B) Normalised confusion matrix for the random forest classifier, values normalised to 1 within each column. (C) Average importance rank for each feature tested across 100 iterations of fitting the classifier, features ordered by mean importance. Error bars indicate 1 s.d.

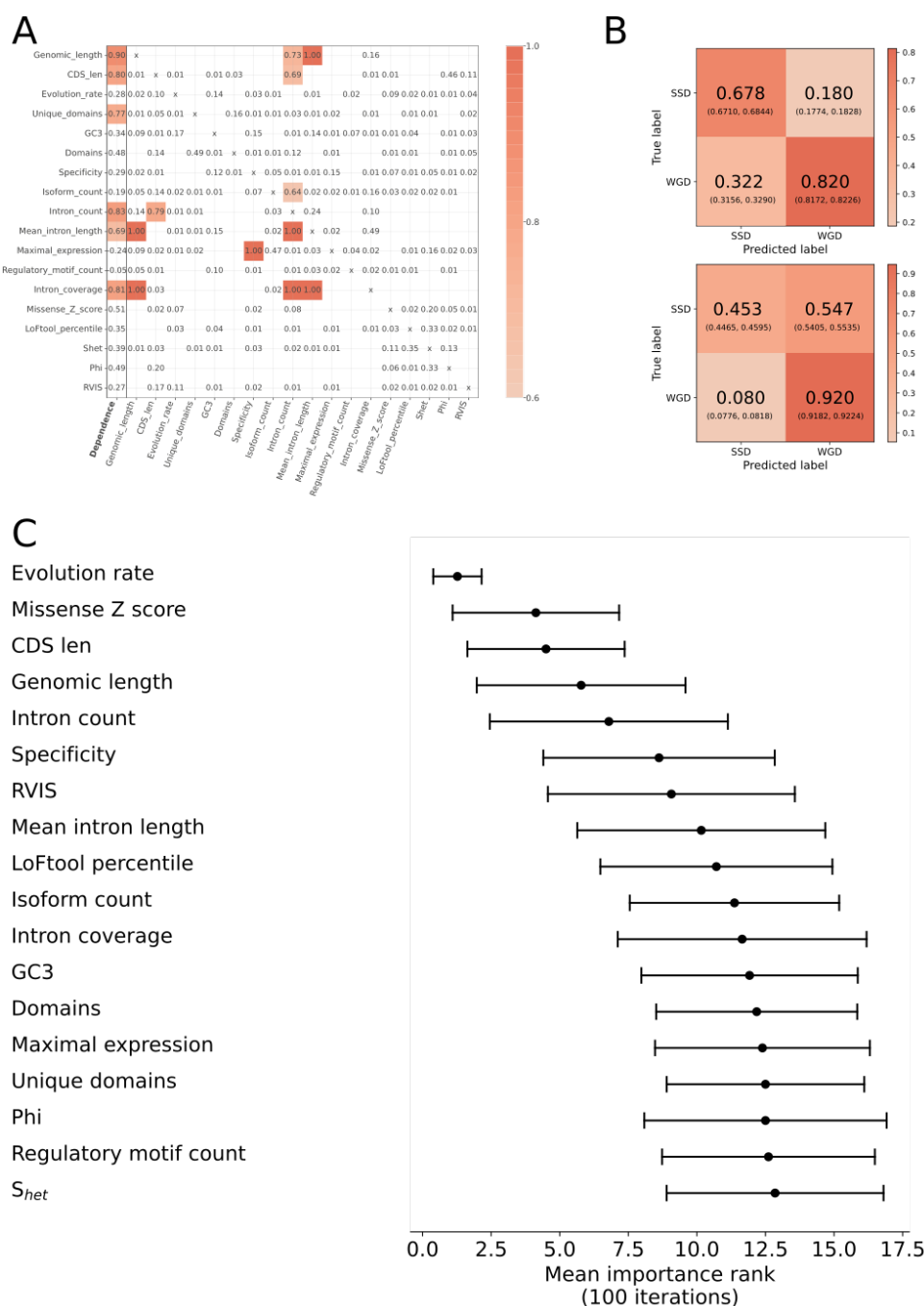


Figure 3.9 | Duplicate type prediction using scaled data. (A) Feature dependencies, based on a fitted classifier with each feature as the target and all other features as predictors. The 'Dependence' column gives the R^2 value for the model i.e. how well the feature can be predicted by the others. Other values give the importance of each variable in the given predictive model. Values are bounded at 0 and 1, zero values not shown. (B) Normalised confusion matrix for the random forest classifier, values normalised to 1 within each column. (C) Average importance rank for each feature tested across 100 iterations of fitting the classifier, features ordered by mean importance. Error bars indicate 1 s.d.

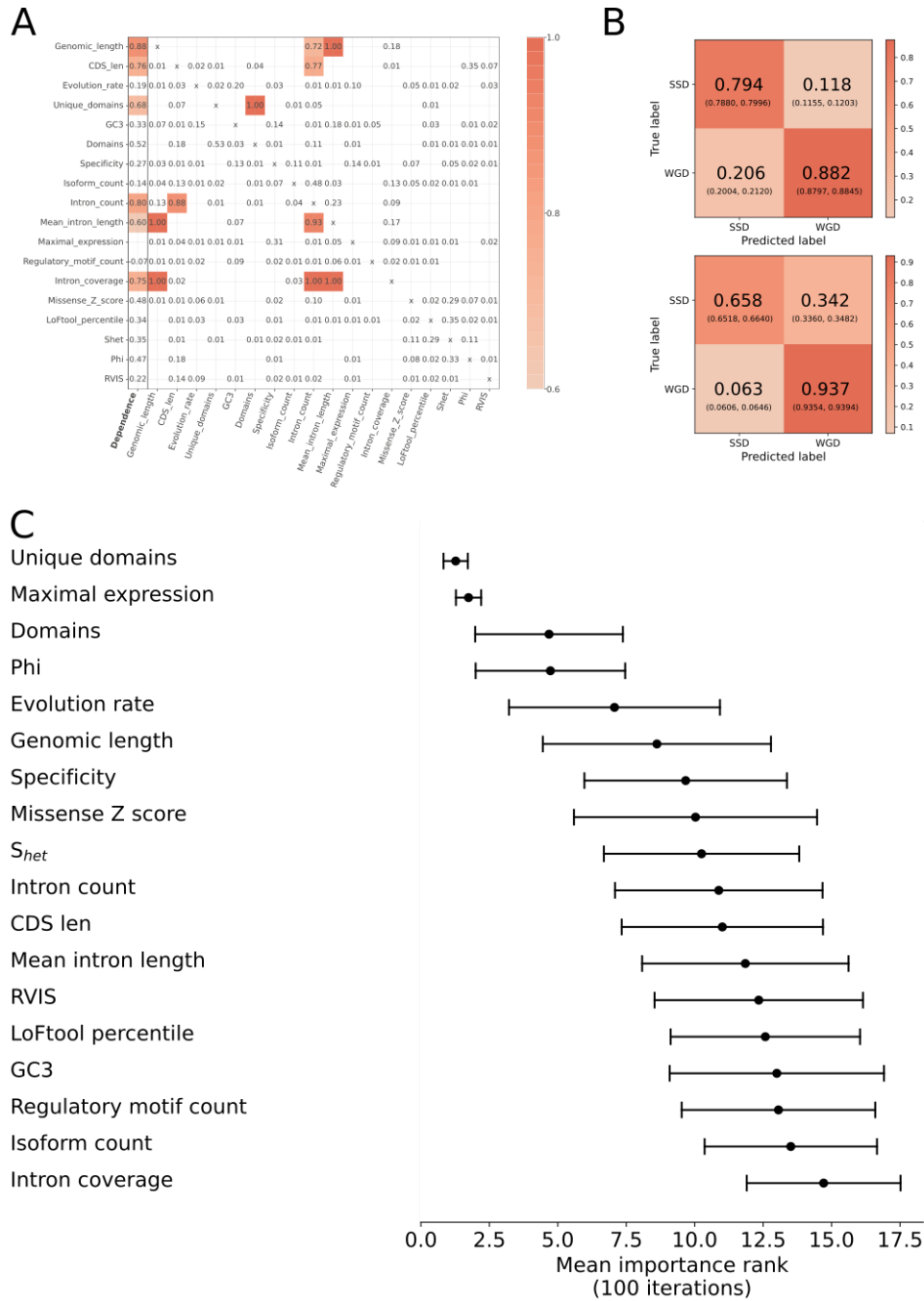


Figure 3.10 | Duplicate type prediction controlling for age differences. (A) Feature dependencies, based on a fitted classifier with each feature as the target and all other features as predictors. The 'Dependence' column gives the R^2 value for the model i.e. how well the feature can be predicted by the others. Other values give the importance of each variable in the given predictive model. Values are bounded at 0 and 1, zero values not shown. (B) Normalised confusion matrix for the random forest classifier, values normalised to 1 within each column. (C) Average importance rank for each feature tested across 100 iterations of fitting the classifier, features ordered by mean importance. Error bars indicate 1 s.d.

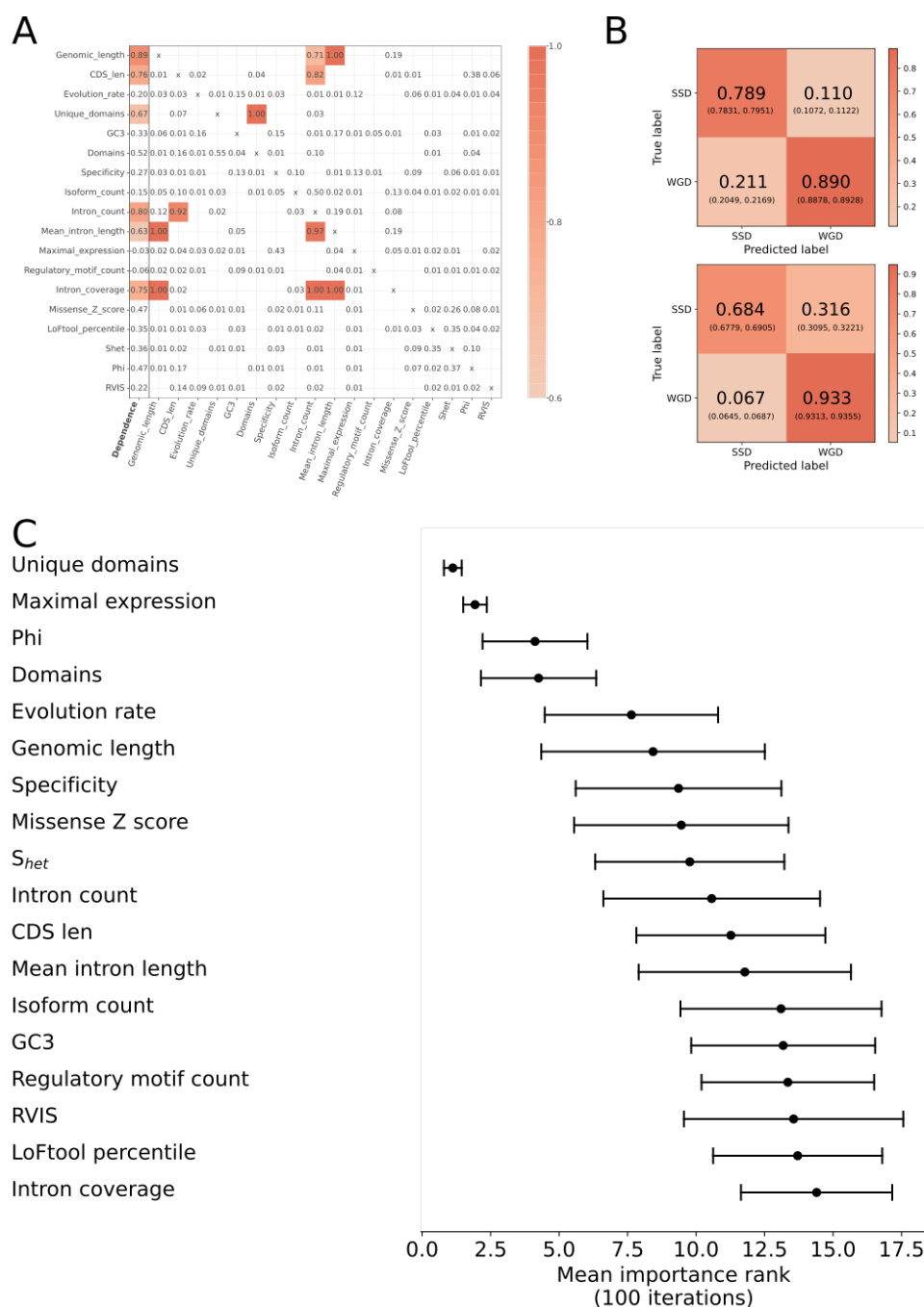


Figure 3.11 | Duplicate type prediction with scaled data, controlling for age differences. (A) Feature dependencies, based on a fitted classifier with each feature as the target and all other features as predictors. The 'Dependence' column gives the R^2 value for the model i.e. how well the feature can be predicted by the others. Other values give the importance of each variable in the given predictive model. Values are bounded at 0 and 1, zero values not shown. (B) Normalised confusion matrix for the random forest classifier, values normalised to 1 within each column. (C) Average importance rank for each feature tested across 100 iterations of fitting the classifier, features ordered by mean importance. Error bars indicate 1 s.d.

Table 3.6 | Change in rank for correlated features considered in isolation for centered data using age controlled model.

Feature	Previous rank	New rank	Previous accuracy	New accuracy
<i>Group 1</i>				
Genomic length	6	6	0.8656	0.8647
CDS length	11	11	0.8656	0.8649
Mean intron length	12	11	0.8656	0.8639
Intron count	9	9	0.8656	0.8640
Intron coverage	17	16	0.8656	0.8632
<i>Group 2</i>				
Unique domains	1	1	0.8656	0.8619
Domains	3	3	0.8656	0.8549

3.4 Discussion

We examined differences between duplication types across a wide range of features covering sequence and gene structure, regulation and constraint. The most extreme differences between SSD and WGD genes generally fall under either features relating to gene length or those relating to constraint on sequence changes. We further showed that these changes can not be fully explained by differences in duplicate lifespan, although they do have a dramatic effect on how strongly these features contribute to accuracy in classifying duplicate type. It is significant that these features demonstrate a stronger relationship to duplicate age than duplication mechanism as, within our dataset, they are also the features with the most accumulated evidence for differing strongly between the duplicate groups. More caution is warranted in future studies seeking to define differences by duplication mechanism if duplicate age differences exist in the dataset.

The features that we find to contribute most strongly to classifier accuracy on controlling for this age affect indicate a strong influence of gene dosage effects in determining successful duplication. Gene expression level and probability of

haploinsufficiency are ranked second and third in our importance rankings, providing evidence that the abundance of gene products has relevance in whether duplicate copies are retained following duplication by either mechanism. The dominance of the number of unique domains as a top predictor of duplicate class is less clear within current models of duplicate retention. The fact that the number of *unique* domains seems to provide information not given by the absolute number of domains (Table 3.5) speaks to diversity of function within the gene perhaps being more relevant than any length or structure related effects of having a greater number of domains. Potentially the presence of multiple types of domain has bearing on retention processes such as subfunctionalization, or performing multiple functions impacts how essential a gene is.

Along with the overarching picture of gene duplication we have defined, some features show particularly interesting or unexpected results in terms of pairwise comparisons and warrant further discussion, for example PPIs and isoform count. While we might have expected that proteins encoded by SSDs would show a lower number of interactions with other proteins due to reasons of dosage sensitivity as well as previous observations showing WGD genes are highly connected/more likely to be complex members, a comparatively weak or non-significant difference was observed between direct comparisons and the regression analysis. While this could perhaps be explained as only a small portion of ohnolog dosage sensitivity being due to interactions with other proteins, it is more likely that this is due to a dataset limitation. The interaction data here, from the Human Reference Interactome, is missing certain interactions which likely impacted WGD genes more heavily than SSD genes. Only binary interactions are covered, meaning indirect interactions due to complex membership will be missed. Additionally, any interactions reliant on post-translational modifications are not included, which again possibly disadvantaged WGD genes as the proteins they encode have been

observed to undergo more of these modifications (Amoutzias et al., 2010).

On the other hand, the difference in isoform count was unexpectedly large, with SSD genes having far fewer than either singletons or WGD genes. One potential explanation for this could be the idea of alternatively spliced transcripts acting as 'internal paralogs' (Modrek and Lee, 2003), with an alternative transcript essentially allowing some novel function to evolve in a similar manner to a copy of the gene. This would be a preferable route to tandem duplication for a highly conserved gene such as an ohnolog as expression of the novel transcript could be partitioned in a tissue-specific manner and need not affect tissues where it might be harmful.

The overall goal of this work was to create an integrated view of differences between duplicate categories, so that these differences could be used to make deductions about what promotes or hinders duplicate retention under each mechanism. With that in mind, what can we now say about the impact of these features on duplicate retention? While it is clear from the single feature comparisons that WGD genes are longer and more broadly expressed as well as being highly constrained and regulated, and the most relevant features implicate dosage and functional diversity as important determinants of duplicate type, what about these features promotes retention post-WGD, or prevents retention post-SSD?

While duplicability can, to a degree, be explained by examining these basic features together, a more holistic explanation requires relating them to the distinctive functional profiles for each duplicate group. Previous work in zebrafish examining paralogs retained from the teleost specific WGD found that many differences between genes with and without retained paralogs are specific to nervous system genes (Roux et al., 2017). Given that the overlap in enriched functions for the two duplicate groups is minimal, it follows that if genes with specific functions display distinctive sets of features then these enrichments will drive differences

in those same features between the groups. Roux et al. focused primarily on the higher constraint on WGD genes, explained as an avoidance of potentially toxic variants of highly expressed genes in the nervous system, however further investigation will be required to identify whether all differences in basic features can be explained through functional profile.

The functional differences between WGD genes and SSD genes likely reflect a fundamental difference in the reasons why duplicates of the two different categories are successfully retained. WGD genes possess functions fundamental to development and organism function which would be dangerous to make major alterations to, while SSD genes are characterised by functions where frequent change is beneficial if not required such as immunity. This view of WGD duplicates retained for preservation and SSD duplicates retained to gain novel functions is supported by recent work in plants showing that WGD duplicates are retained across multiple lineages under purifying selection while SSD duplicates are retained in multiple lineages under positive selection (Ezoe et al., 2021).

One final outstanding question is whether these findings are generalisable. Gene duplication occurs in many diverse lineages and there is no guarantee that duplicate retention follows the same rules in all of them. While work exists comparing duplication types in various species, frequently there is not a comparable study for a given feature across multiple lineages due to differences in the duplicate set studied, in methodology etc. This said, there are many similarities in the functional profiles of WGD and SSD genes across lineages. One striking example is the enrichment of immune functions both in human (this work) and in plants (Qiao et al., 2018), despite huge differences in plant and animal immunity. Parallels along these lines hint that there may be a unifying model of duplicate retention, however further work is necessary to properly define duplicate group differences in a comparable way across species.

Chapter 4

Evidence from *Drosophila* supports higher duplicability of faster evolving genes

4.1 Introduction

Gene duplication is an important process in biological innovation with major roles in the evolution of genome structure and content. Duplication may occur by several mechanisms broadly grouped into small-scale duplication (SSD; which includes tandem duplication and retrocopying) and whole genome duplication (WGD). Genes which successfully duplicate by different mechanisms differ significantly from both singletons and each other in notable ways (Amoutzias et al., 2010; Banerjee et al., 2017; Hakes et al., 2007; Jiang et al., 2013; Qiao et al., 2018). One difference which is frequently seen is that SSDs are less constrained and more dispensable than either WGDs or singletons (Makino et al., 2009).

This lower level of constraint in SSDs is consistently observed as a higher rate of evolution compared to other genes (Defoort et al., 2019; He and Zhang,

2006; Qiao et al., 2018). This feature of duplicated genes is of particular interest as it has frequently been used to form hypotheses on the processes of duplicate retention. Specifically, the higher rate of evolution in tandem duplicates has contributed to the idea that duplicates are redundant at the point of creation and thus can accumulate changes free from evolutionary constraint (Jordan et al., 2004; Lynch and Conery, 2000; Pegueroles et al., 2013). Knowledge of the true relationship between gene duplication and evolutionary rate has important implications for our understanding of duplicate retention and the role of gene duplication in evolutionary processes, as well as informing our general knowledge of how the features of a gene impact on its evolutionary behaviour. Whereas one model implicates a fixation bias in favour of less constrained, fast-evolving genes (He and Zhang, 2006; Woods et al., 2013), the other views these properties as a consequence of duplication and perhaps sometimes instrumental in their successful long-term retention due to sequence and functional divergence of the paralogs, as in subfunctionalization models (Force et al., 1999; Des Marais and Rausher, 2008).

Despite the faster rate of evolution being one of the most reliably observed features of duplicated genes across lineages, and its relevance in our present understanding of duplicate evolution, the relationship between evolutionary rate and duplicability has not been fully resolved. Does duplication release constraint through redundancy and cause faster rates; or is a faster rate of evolution correlated with a higher chance of successful duplication? Put another way, do SSDs show low constraint because they have additional copies, or are more dispensable genes more likely to be successfully duplicated in the first place?

Previous work on the topic of differentiating these two possibilities has focused on rate measurement in outgroups to approximate the ancestral rate of evolution while assessing duplicability in the ingroups, thereby uncoupling the rate

estimation from the duplication event (Davis and Petrov, 2004; O'Toole et al., 2018). Although the methods are similar, these two studies examine different lineages, and estimate evolutionary parameters over vastly different periods of time. O'Toole et al. (2018) use closely-related primate outgroups to approximate the ancestral rate prior to duplication and find that genes which have duplicated within the apes have higher ancestral rates than those which have not. Similarly, Davis and Petrov (2004) use outgroups to measure the evolutionary rate independently of the species in which duplication has taken place, but they concluded that duplicable genes are more conserved. It is worth noting the large evolutionary distance between the outgroup species used in this study (*D. melanogaster* and *A. gambiae*) and the species where duplicate or singleton status was defined (namely *S. cerevisiae* and *C. elegans*).

There are two possible explanations for this contradiction in results. First, there may be genuine biological differences between the lineages which in turn affect the processes governing duplicate retention. Generation time and population size are vastly different when comparing primates (generation times of decades and effective populations in the thousands) to organisms such as yeast (generation time of hours, effective population in the millions). Both of these features have relationships to selection and mutation rate, which may impact on gene duplication and retention as well as evolutionary rates.

Second, it may be the case that the large differences in divergence time of the outgroups used by the two studies has impacted on the genes included in the analysis. Genes with recognisable orthologs between insects and yeasts probably represent a more conserved group than genes with recognisable orthologs within the primate lineage. Additionally, it is known that genes which are duplicable tend to duplicate relatively frequently (Li et al., 2016); this is reflected in the small number of genes which were duplicated in the ingroup but not in the outgroup

that are recovered in O'Toole et al. (2018). Davis and Petrov (2004) do not require that the genes examined are single copy in the outgroup, although it is mentioned that the observed effect holds if this restriction is used. Contrary to this, O'Toole et al (2018) found that applying this restriction leaves no genes in the duplicable category at all. The large span of evolutionary time covered makes it difficult to find an appreciable number of genes which are duplicated in one lineage but not the other and further restricts the set of genes which can be used for the comparison of interest.

Here, we attempt to determine, in lineages other than primates, whether duplicable genes are ancestrally faster evolving or if duplication-induced redundancy accelerates evolutionary rates. We examine this in insect genomes, similarly to Davis and Petrov, but we remove other differences in study design by restricting the analysis to closely-related species within the *Drosophila* lineage. We find that duplicable genes show higher ancestral rates of evolution when we use singleton orthologs in closely related outgroup species to estimate rates, and we do not find much evidence to support the classical view of rate acceleration following duplication. This latter finding may be affected by a subset of rapidly-evolving genes being annotated as singletons when they may actually be genes which have simply suffered from homology detection failure due to extensive sequence divergence.

4.2 Materials & Methods

4.2.1 Inference of orthologies

Proteomes for 37 good quality (contig N50 $\geq$ 200 kb and scaffold N50 $\geq$ 500 kb) *Diptera* genomes were downloaded via the NCBI FTP site (see Table 4.1). Protein sequences were used with a number of orthology inference tools to infer orthologous relationships and groups. The main body of analysis was based on

Table 4.1 | Data location for species used. Locations are given for data directories in the NCBI FTP site (<ftp.ncbi.nlm.nih.gov>) where CDS sequences and associated translations were obtained. Full species names for the abbreviations in this table are given in Table 4.2.

Species	NCBI FTP location
DALB	GCF/009/650/485/GCF_009650485.1_drosAlbom151121751.03v1
DANA	GCF/003/285/975/GCF_003285975.2_DanaRS2.1
DBIA	GCF/000/233/415/GCF_000233415.1_Dbia_2.0
DBUS	GCF/011/750/605/GCF_011750605.1_ASM1175060v1
DELE	GCF/000/224/195/GCF_000224195.1_Dele_2.0
DERE	GCF/003/286/155/GCF_003286155.1_DereRS2
DEUG	GCF/000/236/325/GCF_000236325.1_Deug_2.0
DFIC	GCF/000/220/665/GCF_000220665.1_Dfic_2.0
DGUA	GCF/900/245/975/GCF_900245975.1_DGUA_6
DHYD	GCF/003/285/905/GCF_003285905.1_DhydRS2
DKIK	GCF/000/224/215/GCF_000224215.1_Dkik_2.0
DMAU	GCF/004/382/145/GCF_004382145.1_ASM438214v1
DMIR	GCF/003/369/915/GCF_003369915.1_D.miranda_PacBio2.1
DMEL	GCF/000/001/215/GCF_000001215.4_Release_6_plus_ISO1_MT
DNOV	GCF/003/285/875/GCF_003285875.2_DnovRS2.1
DPER	GCF/003/286/085/GCF_003286085.1_DperRS2
DPSE	GCF/009/870/125/GCF_009870125.1_UCI_Dpse_MV25
DSAN	GCF/016/746/245/GCF_016746245.1_Prin_Dsan_1.0
DSEC	GCF/004/382/195/GCF_004382195.1_ASM438219v1
DSER	GCF/002/093/755/GCF_002093755.1_Dser1.0
DSIM	GCF/016/746/395/GCF_016746395.1_Prin_Dsim_3.0
DSUB	GCF/008/121/235/GCF_008121235.1_UCBerk_Dsub_1.0
DSUP	GCF/014/743/375/GCF_014743375.2_RU_Dsub_v1.1
DSUZ	GCF/013/340/165/GCF_013340165.1_LBDM_Dsuz_2.1.pri
DVIR	GCF/003/285/735/GCF_003285735.1_DvirRS2
DYAK	GCF/016/746/365/GCF_016746365.1_Prin_Dyak_Tai18E2_2.0
AAEG	GCF/002/204/515/GCF_002204515.2_AaegL5.0
AALB	GCF/013/758/885/GCF_013758885.1_VT_AalbS3_pri_1.0
AEAL	GCF/006/496/715/GCF_006496715.1_Aalbo_primary.1
ASTE	GCF/013/141/755/GCF_013141755.1_UCI_ANSTEP_V1.0
BCOP	GCF/014/529/535/GCF_014529535.1_BU_Bcop_v1
BTRY	GCF/016/617/805/GCF_016617805.1_CSIRO_BtryS06_freeze2
CCAP	GCA/905/071/925/GCA_905071925.1_EGII-3.2.1
CPIP	GCF/016/801/865/GCF_016801865.1_TS_Cpip_V1
CQUI	GCF/015/732/765/GCF_015732765.1_VPISU_Cqui_1.0_pri_pater
HILL	GCF/905/115/235/GCF_905115235.1_iHerIll2.2.curated.20191
SLEB	GCF/003/285/725/GCF_003285725.1_SlebRS2

Table 4.2 | Abbreviations used. Full species names for abbreviations used in data processing and in Table 4.1.

Abbreviation	Species
DALB	<i>Drosophila albomicans</i>
DANA	<i>Drosophila ananassae</i>
DBIA	<i>Drosophila biarmipes</i>
DBUS	<i>Drosophila busckii</i>
DELE	<i>Drosophila elegans</i>
DERE	<i>Drosophila erecta</i>
DEUG	<i>Drosophila eugracilis</i>
DFIC	<i>Drosophila ficusphila</i>
DGUA	<i>Drosophila guanche</i>
DHYD	<i>Drosophila hydei</i>
DKIK	<i>Drosophila kikkawai</i>
DMAU	<i>Drosophila mauritiana</i>
DMIR	<i>Drosophila miranda</i>
DMEL	<i>Drosophila melanogaster</i>
DNOV	<i>Drosophila novamexicana</i>
DPER	<i>Drosophila persimilis</i>
DPSE	<i>Drosophila pseudoobscura</i>
DSAN	<i>Drosophila santomea</i>
DSEC	<i>Drosophila sechellia</i>
DSER	<i>Drosophila serrata</i>
DSIM	<i>Drosophila simulans</i>
DSUB	<i>Drosophila subobscura</i>
DSUP	<i>Drosophila subpulchrella</i>
DSUZ	<i>Drosophila suzukii</i>
DVIR	<i>Drosophila virilis</i>
DYAK	<i>Drosophila yakuba</i>
AAEG	<i>Aedes aegypti</i>
AALB	<i>Anopheles albimanus</i>
AEAL	<i>Aedes albopictus</i>
ASTE	<i>Anopheles stephensi</i>
BCOP	<i>Bradysia coprophila</i>
BODO	<i>Bradysia odoriphaga</i>
BTRY	<i>Bactrocera tryoni</i>
CCAP	<i>Ceratitis capitata</i>
CPIP	<i>Culex pipiens pallens</i>
CQUI	<i>Culex quinquefasciatus</i>
HILL	<i>Hermetia illucens</i>
SLEB	<i>Scaptodrosophila lebanonensis</i>

orthologous relationships from Orthofinder with default settings (Emms and Kelly, 2019), with additional checks for duplicate detection capacity using Orthofinder with the ultra-sensitive setting, SonicParanoid (Cosentino and Iwasaki, 2019), and OMA (Train et al., 2017), which was run three times varying the ‘InParalogTol’ parameter.

4.2.2 Selection of singleton groups and tree building

Candidate singletons were defined for *D. suzukii* as genes with no non-self BLAST hit with $E \leq 0.1$ when searching the genome against itself, and orthogroups were then extracted for this set. Each orthogroup was aligned using MUSCLE (Edgar, 2004) and trees were built from these alignments using IQ-TREE with models being selected from WAG, LG and JTT (Hoang et al., 2018; Kalyaanamoorthy et al., 2017; Minh et al., 2020). Trees were processed using the ete3 python package (Huerta-Cepas et al., 2016). The strict initial selection of singletons provides confidence in singleton status in the outgroup species, though it is possible that genes with very ancient duplications that predate our time-frame of interest are excluded at this step. We note that use of a more relaxed cutoff for singletons (genes with no non-self BLAST hit with $E \leq 1 \times 10^{-4}$) does not meaningfully affect our final results (see Figure 4.1 and Figure 4.2 in comparison to Figure 4.7 and Figure 4.8) and so we expect that these results are robust to any such effect.

4.2.3 Dataset filtering and evolutionary rate calculation

A number of filtering steps were carried out to ensure the trees used in the final comparisons did not violate any of our assumptions (Figure 4.3).

Sequence evolution parameters (d_N , d_S and d_N/d_S) were calculated using the codeml module of PAML (Yang, 2007) with model = -2 for pairwise rate calculation and all other parameters set to defaults. Cases where d_S exceeded 4 were excluded

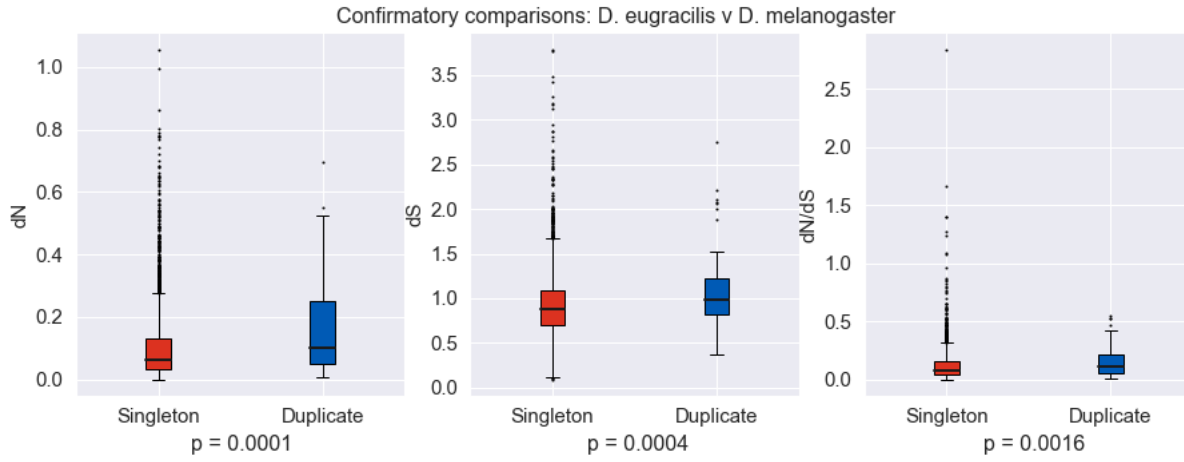


Figure 4.1 | Confirmatory rate comparisons in duplicability assessment clade with more relaxed starting singleton definition. We confirmed the known trend that duplicated genes are faster evolving by comparing rates within the set of species where we assessed duplicability. A dS threshold of 4 is applied and starting set of singletons was defined with an E-value cutoff of 10^{-4} . P-values are given for two-sided Mann-Whitney U-tests.

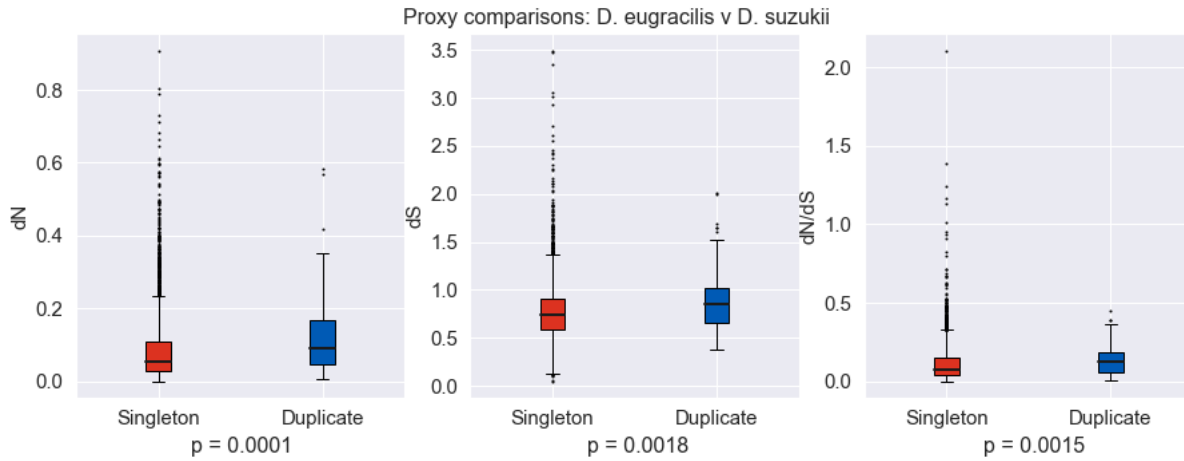


Figure 4.2 | Proxy ancestral rate comparisons with more relaxed starting singleton definition. Proxy ancestral evolutionary rates are faster in duplicable groups on using a more relaxed set of singletons as the initial dataset. A dS threshold of 4 is applied and starting set of singletons was defined with an E-value cutoff of 10^{-4} . P-values are given for two-sided Mann-Whitney U-tests.

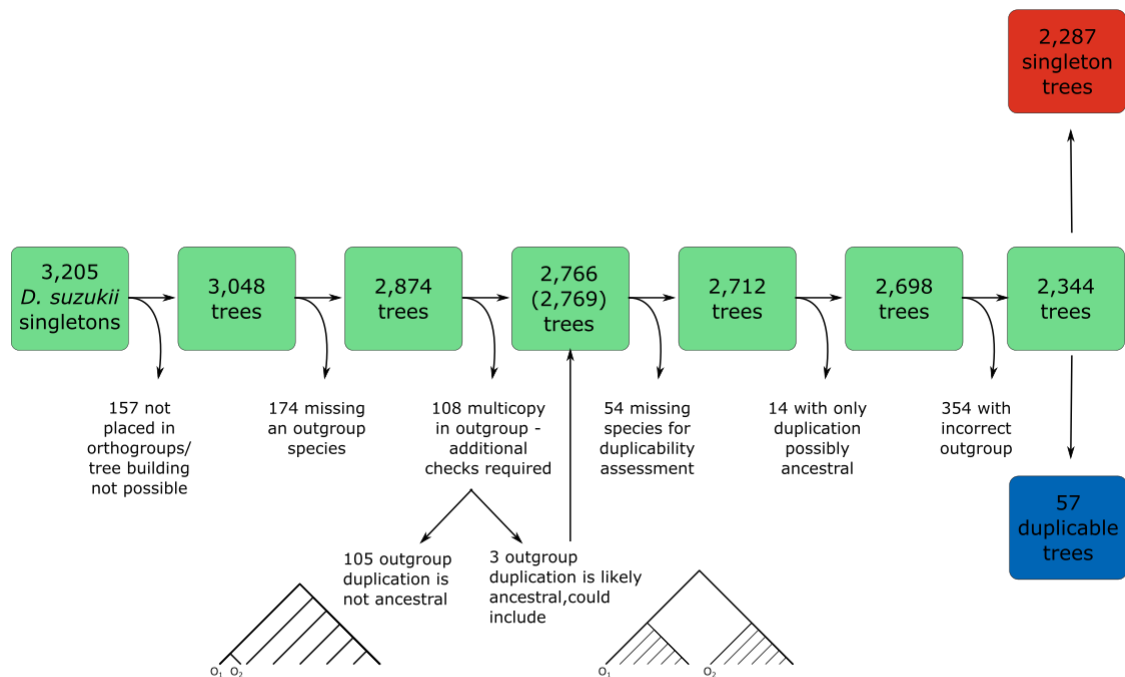


Figure 4.3 | Filtering performed on initial set of singletons. It was required that singletons were present in an orthogroup and that a gene tree could be constructed from the orthogroup, and that the tree produced fulfilled a number of requirements. The numbers shown for the final dataset are prior to rate calculation; additionally we could not calculate evolutionary rates for 130 singleton trees and 5 duplicable trees. Final rate comparisons are thus between 2,157 singleton and 52 duplicable groups. O_1 and O_2 indicate duplicate copies in an outgroup species.

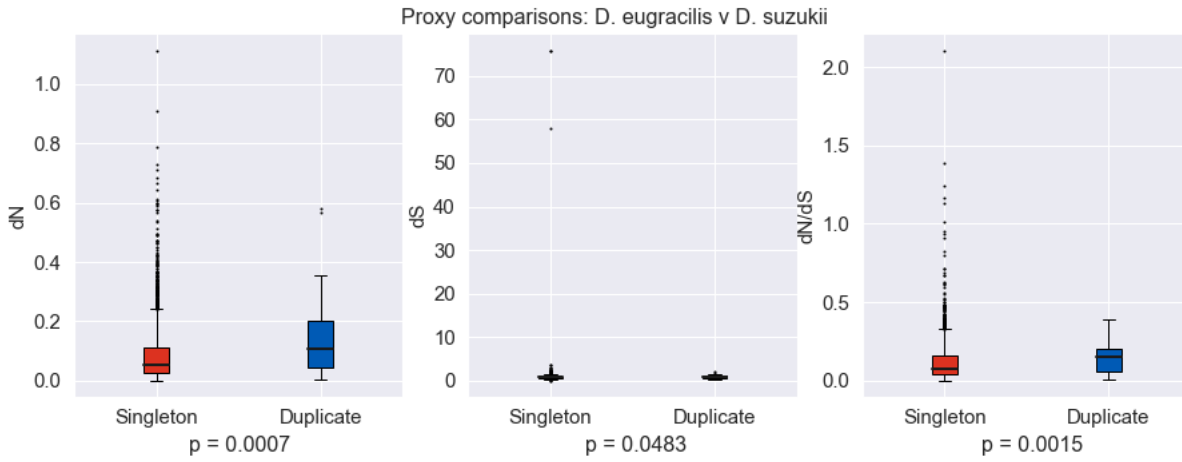


Figure 4.4 | Proxy ancestral rate comparisons with no dS filter applied. Proxy ancestral evolutionary rates differ between singleton and duplicable genes before the removal of cases where the rate of synonymous substitution is saturated. P-values are given for two-sided Mann-Whitney U-tests.

as synonymous substitutions were considered too saturated at this point to give reliable estimates, although we do find significant differences between duplicable and singleton groups even in the absence of this filter (Figure 4.5, Figure 4.4). For ingroups, pairwise rates were calculated between *D. melanogaster* and *D. eugracilis*, while for proxy ancestral rates pairwise rates were calculated between *D. suzukii* and *D. eugracilis*. Although dN/dS does not represent a rate in the strict sense, rather a ratio between two rates, we refer to it as the rate of evolution for simplicity as is conventional. The ratio of non-synonymous to synonymous rate of change is typically used as a measure of the mode of selection operating on a sequence, with values above 1 representing positive selection, below 1 representing negative selection and at 1 representing neutral selection. We also consider it sufficient for our uses as a measure of constraint, as a greater number of non-synonymous changes relative to the same amount of synonymous changes (i.e. a higher dN/dS value) should reflect a less constrained sequence less subject to the effects of purifying selection.

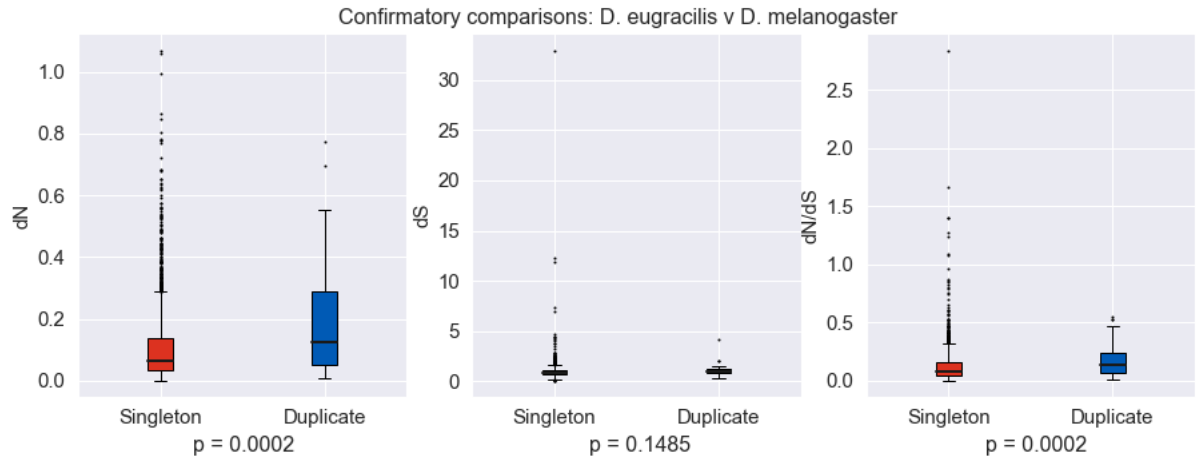


Figure 4.5 | Confirmatory rate comparisons with no dS filter applied. Evolutionary rates as measured within the duplicability assessment clade differ between singleton and duplicable genes before the removal of cases where the rate of synonymous substitution is saturated. P-values are given for two-sided Mann-Whitney U-tests.

4.2.4 PAML branch models

To test different hypotheses regarding rate changes after gene duplication, we estimated evolutionary rates under a number of different models for each duplication event in the gene trees for our duplicable gene set. Duplication events were determined using *ete3* and for each case we estimated rates under 3 models (see Figure 4.14). Firstly, the case where all rates are assumed equal (model=0); secondly, the case where the rate is allowed to differ post-duplication but rates on both branches are assumed equal (model=2 with all branches descending from the duplication labelled with the same rate); thirdly, the case where rates were allowed to differ on the two branches (model=2 with branches descending from the duplication event labelled with two different rates).

For each model, the number of parameters and the log likelihood were extracted and we performed a likelihood ratio test ($\chi = 2(LnL_{max} - LnL_{min})$) to determine if a) allowing the rate to differ from the rest of the tree post-duplication produces a better model than a model with all rates assumed equal; and b) allowing the

duplicates to differ in rate produces a better model than assuming both duplicates share the same rate. The resulting p-values were corrected using FDR.

4.2.5 Quantification of potential confounders

For each gene under consideration we determined values for CDS length, % GC content, % GC3 content based on the longest CDS. Expression values were estimated in transcripts per million (TPM) using RSEM v 1.3.3 (Li and Dewey, 2011). RNA-seq data for *D. suzukii* was obtained from SRA (SRR1002988 and SRR100289), trimmed using Trim Galore (Martin, 2011) and aligned with the STAR aligner v 2.7.7a (Dobin et al., 2013) using default parameters.

Assessment of confounder contribution

In order to assess whether any of our confounding features could be contributing to the observed difference in rate between the singleton and duplicable groups, we endeavoured to compare rate between these groups with influence from a given confounder removed. We achieved this by regressing rate on each feature in turn to obtain residuals which we took as a measure of rate independent of any variance explained by the feature under consideration. We then compared these residuals between singletons and duplicable genes using a Mann-Whitney U test, taking a reduction in p-value compared to the original rate comparison to indicate the feature was masking a difference in rate (i.e. removing the influence of the feature had made the groups less similar). By the same logic, an increase in p-value was taken to indicate the feature contributes to the observed difference in raw rates.

The regression models to determine residuals for each feature were based on ordinary least squares (OLS) regression, i.e. a linear model. We investigated that our data met the assumptions required for such models and find no meaningful deviations save for a minor negative skew in the residual distribution following

log transformation of rate and, in the case of CDS length and expression, of the predictor. As we were considering our results in light of those presented in O’Toole et al. (2018), we additionally performed the same analysis with LOWESS regression as used in that work. LOWESS regression is a smoothing approach that uses local data structure to fit a curve to a dataset. We favour OLS as the local nature of LOWESS makes these models less reliable at values where data is not densely sampled, however we include the LOWESS results to show that the choice of regression approach does not affect our conclusions.

We determined whether the change in p-value between the raw rate comparison and the residual comparison was statistically significant using Monte Carlo simulation. For each feature, we randomly permuted the values and regressed rate on these values 100,000 times in order to create a distribution for the residual comparison between singleton and duplicable genes. Based on this, we estimate the p-value for the change in p-value as the probability of finding a higher value in the case of an increase and the probability of finding a lower value in the case of a decrease. We expect that this approach to judging the significance of the change in p-value should account for any effects non-specific to the contribution of the feature in question, such as a decrease in variance that we would expect to observe in the residuals compared to raw rates.

Previous work examining the determinants of evolutionary rate has shown standard regression methods perform poorly in the presence of noisy, correlated predictors, as noisy predictors can only be imperfectly controlled for given the imprecise measurements, and suggested principle components analysis (PCA) regression was superior (Drummond et al., 2006). We chose not to implement PCA regression in this case as, firstly, we were not considering our potential confounders together in a single model so multicollinearity is not an issue within each model and, secondly, we were not necessarily interested in assessing the

relative contributions of each confounder to explaining rate variation. Rather, we sought to investigate whether any given confounder was related specifically to the difference in rates observed between singleton and duplicable groups. In the case of our method, failure to fully control for a confounder due to noise in measurements should only serve to minimise the change in p-value we observe, as some variation due to the confounder has not been accounted for in the residuals. Thus if this issue exists for any of our measurements we expect it would not lead to spurious significant results.

4.2.6 Reassessment of duplication status for fast evolving singletons

The fast-evolving singletons group was designated as singletons in the top 5% of the final set of genes by evolutionary rate. The level of taxonomic restriction was estimated as the most distant node from the focal clade where orthologs were detected. We used the abSENSE software with *D. melanogaster* as the focal species to confirm that these restriction levels could plausibly be explained by homology detection failure (HDF), using the software as described in Weisman et al. (2020) with the difference that insect BUSCO genes were used for distance estimation between species to increase the number of orthologs recovered. Orthologs included were those defined by Orthofinder. For the bitscores required we ran DIAMOND (Buchfink et al., 2021) searches for each pair of species with a generous E-value cutoff of 10 in order to maximise the ortholog pairs with a recovered bitscore. In cases where the ortholog pair did not produce a hit in the DIAMOND output the ortholog was treated as undetected. For this analysis the restriction level was defined as the most distant node where at least one species has 50% or higher probability of detection at $E \leq 0.0001$.

In order to assess whether we could detect paralogous copies for these assumed

singletons, we also looked at all detected homologies from the similarity search step for Orthofinder (both default and ultra-sensitive searches).

4.2.7 Additional statistical methods

Pairwise comparisons were carried out using the Mann-Whitney U test with Bonferroni correction for multiple testing where appropriate. Regression models for examining the effect of possible confounders were constructed using the statsmodels Python module (Seabold and Perktold, 2010). Residuals from regressing rate on each confounder were compared between duplicate groups and any change in P-value relative to the original comparison was assessed for significance by Monte Carlo simulation with $n = 100,000$.

4.3 Results

4.3.1 Faster evolving genes are more duplicable in the *Drosophila* lineage

Following from the method in O'Toole et al. (2018), we selected a group of four *Drosophila* species as outgroup species to ascertain an ancestral single copy state and provide a proxy measure of the ancestral rate of evolution. A further seven species were designated as the group for duplicability assessment (Figure 4.6). We inferred gene trees for each ancestrally single copy gene family, applied quality filters (Figure 4.3) and assessed duplicability (whether duplication has occurred within the blue clade in Figure 4.6), resulting in 2,157 singleton and 52 duplicable cases. This large difference in sample size between these groups is unsurprising as duplicable genes are expected to duplicate frequently and so our filtering for cases which are single copy in the outgroup species limits the number of these we can

include. For each case we calculated the proxy ancestral rate (between *D. suzukii* and *D. eugracilis*) as well as a rate measured in the clade of the duplication (between *D. suzukii* and *D. melanogaster*) to confirm the higher rate of evolution in duplicates.

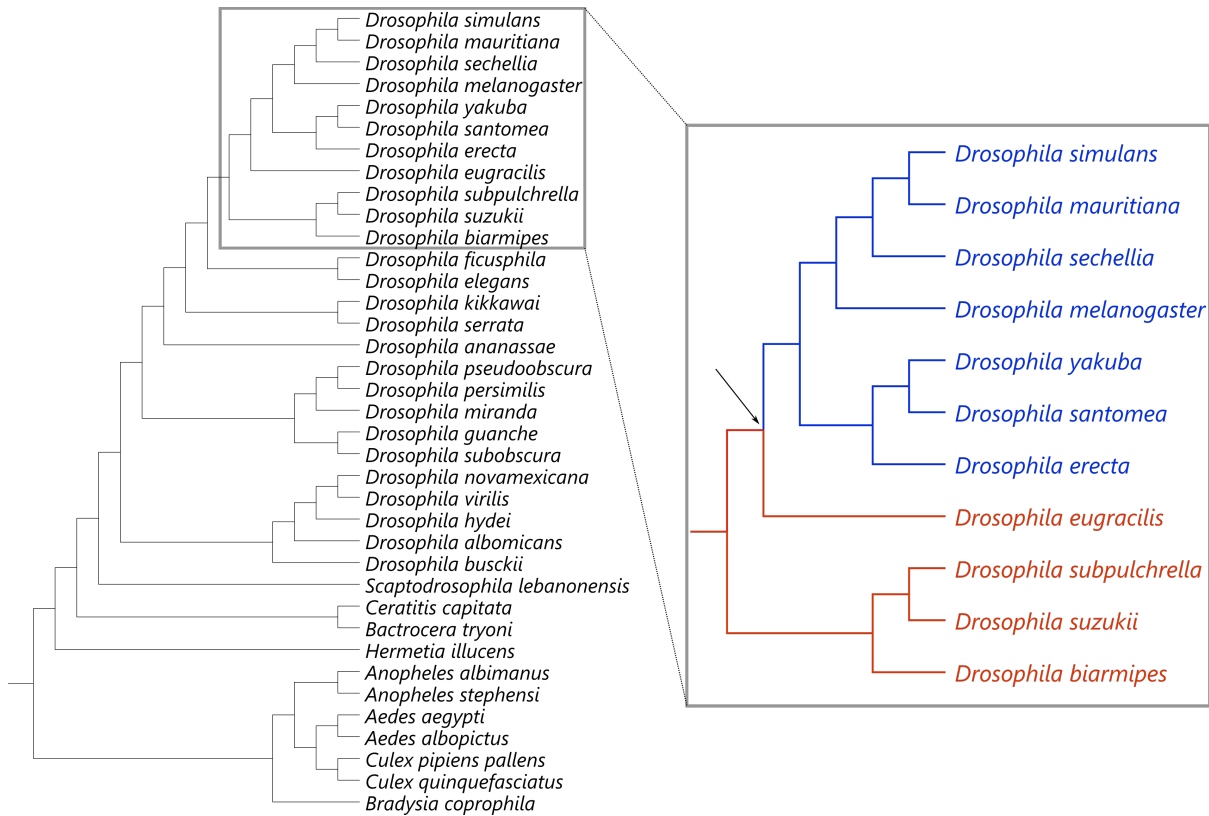


Figure 4.6 | Project strategy. All species included in orthogroup inference and gene tree building are shown. The species considered in our analysis are shown in colour with the outgroup for rate estimation shown below in red and the duplicability assessment group above in blue. The goal is to estimate a proxy for the rate at the point indicated by the arrow where we based on single-copy status in the outgroup that the gene has not yet duplicated.

We find that genes that we have designated as duplicable show faster rates of evolution both within the clade where we assess duplicability (ingroup rates; see Figure 4.7) and in singleton outgroups, which we use as a proxy for the ancestral rate (Figure 4.8). The finding that the higher evolutionary rates predate any gene duplication event supports the conclusions of O'Toole et al. (2018) that

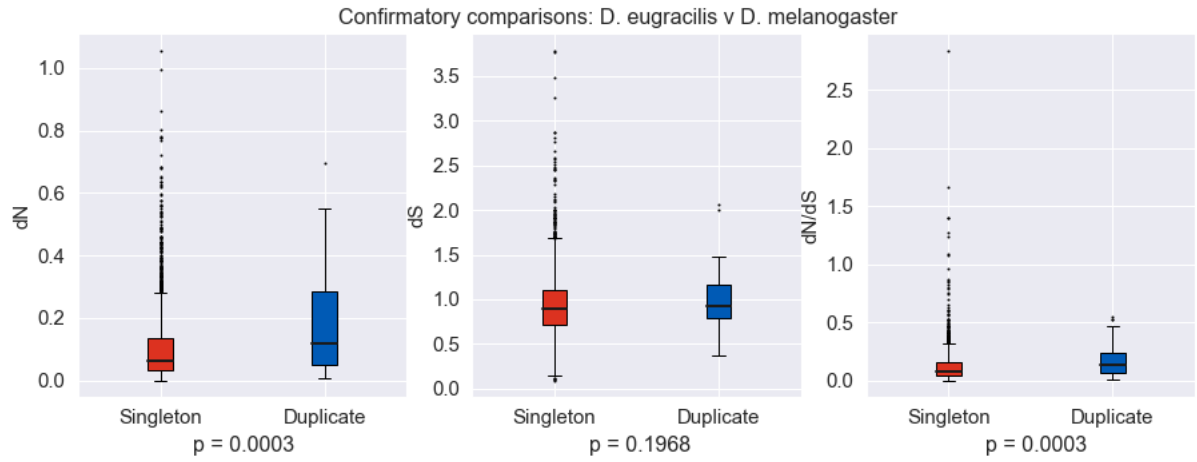


Figure 4.7 | Duplicable genes are faster evolving in a clade with duplications allowed. We confirmed the known trend that duplicated genes are faster evolving by comparing rates within the set of species where we assessed duplicability. P-values are given for two-sided Mann-Whitney U-tests.

biased duplicate retention in favour of ancestrally fast evolving genes contributes to the observed differences in evolutionary rate between single copy and duplicated genes.

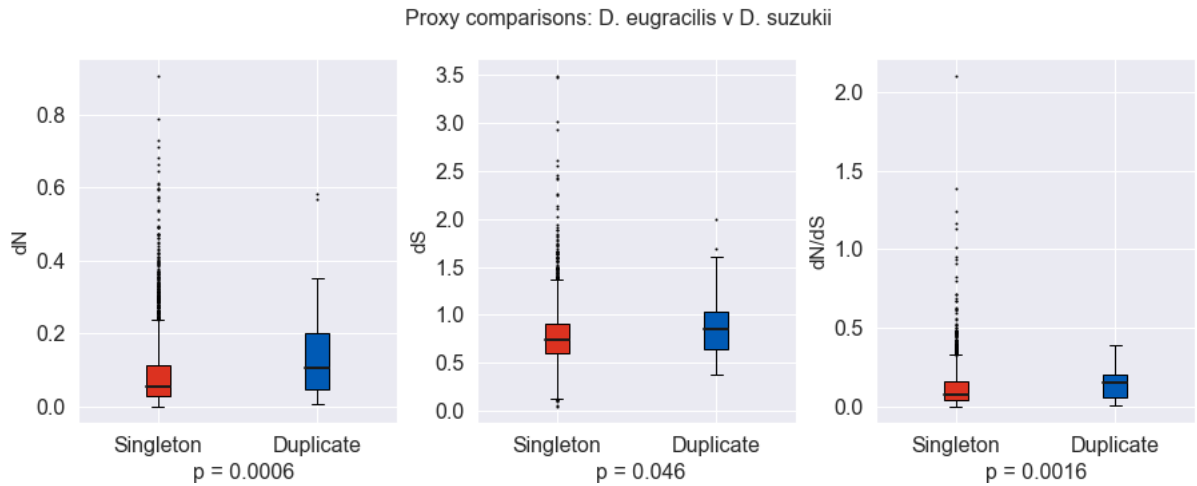


Figure 4.8 | Duplicable genes are ancestrally faster evolving. Comparisons of the amount of non-synonymous substitution, synonymous substitution and the ratio between the two for singleton and duplicable genes. Values were calculated in *D. suzukii* and *D. eugracilis* comparisons. P-values are given for two-sided Mann-Whitney U-tests and are Bonferroni corrected.

4.3.2 Potential confounders do not fully explain the faster evolution rate of duplicable genes

Having confirmed this core result, we next considered possible explanations for the higher rate of evolution in the set of duplicable genes. There are several gene features which are well known to correlate with rate of evolution, most notably expression level (Pál et al., 2001). Many of these features also show differences between groups of differing duplicability. We investigated whether any of these features explained the difference in evolutionary rate between duplicable and singleton genes.

For each of four features examined (CDS length, expression level, % GC content and % GC3 content), we compared our singleton and duplicable groups and also calculated correlations between the feature and evolutionary rate. We do not find significant differences between the groups for any of the features considered (Figure 4.9), though we note duplicates show slightly higher values for CDS length and expression level and lower values for GC content. All also show significant negative correlations with evolutionary rate (Figure 4.10). The GC content features, particularly GC3 content, show a relatively strong correlation ($\text{GC}:\rho = -0.343$, $\text{GC3}:\rho = -0.553$) in comparison to CDS length and expression level ($\text{CDS length}:\rho = -0.158$, $\text{expression}:\rho = -0.189$).

Given these negative correlations, and the fact that O'Toole et al. (2018) had found such features to either mask or explain some of the difference in rates between these groups, we then regressed rate on each feature in order to control for any variation in rate explained by the feature in question and performed another comparison on the residuals. In all cases we find a significant change in p-value for the residual comparison relative to the original rate comparison, suggesting that all features considered are in some way relevant to the observed difference in evolutionary rates. In the case of CDS length and expression level we find that

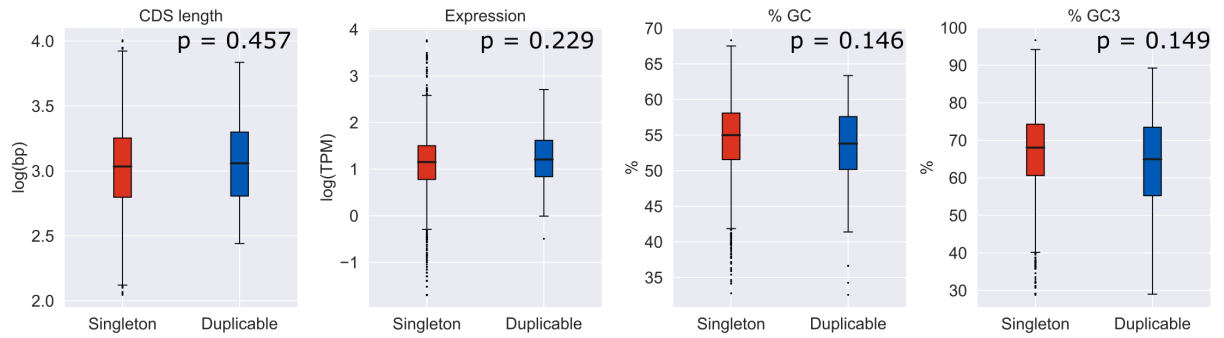


Figure 4.9 | Confounding features show no significant difference between singleton and duplicable genes. P-values given are for two-tailed Mann-Whitney U test and Bonferroni corrected.

differences in these features may be masking the difference in rates, with slightly higher values for duplicable genes combined with negative correlations with rate giving an overall smaller difference than in the comparison controlling for these features (Figure 4.11). On the other hand, it seems that sequence composition may partially explain these differences in rates. Controlling for % GC content or % GC3 content increases the p-value for the comparison, indicating that differences between singleton and duplicate groups are contributing to the rate difference (Figure 4.11). We additionally confirm these results are consistent when LOWESS regression is used (as in O'Toole et al. (2018)) rather than OLS (see Figure 4.12 and Figure 4.13).

Our results for CDS length and GC content differ from those in primates, where CDS length was found to explain some rate difference and GC content was found to have no impact. In the case of CDS length it is possible that misclassification of shorter, faster evolving duplicable genes as singletons has contributed to this difference (see later discussion of homology detection failure). However, in the case of GC (and GC3) content, the difference may be legitimate. Species such as *Drosophila* with large population sizes are expected to show stronger effects of selection and in this case we believe we are seeing evidence of more effective

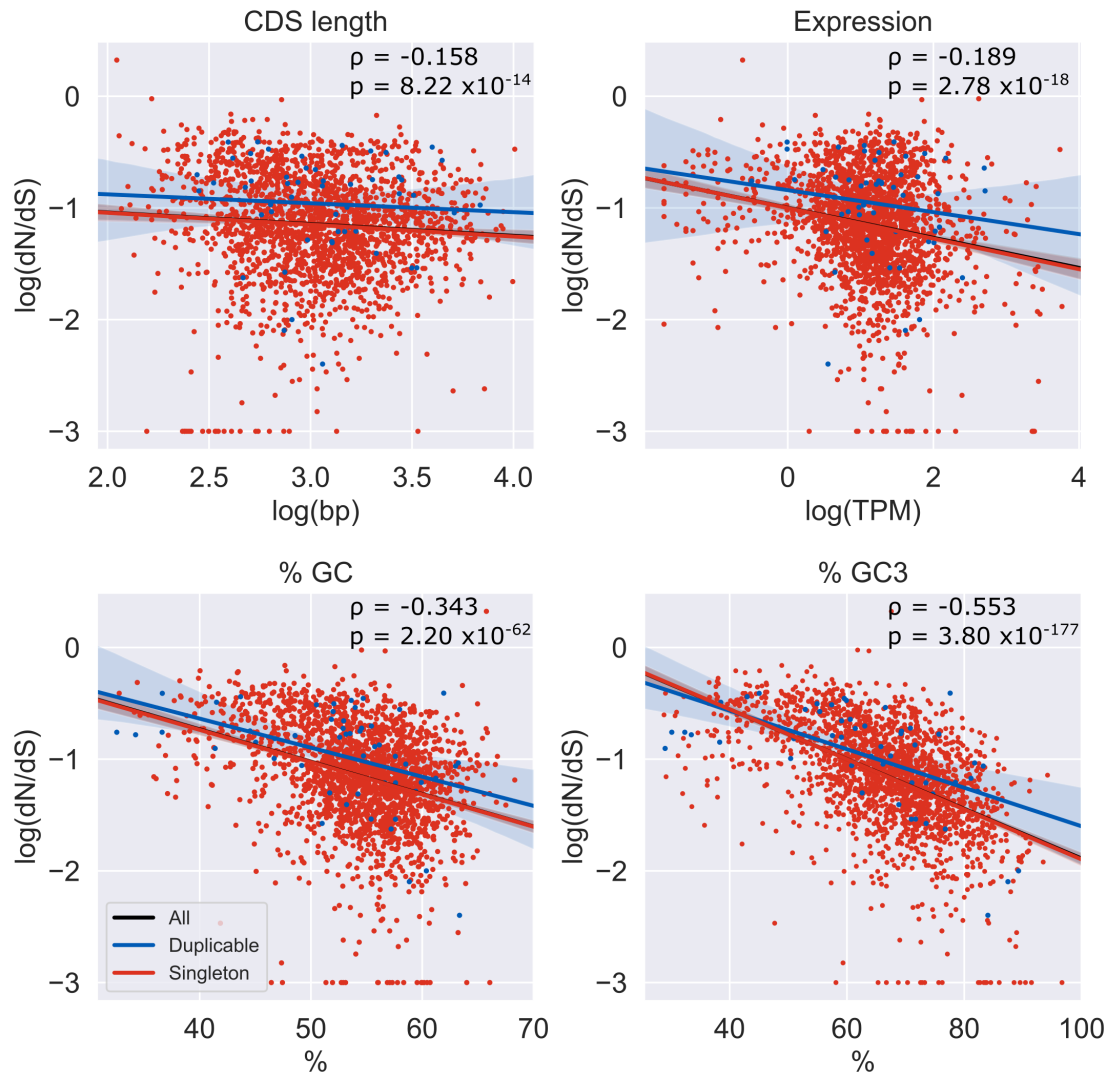


Figure 4.10 | OLS regression of rate on confounding features. P-values and correlation coefficients are shown for Spearman's correlation. Transformations are shown for each feature in axis labels.

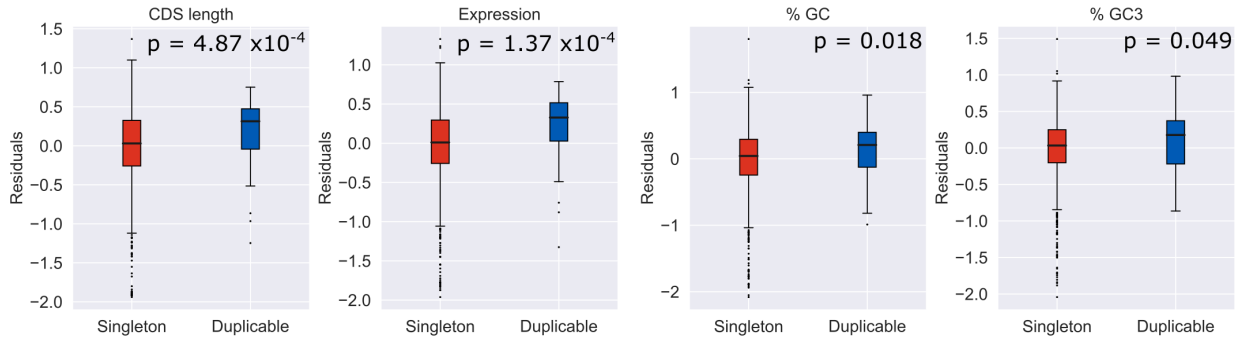


Figure 4.11 | Residual comparisons show significant changes in p-value. Comparisons are shown for residuals from regression rate on each feature. P-values are shown for comparisons using a two-tailed Mann-Whitney U test and also for the change in p-value relative to the original rate comparisons as determined by Monte-Carlo simulation with 100,000 iterations.

selection on codon usage in this lineage relative to primates. Preferred codons in *Drosophila*, both within our group of interest generally and in *D. suzukii* specifically, typically have high GC content and almost exclusively possess G/C in the third position (Athey et al., 2017; Vicario et al., 2007). Thus this difference with primates may reflect selection on codon usage and may be explained by either more effective selection in *Drosophila* as compared to primates, or differences in codon preferences.

4.3.3 Weak evidence for other implications of redundancy-driven rate increase

A bias for retention of duplicates with faster evolving parent genes does not exclude the possibility that gene duplication also allows relaxation of constraint. We therefore additionally tested whether we could find evidence for other implications of the post-duplication redundancy hypothesis, namely asymmetric rates of evolution and overall rate acceleration post-duplication.

Rate asymmetry in duplicates is supposed to originate from differences in functional constraint, with one copy assumed to maintain the functional role

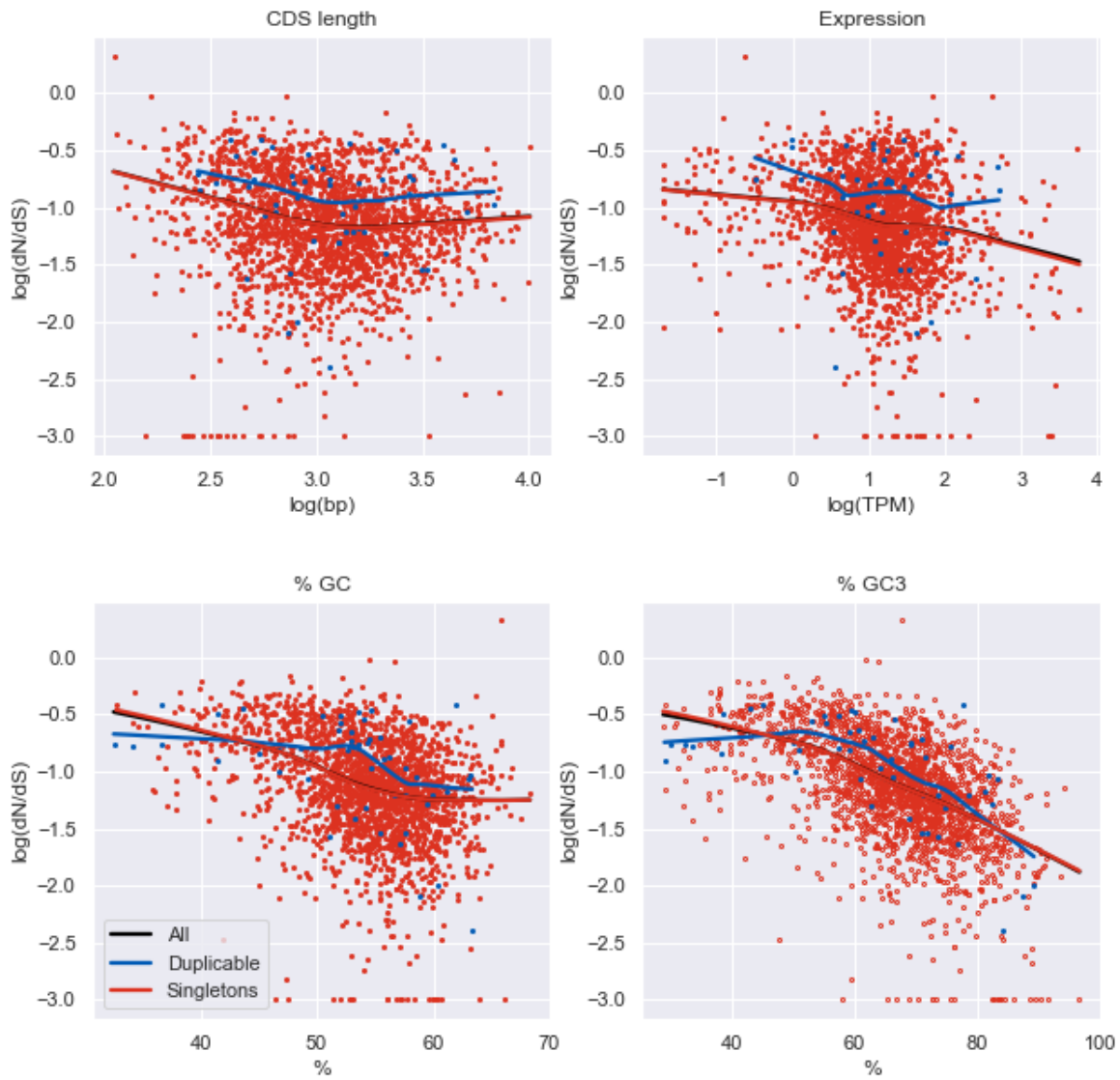


Figure 4.12 | LOWESS regression on confounders. Evolutionary rate was regressed on each of the features shown using LOWESS with $f=3$.

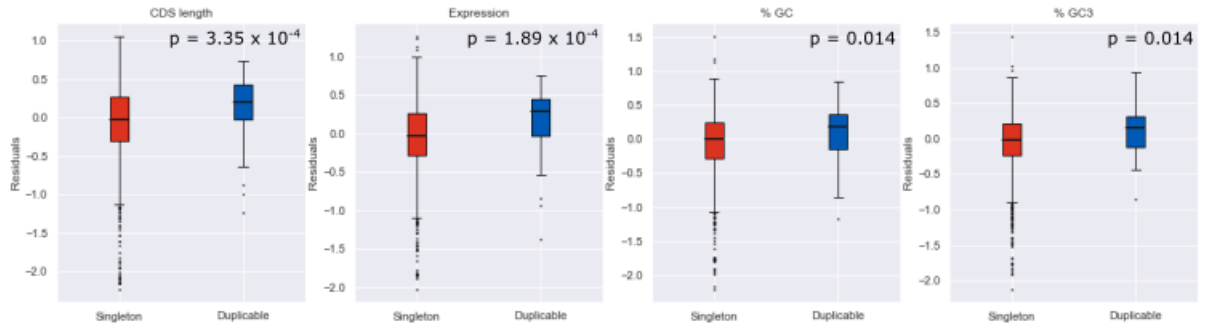


Figure 4.13 | Comparison of LOWESS residuals from regressing rate on counfouders. P-values are given for two-tailed Mann-Whitney U tests.

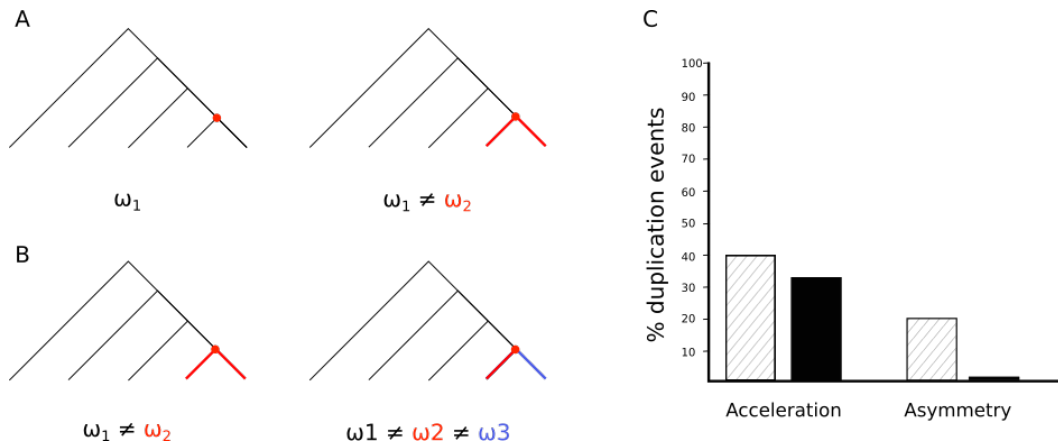


Figure 4.14 | Comparison of differing rate models does not support rate acceleration or rate asymmetry post-duplication. (A) Competing models in the test for rate acceleration. The null model assumes a constant rate across the tree while the alternative allows rate post-duplication to vary. We find that in all cases of the second model performing better the post-duplication rate is higher, and so we refer to these results as acceleration. (B) Competing models in the test for rate asymmetry. Null model assumes equal rates for both duplicate branches, the alternative allows differing rates. (C) Percentage of the 86 duplication events considered where the alternative model is found to be significantly more likely before multiple testing correction (hatched) and after (solid colour).

of the parent gene while the other is free to evolve novel functions due to the redundancy. We investigated this idea and find little evidence for this difference in rate in our dataset (Figure 4.14). When we estimate evolutionary rate using PAML under two models (either assuming the same rate for both branches post-duplication or allowing these two rates to differ), we find that after multiple testing correction only about 1 out of 86 duplication events (1%) show a higher log likelihood under the model that allows for rate asymmetry, indicating that allowing for asymmetry does not produce a superior model. However, as multiple testing correction in this case is actually more generous to the hypothesis of no post-duplication effect, we additionally use Fisher's method for combining p-values to consider all tests together. We find evidence here in favour of asymmetry ($\chi^2 = 292.96$; $p = 2.48 \times 10^{-8}$) This suggests that, while asymmetry can not be discounted as a possibility when duplicates are considered as a group, it is not necessarily a guaranteed outcome on the individual level.

We additionally considered whether there was an overall increase in evolutionary rate post-duplication for any of these genes. Again, we estimate rates under two models allowing for either one rate across the entire tree or for a differing rate post-duplication (Figure 4.14). We find 67 out of 86 duplication events (78%) where the rate of evolution post-duplication is estimated to be higher (a rate acceleration). However, in only 35 (40%) of these does the different rate produce a better model, and in only 28 of the cases (33%) is this model significantly better than the null after multiple testing correction. Similarly to the case of asymmetry, we additionally considered all duplicates as a group in order to be fair to the acceleration hypothesis and again find evidence in favour of this hypothesis ($\chi^2 = 702.30$; $p = 3.43 \times 10^{-65}$; Fisher's method for combining p-values). Nonetheless, we remain cautious in the interpretation of these results due to the possibility that the short time intervals interfere with the d_N/d_S estimates

(see Discussion).

4.3.4 Fast evolving singletons may be explained by limitations of orthology inference

One curious observation is that, despite the overall faster rate of evolution of duplicable genes, there are a considerable number of fast-evolving singleton genes in our dataset; of the top 5% of genes by rate all but four are classified as singletons. While it is possible that this group could represent an interesting exception to faster-evolving genes being more likely to duplicate, we first aimed to rule out technical explanations.

Such technical explanations of this group of fast-evolving singletons could include homology detection failure (HDF) in the case of especially fast-evolving genes. This is because fast evolution of some duplicated genes would result in extensive sequence divergence which may in turn lead to their misclassification as singletons due to the difficulty in detecting the paralogs (Wolfe, 2004). Consistent with this possibility, a majority of fast-evolving singletons in our dataset are taxonomically restricted (Figure 4.15), with 85 out of 107 (79%) lacking identified orthologs outside of the *Drosophilids* and 95 (89%) restricted to within fruit flies. Taxonomic restriction has been shown to be frequently explainable by homology detection failure (Weisman et al., 2020; Vakirlis et al., 2020).

We investigated whether HDF was a plausible explanation for the restriction pattern we observe in these fast-evolving singletons. We used the software *abSENSE* (Weisman et al., 2020) to estimate at what point in the tree we would hypothetically lose the ability to detect orthologs even if there were orthologs present and the gene has been evolving at a constant rate. We find that HDF is a plausible explanation in 54% of cases (58/107), with *abSENSE* predicting loss of detection at or after the point where no orthologs are detected. We also

observe a strong significant correlation between the observed taxonomic restriction level and that predicted by detection failure probabilities from abSENSE ($\rho = 0.756, p = 4.56 \times 10^{-21}$; Figure 4.15B). Thus these data support the idea that the fast evolving singletons observed here may be misclassified due to missed paralogs.

We further considered whether our choice of tool for orthology inference contributed to the presence of cases with potentially missed paralogs and where in the process this failure may have occurred. Firstly, this pattern is not unique to our choice of orthology inference tool or settings, we find that other tools have similar, if not greater, levels of HDF. In the case of SonicParanoid there are 114 genes which meet the criteria of being a fast-evolving singleton (97 in common with Orthofinder), while running Orthofinder with an ultra-sensitive similarity search yielded 111 (102 in common with default parameter Orthofinder). Regardless of the settings used, we fail to recover any duplicable cases at all when OMA is used as the orthology inference tool (i.e., all genes are classified as singletons).

It seems that missed paralogs may involve failures at either the stage of homology detection, or clustering into orthogroups based on the similarity. We attempted to locate potential paralogs based solely on homology search rather than the final clustered groups and find very few cases, however we note a slight increase in groups with possible missed paralogs when comparing default Orthofinder to Orthofinder ultrasensitive (4 cases as opposed to 2). Paralogs which could be detected at the stage of homology searches but which are not called as paralogs are potentially ancestral duplications that occurred prior to the divergence of our species set or could possibly have failed to cluster properly at the clustering step. As we recover very few of these cases, however, we believe the majority of errors would lie with HDF.

We examined whether the genes we designate as fast-evolving singletons are

generally distinct from other groups in the dataset and found that they are a distinct group in terms of all the confounding features we had earlier considered (Figure 4.15C). In all cases, these singletons show significantly lower values than either other singletons or duplicable genes, with the effect particularly pronounced for CDS length and GC3 content. As shorter, less conserved genes are more prone to HDF, and GC3 content is related to conservation level, this provides further support that HDF could be a plausible explanation for the faster evolutionary rates of these singletons

4.4 Discussion

The work presented here extends previous observations on the faster rate of evolution of duplicable genes from primates to insects, and thus supports the interpretation that any apparent contradictions from previous studies (O'Toole et al., 2018; Davis and Petrov, 2004) are not due to biological differences between the lineages (such as generation time and population size), and may instead reflect differences in the study design. Most notably, the vastly different time periods under investigation may inadvertently create a bias for inclusion of slowly-evolving genes because these are the ones where homology remains detectable.

As to why the method used in Davis & Petrov lead to the opposing result, as we discussed briefly previously, the restrictions placed on the set of genes under consideration by the considerable timespan the study species cover may have affected the outcome. To elaborate further, requiring identification of orthologs between yeasts and insects creates a bias towards more conserved sequences. Moreover, no restriction was placed on copy number in the outgroup, which not only seems to defeat the purpose of measuring rate in a different species but also implies that any of the duplication detected within the study species could reflect

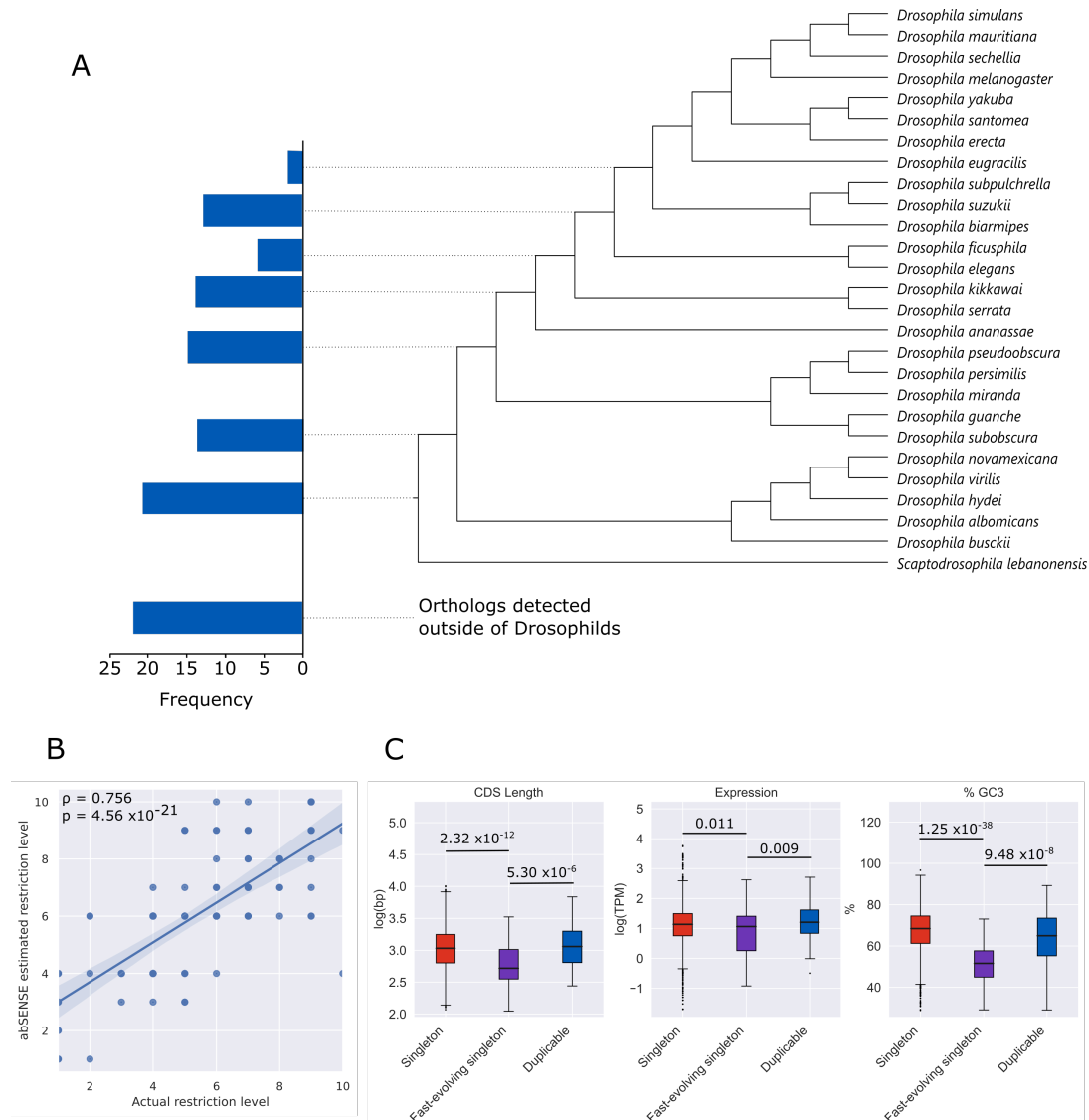


Figure 4.15 | Fast-evolving singletons are a distinct group partially explained by HDF. (A) The distribution of singletons with orthologs identified at the node indicated but no further. (B) Correlation between observed restriction level and that predicted by abSENSE assuming no process other than constant sequence similarity decay for each fast-evolving singleton, Spearman's correlation given. (C) Comparison between fast-evolving singletons and other groups in the analysis. P-values are given for two-tailed Mann-Whitney U tests

any age from ancient duplications shared by both species up to recent yeast or nematode specific duplications.

The level of conservation in duplicated genes varies considerably with the age of duplication (Woods et al., 2013) and given that the initial data set is already predisposed to a higher level of conservation we might assume that the duplications detected may represent an older, more conserved subsample. Such ancient duplications may well be more conserved than singletons, however these paralogs have long since moved beyond the stage of redundancy and initial duplicate evolution which we present here. This said, Davis & Petrov consider duplication age (as measured by Ks) as a possible source of error and find results to be largely consistent across time. Possibly due to the bias created towards more conserved genes in the dataset the younger duplications recovered reflect those which are more likely to be retained in the long term and so their behaviour mirrors that of older duplications.

Prima facie our data lend support to the classical view that after duplication there is also rate acceleration, albeit not typically associated with much rate asymmetry. It is thus possible that both fast evolving genes duplicate more and that after duplication genes can accelerate in their evolution. However, we remain cautious about the rate acceleration. Curiously in all cases the rate of evolution (d_N/d_S) after duplication is higher than that before duplication. Such a uniformity is, in our experience, unusual and more typically associated with a systematic bias. Indeed, it has previously been argued that unresolved polymorphism between recently diverged species can lead to false d_N/d_S signals (Rocha et al., 2006). This is owing to a time lag for a weakly deleterious mutation to be removed from a population. Indeed, if one captures mutations at the point at which they occur (e.g. comparing parents and offspring) d_N/d_S should be unity (or near unity if we allow for rare lethal mutations) as the non-synonymous mutations have yet to be

removed from the population. As Rocha et al. (2006) observe, the expectation is that as the distance between two comparators increases so d_N/d_S declines. For the *Drosophila* species the post duplication branch lengths are small hence there remains the possibility that apparent rate acceleration is an artefact of short branch lengths and unresolved polymorphisms contaminating pairwise analyses. That polymorphism levels are higher in flies than in primates adds to the concern. Masking of polymorphic sites in all genes in all species could enable a possible test of this alternative explanation but is currently not possible.

Additionally, some of our results may be impacted by limitations in homology detection and orthology inference. If particularly quickly evolving duplicable genes are misclassified as singletons this would mean that the association between evolutionary rate and duplicability is under-reported here. By contrast, the impact of such homology detection failure on the estimation of post-duplication rate asymmetry is harder to predict. Asymmetrically evolving duplicates are known to be difficult to place correctly in orthology inference as the faster-evolving paralog can have a much greater sequence distance to other orthologs than its sister and be incorrectly clustered outside of the group as an out-paralog (Train et al., 2017). As such, any missed duplication events may be biased towards cases where rate asymmetry has occurred, leading to a potential under-detection of rate asymmetry here.

It is not clear why reliable duplicate detection appears to be more difficult in *Drosophila* than in primates. There is known to be quick turnover and fast-acting selection on new duplicates in *Drosophila* (Jiang and Assis, 2017; Li et al., 2019), possibly as a result of more efficient selection in a high N_e species. Potentially this could drive faster loss or divergence of paralogs in this lineage, thus making them more difficult to detect even over relatively short stretches of time. We do not address here the possibility that some of the fast-evolving singletons may

be legitimate *de novo* genes (Vakirlis et al., 2020). Our tests with abSENSE do not rule this out as a possibility, only confirm that it is possible the taxonomic restriction pattern has arisen through HDF.

In this work we have resolved previous conflicts in results regarding the source of faster evolutionary rates of duplicable genes. We show that a faster rate in duplicable genes prior to duplications is not limited to the primate lineage and this finding is likely broadly applicable. Overall, our observations support the idea that differences in rate between singleton and duplicated genes have their source before the duplication event occurs and are best explained by biases in duplicability. The nature of the biases is not clear. Previous work in *C. elegans* supports the interpretation of duplication as just another kind of variation (Woods et al., 2013), and genes that are less constrained in sequence evolution are also generally less constrained in terms of copy number evolution. However, it is interesting to speculate whether the faster rate of evolution might contribute to a greater chance of duplicate fixation through functional diversification.

Chapter 5

Consistency of gene duplicability varies across the genome

5.1 Introduction

Despite the large body of existing work on the topic of duplicability, little is known about how duplicability as a property varies over time or interacts with environmental factors i.e. is duplicability an intrinsic property of genes or is it in some way context dependent? Existing knowledge on gene duplication suggests features intrinsic to a gene such as dosage sensitivity and constraint are certainly relevant to duplicability, however there are many ways in which some degree of context dependency could manifest. For example, we might observe that genomic context is relevant as the requirement for dosage balance may diminish with increasing copy number through successive WGDs as the relative impact of losing a single paralog lessens. Losing a paralog following one doubling corresponds to a 2:2 ratio dropping to a 2:1 ratio, a 50% decrease, but losing a single paralog after a second doubling gives a 25% decrease, after a third a 12.5% decrease and so on. This hypothesis is supported by the observation in maize that addition

of an additional copy of a chromosome arm to a haploid produces a more severe phenotype than the same addition in a diploid background (Birchler and Veitia, 2012). The size of the dosage increase relative to the genetic background it occurs in is important in determining the outcome of the increase and we can reasonably hypothesise that this also applies to loss of WGD-derived duplicates in the course of evolution.

The environmental conditions at the time of duplication may also be of relevance, with many WGDs in plants appearing to cluster around times of extreme environmental stress or upheaval (Estep et al., 2014; Vanneste et al., 2014), although it is not clear if changes in environment promote WGDs as a mutation or create a situation where WGD is beneficial. When independent plant WGD events were grouped into 'waves' based on timing, it was found that each wave is associated with convergent retention of the same gene families across different lineages (Wu et al., 2020). These convergently retained families are associated with the specific environmental stresses of the time period, for example families related to water deprivation stress are retained in the most ancient wave studied, which corresponds to a time of particularly arid climate. Thus, environmental conditions at the time of a duplication may impact the set of genes which are retained in duplicate.

The primary body of work examining this question has involved plants, where the majority of known WGD events have taken place, making them a good model for examining interactions between successive duplication events. Evidence has been found for consistencies in gene behaviour across species and duplication events, with previously retained duplicates more likely to be retained in subsequent events and core genes resistant to duplication identified across many lineages (Smet et al., 2013; Paterson et al., 2006; Rong et al., 2010). Work outside of plants is more sparse but does not seem to reflect the same findings.

Fish are the primary candidates for studies on WGD in vertebrates, with multiple independent WGD events (Cheng et al., 2021; Du et al., 2020; Jaillon et al., 2004), some relatively recent such as the 4R duplications in salmonids and cyprinids (Berthelot et al., 2014; Chen et al., 2019; Lien et al., 2016; Xu et al., 2019). In the case of this lineage, there are few dedicated investigations into the consistency of duplicability, although there have been some interesting findings in the characterisation of various WGD events. For example, conflicting results have been reported regarding the interaction between the teleost 3R WGD and the salmonid 4R WGD, with findings that 3R ohnologs are no more likely to be retained following the 4R WGD (Lien et al., 2016) and also that tetrasomy following the 4R aids in preservation of 3R ohnologs (Campbell et al., 2019; Christensen and Davidson, 2017). In the case of rainbow trout, it was found that there is significant overlap between the duplicates retained from the salmonid 4R WGD in this species and those retained in the vertebrate and teleost WGDs (Berthelot et al., 2014). But perhaps the most relevant existing assessment of duplicability in fish is found with regards to the sterlet sturgeon (Du et al., 2020), with a broad comparison of the genes retained in duplicate across the 3R WGD in this species and multiple species across the teleost 3R, salmonid 4R and cyprinid 4R. Only 27 genes were recovered in which retention was consistent across all species considered, leading the authors to conclude that it is unlikely strong adaptive processes are acting at the single gene level and that stochastic and lineage specific processes play a stronger role than any force that acts convergently across events, although they later conclude that glutamate receptors have been convergently over-retained across the sturgeon and teleost events so it is possible that any convergent effects could be restricted to specific functions or families.

With a relatively small body of work encompassing conflicting results, it is difficult to discern a clear answer on whether the same trends regarding consistency

of duplicability observed in plants also apply to fish. Additionally, much of this work has restricted itself to the extremes of duplicability i.e. whether a gene is always retained in duplicate or never (Du et al., 2020), which both fails to define the overall spread of consistency in behaviour and gives little information on whether any hypothesis claiming lineage specificity is viable as such broad analysis does not leave room to assess if retention patterns differ from one event to the next.

In this work, we address this gap in the existing literature by defining groups which are consistent in their post-WGD behaviour using multiple methods and showing the spread of duplicability and duplicability consistency across the genome. We assess whether the retention patterns for specific functions can be best described as convergent or lineage specific and find that examples of functions which are consistent across events outnumber those which could be considered potentially specific to a given event.

5.2 Materials & Methods

5.2.1 Data sources and processing

Genome annotations for 23 bony fish genomes broadly distributed across the tree were obtained from NCBI genome using the FTP site (see Table 5.1), tree shown in Figure 5.1 for the species used. The longest translation for each gene was considered. Gene families were obtained by running Orthofinder (v2.5.4) (Emms and Kelly, 2019) using default settings. Ohnolog pairs were determined by synteny as defined using i-ADHoRe (v3.0) (Proost et al., 2012).

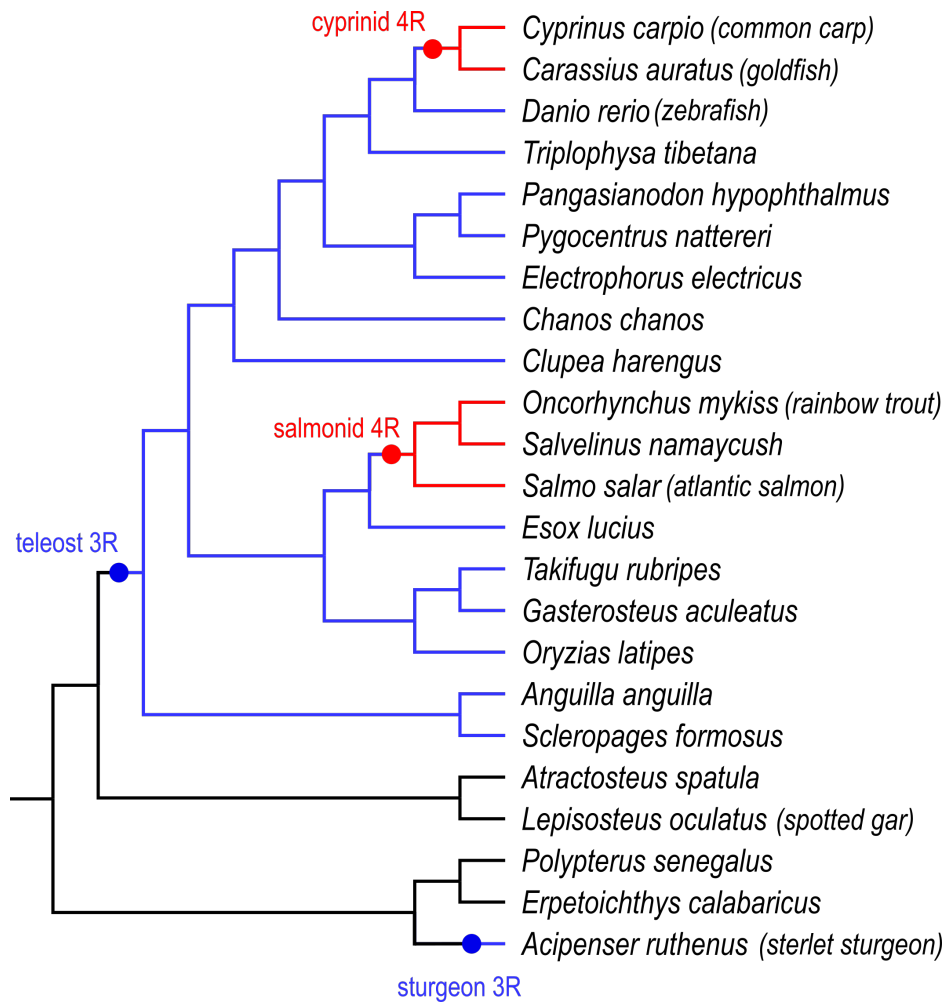


Figure 5.1 | Bony fish species considered in this work.. Species affected by a third round of WGD following the initial two rounds of WGD in jawed vertebrates are shown in blue, those affected by a fourth round are shown in red. Species tree obtained through Orthofinder (Emms and Kelly, 2019) using STAG (Emms and Kelly, 2018) and rooted using STRIDE (Emms and Kelly, 2017)). Branch lengths are not to scale.

Table 5.1 | Genome assemblies used in this analysis.

Species	File location in FTP site
<i>P. senegalus</i>	GCF/016/835/505/GCF_016835505.1_ASM1683550v1
<i>E. calabaricus</i>	/genomes/refseq/vertebrate_other/Erpetoichthys_calabaricus/ latest_assembly_versions/GCF_900747795.1_fErpCal1.1
<i>A. ruthenus</i>	GCF/010/645/085/GCF_010645085.1_ASM1064508v1
<i>L. oculatus</i>	GCF/000/242/695/GCF_000242695.1_LepOcu1
<i>A. spatula</i>	GCA/016/984/175/GCA_016984175.1_BGI_Aspa_1.0
<i>A. anguilla</i>	GCF/013/347/855/GCF_013347855.1_fAngAng1.pri
<i>S. formosus</i>	GCF/900/964/775/GCF_900964775.1_fSclFor1.1
<i>C. harengus</i>	/genomes/refseq/vertebrate_other/Clupea_harengus/ annotation_releases/101/GCF_900700415.1_Ch_v2.0.2
<i>C. carpio</i>	GCF/000/951/615/GCF_000951615.1_common_carp_genome
<i>C. auratus</i>	GCF/003/368/295/GCF_003368295.1_ASM336829v1
<i>D. rerio</i>	GCF/000/002/035/GCF_000002035.6_GRCz11
<i>E. electricus</i>	GCF/013/358/815/GCF_013358815.1_fEleEle1.pri
<i>C. chanos</i>	GCF/902/362/185/GCF_902362185.1_fChaCha1.1
<i>P. hypophthalmus</i>	GCF/009/078/355/GCF_009078355.1_GENO_Phyp_1.0
<i>T. tibetana</i>	GCA/008/369/825/GCA_008369825.1_ASM836982v1
<i>P. nattereri</i>	GCF/015/220/715/GCF_015220715.1_fPygNat1.pri
<i>E. lucius</i>	GCF/011/004/845/GCF_011004845.1_fEsoLuc1.pri
<i>S. namaycush</i>	GCF/016/432/855/GCF_016432855.1_SaNama_1.0
<i>S. salar</i>	GCF/000/233/375/GCF_000233375.1_ICSASG_v2
<i>O. mykiss</i>	GCF/013/265/735/GCF_013265735.2_USDA_OmykA_1.1
<i>O. latipes</i>	GCF/002/234/675/GCF_002234675.1_ASM223467v1
<i>T. rubripes</i>	/genomes/refseq/vertebrate_other/Takifugu_rubripes/ latest_assembly_versions/GCF_901000725.2_fTakRub1.2
<i>G. aculeatus</i>	GCF/016/920/845/GCF_016920845.1_GAculeatus_UGA_version5

5.2.2 Family size distribution

Gene families from Orthofinder were divided into 'WGD' and 'Background' sets based on the ohnolog pairs extracted from i-ADHoRe; any family containing a paralog pair determined to be an ohnolog pair by the software was assigned 'WGD' while any family not containing a putative ohnolog pair was assumed to belong to the background rate of duplication encompassing other sources of paralogs such as SSD and retroduplication. We do not expect to recover families containing paralog pairs from the early rounds of vertebrate WGD as these duplications predate the earliest species divergence in our dataset and so orthogroup definition in Orthofinder should create two families from any such duplications; i.e. Orthofinder creates individual families for each gene present in the ancestor of the species considered, 2R ohnologs would already be present in multiple copies at this time point and therefore give rise to multiple families. Gaussian distributions were fitted to the WGD set of family sizes using a Gaussian mixture model (GMM) in the scikit-learn Python package with the number of WGD events expected in the species as the number of components. GMMs are a type of clustering method that assign data to a pre-determined number of Gaussian distributions, here used to recover the multiple peaks in gene family size created by multiple WGDs.

5.2.3 Cross-event overlap definition

For each gene family, the family was classified as 'retained' for a given species/duplication event if a paralog pair in the family was classed as an ohnolog pair from that event based on the output from i-ADHoRe. Overlap size between pairs of events was tested for significance using the hypergeometric test, which assesses whether the number of 'successes' (in this case, the number of genes in common between events) obtained on drawing a sample from a population is significantly different from random. This test should at least partially account for the issue of differing

retention rates across WGD events as it takes into account both the sample size being drawn (the number of families retained in one event) and the probability of a success (what proportion of total families are retained in the other family), where a success is a family retained in the first event also being retained in the second one. For example, if one of the WGD events under consideration had just occurred and thus had a 100% retention rate then it would be impossible to obtain a significant p-value as the probability of all families retained in any event also being retained in this event could only be 1.

5.2.4 Modelling

Gene family size evolution was modelled using the WGDgc R package with birth and death rates set as equal and WGD retention rates set to 1 in order to replicate the method used by Tasdighian et al. (2017) to estimate reciprocal retention. This method takes gene count data and a tree including WGD event positions as input and estimates the birth-death rate λ (a measure of reciprocal retention) based on gene count at internal nodes determined on a gene birth-death model and gene count doubling with probability of 1 at WGD nodes. A λ value of 0 represents perfect reciprocal retention with an SSD birth rate of 0 and also a death rate for the WGD-created duplicates of zero; the closer this value is to zero for a family, the more strongly reciprocally retained the family. Higher λ values are less straightforward to interpret due to the lack of separate birth/death parameters as they could result from either or both of these parameters having a high value, increasing λ values are thus best interpreted as simply 'less strongly reciprocally retained'.

Tasdighian et al. (2017) chose to use a tree with branch lengths represented in substitutions per codon rather than in time in due to varying evolutionary rates across the angiosperms; with rate variation it may not be realistic to assume similar

Table 5.2 | Source of *Ks* limits used in WGD placement

Species	Upper <i>Ks</i> limit	Source
<i>D. rerio</i>	2.2	Berthelot et al., 2014
<i>A. ruthenus</i>	0.05	Du et al., 2020
<i>O. mykiss</i>	0.5	Berthelot et al., 2014
<i>C. carpio</i>	0.3	Chen et al., 2019

gene duplication/loss rates across absolute time. As there are also variances in evolutionary rate within our species set (e.g. sturgeon is known to be particularly slow evolving, see Du et al. (2020)), we also took this approach and followed a similar method for correct placement of WGD events along their respective branches. WGD events were placed within the tree by estimating the number of substitutions per codon (t) per ohnolog pair with the codeml module of PAML (Yang, 2007). For the final figure used in positioning the WGD event, we estimated the mean value for the duplication event from the obtained distribution of the log of values for t using Gaussian mixture models with the number of components fixed at the number of WGD events the species is expected to have experienced in this timeframe. As in Tasdighian et al. (2017), we limited pairs included in the curve fitting to those with $t < 6$ and $Ks < 5$, additionally excluding pairs which showed *Ks* higher than the likely range for the event as determined in previous studies Table 5.2. *O. mykiss* was used in this estimation in the case of the salmonid 4R as there was no clear *Ks* estimate available for *S. salar*. Estimation of reciprocal retention was limited to families with orthologs present in all species and no more than 12 copies of a gene per species.

5.2.5 Investigation into relationship between reciprocal retention and other gene features

Protein-protein interactions (PPIs) were obtained from the STRING database (Szklarczyk et al., 2021), 2R ohnolog status was obtained from Singh et al. (2019) for *L. oculatus*, both relaxed and strict definitions were considered. A χ^2 test was used to test if families in the lowest 5% by lambda value or families identified in the overlap analysis as consistently retained were also more likely to be 2R families or to be single copy in species not expected to have undergone a WGD in this timeframe. PPI count was retrieved for zebrafish orthologs and was compared between groups using a two-tailed Mann-Whitney U test.

5.2.6 GO function analysis

Functional enrichment/depletion analysis was generally performed by extracting the zebrafish orthologs of each family and testing this gene set using gProfiler (Raudvere et al., 2019). When considering potentially lineage specific functions, we additionally took a different approach to assess the relative retention of a given function in the case of younger WGD events where enrichment may be more difficult to detect due to the earlier stage of WGD resolution contributing to noise. In the case of the second approach, we defined the groups of functional categories for which the retention percentage fell beyond 1.96σ from the mean as over/under retained. Functions which were assigned to fewer than 10 families as very slight differences would have extreme effects on the percentage retained. Groups which were ranked in the top 25 functions by retention percentage were also defined following recovery of a very small number of functions using the cutoff method. In older events it is challenging to identify lowly retained functions in a background of high likelihood of loss and vice versa for younger events, hence the low quantity

of functions recovered using the statistical cutoff.

5.3 Results

5.3.1 Gene family size distribution is consistent with a broad spread of intragenomic duplicability

We initially sought to characterise the distribution of gene family sizes in species with varying WGD histories as a proxy for duplicability within the fish. We note that, in the absence of WGD, gene family size is heavily biased towards lower values (Figure 5.2). Distributions have a long tail present but most genes are present in low copy number of 1 or 2 in any given species. WGD families (those which contained paralogs created by a WGD) can be modelled as a set of Gaussian distributions, giving peaks at or slightly below the expected maximum copy number for each event given the WGD history.

Overall, the distribution of gene family sizes indicates that these genomes comprise heterogeneous groupings with many subsets of genes possessing differing duplicability. Such a spread in behaviour is relevant to keep in mind in considering duplicability in the broader sense across events.

5.3.2 Significant overlaps exist in retention following independent WGD events

We next investigated whether the retention patterns for independent WGD events are lineage-specific by means of testing for significant overlap between families showing retention for each WGD event. In each pairwise comparison between events for a representative species, we find the overlap is significantly higher than would be expected by chance for all pairwise comparisons (Figure 5.3).

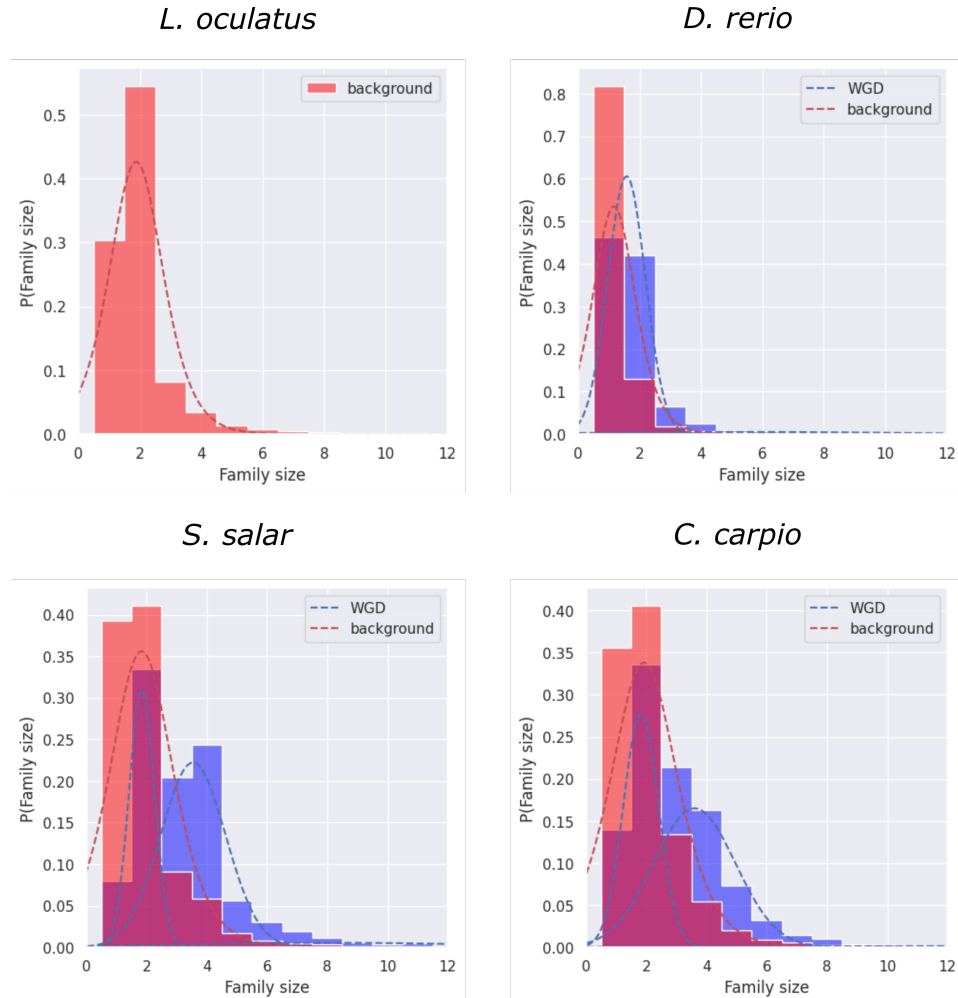


Figure 5.2 | Family size distributions show intragenome duplicability variation. Background (non-WGD) families are shown in red, families showing WGD retention in blue, overlap appears in purple. *L. oculatus* (spotted gar) represents a species with no WGDs in this timeframe, *D. rerio* (zebrafish) is acting as a representative species for the teleost specific 3R, and *S. salar* (Atlantic salmon) and *C. carpio* (common carp) for the salmonid and cyprinid specific duplications, respectively. Distributions for the salmonid and cyprinid species show two duplications.

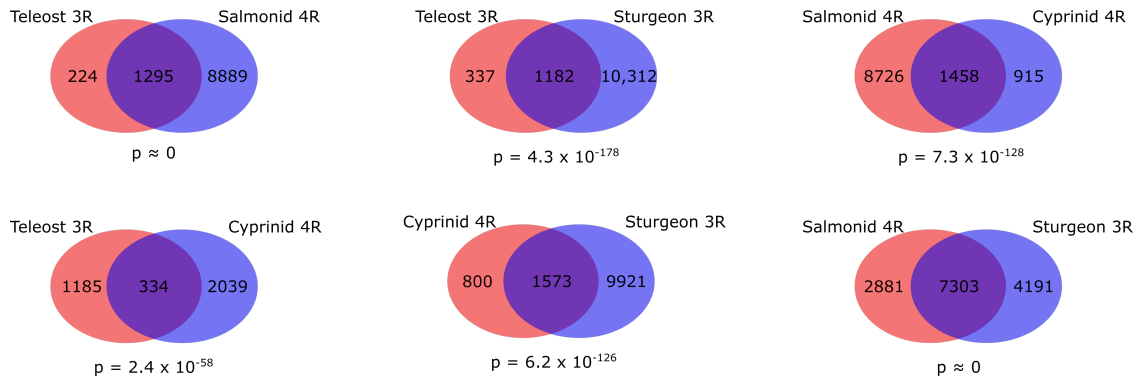


Figure 5.3 | Significant overlaps exist across ohnolog retention sets. Overlap between families showing retention for the events indicated, species used as representatives of each event are as in Figure 5.2. P values given from a hypergeometric test.

Delving deeper into these overlaps, we find that there are also substantial commonalities across events when considering the entire group Figure 5.4. It is clear that the overall retention rate is driving much of the observed patterns here, with the largest overlap (5390 families) belonging to the overlap between the two largest sets of families (the salmonid 4R and sterlet sturgeon 3R) and all overlaps involving these events tending to be larger (e.g. the largest overlaps for both of the other events are those with both of these events). This is clearly not meaningful information in terms of whether these families show consistent retention. This said, we can identify that there are a substantial number of families (253 families) which show retention across all events and a much greater number that show retention across 3 of the 4 events (1986 families total showing retention in any 3 events). While event-specific features, such as the time since the duplication occurred and the speed at which the event resolves, are likely to be the main determinant of set size and therefore the absolute size of these overlaps, it is clear that these relationships are not random and not broadly reflective of lineage-specific factors driving retention. We defined the 253 families showing retention across all 4 events as a group showing consistency in duplicability retained this group for further analysis (referred to as the 'overlap-defined group' going forward).

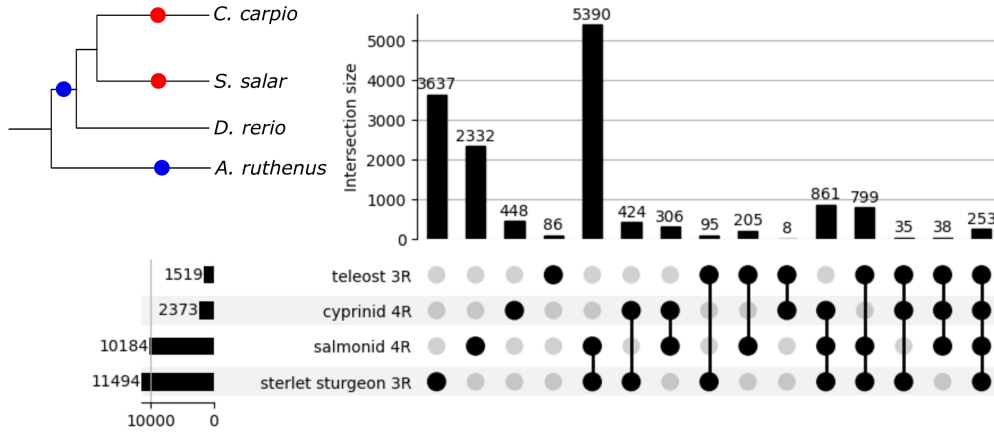


Figure 5.4 | Identification of highly overlapping retention. Total number of families showing retention is given on the far left. Events considered in each overlap are indicated in the bottom panel with overlap size above. Each overlap is exclusive, i.e. the overlap showing 'sterlet sturgeon 3R' and 'salmonid 4R' only included families which show retention for these two events and no others, 5390 families in this case. The relationship between the events is shown in the top left with the sterlet sturgeon 3R (in *A. ruthenus*) and teleost 3R shown as blue points, and the cyprinid 4R and salmonid 4R in *C. carpio* and *S. salar* shown as red points.

5.3.3 Gene duplicability can be modelled as a continuous value

Although the straightforward investigation of overlap sizes gives us some indication of consistency across duplication events, it does present some issues. As mentioned above, the age of the duplication event impacts on the current number of retained duplicates, which may obscure signal as many of these duplicates are likely not destined for long-term retention. Additionally, this fails to take into account small-scale duplicability. Although generally they are not properties observed to co-exist within the same family (see Chapter 3), small-scale duplicability and retention post-WGD are occasionally observed together and ignoring this could be failing to account for differences between this group and 'pure' ohnologs. Moreover, this analysis is only as accurate as our ohnolog assignments, with any missed ohnologs or misclassified paralogs contributing to what is likely to be considerable

randomness in the retention patterns. Due to these limitations, we chose to further investigate consistency of duplication patterns by modelling duplicability as a continuous trait of gene families. This approach lessens the effects of these issues, with both time since duplication and SSD duplicability taken into account and the need to define ohnologs removed as raw gene count can be used as input.

Using a method previously applied in defining reciprocally retained genes in angiosperms (Tasdighian et al., 2017), we estimated a parameter (λ) which captures how closely the observed gene counts correspond to those expected under perfectly consistent retention of WGD duplicates across any WGD events present and complete loss or non-occurrence of SSD duplicates. As this parameter represents both the birth rate of genes by SSD and loss rate of all duplicates (see Materials & Methods), lower values represent cases where SSD is rarely successful but any duplicates that have occurred are typically retained and correspond to gene counts closer to reciprocal retention.

For the 8,206 families we assessed, we find a skew towards lower values Figure 5.5. Although no family reaches a λ of zero for perfect reciprocal retention, our minimum observed value of 0.084 represents a family where only one species (*A. anguilla*) deviates from the expected gene counts, having 1 copy rather than the 2 expected from retention of a paralog from the teleost event. The average value of 0.757 corresponds to more widespread deviation from the expected values with most species deviating up or down by one copy, however on the whole it seems that the majority of families are quite constrained in their duplicability. We defined the bottom 5% of families by λ value as a group of consistently duplicable genes to assess further (see later), creating a group of 650 families with λ values falling below the 5th percentile value (0.312).

The λ distribution recovered here for fish is not dissimilar from that of angiosperms ($\mu = 0.827$ compared to $\mu = 0.757$ here). From the point of view of

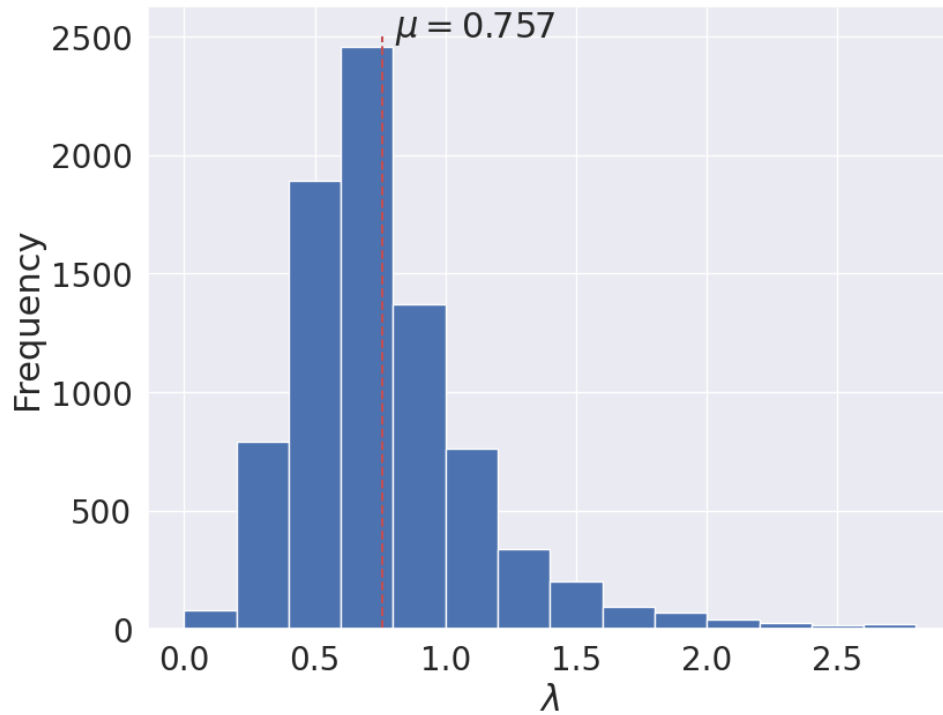


Figure 5.5 | Distribution of reciprocal retention. The distribution of λ values across the 8,206 families assessed, encompassing all genomes as listed in Figure 5.1. Values are truncated at 3 for clarity, mean value for the values presented is shown with a red dashed line. Lower values of λ correspond to families with gene counts more closely following a pattern of reciprocal retention i.e. families where all possible WGD duplicates have been retained but no SSDs have been retained.

duplicability consistency, it is interesting to note that such a property is similarly distributed when considering such long-diverged lineages as plants and fish. Although the specifics of duplication may differ, it is possible that the relative frequencies of different classes of duplicability are relatively consistent in different lineages.

5.3.4 Families showing retention across all events are not strongly reciprocally retained

We next investigated the relationship between the group defined through overlaps between events with our reciprocal retention estimate and do not find that families

retained in duplicate across all the events considered to be particularly strongly reciprocally retained relative to other families. Although there is a visible shift towards lower values in the overlapping families (Figure 5.6), this is not significant ($p = 0.7$; Mann-Whitney U test) and there is substantial overlap between the two groups.

We might have expected the overlap-defined group to show evidence of reciprocal retention. Reciprocal retention as a pattern over time would be heavily indicative of dosage balance, the most obvious explanation for consistent retention across successive WGD events. Thus, it would have been reasonable to hypothesise that this group which is consistently retained across the events we have considered might show such a pattern of duplication retention and loss. The fact that we do not observe a strong reciprocal retention pattern within this group indicates that dosage is not playing a role in the consistent retention of these families, at least not in the sense of dosage balance, and any convergent force acting across events to drive their retention remains to be defined.

5.3.5 Families with consistent duplicability show indications of dosage sensitivity

As the retention pattern shown by reciprocally retained families is most likely related to dosage balance (Tasdighian et al., 2017), we investigated whether our groups of consistently duplicable families showed signs of dosage sensitivity.

Genes under dosage balance are unlikely to duplicate outside of a WGD context and so we would expect such genes to remain in single copy in species which have not undergone WGD. This is true in the reciprocally retained group, with families in this group more likely to be single copy in all non-WGD species considered in the analysis ($p=8.9 \times 10^{-12}$, χ^2 test). However, this is to be expected by definition as gene count of all species was taken into account in the estimation of the λ

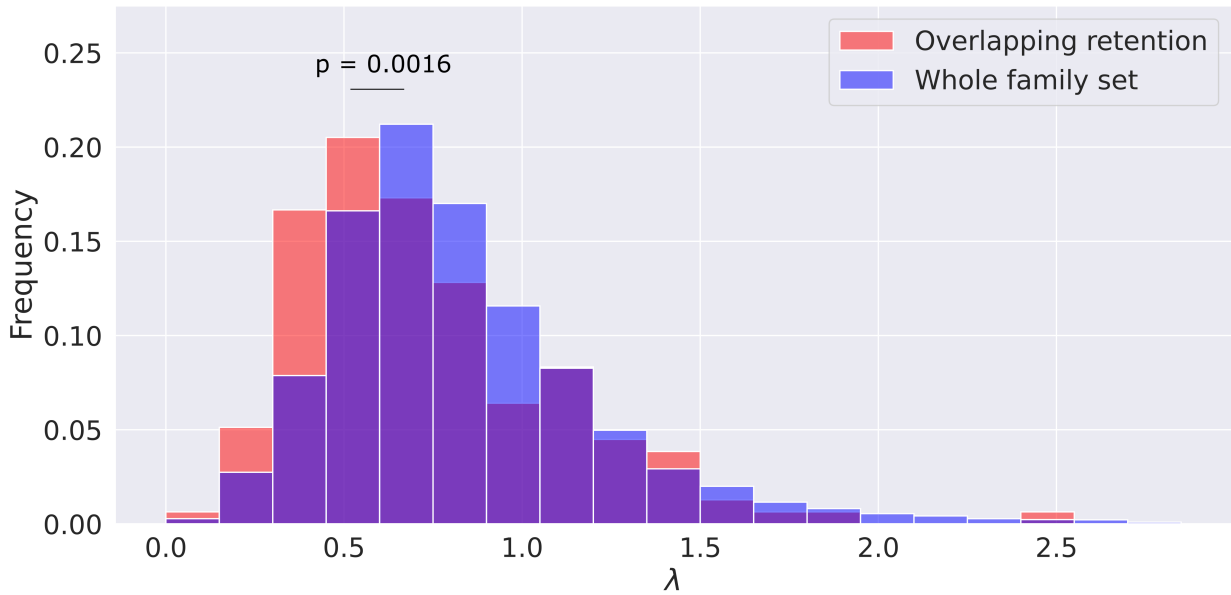


Figure 5.6 | Comparison of lambda values between overlapping set and whole set. Lambda values for families which showed retention across all four events are shown in red with the distribution for the whole genome in blue. P value given for a two-sided Mann-Whitney U test.

value. The overlap-defined group does not not show this pattern, with families in this group not more likely to be single copy in non-WGD species than other families ($p=0.089, \chi^2$ test). This is again consistent with these genes not being under dosage balance constraints as these families do not seem to be refractory to duplication in the absence of WGD.

Another indicator of dosage sensitivity is the quantity of interaction partners a protein has, likely at least partially driven by proteins that are complex members existing in dosage balance with other members of the same complex. We find that both the reciprocally retained and overlapping groups are more likely to have interactions listed in the STRING database (reciprocally retained, $p = 0.002$; overlap-defined group, $p = 3.9 \times 10^{-12}$; χ^2 test) and have significantly more interactions than other families (reciprocally retained, $p = 0.006$; overlap-defined group, $p = 2.7 \times 10^{-13}$; Mann-Whitney U test). This is the first sign of

dosage sensitivity shown by the overlap-defined group, however it may also reflect genes that occupy a more central position in the interaction network and so are more constrained, with constraint being another feature of duplicates retained post-WGD (Chapter 3).

Certain functions are more prone to dosage sensitivity than others, for example those relating to development and regulation where dosage thresholds are relevant or in neural functions, where over-expression can lead to dangerous aggregates in neurons (Foster et al., 2021; Lee et al., 2012). Reciprocally retained families are enriched for such functions (Table 5.3), with significant enrichments for developmental functions (GO:0048869, cellular developmental process; GO:0007275, multicellular organism development), regulatory functions (GO:0005515, protein binding; GO:0032422, purine-rich negative regulatory element binding) and neural functions (GO:0007409, axonogenesis; GO:0097485, neuron projection guidance). The group defined by overlaps is not as clear cut, with no immediately apparent pattern Table 5.4. However, on closer inspection a number of the top enrichments are brain function related, with a number related to neurotransmitter receptors (GO:0097112, gamma-aminobutyric acid receptor clustering; GO:0072579, glycine receptor clustering), ions which are present at high concentration in the brain (Pfeiffer and Braverman, 1982) (GO:0008270, zinc ion binding), and ketone body formation (GO:000441, hydroxymethylglutaryl-CoA lyase activity). In light of findings that glutamate receptors are convergently over-retained in teleosts and sturgeon, it is possible that these enrichments in a group showing retention across all fish WGDs could reflect a broader pattern relating to cognitive function (Du et al., 2020; Scharl et al., 2013).

We finally considered WGDs which had not been included in our analysis, investigating whether the groups we have defined are also more likely to be families originating from the two rounds of duplication early in the vertebrate lineage (2R

WGD). This serves as both another indicator of potential dosage sensitivity but also as an independent check on whether we have successfully recovered groups which are more likely to be retained in a given WGD. In both the reciprocally retained group and the overlap defined group, we find these families are more likely to be 2R ohnolog families, regardless of whether 'strict' (reciprocally retained, $p = 2.6 \times 10^{-46}$; overlap-defined group $p = 4.0 \times 10^{-22}$; χ^2 test) or 'relaxed' (reciprocally retained, $p = 7.4 \times 10^{-58}$; overlap-defined group $p = 3.1 \times 10^{-32}$; χ^2 test) definition ohnologs were considered. This confirms that these methods do successfully identify gene families which are consistent in their post-WGD behaviour as these definitions have relevance to other WGD events which are far removed in time from the lineage under study.

We observe some features expected in genes showing dosage sensitivity or in ohnologs in both groups, though this pattern is much stronger in the case of reciprocally retained genes. The overlap defined group does not necessarily show convincing evidence of dosage balance, however this group does possess interesting functional enrichments

5.3.6 Weak evidence for lineage-specific preferential retention/loss of WGD duplicates

If some duplicate retention post-WGD is lineage-specific then we might expect to see event-specific functional enrichments corresponding to the selective forces driving retention in a given lineage. We considered this by testing for functional enrichments and depletions in the set of retained families for each event (Table 5.5-Table 5.8).

The most obvious pattern that emerges from this analysis is a clear correspondence between the sturgeon and salmonid events, with almost total agreement between them. Both events primarily show enrichments for developmental func-

Table 5.3 | Top 20 enriched GO terms for most strongly reciprocally retained families.

GO id	GO description	P value
GO:0005515	protein binding	0.00018
GO:0032422	purine-rich negative regulatory element binding	0.0011
GO:0005488	binding	0.00149
GO:0007409	axonogenesis	0.00305
GO:0007411	axon guidance	0.00305
GO:0032990	cell part morphogenesis	0.00305
GO:0000904	cell morphogenesis involved in differentiation	0.00305
GO:0048869	cellular developmental process	0.00305
GO:0007399	nervous system development	0.00305
GO:0007275	multicellular organism development	0.00305
GO:0097485	neuron projection guidance	0.00305
GO:0048858	cell projection morphogenesis	0.00312
GO:0009653	anatomical structure morphogenesis	0.00312
GO:0032989	cellular component morphogenesis	0.00312
GO:0030182	neuron differentiation	0.00312
GO:0061564	axon development	0.00415
GO:0072583	clathrin-dependent endocytosis	0.00414
GO:0000902	cell morphogenesis	0.00415
GO:0030154	cell differentiation	0.00415
GO:0031175	neuron projection development	0.00432

Table 5.4 | Top 20 enriched GO terms for overlap families.

GO id	GO description	P value
GO:0008270	zinc ion binding	7.9×10^{-7}
GO:0046914	transition metal ion binding	0.00051
GO:0015020	glucuronosyltransferase activity	0.00168
GO:0008301	DNA binding, bending	0.00736
GO:0008194	UDP-glycosyltransferase activity	0.00736
GO:0001671	ATPase activator activity	0.00736
GO:0004990	oxytocin receptor activity	0.00771
GO:0005518	collagen binding	0.00771
GO:0061599	molybdopterin molybdotransferase activity	0.00771
GO:0004832	valine-tRNA ligase activity	0.00771
GO:0061598	molybdopterin adenylyltransferase activity	0.00771
GO:0004419	hydroxymethylglutaryl-CoA lyase activity	0.00771
GO:0071944	cell periphery	0.00795
GO:0005886	plasma membrane	0.00795
GO:0032501	multicellular organismal process	0.01238
GO:0097112	gamma-aminobutyric acid receptor clustering	0.01238
GO:0072579	glycine receptor clustering	0.01238
GO:0072578	neurotransmitter-gated ion channel clustering	0.01238
GO:0032689	negative regulation of interferon-gamma production	0.01238
GO:0060137	maternal process involved in parturition	0.01238

tions, with depletions mainly covering functions relating to the regulation of the cell cycle and cell death. On the other hand, both the teleost and cyprinid events show many unique enrichments and depletions. Families showing retention in the teleost event are enriched for muscle related functions (GO:0030017, sarcomere; GO:0043292, contractile fiber) and regulatory functions (GO:0065007, biological regulation; GO:0000981, DNA-binding transcriptionfactor activity, RNA polymerase II-specific), while those showing retention from the cyprinid event are chiefly concerned with vision-related functions (GO:0005212, structural constituent of eye lens; GO:0007601, visual perception; GO:0050953, sensory perception of light stimulus) and immunity (GO:0045916, negative regulation of complement activation; GO:0002921, negative regulation of humoral immune response).

These unique functional patterns for the teleost and cyprinid events may represent lineage-specific adaptation following the WGD. This is corroborated by the fact that teleost muscles are known to differ in structure from the majority of vertebrates (Luther et al., 1995), although recent work on ohnologs relating to myogenesis suggests broad redundancy and recovers evidence of duplicate divergence in only a small minority of cases (Duran et al., 2021). In the case of cyprinids, there is existing evidence of duplication and adaptation in vision-related genes within this lineage, with existing work having investigated divergence between opsin paralogs in this lineage leading to shifts in photosensitivity that may correspond to differences in fish colouration or environment (Wang et al., 2008).

This analysis does not provide much material to speculate on possible lineage-specific retention in the salmonid or sturgeon lineages, though it is informative from the standpoint of consistent retention that both events show such similar functional profiles. As both of these events are early in the WGD resolution

process, these functions may reflect the earliest state of retention, with these functions representing those that are the most decisively retained or not retained, at least in the context of the earlier stages of post-WGD evolution.

It is also possible that these events where the WGD has yet to fully resolve create difficulties in detecting lineage-specific retention within the noise created by the many duplicate copies which will not be retained in the long term. In an attempt to circumvent this issue, we took an alternative approach to searching for unusually highly or lowly retained functional groups. In this case we found the percentage of families assigned to a given function that are retained for a given event, with the aim being to find functions which have an unusually high retention rate even relative to the high retention across the entire genome for the salmonid and sturgeon events.

Ultimately, this method also proved not sensitive enough to detect any lineage-specific retention for these events. Retention percentages cluster tightly at the high end for the less-resolved events (see Figure 5.7), making it difficult to distinguish any functional groups that are unusually highly retained. We endeavored to use a cutoff-based approach to define such highly retained groups, taking those more than 1.96σ from the mean to be unusual in their retention percentage. This recovered an extremely small number of highly retained functions across the board (Table 5.9, Table 5.11, Table 5.13, Table 5.15), with none recovered for the salmonid event and only one for the sturgeon event (GO:0005768, endosome). Unusually lowly retained functions are more easily detected for the higher-retention events as they stand out in the background of retention. The functions recovered in this case reflect low retention of functions related to cellular components and metabolism (e.g. GO:0042254, ribosome biogenesis; GO:0051276, chromosome organization), functions which are known to be enriched in single copy genes in vertebrates (see Chapter 3).

Table 5.5 | Functional enrichment and depletion for families showing teleost 3R retention.

GO id	GO name	Enriched/depleted	P value
GO:0030017	sarcomere	Enriched	2.28×10^{-12}
GO:0030016	myofibril	Enriched	6.39×10^{-12}
GO:0043292	contractile fiber	Enriched	6.39×10^{-12}
GO:0065007	biological regulation	Enriched	1.87×10^{-11}
GO:0050794	regulation of cellular process	Enriched	2.95×10^{-11}
GO:0015629	actin cytoskeleton	Enriched	9.13×10^{-11}
GO:0050789	regulation of biological process	Enriched	1.47×10^{-10}
GO:0099512	supramolecular fiber	Enriched	2.06×10^{-10}
GO:0099081	supramolecular polymer	Enriched	2.53×10^{-10}
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	Enriched	2.56×10^{-10}
GO:0006396	RNA processing	Depleted	8.09×10^{-12}
GO:0006259	DNA metabolic process	Depleted	3.61×10^{-9}
GO:0044265	cellular macromolecule catabolic process	Depleted	3.65×10^{-8}
GO:0006974	cellular response to DNA damage stimulus	Depleted	5.20×10^{-8}
GO:0006281	DNA repair	Depleted	5.20×10^{-8}
GO:0009057	macromolecule catabolic process	Depleted	6.70×10^{-8}
GO:0016071	mRNA metabolic process	Depleted	1.14×10^{-7}
GO:0034660	ncRNA metabolic process	Depleted	2.57×10^{-6}
GO:0005840	ribosome	Depleted	1.84×10^{-5}
GO:0003723	RNA binding	Depleted	2.71×10^{-5}

Table 5.6 | Functional enrichment and depletion for families showing sturgeon 3R retention.

GO id	GO name	Enriched/depleted	P value
GO:0048856	anatomical structure development	Enriched	4.39×10^{-38}
GO:0007275	multicellular organism development	Enriched	3.35×10^{-37}
GO:0032502	developmental process	Enriched	3.12×10^{-36}
GO:0048731	system development	Enriched	6.20×10^{-32}
GO:0032501	multicellular organismal process	Enriched	3.72×10^{-31}
GO:0009653	anatomical structure morphogenesis	Enriched	9.46×10^{-24}
GO:0048869	cellular developmental process	Enriched	1.83×10^{-21}
GO:0030154	cell differentiation	Enriched	8.56×10^{-21}
GO:0007399	nervous system development	Enriched	7.92×10^{-20}
GO:0048513	animal organ development	Enriched	7.06×10^{-19}
GO:0046777	protein autophosphorylation	Depleted	1.69×10^{-14}
GO:0000977	RNA polymerase II transcription regulatory region sequence-specific DNA binding	Depleted	2.62×10^{-14}
GO:0000976	transcription cis-regulatory region binding	Depleted	5.44×10^{-13}
GO:0001067	transcription regulatory region nucleic acid binding	Depleted	5.44×10^{-13}
GO:1990837	sequence-specific double-stranded DNA binding	Depleted	7.41×10^{-13}
GO:0043069	negative regulation of programmed cell death	Depleted	1.61×10^{-12}
GO:0043066	negative regulation of apoptotic process	Depleted	1.61×10^{-12}
GO:0060548	negative regulation of cell death	Depleted	1.61×10^{-12}
GO:0003690	double-stranded DNA binding	Depleted	1.73×10^{-12}
GO:0044815	DNA packaging complex	Depleted	5.39×10^{-12}

Table 5.7 | Functional enrichment and depletion for families showing salmonid 4R retention.

GO id	GO name	Enriched/depleted	P value
GO:0048856	anatomical structure development	Enriched	4.24×10^{-58}
GO:0007275	multicellular organism development	Enriched	1.45×10^{-57}
GO:0032502	developmental process	Enriched	5.38×10^{-56}
GO:0048731	system development	Enriched	1.51×10^{-48}
GO:0032501	multicellular organismal process	Enriched	8.45×10^{-42}
GO:0009653	anatomical structure morphogenesis	Enriched	1.54×10^{-32}
GO:0048869	cellular developmental process	Enriched	4.94×10^{-32}
GO:0030154	cell differentiation	Enriched	3.41×10^{-31}
GO:0048513	animal organ development	Enriched	3.88×10^{-29}
GO:0007399	nervous system development	Enriched	9.80×10^{-26}
GO:0046777	protein autophosphorylation	Depleted	6.24×10^{-16}
GO:0060548	negative regulation of cell death	Depleted	6.24×10^{-16}
GO:0043066	negative regulation of apoptotic process	Depleted	9.30×10^{-16}
GO:0043069	negative regulation of programmed cell death	Depleted	1.04×10^{-15}
GO:0007346	regulation of mitotic cell cycle	Depleted	1.37×10^{-13}
GO:0044815	DNA packaging complex	Depleted	4.48×10^{-13}
GO:0000786	nucleosome	Depleted	3.43×10^{-12}
GO:0032993	protein-DNA complex	Depleted	6.25×10^{-11}
GO:0006259	DNA metabolic process	Depleted	7.06×10^{-7}
GO:0051726	regulation of cell cycle	Depleted	1.69×10^{-6}

Table 5.8 | Functional enrichment and depletion for families showing cyprinid 4R retention.

GO id	GO name	Enriched/depleted	P value
GO:0005212	structural constituent of eye lens	Enriched	8.88×10^{-12}
GO:0002088	lens development in camera-type eye	Enriched	9.23×10^{-9}
GO:0007601	visual perception	Enriched	5.92×10^{-5}
GO:0032502	developmental process	Enriched	5.92×10^{-5}
GO:0050953	sensory perception of light stimulus	Enriched	5.92×10^{-5}
GO:0048856	anatomical structure development	Enriched	7.41×10^{-5}
GO:0045916	negative regulation of complement activation	Enriched	1.13×10^{-4}
GO:0002921	negative regulation of humoral immune response	Enriched	1.13×10^{-4}
GO:0051234	establishment of localization	Enriched	2.11×10^{-4}
GO:0007275	multicellular organism development	Enriched	3.47×10^{-4}
GO:0061630	ubiquitin protein ligase activity	Depleted	0.004
GO:0061659	ubiquitin-like protein ligase activity	Depleted	0.004
GO:0038023	signaling receptor activity	Depleted	0.035
GO:0060089	molecular transducer activity	Depleted	0.035

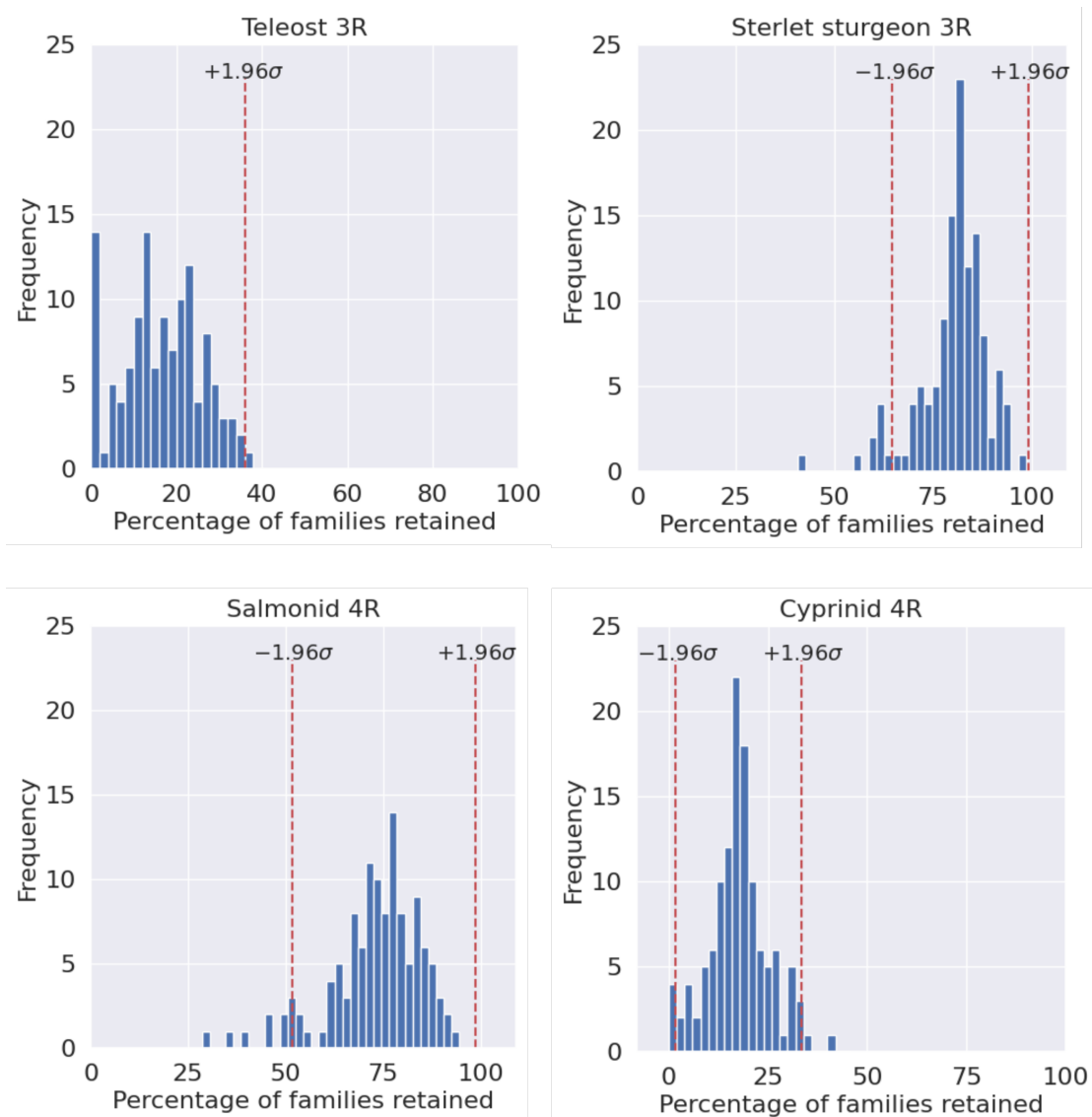


Figure 5.7 | Distribution of GO terms by percentage of families retained. Distributions of the proportion of families showing retention for the WGD event shown grouped by the function assigned. Higher values indicated that more families with a certain function have retained duplicates. Values ± 1.96 standard deviations from the mean are shown as red dashed lines. In the case of the teleost 3R the lower bound falls below zero and is not shown. Species representing each event are as in Figure 5.2.

We also considered the top/bottom 25 functions as ranked by percentage retention as a less strict set of potentially highly/lowly retained functions. While we do see some broad consistencies across events, such as higher retention of development/morphogenesis related functions, cellular signalling and transport, and lower retention of cellular components and cell cycle functions (Table 5.9-Table 5.16), it is debatable how meaningful such rankings are across a narrow distribution with no cutoff applied. These ranking based sets also do little in the way of identifying any lineage-specific retention, though one possible candidate is pigmentation in salmonids. This function (GO:0043473, pigmentation) ranks in the highest retained in salmonids despite ranking as one of the lowest in both sturgeon and cyprinids. Pigmentation genes have previously been identified as over-retained in both the teleost 3R and the salmonid 4R (Lorin et al., 2018), though no adaptive reason or correspondence to changes in pigmentation in salmonids was identified.

On the whole, our analysis of the functional profiles of these WGD events does more to confirm consistencies across events regarding functions such as development and regulation than to identify lineage-specific behaviour. Nevertheless, we do find some candidate functional categories for lineage-specific retention which are supported by the existing literature, namely muscle structure in teleosts, vision in cyprinids and pigmentation in salmonids.

5.4 Discussion

This work is the first dedicated study into consistency of duplicability in vertebrates and aimed to provide a broad characterisation of how duplicability as a property varies across WGD events within fish. We find that there is substantial overlap in the gene families showing retention across multiple WGD events despite intragenomic variance in duplicability. We additionally define groups

Table 5.9 | Highest retention functions for teleost 3R.

GO id	GO name
<i>Highest retained by rank:</i>	
GO:0043167	ion binding
GO:0008283	cell population proliferation
GO:0031410	cytoplasmic vesicle
GO:0065003	protein-containing complex assembly
GO:0008233	peptidase activity
GO:0009790	embryo development
GO:0006913	nucleocytoplasmic transport
GO:0042592	homeostatic process
GO:0021700	developmental maturation
GO:0007009	plasma membrane organization
GO:0005737	cytoplasm
GO:0040007	growth
GO:0003677	DNA binding
GO:0008134	transcription factor binding
GO:0055085	transmembrane transport
GO:0005856	cytoskeleton
GO:0008219	cell death
GO:0006914	autophagy
GO:0007010	cytoskeleton organization
GO:0008092	cytoskeletal protein binding
GO:0005576	extracellular region
GO:0007267	cell-cell signaling
GO:0016301	kinase activity
GO:0030674	protein-macromolecule adaptor activity
GO:0003700	DNA-binding transcription factor activity
<i>Highest retained by cutoff:</i>	
GO:0003700	DNA-binding transcription factor activity
GO:0030674	protein-macromolecule adaptor activity

Table 5.10 | Lowest retention functions for teleost 3R.

GO id	GO name
<i>Lowest retained by rank</i>	
GO:0016853	isomerase activity
GO:0006399	tRNA metabolic process
GO:0061024	membrane organization
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups
GO:0016874	ligase activity
GO:0006605	protein targeting
GO:0005840	ribosome
GO:0016798	hydrolase activity, acting on glycosyl bonds
GO:0005777	peroxisome
GO:0016829	lyase activity
GO:0005815	microtubule organizing center
GO:0019843	rRNA binding
GO:0004386	helicase activity
GO:0003735	structural constituent of ribosome
GO:0051604	protein maturation
GO:0140014	mitotic nuclear division
GO:0006397	mRNA processing
GO:0042254	ribosome biogenesis
GO:0006259	DNA metabolic process
GO:0003729	mRNA binding
GO:0007049	cell cycle
GO:0051301	cell division
GO:0004518	nuclease activity
GO:0005768	endosome
GO:0005773	vacuole

Table 5.11 | Highest retention functions for sturgeon 3R.

GO id	GO name
<i>Highest retained by rank:</i>	
GO:0005829	cytosol
GO:0006412	translation
GO:0031012	extracellular matrix
GO:0005773	vacuole
GO:0016887	ATPase activity
GO:0006790	sulfur compound metabolic process
GO:0022607	cellular component assembly
GO:0008289	lipid binding
GO:0048646	anatomical structure formation involved in morphogenesis
GO:0022618	ribonucleoprotein complex assembly
GO:0030705	cytoskeleton-dependent intracellular transport
GO:0022857	transmembrane transporter activity
GO:0061024	membrane organization
GO:0021700	developmental maturation
GO:0005730	nucleolus
GO:0003013	circulatory system process
GO:0008134	transcription factor binding
GO:0006397	mRNA processing
GO:0006913	nucleocytoplasmic transport
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups
GO:0015031	protein transport
GO:0005635	nuclear envelope
GO:0003729	mRNA binding
GO:0016757	transferase activity, transferring glycosyl groups
GO:0005768	endosome
<i>Highest retained by cutoff:</i>	
GO:0005768	endosome

Table 5.12 | Lowest retention functions for sturgeon 3R.

GO id	GO name
<i>Lowest retained by rank</i>	
GO:0003700	DNA-binding transcription factor activity
GO:0007155	cell adhesion
GO:0007034	vacuolar transport
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0016829	lyase activity
GO:0019843	rRNA binding
GO:0008219	cell death
GO:0005634	nucleus
GO:0051276	chromosome organization
GO:0032182	ubiquitin-like protein binding
GO:0006399	tRNA metabolic process
GO:0005576	extracellular region
GO:0005615	extracellular space
GO:0000228	nuclear chromosome
GO:0002376	immune system process
GO:0140014	mitotic nuclear division
GO:0004518	nuclease activity
GO:0007059	chromosome segregation
GO:0016874	ligase activity
GO:0006810	transport
GO:0006259	DNA metabolic process
GO:0043473	pigmentation
GO:0000003	reproduction
GO:0005575	cellular component
GO:0005840	ribosome
<i>Lowest retained by cutoff</i>	
GO:0003700	DNA-binding transcription factor activity
GO:0007155	cell adhesion
GO:0008219	cell death
GO:0051276	chromosome organization
GO:0007034	vacuolar transport
GO:0016829	lyase activity
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0019843	rRNA binding
GO:0005634	nucleus

Table 5.13 | Highest retention functions for salmonid 4R.

GO id	GO name
<i>Highest retained by rank:</i>	
GO:0009056	catabolic process
GO:0005886	plasma membrane
GO:0007010	cytoskeleton organization
GO:0005794	Golgi apparatus
GO:0034655	nucleobase-containing compound catabolic process
GO:0016746	transferase activity, transferring acyl groups
GO:0030154	cell differentiation
GO:0005856	cytoskeleton
GO:0051604	protein maturation
GO:0007059	chromosome segregation
GO:0048646	anatomical structure formation involved in morphogenesis
GO:0030198	extracellular matrix organization
GO:0048856	anatomical structure development
GO:0007034	vacuolar transport
GO:0030674	protein-macromolecule adaptor activity
GO:0042592	homeostatic process
GO:0008283	cell population proliferation
GO:0008134	transcription factor binding
GO:0007267	cell-cell signaling
GO:0000902	cell morphogenesis
GO:0007009	plasma membrane organization
GO:0040007	growth
GO:0009790	embryo development
GO:0034330	cell junction organization
GO:0043473	pigmentation

Table 5.14 | Lowest retention functions for salmonid 4R.

GO id	GO name
<i>Lowest retained by rank</i>	
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups
GO:0006399	tRNA metabolic process
GO:0004386	helicase activity
GO:0016798	hydrolase activity, acting on glycosyl bonds
GO:0004518	nuclease activity
GO:0005777	peroxisome
GO:0032182	ubiquitin-like protein binding
GO:0042254	ribosome biogenesis
GO:0050877	nervous system process
GO:0016757	transferase activity, transferring glycosyl groups
GO:0019843	rRNA binding
GO:0003735	structural constituent of ribosome
GO:0005840	ribosome
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0006520	cellular amino acid metabolic process
GO:0005929	cilium
GO:0000228	nuclear chromosome
GO:0005634	nucleus
GO:0008219	cell death
GO:0008233	peptidase activity
GO:0005815	microtubule organizing center
GO:0005694	chromosome
GO:0005635	nuclear envelope
GO:0008168	methyltransferase activity
GO:0006259	DNA metabolic process
<i>Lowest retained by cutoff</i>	
GO:0006399	tRNA metabolic process
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups
GO:0042254	ribosome biogenesis
GO:0004518	nuclease activity
GO:0016798	hydrolase activity, acting on glycosyl bonds
GO:0005777	peroxisome
GO:0004386	helicase activity
GO:0032182	ubiquitin-like protein binding

Table 5.15 | Highest retention functions for cyprinid 4R.

GO id	GO name
<i>Highest retained by rank:</i>	
GO:0016491	oxidoreductase activity
GO:0055085	transmembrane transport
GO:0030234	enzyme regulator activity
GO:0005576	extracellular region
GO:0002376	immune system process
GO:0030674	protein-macromolecule adaptor activity
GO:0003700	DNA-binding transcription factor activity
GO:0005654	nucleoplasm
GO:0005198	structural molecule activity
GO:0040007	growth
GO:0030198	extracellular matrix organization
GO:0005730	nucleolus
GO:0000902	cell morphogenesis
GO:0007059	chromosome segregation
GO:0005815	microtubule organizing center
GO:0007155	cell adhesion
GO:0016757	transferase activity, transferring glycosyl groups
GO:0034330	cell junction organization
GO:0048856	anatomical structure development
GO:0000228	nuclear chromosome
GO:0007034	vacuolar transport
GO:0005694	chromosome
GO:0005777	peroxisome
GO:0032182	ubiquitin-like protein binding
GO:0005768	endosome
<i>Highest retained by cutoff:</i>	
GO:0005694	chromosome
GO:0005768	endosome
GO:0005777	peroxisome
GO:0032182	ubiquitin-like protein binding

Table 5.16 | Lowest retention functions for cyprinid 4R.

GO id	GO name
<i>Lowest retained by rank</i>	
GO:0022618	ribonucleoprotein complex assembly
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0005634	nucleus
GO:0031012	extracellular matrix
GO:0006913	nucleocytoplasmic transport
GO:0006605	protein targeting
GO:0043473	pigmentation
GO:0003729	mRNA binding
GO:0016853	isomerase activity
GO:0004386	helicase activity
GO:0019843	rRNA binding
GO:0006397	mRNA processing
GO:0003735	structural constituent of ribosome
GO:0005773	vacuole
GO:0008168	methyltransferase activity
GO:0007009	plasma membrane organization
GO:0005929	cilium
GO:0007049	cell cycle
GO:0050877	nervous system process
GO:0008092	cytoskeletal protein binding
GO:0042254	ribosome biogenesis
GO:0061024	membrane organization
GO:0006399	tRNA metabolic process
GO:0051604	protein maturation
<i>Lowest retained by cutoff</i>	
GO:0022618	ribonucleoprotein complex assembly
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0005634	nucleus

of genes which are consistent in their post-WGD behaviour, both through raw overlap between events and a modelling based approach that estimates reciprocal retention based on gene count data, and show that their consistency may be due to dosage effects or a convergent pressure across fish related to cognition. The functional profiles of genes retained in duplicate following WGD were found to be broadly consistent with a limited number of detectable candidates for lineage specific patterns.

Despite the contradictions in previous work which has considered consistency across multiple WGD events in fish, we find that the evidence overall supports at least some degree of commonality across events. Of gene families showing retention in any of the four events covered, 13% of them show retention in at least three events (Figure 5.4). This is likely to be a conservative estimate of how consistent duplicate retention is, given the total families are inflated somewhat by events in the earlier stages of post-WGD evolution which remain at very high retention frequencies. Agreement across fully resolved events would likely be higher. The functional profile of genes retained/lost also shows clear similarities across events, indicating that the overall process or pressures in duplicate retention or loss are likely to be similar.

Previous work had argued that a small number of genes found to be fully consistent in their retention pattern indicated there are not consistent pressures across events at the single gene level (Du et al., 2020). We contend that these pressures do exist, but apply only to specific groups of genes i.e. consistency of duplicability varies on a gene to gene basis. We have here identified groups of such genes, with one groups consistently retained across the events studied here and the other with the additional constraint of lacking duplication by SSD (reciprocal retention). These groups are distinct from the rest of the genome in their number of interaction partners and specific functional profiles. In the case of reciprocally

retained families, it is likely that selection for dosage balance is the primary force acting to constrain their duplication behaviour, as confirmed by the enrichment for functions known to have relationships to dosage.

The exact force driving consistency for the second group remains unclear, although they show elevated numbers of interactions they are not constrained to single copy outside of WGDs and so dosage balance seems an unlikely explanation. The higher number of PPIs may be reflective of genes that occupy more central positions in the interaction network and which are more conserved, a known feature of ohnologs (Chapter 3). Quantifying the level sequence and expression conservation for these families would be informative in determining if overall constraint relates to the consistent retention patterns of this group. Function-driven processes are also a possible explanation, with this group enriched for certain neural functions and neurotransmitter receptors previously identified to be convergently retained across fish WGD events (Du et al., 2020).

Both of these groups show a relationship to much earlier WGD events in the vertebrate lineage, with both significantly more likely to be 2R ohnologs. This suggests that these classifications of consistent duplicability have a wide relevance in likelihood of duplicate retention. This also adds to the evidence of consistent processes acting across WGD events. For our analysis of consistency across these fish WGDs to yield groups more likely to have shown retention in events separated from our study group by hundreds of millions of years indicates that there must be some convergent force acting on duplicate retention post-WGD.

Finally, we endeavored to investigate whether there were clear indications of lineage specificity in the events under consideration. Although we have found potential cases of this, further investigation would be required to fully determine if duplicate retention is truly behaving in a lineage specific way with respect to these functions and if there is some factor unique to either the lineage in question

or the timing of the duplication which has prompted this.

A consistent issue in many of the analyses presented here is the need to account for differences in retention rate/state of WGD resolution. This presented an issue in the overlap analysis, though we were able to use a modelling based approach to account for this with the species tree and relative placement of WGD events therein taken into account in considering retention patterns. A similar pivot in method may be required to define lineage specific retention patterns. For example, there exist methods, not dissimilar to the basis of the model used here to identify reciprocal retention, which model gene family size evolution and determine families that have had a significant shift on a given branch (CAFE (Bie et al., 2006), for example). Another advantage to a more probabilistic approach would be the ability to consider stochasticity. Though we have largely not considered this factor in duplicate retention within this work, it is likely to play a substantial role in the process, in particular in the case of genes which are inconsistent in their retention behaviour.

Going forward, there remains at least one group of consistently duplicable genes that remains to be defined for this lineage. Duplication resistant (DR) genes have been examined in plants previously (Smet et al., 2013), defined as a group of genes which experience convergent duplicate loss after both WGD and SSD which are single copy across many plant species. Such genes are distinct from the genes which show consistent retention rather than loss, showing enrichments for housekeeping functions, and are relevant to consider in the study of duplicability, offering a chance to examine the opposite pattern to the families always retained post-WGD. Interestingly, the less-resolved events which have presented difficulties in this study are likely better suited to the study of these genes, as any genes reliably driven back to single copy should presumably lose their duplicate copy relatively soon after a WGD event. We observe hints of this in Figure 5.7, with

both the salmonid and sturgeon events showing a slight tail of lowly-retained functions that can not be recovered in the other events within an overall low rate of retention across the genome.

Another potentially useful addition to this work would be additional WGD events. Another WGD event does exist within *Cypriniformes*, specific to the Catostomids (suckers). The first chromosome-level genome assembly for a member of this family (the Chinese sucker, *M. asiaticus* has recently been published (Krabbenhoft et al., 2021), allowing genomic characterisation of this event for the first time. This adds another 4R event to be compared with the salmonid and cyprinid events, improving our ability to discern the frequency of lineage-specific effects. Many duplications have also taken place within the paddlefish and sturgeons lineage, with some species highly polyploid (Havelka et al., 2013). Such cases would be more suited to testing the hypothesis of diminishing dosage sensitivity with increasing genome copy number as, if such an effect exists, we are more likely to observe it in more dramatically polyploid species. From this point of view, greater availability of genomic resources within this group would be invaluable to the study of duplicability in vertebrates.

Overall, many consistencies exist across duplication events, suggestive of broad trends in retention. On the individual gene level however, consistent post-duplication behaviour varies across the genome, with gene families falling into one of several potential groups (hypothetical groups summarised in Figure 5.8). Further characterisation of these groups will be relevant in unravelling the factors which contribute to duplicability and determining the role of randomness and environmental variation versus intrinsic features of a gene. Most likely, these are not questions with meaningful answers on the whole genome level. Rather, we expect that distinct groups exist for which these trends differ, as initially investigated here.

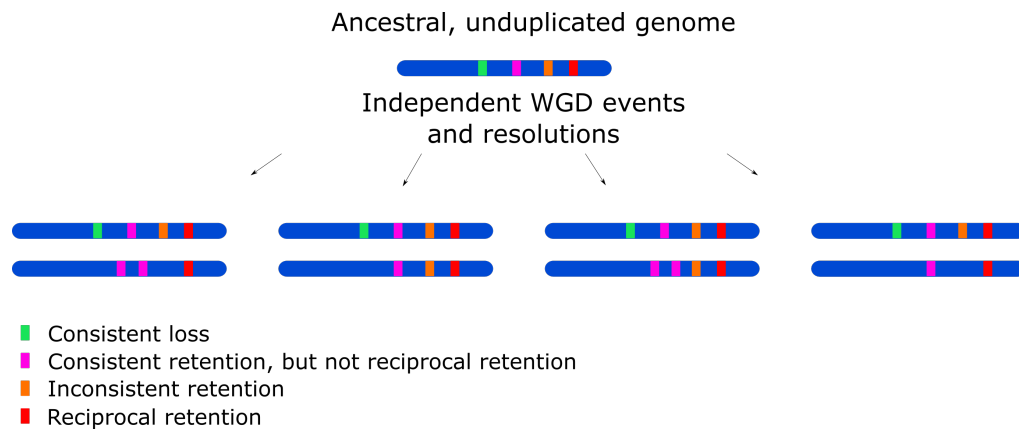


Figure 5.8 | Hypothetical groupings for consistency of duplicability. Hypothetical outcomes are shown for four independent duplications of the same ancestral gene set. Different coloured bands represent different genes, multiple bands of the same colour side by side indicate a tandem duplication.

Chapter 6

Conclusions

Gene duplication is a crucial process in evolution, impacting strongly on the structure and content of genomes. While it is widely known that differences exist across genes in the likelihood of successful duplication, we have not necessarily had an in depth knowledge regarding the determinants of duplicability in any holistic sense up to now. Gaps in our knowledge existed in terms of a comprehensive characterisation of duplicability outside of plants, the role of observed differences in the process of duplication and the consistency of duplicability as a property. In this thesis, the three studies presented attempted to answer these outstanding questions within the topic of duplicability.

An extensive array of features have been associated to differences in duplicability between genes, but few studies have attempted to perform any synthesis of these differences or determine their relative importance in gene duplicability. In chapter 3 we made the first attempt in vertebrates to create a full and cohesive narrative on the features that relate to gene duplicability and rank their importance in determining duplicability within a genome. We confirmed known trends in the features associated with duplicability, showing that ohnologs are larger, more complex and more constrained in contrast to shorter, faster evolving

SSDs. Genes belonging to different classes of duplicability additionally show specific functional profiles. Ohnologs enriched for developmental, neural, signaling and regulatory functions, while SSDs are heavily involved in interactions with external stimulus and singletons are enriched for more base cellular functions such as translation and organelle-based functions. We combined the observed patterns of functional enrichment and depletion across these categories to form hypotheses on how function interacts with duplicability, finding that the functional profile of SSDs is primarily shaped by the functions which are permitted to duplicate by SSD with many functions depleted in this group. In contrast, WGD enrichments are shaped primarily by the functions which promote retention post-WGD. In short, gene retention post-SSD requires a lack of pressure for loss and retention post-WGD is driven by presence of pressure for retention. Ranking the importance of features in distinguishing between duplicates of the two types (and controlling for the effect of age differences) yields high importances for the number of domains present, expression level and constraint features, in particular relating to haploinsufficiency.

Our characterisation of the intrinsic gene features contributing to the duplicability of a gene creates a framework for understanding how the differences between genes of each category contribute to the duplication process. The feature importance rankings we have obtained suggest roles of particular importance for structural complexity (number of domains), expression level and dosage sensitivity (probability of haploinsufficiency). We can interpret this in light of the driving forces of retention revealed by the comparison of functional profiles; features which distinguish WGD from SSD must be related to either a lack of pressure to lose an SSD duplicate or a pressure to retain a WGD duplicate. This is straightforward for haploinsufficiency as, under the DBH, we expect dosage sensitive genes to be both preferentially retained after WGD and preferentially lost following

SSD. The other features are less clear, expression differences may be driven by the high relative cost of duplicating a highly-expressed gene by SSD without the corresponding duplication of translation and protein processing machinery that occurs in WGD. The difference in number of domains may reflect that larger, more complex genes are more difficult to safely duplicate, as the genes which are permitted to duplicate by SSD tend to be particularly lowly constrained. Taking into consideration that the SSD functional profile is primarily shaped by what *can't* duplicate by SSD, it is reasonable to suppose that the set of genes which are SSD duplicable are of particularly low impact: the shortest, least constrained, least complex genes.

To our knowledge, there has not been such a comprehensive study of duplicability in other lineages. As such, we can not be sure that our findings generalise well to species outside of human. Corroboration of these results in other species would be informative in this regard, particularly those which differ meaningfully from human in select ways. For example, lineages where a WGD has not occurred, lineages with many more ancestral WGDs or single-celled organisms would all present unique contexts to further illuminate consistencies and contrasts in gene features impacting duplicability across the tree of life.

The low constraint observed in SSDs is above suggested to relate to a tendency for higher constraint genes to be refractory to SSD. However, this observation has often been assigned an alternative explanation. While a duplicability driven hypothesis would posit that biases applied to pre-existing levels of constraint create the observed higher rates of evolution, a long-standing hypothesis argues that the act of duplication itself causes a release from constraint that accelerates evolution in SSDs. In chapter 4 we examine this problem by comparing ancestral rates of evolution for singleton and duplicable genes. We find that duplicable genes are faster evolving prior to the duplication event, supporting the idea that gene

duplicability is a contributing force in the observed differences in evolutionary rate between duplicates and singletons. Testing for evidence of the duplication process contributing to differences in rate gave uncertain results. Although the majority of duplications did not show evidence for either post-duplication rate acceleration or an asymmetrical rate acceleration, considering all hypothesis tests together provides evidence in favour of both scenarios. Overall, it is possible that both duplicability and post-duplication effects contribute to the faster rate of evolution we see in duplicated genes.

The finding that faster evolving genes are more likely to duplicate raises the question of why this is. It could be that the lowest constraint genes are simply the least problematic to duplicate, in much the same way that they are the least problematic to mutate. Another possibility which could be tested in future work, is that there is some advantage as a newly created duplicate to being fast evolving. This could relate to the kind of functions we typically see in sets of duplicated genes i.e. immunity and other functions relating to external stimulus. Such functions would benefit from faster-evolution in the face of shifting environments and competing adaptive forces in pathogens, and also from the opportunity provided by duplication to diversify. Thus, faster-evolving genes may be at an advantage for retention due to possessing specific functions which benefit from duplication.

There has been extensive work considering duplicability as a property of genes in the context of single species or single duplication events. Comparatively little exists in the way of considering duplicability as a property that may vary across time. In chapter 5, we considered duplicability in a broader context, seeking to quantify how the biases we identified in the previous chapters might play into duplication across time and multiple duplication events. We find that there are significant overlaps across WGD events in the bony fish lineage in both the

specific families retained and the overall functional profile of retained genes. At the single gene family level, we find there are specific groups of genes which are under consistent pressure for retention in duplicate post-WGD. We define and characterised two such groups. The first is based solely on fully consistent retention across the events covered, while the second was a group of genes showing a pattern of reciprocal retention: high post-WGD combined with low post-SSD retention. In the case of the reciprocally retained set, these families are likely under dosage balance constraint, given their pattern of duplicate retention and functional enrichment in areas likely to show dosage sensitivity such as development. It is less clear what pressure might be driving consistent behaviour in the group defined through overlapping retention across WGD events. We identify functional enrichments in this group relating to cognition, an area which has previously been found to be convergently retained across the teleost and sturgeon WGDs, suggesting that this may be a pattern which is more broadly applicable across this family of species. On the other end of the question of consistency in duplicability, we attempted to identify any instances of lineage-specific retention patterns in the events we had considered. Although there are difficulties in discerning such patterns for less resolved events, for both the teleost and cyprinid events we were able to identify muscle structure and vision respectively as candidate functions for lineage specific effects. It remains to be investigated whether these specific functional enrichments are linked to environmental conditions at the time of these WGD events.

This work only begins to investigate how consistent the property of duplicability is over time. There remains much to be done in elucidating how duplicability as a property of genes interacts with external factors, for example in determining how retention in *inconsistently* retained families is determined by extrinsic factors and random chance. We can speculate on how such families differ from

the consistently retained families we have characterised here, which we believe represent families where factors external to the gene itself hold very little sway over the retention of duplicates. To elaborate further on that, these are genes which are under such strong constraints relating to gene-specific factors such as dosage sensitivity that other factors have comparatively small contributions to the probability of retention. Inconsistently retained genes are more likely to represent groups which can be retained, but do not necessarily have reason to be unless there is a push from some environmental factor. This relates back to the conclusions we reached in Chapter 3, where SSDs are not depleted for functions associated with WGD. Any gene which can be safely duplicated by SSD can presumably also duplicate through WGD, although there are not strong pressures promoting such retention. We hypothesise that, in the absence of consistent pressures such as dosage balance, or event specific pressures relating to environment, the retention of such genes would be for the most part driven by random chance.

A valuable addition to this work would be the definition of duplication resistant genes. These are genes which are under consistent pressures to return to single copy, in contrast to genes which are under pressure to retain WGD duplicates. With this last group defined, a more complete model of duplicability consistency would be possible, with a gradient centering on cases where intrinsic duplicability contributes little to the outcome of duplication and consistently retained and single copy genes at either end, both having their behaviour heavily impacted by gene-specific features Figure 6.1.

Here, we have presented an exploration of the determinants of gene duplicability. We set out to explore outstanding questions regarding the gene-specific determinants of duplicability, their role in the duplication process and their broader relevance to duplicability over time. We have made substantial progress in understanding the relative importance of gene features both with relation to each

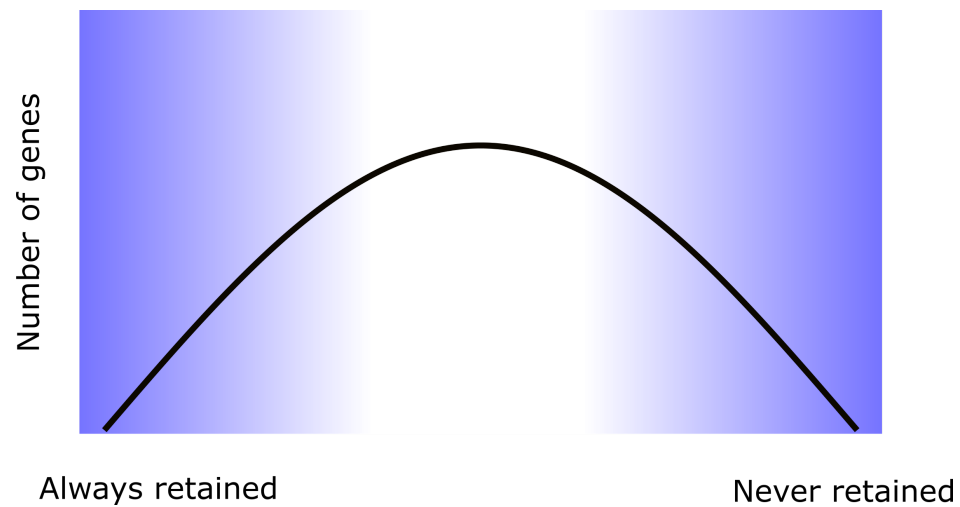


Figure 6.1 | A proposed model for consistency of duplicability. Strength of background colour indicates the strength of intrinsic duplicability for a given gene in determining retention outcomes. Strong intrinsic forces exist for both genes which are always retained and those which are always lost.

other for a given duplication and to the likelihood of the gene's duplication across time. Better understanding of duplicability is key to understanding the process of duplication and its role in adaptation and investigations of the factors influencing this property are crucial to this understanding.

Bibliography

- Altenhoff, Adrian M., Natasha M. Glover, and Christophe Dessimoz (2019). *Inferring Orthology and Paralogy*. DOI: 10.1007/978-1-4939-9074-0_5.
- Altenhoff, Adrian M., Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz (2012). “Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs”. In: *PLoS Computational Biology* 8 (5), e1002514. DOI: 10.1371/journal.pcbi.1002514.
- Alvarez-Ponce, David and Mario A. Fares (2012). “Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network”. In: *Genome Biology and Evolution* 4 (12), pp. 1263–1274. DOI: 10.1093/gbe/evs101.
- Amoutzias, G. D., Y. He, J. Gordon, D. Mossialos, S. G. Oliver, and Y. Van de Peer (2010). “Posttranslational regulation impacts the fate of duplicated genes”. In: *Proceedings of the National Academy of Sciences* 107 (7), pp. 2967–2971. DOI: 10.1073/pnas.0911603107.
- Arrigo, Nils and Michael S. Barker (2012). “Rarely successful polyploids and their legacy in plant genomes”. In: *Current Opinion in Plant Biology* 15 (2), pp. 140–146. DOI: 10.1016/j.pbi.2012.03.010.
- Athey, John, Aikaterini Alexaki, Ekaterina Osipova, Alexandre Rostovtsev, Luis V. Santana-Quintero, Upendra Katneni, Vahan Simonyan, and Chava Kimchi-

- Sarfaty (2017). “A new and updated resource for codon usage tables”. In: *BMC Bioinformatics* 18 (1), p. 391. DOI: 10.1186/s12859-017-1793-7.
- Aury, Jean-Marc et al. (2006). “Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*”. In: *Nature* 444 (7116), pp. 171–178. DOI: 10.1038/nature05230.
- Bailey, Jeffrey A, Amy M Yavor, Hillary F Massa, Barbara J Trask, and Evan E Eichler (2001). “Segmental duplications: organization and impact within the current human genome project assembly.” In: *Genome research* 11 (6), pp. 1005–17. DOI: 10.1101/gr.187101.
- Bailey, Jeffrey A. and Evan E. Eichler (2006). “Primate segmental duplications: crucibles of evolution, diversity and disease”. In: *Nature Reviews Genetics* 7 (7), pp. 552–564. DOI: 10.1038/nrg1895.
- Banerjee, Sanghita, Felix Feyertag, and David Alvarez-Ponce (2017). “Intrinsic protein disorder reduces small-scale gene duplicability”. In: *DNA Research* 24 (4), pp. 435–444. DOI: 10.1093/dnares/dsx015.
- Bartha, István, Antonio Rausell, Paul J. McLaren, Pejman Mohammadi, Manuel Tardaguila, Nimisha Chaturvedi, Jacques Fellay, and Amalio Telenti (2015). “The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome”. In: *PLOS Computational Biology* 11 (12), e1004647. DOI: 10.1371/journal.pcbi.1004647.
- Bennett, Richard J. and Alexander D. Johnson (2003). “Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains”. In: *The EMBO Journal* 22 (10), pp. 2505–2515.
- Berthelot, Camille et al. (2014). “The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates”. In: *Nature Communications* 5 (1), p. 3657. DOI: 10.1038/ncomms4657.

- Bickhart, Derek M., Yali Hou, Steven G. Schroeder, Can Alkan, Maria Francesca Cardone, Lakshmi K. Matukumalli, Jiuzhou Song, Robert D. Schnabel, Mario Ventura, Jeremy F. Taylor, Jose Fernando Garcia, Curtis P. Van Tassell, Tad S. Sonstegard, Evan E. Eichler, and George E. Liu (2012). “Copy number variation of individual cattle genomes using next-generation sequencing”. In: *Genome Research* 22 (4), pp. 778–790. DOI: 10.1101/gr.133967.111.
- Bie, Tijl De, Nello Cristianini, Jeffery P. Demuth, and Matthew W. Hahn (2006). “CAFE: A computational tool for the study of gene family evolution”. In: *Bioinformatics* 22 (10), pp. 1269–1271. DOI: 10.1093/bioinformatics/bt1097.
- Birchler, J. A. and R. A. Veitia (2012). “Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines”. In: *Proceedings of the National Academy of Sciences* 109 (37), pp. 14746–14753. DOI: 10.1073/pnas.1207726109.
- Bird, Kevin A., Robert VanBuren, Joshua R. Puzey, and Patrick P. Edger (2018). *The causes and consequences of subgenome dominance in hybrids and recent polyploids*. DOI: 10.1111/nph.15256.
- Blanc, Guillaume and Kenneth H. Wolfe (2004). “Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution[W]”. In: *The Plant Cell* 16 (7), pp. 1679–1691. DOI: 10.1105/tpc.021410.
- Blomme, Tine, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer (2006). “The gain and loss of genes during 600 million years of vertebrate evolution”. In: *Genome Biology* 7, R43. DOI: 10.1186/gb-2006-7-5-r43.
- Bomblies, Kirsten and Andreas Madlung (2014). “Polyploidy in the Arabidopsis genus”. In: *Chromosome Research* 22 (2), pp. 117–134. DOI: 10.1007/s10577-014-9416-x.

- Braasch, Ingo et al. (2016). “The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons”. In: *Nature Genetics* 48 (4), pp. 427–437. DOI: 10.1038/ng.3526.
- Brunet, Frédéric G., Hugues Roest Crollius, Mathilde Paris, Jean-Marc Aury, Patricia Gibert, Olivier Jaillon, Vincent Laudet, and Marc Robinson-Rechavi (2006). “Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes”. In: *Molecular Biology and Evolution* 23 (9), pp. 1808–1816. DOI: 10.1093/molbev/msl049.
- Bu, Lijing, Ulfar Bergthorsson, and Vaishali Katju (2011). “Local synteny and codon usage contribute to asymmetric sequence divergence of *Saccharomyces cerevisiae* gene duplicates”. In: *BMC Evolutionary Biology* 11 (1), p. 279. DOI: 10.1186/1471-2148-11-279.
- Buchfink, Benjamin, Klaus Reuter, and Hajk Georg Drost (2021). “Sensitive protein alignments at tree-of-life scale using DIAMOND”. In: *Nature Methods* 18 (4), pp. 366–368. DOI: 10.1038/s41592-021-01101-x.
- Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden (2009). “BLAST + : architecture and applications”. In: *BMC Bioinformatics* 10 (421), p. 421. DOI: 10.1186/1471-2105-10-421.
- Campbell, Matthew A., Matthew C. Hale, Garrett J. McKinney, Krista M. Nichols, and Devon E. Pearse (2019). “Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in Salmonidae”. In: *G3: Genes, Genomes, Genetics* 9 (6), pp. 2017–2028. DOI: 10.1534/g3.119.400070.
- Cassa, Christopher A., Donate Weghorn, Daniel J. Balick, Daniel M. Jordan, David Nusinow, Kaitlin E. Samocha, Anne O’Donnell-Luria, Daniel G. MacArthur, Mark J. Daly, David R. Beier, and Shamil R. Sunyaev (2017). “Estimating the

- selective effects of heterozygous protein-truncating variants from human exome data". In: *Nature Genetics* 49 (5), pp. 806–810. DOI: 10.1038/ng.3831.
- Chao, Dai Yin, Brian Dilkes, Hongbing Luo, Alex Douglas, Elena Yakubova, Brett Lahner, and David E. Salt (2013). "Polyploids exhibit higher potassium uptake and salinity tolerance in *Arabidopsis*". In: *Science* 341 (6146), pp. 658–659. DOI: 10.1126/science.1240561.
- Chapman, Brad A, John E Bowers, Frank A Feltus, and A. H. Paterson (2006). "Buffering of crucial functions by paleologous duplicated genes may contribute cyclical to angiosperm genome duplication". In: *Proceedings of the National Academy of Sciences* 103 (8), pp. 2730–2735. DOI: 10.1073/pnas.0507782103.
- Charlesworth, Brian and Deborah Charlesworth (1998). *Some evolutionary consequences of deleterious mutations*, pp. 3–19.
- Chen, L., A. L. DeVries, and C.-H. C. Cheng (1997). "Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish". In: *Proceedings of the National Academy of Sciences* 94 (8), pp. 3811–3816. DOI: 10.1073/pnas.94.8.3811.
- Chen, Zelin, Yoshihiro Omori, Sergey Koren, Takuya Shirokiya, Takuo Kuroda, Atsushi Miyamoto, Hironori Wada, Asao Fujiyama, Atsushi Toyoda, Suiyuan Zhang, Tyra G. Wolfsberg, Koichi Kawakami, Adam M. Phillippy, James C. Mullikin, and Shawn M. Burgess (2019). "De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication". In: *Science Advances* 5 (6), pp. 1–13. DOI: 10.1126/sciadv.aav0547.
- Cheng, Peilin, Yu Huang, Yunyun Lv, Hao Du, Zhiqiang Ruan, Chuangju Li, Huan Ye, Hui Zhang, Jinming Wu, Chengyou Wang, Rui Ruan, Yanping Li, Chao Bian, Xinxin You, Chengcheng Shi, Kai Han, Junming Xu, Qiong

- Shi, and Qiwei Wei (2021). “The American Paddlefish Genome Provides Novel Insights into Chromosomal Evolution and Bone Mineralization in Early Vertebrates”. In: *Molecular biology and evolution* 38 (4), pp. 1595–1607. DOI: 10.1093/molbev/msaa326.
- Christensen, Kris A. and William S. Davidson (2017). “Autopolyploidy genome duplication preserves other ancient genome duplications in Atlantic salmon (*Salmo salar*)”. In: *PLoS ONE* 12 (2), pp. 1–16. DOI: 10.1371/journal.pone.0173053.
- Clarke, John T., Graeme T. Lloyd, and Matt Friedman (2016). “Little evidence for enhanced phenotypic evolution in early teleosts relative to their living fossil sister group”. In: *Proceedings of the National Academy of Sciences of the United States of America* 113 (41), pp. 11531–11536. DOI: 10.1073/pnas.1607237113.
- Comai, Luca (2005). “The advantages and disadvantages of being polyploid”. In: *Nature Reviews Genetics* 6 (11), pp. 836–846. DOI: 10.1038/nrg1711.
- Conant, Gavin C. (2020). “The lasting after-effects of an ancient polyploidy on the genomes of teleosts”. In: *PLOS ONE* 15 (4), e0231356. DOI: 10.1371/journal.pone.0231356.
- Conant, Gavin C., James A. Birchler, and J. Chris Pires (2014). “Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time”. In: *Current Opinion in Plant Biology* 19 (June), pp. 91–98. DOI: 10.1016/j.pbi.2014.05.008.
- Conrad, Bernard and Stylianos E. Antonarakis (2007). “Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease”. In: *Annual Review of Genomics and Human Genetics* 8 (1), pp. 17–35. DOI: 10.1146/annurev.genom.8.021307.110233.

- Cosentino, Salvatore and Wataru Iwasaki (2019). “SonicParanoid: Fast, accurate and easy orthology inference”. In: *Bioinformatics* 35 (1), pp. 149–151. DOI: 10.1093/bioinformatics/bty631.
- Coward, Jermaine and Angus Harding (2014). “Size does matter: Why polyploid tumor cells are critical drug targets in the war on cancer”. In: *Frontiers in Oncology* 4. DOI: 10.3389/fonc.2014.00123.
- Crow, Karen D. and Günter P. Wagner (2006). “What is the role of genome duplication in the evolution of complexity and diversity?” In: *Molecular Biology and Evolution* 23 (5), pp. 887–892. DOI: 10.1093/molbev/msj083.
- Davis, Jerel C. and Dmitri A. Petrov (2004). “Preferential Duplication of Conserved Proteins in Eukaryotic Genomes”. In: *PLoS Biology* 2 (3), e55. DOI: 10.1371/journal.pbio.0020055.
- De Storme, Nico and Danny Geelen (2020). “High temperatures alter cross-over distribution and induce male meiotic restitution in *Arabidopsis thaliana*”. In: *Communications Biology* 3 (1), p. 187. DOI: 10.1038/s42003-020-0897-1.
- Defoort, Jonas, Yves Van De Peer, Lorenzo Carretero-Paulet, and Brian Golding (2019). “The Evolution of Gene Duplicates in Angiosperms and the Impact of Protein-Protein Interactions and the Mechanism of Duplication”. In: *Genome Biology and Evolution* 11 (8), pp. 2292–2305. DOI: 10.1093/gbe/evz156.
- DeGiorgio, Michael and Raquel Assis (2021). “Learning Retention Mechanisms and Evolutionary Parameters of Duplicate Genes from Their Expression Data”. In: *Molecular Biology and Evolution* 38 (3), pp. 1209–1224. DOI: 10.1093/molbev/msaa267.
- Dehal, Paramvir and Jeffrey L. Boore (2005). “Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate”. In: *PLoS Biology* 3 (10), e314. DOI: 10.1371/journal.pbio.0030314.

- Des Marais, David L. and Mark D. Rausher (2008). “Escape from adaptive conflict after duplication in an anthocyanin pathway gene”. In: *Nature* 454 (7205), pp. 762–765. DOI: 10.1038/nature07092.
- Dewhurst, Sally M., Nicholas McGranahan, Rebecca A. Burrell, Andrew J. Rowan, Eva Grönroos, David Endesfelder, Tejal Joshi, Dmitri Mouradov, Peter Gibbs, Robyn L. Ward, Nicholas J. Hawkins, Zoltan Szallasi, Oliver M. Sieber, and Charles Swanton (2014). “Tolerance of whole- genome doubling propagates chromosomal instability and accelerates cancer genome evolution”. In: *Cancer Discovery* 4 (2), pp. 175–185. DOI: 10.1158/2159-8290.CD-13-0285.
- Dittwald, Piotr, Tomasz Gambin, Claudia Gonzaga-Jauregui, Claudia M.B. Carvalho, James R. Lupski, Paweł Stankiewicz, and Anna Gambin (2013). “Inverted Low-Copy Repeats and Genome Instability-A Genome-Wide Analysis”. In: *Human Mutation* 34 (1), pp. 210–220. DOI: 10.1002/humu.22217.
- Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras (2013). “STAR: Ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29 (1), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.
- Dong, H, L Nilsson, and C G Kurland (1995). “Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction”. In: *Journal of Bacteriology* 177 (6), pp. 1497–1504. DOI: 10.1128/jb.177.6.1497-1504.1995.
- Donne, Romain, Maëva Saroul-Aïnama, Pierre Cordier, Séverine Celton-Morizur, and Chantal Desdouets (2020). “Polyploidy in liver development, homeostasis and disease”. In: *Nature Reviews Gastroenterology and Hepatology* 17 (7), pp. 391–405. DOI: 10.1038/s41575-020-0284-x.
- Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold (2005). “Why highly expressed proteins evolve slowly”. In: *Proceedings of the*

- National Academy of Sciences* 102 (40), pp. 14338–14343. DOI: 10.1073/pnas.0504070102.
- Drummond, D. Allan, Alpan Raval, and Claus O. Wilke (2006). “A Single Determinant Dominates the Rate of Yeast Protein Evolution”. In: *Molecular Biology and Evolution* 23 (2), pp. 327–337. DOI: 10.1093/molbev/msj038.
- Du, Kang et al. (2020). “The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization”. In: *Nature Ecology & Evolution* 4 (6), pp. 841–852. DOI: 10.1038/s41559-020-1166-x.
- Duran, Bruno Oliveira Silva, Daniel Garcia de la serrana, Bruna Tereza Thomazini Zanella, Erika Stefani Perez, Edson Assunção Mareco, Vander Bruno Santos, Robson Francisco Carvalho, and Maeli Dal-Pai-Silva (2021). “An insight on the impact of teleost whole genome duplication on the regulation of the molecular networks controlling skeletal muscle growth”. In: *PLOS ONE* 16 (7), e0255006. DOI: 10.1371/journal.pone.0255006.
- Edgar, Bruce A., Norman Zielke, and Crisanto Gutierrez (2014). “Endocycles: A recurrent evolutionary innovation for post-mitotic cell growth”. In: *Nature Reviews Molecular Cell Biology* 15 (3), pp. 197–210. DOI: 10.1038/nrm3756.
- Edgar, Robert C. (2004). “MUSCLE: Multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32 (5), pp. 1792–1797. DOI: 10.1093/nar/gkh340.
- Emms, D. M. and S. Kelly (2018). “STAG: Species Tree Inference from All Genes”. In: *bioRxiv*. DOI: 10.1101/267914.
- Emms, David M. and Steven Kelly (2019). “OrthoFinder: phylogenetic orthology inference for comparative genomics”. In: *Genome Biology* 20 (1), p. 238. DOI: 10.1186/s13059-019-1832-y.

- (2017). “STRIDE: Species tree root inference from gene duplication events”. In: *Molecular Biology and Evolution* 34 (12), pp. 3267–3278. DOI: 10.1093/molbev/msx259.
- Estep, Matt C., Michael R. McKain, Dilys Vela Diaz, Jinshun Zhong, John G. Hodge, Trevor R. Hodgkinson, Daniel J. Layton, Simon T. Malcomber, Rémy Pasquet, and Elizabeth A. Kellogg (2014). “Allopolyploidy, diversification, and the Miocene grassland expansion”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111 (42), pp. 15149–15154. DOI: 10.1073/pnas.1404177111.
- Ezoe, Akihiro, Kazumasa Shirai, and Kousuke Hanada (2021). “Degree of Functional Divergence in Duplicates Is Associated with Distinct Roles in Plant Evolution”. In: *Molecular Biology and Evolution* 38 (4), pp. 1447–1459. DOI: 10.1093/molbev/msaa302.
- Fadista, João, Nikolay Oskolkov, Ola Hansson, and Leif Groop (2016). “LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals”. In: *Bioinformatics* 33 (4), btv602. DOI: 10.1093/bioinformatics/btv602.
- Force, Allan, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait (1999). “Preservation of Duplicate Genes by Complementary, Degenerative Mutations”. In: *Genetics* 151 (4), pp. 1531–1545. DOI: 10.1093/genetics/151.4.1531.
- Foster, A. D., L. L. Flynn, C. Cluning, F. Cheng, J. M. Davidson, A. Lee, N. Polain, R. Mejzini, N. Farrawell, J. J. Yerbury, R. Layfield, P. A. Akkari, and S. L. Rea (2021). “p62 overexpression induces TDP-43 cytoplasmic mislocalisation, aggregation and cleavage and neuronal death”. In: *Scientific Reports* 11 (1), p. 11474. DOI: 10.1038/s41598-021-90822-2.

- Fowler, Norma L. and Donald A. Levin (2016). “Critical factors in the establishment of allopolyploids”. In: *American Journal of Botany* 103 (7), pp. 1236–1251. DOI: 10.3732/ajb.1500407.
- Freeling, Michael (2009). “Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition”. In: *Annual Review of Plant Biology* 60 (1), pp. 433–453. DOI: 10.1146/annurev.arplant.043008.092122.
- Freeling, Michael, Margaret R. Woodhouse, Shabarinath Subramaniam, Gina Turco, Damon Lisch, and James C. Schnable (2012). *Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants*. DOI: 10.1016/j.pbi.2012.01.015.
- Gerstein, Aleeza C., Hye-Jung E. Chun, Alex Grant, and Sarah P. Otto (2006). “Genomic Convergence toward Diploidy in *Saccharomyces cerevisiae*”. In: *PLoS Genetics* 2 (9), e145. DOI: 10.1371/journal.pgen.0020145.
- Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre (2017). “Correlation and variable importance in random forests”. In: *Statistics and Computing* 27 (3), pp. 659–678. DOI: 10.1007/s11222-016-9646-1.
- GTEX Consortium (2020). “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369 (6509), pp. 1318–1330. DOI: 10.1126/science.aaz1776.
- Gu, Xun (2003). “Evolution of duplicate genes versus genetic robustness against null mutations”. In: *Trends in Genetics* 19 (7), pp. 354–356. DOI: 10.1016/S0168-9525(03)00139-2.
- Guan, Yuanfang, Maitreya J. Dunham, and Olga G. Troyanskaya (2007). “Functional Analysis of Gene Duplications in *Saccharomyces cerevisiae*”. In: *Genetics* 175 (2), pp. 933–943. DOI: 10.1534/genetics.106.064329.

- Hakes, Luke, John W Pinney, Simon C Lovell, Stephen G Oliver, and David L Robertson (2007). “All duplicates are not equal: the difference between small-scale and genome duplication”. In: *Genome Biology* 8 (10), R209. DOI: 10.1186/gb-2007-8-10-r209.
- Havelka, M., M. Hulák, D. A. Bailie, P. A. Prodöhl, and M. Flajšhans (2013). “Extensive genome duplications in sturgeons: New evidence from microsatellite data”. In: *Journal of Applied Ichthyology* 29 (4), pp. 704–708. DOI: 10.1111/jai.12224.
- He, Xionglei and Jianzhi Zhang (2005). “Gene Complexity and Gene Duplicability”. In: *Current Biology* 15 (11), pp. 1016–1021. DOI: 10.1016/j.cub.2005.04.035.
- (2006). “Higher Duplicability of Less Important Genes in Yeast Genomes”. In: *Molecular Biology and Evolution* 23 (1), pp. 144–151. DOI: 10.1093/molbev/msj015.
- Hittinger, Chris Todd and Sean B. Carroll (2007). “Gene duplication and the adaptive evolution of a classic genetic switch”. In: *Nature* 449 (7163), pp. 677–681. DOI: 10.1038/nature06151.
- Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh (2018). “UFBoot2: Improving the Ultrafast Bootstrap Approximation”. In: *Molecular Biology and Evolution* 35 (2), pp. 518–522. DOI: 10.1093/molbev/msx281.
- Holt, Richard, Nuala H. Sykes, Inês C. Conceição, Jean Baptiste Cazier, Richard J.L. Anney, Guiomar Oliveira, Louise Gallagher, Astrid Vicente, Anthony P. Monaco, and Alistair T. Pagnamenta (2012). “CNVs leading to fusion transcripts in individuals with autism spectrum disorder”. In: *European Journal of Human Genetics* 20 (11), pp. 1141–1147. DOI: 10.1038/ejhg.2012.73.

- Huerta-Cepas, Jaime, François Serra, and Peer Bork (2016). “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data”. In: *Molecular Biology and Evolution* 33 (6), pp. 1635–1638. DOI: 10.1093/molbev/msw046.
- Inoue, Jun, Yukuto Sato, Robert Sinclair, Katsumi Tsukamoto, and Mutsumi Nishida (2015). “Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling”. In: *Proceedings of the National Academy of Sciences* 112 (48), pp. 14918–14923. DOI: 10.1073/pnas.1507669112.
- Ishikawa, Koji, Koji Makanae, Shintaro Iwasaki, Nicholas T. Ingolia, and Hisao Moriya (2017). “Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes”. In: *PLOS Genetics* 13 (1), e1006554. DOI: 10.1371/journal.pgen.1006554.
- Jaillon, Olivier et al. (2004). “Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype”. In: *Nature* 431 (7011), pp. 946–957. DOI: 10.1038/nature03025.
- Jiang, W.-k., Y.-l. Liu, E.-h. Xia, and L.-z. Gao (2013). “Prevalent Role of Gene Features in Determining Evolutionary Fates of Whole-Genome Duplication Duplicated Genes in Flowering Plants”. In: *PLANT PHYSIOLOGY* 161 (4), pp. 1844–1861. DOI: 10.1104/pp.112.200147.
- Jiang, Xueyuan and Raquel Assis (2017). “Natural Selection Drives Rapid Functional Evolution of Young *Drosophila* Duplicate Genes”. In: *Molecular Biology and Evolution* 34 (12), pp. 3089–3098. DOI: 10.1093/molbev/msx230.
- Jordan, I. King, Yuri I. Wolf, and Eugene V. Koonin (2004). “Duplicated genes evolve slower than singletons despite the initial rate increase”. In: *BMC Evolutionary Biology* 4, p. 22. DOI: 10.1186/1471-2148-4-22.
- Jun, Jin, Paul Ryvkin, Edward Hemphill, and Craig Nelson (2009). “Duplication Mechanism and Disruptions in Flanking Regions Determine the Fate of

- Mammalian Gene Duplicates”. In: *Journal of Computational Biology* 16 (9), pp. 1253–1266. DOI: 10.1089/cmb.2009.0074.
- Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K.F. Wong, Arndt Von Haeseler, and Lars S. Jermiin (2017). “ModelFinder: Fast model selection for accurate phylogenetic estimates”. In: *Nature Methods* 14 (6), pp. 587–589. DOI: 10.1038/nmeth.4285.
- Kassahn, Karin S., Vinh T. Dang, Simon J. Wilkins, Andrew C. Perkins, and Mark A. Ragan (2009). “Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates”. In: *Genome Research* 19 (8), pp. 1404–1418. DOI: 10.1101/gr.086827.108.
- Keller, T. E. and S. V. Yi (2014). “DNA methylation and evolution of duplicate genes”. In: *Proceedings of the National Academy of Sciences* 111 (16), pp. 5932–5937. DOI: 10.1073/pnas.1321420111.
- Kellis, Manolis, Bruce W Birren, and Eric S Lander (2004). “Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*”. In: *Nature* 428, pp. 617–624.
- Kenny, N J, K W Chan, W Nong, Z Qu, I Maeso, H Y Yip, T F Chan, H S Kwan, P W H Holland, K H Chu, and J H L Hui (2016). “Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs”. In: *Heredity* 116 (2), pp. 190–199. DOI: 10.1038/hdy.2015.89.
- Klüver, Nils, Mariko Kondo, Amaury Herpin, Hiroshi Mitani, and Manfred Schartl (2005). “Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization”. In: *Development Genes and Evolution* 215 (6), pp. 297–305. DOI: 10.1007/s00427-005-0477-x.
- Kondrashov, Fyodor A. (2012). “Gene duplication as a mechanism of genomic adaptation to a changing environment”. In: *Proceedings of the Royal Society*

- B: Biological Sciences* 279 (1749), pp. 5048–5057. DOI: 10.1098/rspb.2012.1108.
- Krabbenhof, Trevor J, Daniel J MacGuigan, Nathan J C Backenstose, Hannah Waterman, Tianying Lan, Jessie A Pelosi, Milton Tan, and Simen R Sandve (2021). “Chromosome-Level Genome Assembly of Chinese Sucker (*Myxocyprinus asiaticus*) Reveals Strongly Conserved Synteny Following a Catostomid-Specific Whole-Genome Duplication”. In: *Genome Biology and Evolution* 13 (9). DOI: 10.1093/gbe/evab190.
- Kreiner, Julia M., Paul Kron, and Brian C. Husband (2017). “Frequency and maintenance of unreduced gametes in natural plant populations: associations with reproductive mode, life history and genome size”. In: *New Phytologist* 214 (2), pp. 879–889. DOI: 10.1111/nph.14423.
- Kryuchkova-Mostacci, Nadezda and Marc Robinson-Rechavi (2016). “A benchmark of gene expression tissue-specificity metrics”. In: *Briefings in Bioinformatics* 18 (2), bbw008. DOI: 10.1093/bib/bbw008.
- Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges (2017). “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times”. In: *Molecular biology and evolution* 34 (7), pp. 1812–1819. DOI: 10.1093/molbev/msx116.
- Lan, X. and J. K. Pritchard (2016). “Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals”. In: *Science* 352 (6288), pp. 1009–1013. DOI: 10.1126/science.aad8411.
- Laurent, Sacha, Nicolas Salamin, and Marc Robinson-Rechavi (2017). “No evidence for the radiation time lag model after whole genome duplications in Teleostei”. In: *PLoS ONE* 12 (4), e0176384. DOI: 10.1371/journal.pone.0176384.

- Lee, Wen-Ching, Yun-Yu Chen, Daphne Kan, and Chung-Liang Chien (2012). “A neuronal death model: overexpression of neuronal intermediate filament protein peripherin in PC12 cells”. In: *Journal of Biomedical Science* 19 (8).
- Lek, Monkol et al. (2016). “Analysis of protein-coding genetic variation in 60,706 humans”. In: *Nature* 536 (7616), pp. 285–291. DOI: 10.1038/nature19057.
- Levin, Donald A. and Douglas E. Soltis (2018). “Factors promoting polyploid persistence and diversification and limiting diploid speciation during the K–Pg interlude”. In: *Current Opinion in Plant Biology* 42, pp. 1–7. DOI: 10.1016/j.pbi.2017.09.010.
- Li, Bo and Colin N Dewey (2011). “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC Bioinformatics* 12 (323).
- Li, Juan, Lan Jiang, Chung I. Wu, Xuemei Lu, Shu Fang, and Chau Ti Ting (2019). “Small segmental duplications in *Drosophila*-high rate of emergence and elimination”. In: *Genome Biology and Evolution* 11 (2), pp. 486–496. DOI: 10.1093/gbe/evz011.
- Li, Zhen, Jonas Defoort, Setareh Tasdighian, Steven Maere, Yves Van de Peer, and Riet De Smet (2016). “Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms”. In: *The Plant Cell* 28 (2), pp. 326–344. DOI: 10.1105/tpc.15.00877.
- Li, Zheng, Anthony E. Baniaga, Emily B. Sessa, Moira Scascitelli, Sean W. Graham, Loren H. Rieseberg, and Michael S. Barker (2015). “Early genome duplications in conifers and other seed plants”. In: *Science Advances* 1 (10). DOI: 10.1126/sciadv.1501084.
- Li, Zheng, George P. Tiley, Sally R. Galuska, Chris R. Reardon, Thomas I. Kidder, Rebecca J. Rundell, and Michael S. Barker (2018). “Multiple large-scale gene and genome duplications during the evolution of hexapods”. In: *Proceedings*

- of the National Academy of Sciences of the United States of America* 115 (18), pp. 4713–4718. DOI: 10.1073/pnas.1710791115.
- Lien, Sigbjørn et al. (2016). “The Atlantic salmon genome provides insights into rediploidization”. In: *Nature* 533 (7602), pp. 200–205. DOI: 10.1038/nature17164.
- Lipinski, Kendra J., James C. Farslow, Kelly A. Fitzpatrick, Michael Lynch, Vaishali Katju, and Ulfar Bergthorsson (2011). “High spontaneous rate of gene duplication in *Caenorhabditis elegans*”. In: *Current Biology* 21 (4), pp. 306–310. DOI: 10.1016/j.cub.2011.01.026.
- Liu, Zhong Jian (2015). “The genome sequence of the orchid *Phalaenopsis equestris*”. In: *Nature genetics* 47 (1), pp. 65–72. DOI: 10.1038/ng.3149.
- Lorin, Thibault, Frédéric G. Brunet, Vincent Laudet, and Jean Nicolas Volff (2018). “Teleost fish-specific preferential retention of pigmentation gene-containing families after whole genome duplications in vertebrates”. In: *G3: Genes, Genomes, Genetics* 8 (5), pp. 1795–1806. DOI: 10.1534/g3.118.200201.
- Luck, Katja et al. (2020). “A reference map of the human binary protein interactome”. In: *Nature* 580 (7803), pp. 402–408. DOI: 10.1038/s41586-020-2188-x.
- Luther, Pradeep K, Peter M G Munrot, and John M Squire (1995). “Muscle Ultrastructure in the Teleost Fish”. In: *Micron* 26 (5), pp. 431–459.
- Lynch, M., W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas (2008). “A genome-wide view of the spectrum of spontaneous mutations in yeast”. In: *Proceedings of the National Academy of Sciences* 105 (27), pp. 9272–9277. DOI: 10.1073/pnas.0803466105.

- Lynch, Michael and John S Conery (2000). “The Evolutionary Fate and Consequences of Duplicate Genes”. In: *Science* 290 (5494), pp. 1151–1155. DOI: 10.1126/science.290.5494.1151.
- (2003). “The evolutionary demography of duplicate genes”. In: *Journal of Structural and Functional Genomics* 3, pp. 35–44. DOI: 10.1023/A:1022696612931.
- Mable, B. K., M. A. Alexandrou, and M. I. Taylor (2011). “Genome duplication in amphibians and fish: An extended synthesis”. In: *Journal of Zoology* 284 (3), pp. 151–182. DOI: 10.1111/j.1469-7998.2011.00829.x.
- Makino, Takashi, Karsten Hokamp, and Aoife McLysaght (2009). “The complex relationship of gene duplication and essentiality”. In: *Trends in Genetics* 25 (4), pp. 152–155. DOI: 10.1016/j.tig.2009.02.001.
- Makino, Takashi and A. McLysaght (2010). “Ohnologs in the human genome are dosage balanced and frequently associated with disease”. In: *Proceedings of the National Academy of Sciences* 107 (20), pp. 9270–9274. DOI: 10.1073/pnas.0914697107.
- Marsit, S., M. Hénault, G. Charron, A. Fijarczyk, and C. R. Landry (2021). “The neutral rate of whole-genome duplication varies among yeast species and their hybrids”. In: *Nature Communications* 12 (1), pp. 1–11. DOI: 10.1038/s41467-021-23231-8.
- Martin, Marcel (2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17 (1), p. 10. DOI: 10.14806/ej.17.1.200.
- Matsumoto, Tomonori, Leslie Wakefield, Alexander Peters, Myron Peto, Paul Spellman, and Markus Grompe (2021). “Proliferative polyploid cells give rise to tumors via ploidy reduction”. In: *Nature Communications* 12 (1), p. 646. DOI: 10.1038/s41467-021-20916-y.

- Mayrose, Itay, Shing H. Zhan, Carl J. Rothfels, Karen Magnuson-Ford, Michael S. Barker, Loren H. Rieseberg, and Sarah P. Otto (2015). “Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al.(2014)”. In: *New Phytologist* 206, pp. 27–35.
- (2011). “Recently Formed Polyploid Plants Diversify at Lower Rates”. In: *Science* 333, p. 1257. DOI: 10.5061/dryad.6hf21.
- McGrath, Casey L., Jean Francois Gout, Thomas G. Doak, Akira Yanagi, and Michael Lynch (2014). “Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence”. In: *Genetics* 197 (4), pp. 1417–1428. DOI: 10.1534/genetics.114.163287.
- Mefford, Heather C. and Evan E. Eichler (2009). “Duplication hotspots, rare genomic disorders, and common disease”. In: *Current Opinion in Genetics and Development* 19 (3), pp. 196–204. DOI: 10.1016/j.gde.2009.04.003.
- Mészáros, Bálint, Gábor Erdős, and Zsuzsanna Dosztányi (2018). “IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding”. In: *Nucleic Acids Research* 46 (W1), W329–W337. DOI: 10.1093/nar/gky384.
- Meyer, Axel and Yves Van De Peer (2005). “From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD)”. In: *BioEssays* 27 (9), pp. 937–945. DOI: 10.1002/bies.20293.
- Meyers, Lauren Ancel and Donald A Levin (2006). “ON THE ABUNDANCE OF POLYPLOIDS IN FLOWERING PLANTS”. In: *Evolution* 60 (6), pp. 1198–1206.
- Ming, Ray et al. (2015). “The pineapple genome and the evolution of CAM photosynthesis”. In: *Nature Genetics* 47 (12), pp. 1435–1442. DOI: 10.1038/ng.3435.

- Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt Von Haeseler, Robert Lanfear, and Emma Teeling (2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. In: *Molecular Biology and Evolution* 37 (5), pp. 1530–1534. DOI: 10.1093/molbev/msaa015.
- Modrek, Barmak and Christopher J. Lee (2003). “Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss”. In: *Nature Genetics* 34 (2), pp. 177–180. DOI: 10.1038/ng1159.
- Morgan, Chris, Huakun Zhang, Clare E. Henry, F. Chris H. Franklin, and Kirsten Bomblies (2020). “Derived alleles of two axis proteins affect meiotic traits in autotetraploid *Arabidopsis arenosa*”. In: *Proceedings of the National Academy of Sciences* 117 (16), pp. 8980–8988. DOI: 10.1073/pnas.1919459117.
- Morrill, Summer A. and Angelika Amon (2019). “Why haploinsufficiency persists”. In: *Proceedings of the National Academy of Sciences*, p. 201900437. DOI: 10.1073/pnas.1900437116.
- Nakatani, Yoichiro and Aoife McLysaght (2019). “Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116 (6), pp. 1816–1818. DOI: 10.1073/pnas.1817937116.
- Nakatani, Yoichiro, Prashant Shingate, Vydianathan Ravi, Nisha E. Pillai, Aravind Prasad, Aoife McLysaght, and Byrappa Venkatesh (2021). “Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution”. In: *Nature Communications* 12 (1). DOI: 10.1038/s41467-021-24573-z.

- Nakatani, Yoichiro, Hiroyuki Takeda, Yuji Kohara, and Shinichi Morishita (2007). “Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates”. In: *Genome Research* 17 (9), pp. 1254–1265. DOI: 10.1101/gr.6316407.
- Natsidis, Paschalis, Paschalia Kapli, Philipp H. Schiffer, and Maximilian J. Telford (2021). “Systematic errors in orthology inference and their effects on evolutionary analyses”. In: *iScience* 24 (2), p. 102110. DOI: 10.1016/j.isci.2021.102110.
- Nong, Wenyan et al. (2021). “Horseshoe crab genomes reveal the evolution of genes and microRNAs after three rounds of whole genome duplication”. In: *Communications Biology* 4 (1), pp. 1–11. DOI: 10.1038/s42003-020-01637-2.
- O’Toole, Áine N, Laurence D Hurst, and Aoife McLysaght (2018). “Faster Evolving Primate Genes Are More Likely to Duplicate”. In: *Molecular Biology and Evolution* 35 (1), pp. 107–118. DOI: 10.1093/molbev/msx270.
- Oberdorf, Richard and Tanja Kortemme (2009). “Complex topology rather than complex membership is a determinant of protein dosage sensitivity”. In: *Molecular Systems Biology* 5 (1), p. 253. DOI: 10.1038/msb.2009.9.
- Ohno, S (1970). *Evolution by gene duplication*. Springer, New York.
- Oswald, Benjamin P. and Scott L. Nuismer (2007). “Neopolyploidy and pathogen resistance”. In: *Proceedings of the Royal Society B: Biological Sciences* 274 (1624), pp. 2393–2397. DOI: 10.1098/rspb.2007.0692.
- Pál, Csaba, Balázs Papp, and Laurence D Hurst (2001). “Highly Expressed Genes in Yeast Evolve Slowly”. In: *Genetics* 158 (2), pp. 927–931. DOI: 10.1093/genetics/158.2.927.

- Papp, Balázs, Csaba Pál, and Laurence D. Hurst (2003). “Dosage sensitivity and the evolution of gene families in yeast”. In: *Nature* 424 (6945), pp. 194–197. DOI: 10.1038/nature01713.
- Parr, Terence and Kerem Turgutlu (2020). *Feature importances for scikit-learn machine learning models*.
- Paterson, Andrew H., Brad A. Chapman, Jessica C. Kissinger, John E. Bowers, Frank A. Feltus, and James C. Estill (2006). “Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon”. In: *Trends in Genetics* 22 (11), pp. 597–602. DOI: 10.1016/j.tig.2006.09.003.
- Pécix, Yann, Géraldine Rallo, Hélène Folzer, Mireille Cigna, Serge Gudín, and Manuel Le Bris (2011). “Polyploidization mechanisms: Temperature environment can induce diploid gamete formation in *Rosa* sp.” In: *Journal of Experimental Botany* 62 (10), pp. 3587–3597. DOI: 10.1093/jxb/err052.
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau (2011). “Journal of Machine Learning Research: Preface”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Pegueroles, Cinta, Steve Laurie, and M. Mar Albà (2013). “Accelerated evolution after gene duplication: A time-dependent process affecting just one copy”. In: *Molecular Biology and Evolution* 30 (8), pp. 1830–1842. DOI: 10.1093/molbev/mst083.
- Perry, George H, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, Fernando A Villanea, Joanna L Mountain, Rajeev Misra, Nigel P Carter, Charles Lee, and Anne C Stone (2007).

- “Diet and the evolution of human amylase gene copy number variation”. In: *Nature Genetics* 39 (10), pp. 1256–1260.
- Petrovski, Slavé, Quanli Wang, Erin L. Heinzen, Andrew S. Allen, and David B. Goldstein (2013). “Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes”. In: *PLoS Genetics* 9 (8), e1003709. DOI: 10.1371/journal.pgen.1003709.
- Pfeiffer, C C and E R Braverman (1982). “Zinc, the brain and behavior.” In: *Biological psychiatry* 17 (4).
- Proost, Sebastian, Jan Fostier, Dieter De Witte, Bart Dhoedt, Piet Demeester, Yves Van De Peer, and Klaas Vandepoele (2012). “i-ADHoRe 3.0-fast and sensitive detection of genomic homology in extremely large data sets”. In: *Nucleic Acids Research* 40 (2), e11. DOI: 10.1093/nar/gkr955.
- Puigbò, Pere, Ignacio G. Bravo, and Santiago Garcia-Vallve (2008). “CAIcal: A combined set of tools to assess codon usage adaptation”. In: *Biology Direct* 3 (1), p. 38. DOI: 10.1186/1745-6150-3-38.
- Putnam, Nicholas H. et al. (2008). “The amphioxus genome and the evolution of the chordate karyotype”. In: *Nature* 453 (7198), pp. 1064–1071. DOI: 10.1038/nature06967.
- Qiao, Xin, Hao Yin, Leiting Li, Runze Wang, Juyou Wu, Jun Wu, and Shaoling Zhang (2018). “Different Modes of Gene Duplication Show Divergent Evolutionary Patterns and Contribute Differently to the Expansion of Gene Families Involved in Important Fruit Traits in Pear (*Pyrus bretschneideri*)”. In: *Frontiers in Plant Science* 9. DOI: 10.3389/fpls.2018.00161.
- Rackham, Owen J. L., Hashem A. Shihab, Michael R. Johnson, and Enrico Petretto (2015). “EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization”. In: *Nucleic Acids Research* 43 (5), e33–e33. DOI: 10.1093/nar/gku1322.

- Ramsey, Justin and Douglas W Schemske (1998). “PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS”. In: *Annu. Rev. Ecol. Syst* 29, pp. 467–501.
- Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo (2019). “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic Acids Research* 47 (W1), W191–W198. DOI: 10.1093/nar/gkz369.
- Ren, Dani, Juan Song, Ming Ni, Le Kang, and Wei Guo (2020). “Regulatory Mechanisms of Cell Polyploidy in Insects”. In: *Frontiers in Cell and Developmental Biology* 8. DOI: 10.3389/fcell.2020.00361.
- Rice, Alan M and Aoife McLysaght (2017a). “Dosage sensitivity is a major determinant of human copy number variant pathogenicity”. In: *Nature Communications* 8 (1), p. 14366. DOI: 10.1038/ncomms14366.
- (2017b). “Dosage-sensitive genes in evolution and disease”. In: *BMC Biology* 15 (1), p. 78. DOI: 10.1186/s12915-017-0418-y.
- Robertson, Fiona M., Manu Kumar Gundappa, Fabian Grammes, Torgeir R. Hvidsten, Anthony K. Redmond, Sigbjørn Lien, Samuel A.M. Martin, Peter W.H. Holland, Simen R. Sandve, and Daniel J. Macqueen (2017). “Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification”. In: *Genome Biology* 18 (1), p. 111. DOI: 10.1186/s13059-017-1241-z.
- Rocha, Eduardo P.C., John Maynard Smith, Laurence D. Hurst, Matthew T.G. Holden, Jessica E. Cooper, Noel H. Smith, and Edward J. Feil (2006). “Comparisons of dN/dS are time dependent for closely related bacterial genomes”. In: *Journal of Theoretical Biology* 239 (2), pp. 226–235. DOI: 10.1016/j.jtbi.2005.08.037.

- Roelofs, Dick, Arthur Zwaenepoel, Tom Sistermans, Joey Nap, Andries A. Kampfraath, Yves Van De Peer, Jacintha Ellers, and Ken Kraaijeveld (2020). “Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution”. In: *BMC Biology* 18 (1), p. 57. DOI: 10.1186/s12915-020-00789-1.
- Rong, J., F. A. Feltus, L. Liu, L. Lin, and A. H. Paterson (2010). “Gene copy number evolution during tetraploid cotton radiation”. In: *Heredity* 105 (5), pp. 463–472. DOI: 10.1038/hdy.2009.192.
- Roux, Julien, Jialin Liu, and Marc Robinson-Rechavi (2017). “Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates”. In: *Molecular Biology and Evolution* 34 (11), pp. 2773–2791. DOI: 10.1093/molbev/msx199.
- Roy, Scott William and David Penny (2007). “On the Incidence of Intron Loss and Gain in Paralogous Gene Families”. In: *Molecular Biology and Evolution* 24 (8), pp. 1579–1581. DOI: 10.1093/molbev/msm082.
- Ruiz, M., A. Quiñones, M.R. Martínez-Cuenca, P. Aleza, R. Morillon, L. Navarro, E. Primo-Millo, and B. Martínez-Alcántara (2016). “Tetraploidy enhances the ability to exclude chloride from leaves in carrizo citrange seedlings”. In: *Journal of Plant Physiology* 205, pp. 1–10. DOI: 10.1016/j.jplph.2016.08.002.
- Samocha, Kaitlin E. et al. (2014). “A framework for the interpretation of de novo mutation in human disease”. In: *Nature Genetics* 46 (9), pp. 944–950. DOI: 10.1038/ng.3050.
- Schartl, Manfred, Ronald B. Walter, Yingjia Shen, Tzintzuni Garcia, Julian Catchen, Angel Amores, Ingo Braasch, Domitille Chalopin, Jean Nicolas Volff, Klaus Peter Lesch, Angelo Bisazza, Pat Minx, Ladeana Hillier, Richard K. Wilson, Susan Fuerstenberg, Jeffrey Boore, Steve Searle, John H. Postlethwait,

- and Wesley C. Warren (2013). “The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits”. In: *Nature Genetics* 45 (5), pp. 567–572. DOI: 10.1038/ng.2604.
- Schranz, M. Eric, Setareh Mohammadin, and Patrick P. Edger (2012). “Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model”. In: *Current Opinion in Plant Biology* 15 (2), pp. 147–153. DOI: 10.1016/j.pbi.2012.03.011.
- Seabold, Skipper and Josef Perktold (2010). “Statsmodels: Econometric and Statistical Modeling with Python”. In: *Proceedings of the 9th Python in Science Conference (Scipy)*, pp. 92–96. DOI: 10.25080/majora-92bf1922-011.
- Sémon, Marie and Kenneth H. Wolfe (2007). “Rearrangement rate following the whole-genome duplication in teleosts”. In: *Molecular Biology and Evolution* 24 (3), pp. 860–867. DOI: 10.1093/molbev/msm003.
- Seoighe, Cathal and Kenneth H Wolfe (1999). “Yeast genome evolution in the post-genome era”. In: *Current Opinion in Microbiology* 2 (5), pp. 548–554. DOI: 10.1016/S1369-5274(99)00015-6.
- Session, Adam M. et al. (2016). “Genome evolution in the allotetraploid frog *Xenopus laevis*”. In: *Nature* 538 (7625), pp. 336–343. DOI: 10.1038/nature19840.
- Shi, Xiaoli, Changqing Zhang, Dae Kwan Ko, and Z. Jeffrey Chen (2015). “Genome-wide dosage-dependent and-independent regulation contributes to gene expression and evolutionary novelty in plant polyploids”. In: *Molecular Biology and Evolution* 32 (9), pp. 2351–2366. DOI: 10.1093/molbev/msv116.
- Simakov, Oleg, Ferdinand Marlétaz, Jia Xing Yue, Brendan O’Connell, Jerry Jenkins, Alexander Brandt, Robert Calef, Che Huang Tung, Tzu Kai Huang, Jeremy Schmutz, Nori Satoh, Jr Kai Yu, Nicholas H. Putnam, Richard E. Green, and Daniel S. Rokhsar (2020). “Deeply conserved synteny resolves

- early events in vertebrate evolution". In: *Nature Ecology and Evolution* 4 (6), pp. 820–830. DOI: 10.1038/s41559-020-1156-z.
- Singh, Param Priya, Séverine Affeldt, Ilaria Cascone, Rasim Selimoglu, Jacques Camonis, and Hervé Isambert (2012). "On the Expansion of " Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates". In: *Cell Reports* 2 (5), pp. 1387–1398. DOI: 10.1016/j.celrep.2012.09.034.
- Singh, Param Priya, Jatin Arora, and Hervé Isambert (2015). "Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes". In: *PLOS Computational Biology* 11 (7), e1004394. DOI: 10.1371/journal.pcbi.1004394.
- Singh, Param Priya and Hervé Isambert (2019). "OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates". In: *Nucleic Acids Research* 48 (D1), pp. D724–D730. DOI: 10.1093/nar/gkz909.
- Smet, R. De, K. L. Adams, K. Vandepoele, M. C. E. Van Montagu, S. Maere, and Y. Van de Peer (2013). "Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants". In: *Proceedings of the National Academy of Sciences* 110 (8), pp. 2898–2903. DOI: 10.1073/pnas.1300127110.
- Soltis, D. E. and J. G. Burleigh (2009). "Surviving the K-T mass extinction: New perspectives of polyploidization in angiosperms". In: *Proceedings of the National Academy of Sciences* 106 (14), pp. 5455–5456. DOI: 10.1073/pnas.0901994106.
- Soltis, Pamela S., Xiaoxian Liu, D. Blaine Marchant, Clayton J. Visger, and Douglas E. Soltis (2014). "Polyploidy and novelty: Gottlieb's legacy". In:

- Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1648). DOI: 10.1098/rstb.2013.0351.
- Soo, Valerie W.C., Paulina Hanson-Manful, and Wayne M. Patrick (2011). “Artificial gene amplification reveals an abundance of promiscuous resistance determinants in *Escherichia coli*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (4), pp. 1484–1489. DOI: 10.1073/pnas.1012108108.
- Soshnikova, Natalia, Romain Dewaele, Philippe Janvier, Robb Krumlauf, and Denis Duboule (2013). “Duplications of hox gene clusters and the emergence of vertebrates”. In: *Developmental Biology* 378 (2), pp. 194–199. DOI: 10.1016/j.ydbio.2013.03.004.
- Stebbins, G.L. (1971). *Chromosomal evolution in higher plants*. Edward Arnold, London, UK.
- Storz, Jay F., Juan C. Opazo, and Federico G. Hoffmann (2013). “Gene duplication, genome duplication, and the functional diversification of vertebrate globins”. In: *Molecular Phylogenetics and Evolution* 66 (2), pp. 469–478. DOI: 10.1016/j.ympev.2012.07.013.
- Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn (2007). “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC Bioinformatics* 8 (1), p. 25. DOI: 10.1186/1471-2105-8-25.
- Szklarczyk, Damian, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering (2021). “The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets”. In: *Nucleic Acids Research* 49 (D1), pp. D605–D612. DOI: 10.1093/nar/gkaa1074.

- Tank, David C., Jonathan M. Eastman, Matthew W. Pennell, Pamela S. Soltis, Douglas E. Soltis, Cody E. Hinchliff, Joseph W. Brown, Emily B. Sessa, and Luke J. Harmon (2015). “Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications”. In: *New Phytologist* 207 (2), pp. 454–467. DOI: 10.1111/nph.13491.
- Tasdighian, Setareh, Michiel Van Bel, Zhen Li, Yves Van De Peer, Lorenzo Carretero-Paulet, and Steven Maere (2017). “Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity”. In: *Plant Cell* 29 (11), pp. 2766–2785. DOI: 10.1105/tpc.17.00313.
- Thompson, Ammon, Derek Vo, Caitlin Comfort, and Harold H. Zakon (2014). “Expression evolution facilitated the convergent neofunctionalization of a sodium channel gene”. In: *Molecular Biology and Evolution* 31 (8), pp. 1941–1955. DOI: 10.1093/molbev/msu145.
- Tiley, George P., Michael S. Barker, and J. Gordon Burleigh (2018). “Assessing the performance of KS plots for detecting ancient whole genome duplications”. In: *Genome Biology and Evolution* 10 (11), pp. 2882–2898. DOI: 10.1093/gbe/evy200.
- Tomala, Katarzyna and Ryszard Korona (2013). “Evaluating the fitness cost of protein expression in *saccharomyces cerevisiae*”. In: *Genome Biology and Evolution* 5 (11), pp. 2051–2060. DOI: 10.1093/gbe/evt154.
- Train, Clément Marie, Natasha M. Glover, Gaston H. Gonnet, Adrian M. Altenhoff, and Christophe Dessimoz (2017). “Orthologous Matrix (OMA) algorithm 2.0: More robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference”. In: *Bioinformatics* 33 (14), pp. i75–i82. DOI: 10.1093/bioinformatics/btx229.

- Vakirlis, Nikolaos, Anne-Ruxandra Carvunis, and Aoife McLysaght (2020). “Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes”. In: *eLife* 9, pp. 1–23. DOI: 10.7554/eLife.53500.
- Van De Peer, Yves, Eshchar Mizrachi, and Kathleen Marchal (2017). “The evolutionary significance of polyploidy”. In: *Nature Reviews Genetics* 18 (7), pp. 411–424. DOI: 10.1038/nrg.2017.26.
- Vanneste, Kevin, Guy Baele, Steven Maere, and Yves Van De Peer (2014). “Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary”. In: *Genome Research* 24 (8), pp. 1334–1347. DOI: 10.1101/gr.168997.113.
- Vanneste, Kevin, Lieven Sterck, Alexander Andrew Myburg, Yves Van de Peer, and Eshchar Mizrachi (2015). “Horsetails are ancient polyploids: Evidence from *Equisetum giganteum*”. In: *Plant Cell* 27 (6), pp. 1567–1578. DOI: 10.1105/tpc.15.00157.
- Veitia, Reiner A (2003). “Nonlinear Effects in Macromolecular Assembly and Dosage Sensitivity”. In: *J. theor. Biol* 220, pp. 19–25. DOI: 10.1006.
- (2009). “On gene dosage balance in protein complexes: a comment on Semple JI, Vavouri T, Lehner B. A simple principle concerning the robustness of protein complex activity to changes in gene expression.” In: *BMC Systems Biology* 3 (1), p. 16. DOI: 10.1186/1752-0509-3-16.
- (2002). “Exploring the etiology of haploinsufficiency”. In: *BioEssays* 24 (2), pp. 175–184. DOI: 10.1002/bies.10023.
- Veitia, Reiner A. and James A. Birchler (2015). “Models of buffering of dosage imbalances in protein complexes”. In: *Biology Direct* 10 (1), p. 42. DOI: 10.1186/s13062-015-0063-8.

- Vicario, Saverio, Etsuko N Moriyama, and Jeffrey R Powell (2007). “Codon usage in twelve species of *Drosophila*”. In: *BMC Evolutionary Biology* 7 (1), p. 226. DOI: 10.1186/1471-2148-7-226.
- Wagner, G. P., C. Amemiya, and F. Ruddle (2003). “Hox cluster duplications and the opportunity for evolutionary novelties”. In: *Proceedings of the National Academy of Sciences* 100 (25), pp. 14603–14606. DOI: 10.1073/pnas.2536656100.
- Wang, Feng Yu, Wen Sung Chung, Hong Young Yan, and Chyng Shyan Tzeng (2008). “Adaptive evolution of cone opsin genes in two colorful cyprinids, *Opsariichthys pachycephalus* and *Candidia barbatus*”. In: *Vision Research* 48 (16), pp. 1695–1704. DOI: 10.1016/j.visres.2008.04.026.
- Wang, T., K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, and D. M. Sabatini (2015). “Identification and characterization of essential genes in the human genome”. In: *Science* 350 (6264), pp. 1096–1101. DOI: 10.1126/science.aac7041.
- Wang, Zhong-Yi, Evgeny Leushkin, Angélica Liechti, Svetlana Ovchinnikova, Katharina Mößinger, Thoomke Brüning, Coralie Rummel, Frank Grützner, Margarida Cardoso-Moreira, Peggy Janich, David Gatfield, Boubou Diagouraga, Bernard de Massy, Mark E. Gill, Antoine H. F. M. Peters, Simon Anders, and Henrik Kaessmann (2020). “Transcriptome and translome co-evolution in mammals”. In: *Nature* 588 (7839), pp. 642–647. DOI: 10.1038/s41586-020-2899-z.
- Weisman, Caroline M, Andrew W Murray, and Sean R Eddy (2020). “Many, but not all, lineage-specific genes can be explained by homology detection failure.” In: *PLoS biology* 18 (11), e3000862. DOI: 10.1371/journal.pbio.3000862.

- Wolfe, K H and D C Shields (1997). “Molecular evidence for an ancient duplication of the entire yeast genome.” In: *Nature* 387 (6634), pp. 708–13. DOI: 10.1038/42711.
- Wolfe, Ken (2004). “Evolutionary Genomics: Yeasts Accelerate beyond BLAST”. In: *Current Biology* 14 (10), R392–R394. DOI: 10.1016/j.cub.2004.05.015.
- Woods, Shane, Avril Coghlan, David Rivers, Tobias Warnecke, Sean J. Jeffries, Taejoon Kwon, Anthony Rogers, Laurence D. Hurst, and Julie Ahringer (2013). “Duplication and Retention Biases of Essential and Non-Essential Genes Revealed by Systematic Knockdown Analyses”. In: *PLoS Genetics* 9 (5), e1003330. DOI: 10.1371/journal.pgen.1003330.
- Wu, Shengdan, Baocai Han, and Yuannian Jiao (2020). “Genetic Contribution of Paleopolyploidy to Adaptive Evolution in Angiosperms”. In: *Molecular Plant* 13 (1), pp. 59–71. DOI: 10.1016/j.molp.2019.10.012.
- Xu, Peng et al. (2019). “The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*”. In: *Nature Communications* 10 (1), p. 4625. DOI: 10.1038/s41467-019-12644-1.
- Yang, Fan, David A. Pecina, Shelly D. Kelly, Sang Hoon Kim, Kenneth M. Kemner, David T. Long, and Terence L. Marsh (2010). “Biosequestration via cooperative binding of copper by *Ralstonia pickettii*”. In: *Environmental Technology* 31 (8-9), pp. 1045–1060. DOI: 10.1080/09593330.2010.487290.
- Yang, P. M., Q. C. Huang, G. Y. Qin, S. P. Zhao, and J. G. Zhou (2014). “Different drought-stress responses in photosynthesis and reactive oxygen metabolism between autotetraploid and diploid rice”. In: *Photosynthetica* 52 (2), pp. 193–202. DOI: 10.1007/s11099-014-0020-2.
- Yang, Ziheng (2007). “PAML 4: Phylogenetic analysis by maximum likelihood”. In: *Molecular Biology and Evolution* 24 (8), pp. 1586–1591. DOI: 10.1093/molbev/msm088.

190 BIBLIOGRAPHY

- Yates, Andrew D. et al. (2019). “Ensembl 2020”. In: *Nucleic Acids Research* 48 (D1), pp. D682–D688. DOI: 10.1093/nar/gkz966.
- Zhang, Feng, Wenli Gu, Matthew E. Hurles, and James R. Lupski (2009). “Copy number variation in human health, disease, and evolution”. In: *Annual Review of Genomics and Human Genetics* 10, pp. 451–481. DOI: 10.1146/annurev.genom.9.081307.164217.
- Zhu, Yun, Zhenguo Lin, and Luay Nakhleh (2013). “Evolution After Whole-Genome Duplication: A Network Perspective”. In: *G3: Genes/Genomes/Genetics* 3 (11), pp. 2049–2057. DOI: 10.1534/g3.113.008458.