Using Micro-Synteny for Phylogenetic Inference and Analysis

by

Dearbhaile Casey

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

Doctor of Philosophy

Smurfit Institute of Genetics
Trinity College Dublin
The University of Dublin

May, 2024

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.

Dearbhaile Casey — *Signature*

09/05/2024 — *Date*

This thesis is dedicated to my father, Raymond Casey. Rest in peace.

Acknowledgements

I firstly want to thank my friends. One thing is certain, being my friend over the past 5 years has not been an easy task. I thank you all deeply for sticking with me through the good and bad. I must be some craic!

Thank you to Oisín, while you encouraged me to take a break post bachelors degree you supported me through this four year toil none the less. I'll take a break soon. Maybe. Sophie, Orla, Laura, Joe, Ciaran and Kristina - we started this journey together 9 years ago and half of us are, or nearly are, doctors and the other half extremely successful in other fields. Thank you for all your support throughout this time. Laura, I will never forget how incredible you were to me during my mental health struggles. Hopefully there won't be anymore crying sessions over black coffee and hot chocolate in *cafe tri va*. Those poor baristas were certain we were in an extremely tumultuous relationship. You are a safety net for me and I love you. To Brendan, thank you for your endless support and constant hype. You always lifted my spirits during the challenging moments of my PhD journey. The looming pandemic added an extra layer of difficulty, making 'real-life' and family feel distant. Your support has been a valuable source of comfort. Thank you to Sian, Clara, Cissy, Aisling, Holly, Alison, Jess, Yvonne and Aoife (my Mummas) for wine and chats soundtracked by Robbie Williams. Thank you to Stephen and Pearse for welcoming me into Moy Elta during the final months of this PhD journey. It was a challenging and emotional period for me, and your

unwavering support and kindness will forever be etched in my memory. When I started my PhD, I never thought I would move to London, work in UCL and manage to wrangle together a bunch of incredible friends along the way. Laura, Kevin, Giorgia, Eddie, Paris, Lena and Lou; thank you for making me feel at home in London. What can I say? I love my friends!

To Max and Paschalia, I came to UCL with very little experience in phylogenetics and you allowed me the space to grow and learn from you both. Max, your fervor for delving into unanswered questions with a unique perspective has been truly inspiring. I appreciate the opportunity you provided me to learn from you and the EvoCELL group. Paschalia, and Tomas, thank you for welcoming me into your home and Paschalia, your mentorship means a lot to me. As a female scientist, I admire and aspire to emulate your excellence in the field.

To Anthony, not sure I would have had a thesis had you not tied me to your paddlefish wagon. Thank you for all the phylogenetics wisdom you have imparted on me. While I believe I embraced most of it willingly, we both recognize that deep down, you just wanted someone else in the lab to talk about trees with. You are a huge inspiration to me and your mentorship and friendship has made me a far better scientist and made this PhD experience a whole lot easier. I am excited to see what you will achieve in the next few years and I hope we will cross paths many times again in the future.

To Aoife, Wow. My mentor. My Dublin Mammy. My friend. Over the past 8 years you have guided and supported me through both exciting and difficult times. You are the reason I fell in love with genetics, did this PhD and I aspire to become half as good a scientist as you are and quarter as good a person. Thank you for believing in me and helping me grow. It's true, we are the best lab and that is because you are the best PI.

To my family, thank you for your love and support not only over the past four

years but my whole life. I didn't make it easy for you all, that is for sure! Grainne, thank you for always sticking with me and believing in me. You're kindness knows no bounds and my daily goal is to be half as decent a human as you are. Aisling and Sean (and the boys), you were always there if I needed an office, a laptop or a bit of money to see me to the end of the month. Thank you for your generosity. Blathin, growing up, I often felt like I was in your shadow, thinking I could never match your intelligence or accomplishments. However, I've come to realize that your achievements are yours, and I have my own unique strengths and we are both remarkable in our own ways. Thank you paving the path and giving me guidance. Doireann, my baby sister, we have both been through so much together and thank you for believing in me and supporting me through the good and bad times. Raymond, thank you for paying for those coding lessons when I was 18. Not sure I would be doing this particular PhD if it wasn't for that. Thank you for always worrying about me and I think, in most instances, you were right to do so. Growing up, I always hoped to impress you Ray, but God have I tripped up many times in my 27 years. I hope you are proud of me now.

To Dylan, the past three years falling in love with you have been magical. You met me on the first year of the PhD journey and so don't actually know me not stressed and you still fell for me! Now, I am not sure I will be any less different post-doctoral but here's hoping you still like the more calm Dearbhaile I hope to be. Your love and unwavering support has kept me going through this tough period in my life and I am endlessly grateful for you and excited to see how our lives will continue to unfold together.

Finally Mam and Dad, thank you for your endless support throughout my life. You had a few tough years with me but you got me through it and I know you are proud of me now. Mam, I will never be able to understand how you are so selfless. I hope one day I can show you how much you mean to me. I probably

wouldn't have made it this far in life if it wasn't for you and I will be forever in your debt for that. Dad, I wish you were here to see me graduate with a PhD from Trinity. I know you are up there somewhere bragging to your friends about me. I miss you dearly but there is no doubt about it, you are well and truly alive in all of us. I dedicate this thesis and all the work I have done over the past four years to you. Rest in peace Dad.

Contents

List of Tables

List of Figures

Abbreviations

DNA	Deoxyribonucleic acid
NGS	Next Generation Sequencing
AORe	Ancestral Ohnolog Resolution
SSD	Small Scale Duplication
WGD	Whole genome duplication
BL	Branch lengths
BUSCO	Benchmarking Universal Single Copy Orthologs
CCS	Circular Consensus Sequencing
LORe	Lineage-specific Ohnolog Resolution
OLC	Overlap-Layout-Consensus
PacBio	Pacific Biosciences, Inc
PHOGS	Phylogenetic Hierarchical Orthogroups
PostSpec	Post-Speciation
PreSpec	Pre-Speciation
RNA	Ribonucleic acid

RF	Robinson-Foulds
SA	Sequence Alignment
SMS	Single-Molecule Sequencing
TE	Transposable-element
TsGD	Teleost-specific Genome Duplication
VCF	Variant Call Format
ZMW	Zero-Mode Waveguide

A Metazoan puzzle

A shallow blue,
dappled in pockets of a translucent mass.
Iridescent shivering strings emanate from a pellucid core,
swaying.

Signs of life.

Who is their brother, sister, ancestor?
We zoom out and within . . .

The A, T, G, C's aligned in their billions, as a jumbled pile of puzzle pieces is picked apart and pieced together.
Patterns emerge, similarities unveiled and relatives reunited.

Yet still, secrets remain.

We toil together.
Assemble, disassemble, reassemble.

Perhaps this puzzle will forever remain undone.

— Dearbhaile Casey

Summary

Phylogenetics is the study of the evolutionary relationships and history of groups of organisms. Over the past twenty years, the burgeoning number of sequenced genomes has revolutionised the field of phylogenetics taking it from morphological cladistics, in which organisms are grouped together based on shared, derived morphological characteristics (synapomorphies) to more exact molecular methods; comparing the percent identity of amino acid, DNA or RNA sequences between species. Most commonly, phylogenetic inference utilises molecular alignments, arranging the genetic sequences in such a way that homologous positions are matched, allowing similarities and differences to be quantified. In chapter 3 we will use a novel method for inference, inspecting the changes in local gene position, or micro-synteny, between species as a phylogenetic character.

By characterising the spatial organisation of genes across evolutionary timescales we investigate how this can offer an alternative perspective on species relationships, juxtaposing it to standard sequence alignment methods. Micro-synteny, or micro-collinearity, can provide a novel and powerful framework for understanding the shared positional ancestry of genes; help to unravel inconclusive orthology assignments of large multi-gene families; be used to delineate and map ancestral rearrangments as well as here, being used for phylogenetic inference.

The genomic architecture of different lineages has been shaped by ancestral Whole genome duplications(WGD), Small Scale Duplications (SSD), inversions

and translocations. Micro-collinearity, therefore, may not be highly conserved across all taxa with such complex histories. However, recent work looking at conserved micro-synteny in angiosperm and mammals showed that while genome organisation is more conserved in the latter group, the complex polyploid histories of flowering plants contained enough phylogenetic information for highly accurate phylogenies to be built. In chapter 3, we carry out a comparative analysis, contrasting gene order evolution with sequence evolution, revealing the interplay between the two characters throughout the evolution of mammals and angiosperms, bolstering recent evidence which highlighted the strength of genome organisation as a character to explain lineage evolution.

In chapter 4 we move away from synteny and focus on improving the genome assembly of *Polyodon spathula*, the American paddlefish, which is then used for analysis in chapter 5. The paddlefish is a member of the Acipenseriforme lineage, a group which occupies the basal position of ray-finned fishes. The group have complex genomes that appear to be highly repetitive, probably a repercussion of the WGD event that took place in the history of the lineage. As a result of this repetitiveness, there are many regions that failed to assemble correctly in the initial sequencing attempt with some duplicates collapsing and presenting as a single read rather than two separate reads. Many of these collapsed regions may contain genes important for evolutionary studies and so in this chapter we attempt to "uncollapse" or "reduplicate" these regions of the genome and improve the overall assembly and annotation of the paddlefish.

In chapter 5, we again use synteny, this time not as a character for phylogenetic inference, but rather to identify blocks of conserved gene order between two animal genomes, the paddlefish and sturgeon, which experienced a shared ancestral-WGD. With our improved paddlefish genome from chapter 4, we show that the positional conservation of groups of genes between and within these fishes can be used as

identifiers of ancestral genome rearrangements that facilitated the rediploidization process in their genomes. Rediploidization is the process of returning to bivalent chromosome pairing after a WGD event. Following the self-doubling of a diploid genome in a process known as auto-polyploidisation, there will be homologous recombination across four chromosomes which will continue until recombination is suppressed. Tetraploid inheritance is unstable and it is thought that the halting of recombination is caused by asynchronous genomic rearrangements and an accumulation of mutations which facilitate the return to stable, bivalent pairing. Once the loci have been uncoupled from the tetraploid conformation, they can begin functionally diverging and be thus, considered duplicated loci known as ohnologs. In chapter 5 we show that there are blocks of contiguous ohnolog loci, purported to have originated from genome rearrangements at different time points, either in the ancestor of the Sturgeon and Paddlefish or independently in each lineage.

The information contained within this work highlights the importance of gene collinearity, its conservation and movement, in aiding in determination of evolutionary relationships. Genomes are dynamic and appreciation of this in our research is paramount for unravelling the true genetic history of lineages. In this thesis I hope to show that incorporating synteny and synteny evolution into phylogenetic analyses can build upon previous frameworks and help in solving many unanswered questions in animal evolution.

Chapter 1

Introduction

1.1 Micro-synteny as a character for phylogenetic inference

Phylogenetics is the study of evolutionary relationships among and between different species or groups of organisms. From the development of cladistics in the early 20th century which was based on the idea that organisms are grouped together by shared, derived morphological and some early molecular characteristics (synapomorphies) to the ever-advancing field of molecular phylogenetics, the study of relationships has become a fundamental discipline that underpins our understanding of the evolutionary history of all organisms, extant and extinct. Its applications extend across many fields ranging from basic research to practical applications in medicine, conservation, and biotechnology.

Phylogenetic reconstruction can be achieved through several different methods via the profiling of different characters; molecular and morphological. Most commonly we compare changes in the genetic sequence between two or more organisms. Less attention however, has been dedicated to other characters such as morphological changes and gene order changes, in part because they have been shown to be less

accurate. In this work, we will look at the use of micro-synteny in phylogenetic inference. Local conservation of gene order or micro-synteny are terms used to describe the shared spatial ancestry of groups of genes within and between genomes. Terms such as collinearity or micro-collinearity are also used interchangeably to define this same phenomenon. The practice of incorporating micro-synteny information into phylogenetics spans decades, but many of these older methods were computationally expensive, relied on well developed models and worked only on small, simple genomes (Belda et al., 2005; Tang and Moret, 2003; Luo et al., 2008; Feng et al., 2017). In the past five years there has been a wave of new techniques which use a distance-based approach based on breakpoints in syntenic blocks and combine synteny networks with ML based phylogenetics tools (Drillon et al., 2020; Zhao et al., 2021). These approaches have succeeded in building phylogenies of large whole-genome datasets. In this thesis I take a closer look at this and other alternative approaches to phylogenetic inference and compare characters, like branch length, from these methods to characters in sequence alignment phylogenies. We also investigate the evolution of these novel characters through time.

1.1.1 An overview of phylogenetic reconstruction

Most commonly we use sequence alignments to construct phylogenies. This technique aligns two or more biological sequences to identify regions of similarity and difference. Alignments are a fundamental tool in bioinformatics and molecular biology because they allow researchers to infer evolutionary relationships, identify conserved regions and predict functional elements.

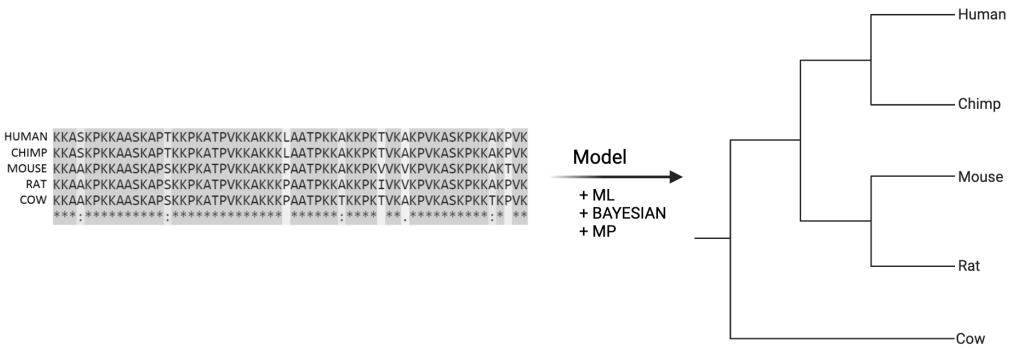

Figure 1.1 | Alignments for phylogenetic inference Figure shows how aligning information from five species can be transformed into phylogenetic information by using models based off mathematical frameworks like Maximum Likelihood (ML), Bayesian statistics or Maximum Parsimony (MP).

The use of molecular alignments can be traced back to the middle of the 20th century, with the advent of early biological sequence data and recognition of how aligning sequences can highlight their similarities and differences (Needleman and Wunsch, 1970; Smith and Waterman, 1981). Early methods like Needleman-Wunsch Algorithm (Needleman and Wunsch, 1970) and Smith-Waterman Algorithm (Smith and Waterman, 1981) were pairwise techniques, still in use today for local alignments but nowadays, for phylogenetic studies, we use multiple sequence aligners like ClustalW and MAFFT which are computationally efficient and more accurate for larger datasets (Higgins and Sharp, 1988; Katoh et al., 2002). Once you've aligned your sequences and checked their quality, the next step is inferring the relationships between the sequences which can be achieved with mathematical approaches such as ML, Bayesian methods or Maximum parsimony (Fig.1.1). Model choice is often the most important decision you'll make at this point in your inference (Abadi et al., 2019; Posada and Crandall, 2001). Different approaches make different assumptions about the underlying evolutionary processes. While, no evolutionary model can fully capture the genuine complexity of the evolutionary

process, such that even the most adequate one merely provides an approximation of reality. To draw accurate evolutionary trees, it's essential to choose a model that closely mirrors the true biological processes that shaped the genetic data because inappropriate model selection can lead to erroneous conclusion.

For more simple situations where you have high quality data, in general, a simple model is adequate (Felsenstein, 1981; Hasegawa et al., 1985; Kimura, 1980; Zharkikh, 1994; Tamura and Nei, 1993). Complex situations call for complex models, like CAT+GTR[1.1.1] , and have the potential to offer a more nuanced view of evolution. Complex models can better capture the heterogeneity in sequence evolution across sites and branches, which can lead to more accurate tree topologies. However, complex models often require significantly more computational resources, have issues with overfitting the data and results can often be uninterruptible. Even with the most complex models some species relationships remain unsolved. Notably, the ongoing debate on whether the sister group to all animals are sponges or ctenophores as well as controversy over the monophyletic grouping of deuterostomes (Dunn et al., 2015; Redmond et al., 2023; Pisani et al., 2015; Kapli et al., 2021). For these reasons, scientists have looked elsewhere for answers including exploring different types of signals representing rare genomic changes as well as using synteny information (Rokas and Holland, 2000; Parey et al., 2023; Zhao et al., 2021; Drillon et al., 2014).

While typically, molecular sequence alignments have proven to be the most accurate tool for reconstructing relationships, less attention has been devoted to understanding other characters which could enhance or even surpass the infor-

[1]**CAT+GTR:** This model is a sophisticated and computationally intensive phylogenetic model that combines the CAT (Mixture Model of Substitution) model with the GTR (General Time Reversible) model. CAT accounts for rate heterogeneity among sites, while GTR models the substitution rates between different nucleotides or amino acids. This model is particularly useful for capturing complex evolutionary processes and improving the accuracy of phylogenetic tree inferences. However, it comes with high computational demands and requires careful parameter estimation, making it suitable for well-resourced research projects.

mation extracted from standard methods. The use of synteny in evolutionary analysis has advanced in the past few years and brought with it new tools that can be used for orthology inference, ancestral reconstructions of polyploidy events, phylogenetic profiling and much more (Lechner et al., 2014; Walden and Schranz, 2023; Simakov et al., 2020; Nakatani et al., 2021; Zhao et al., 2021). In this work, we look at how micro-synteny can be used as a character for phylogenetic inference. We assess its effectiveness in resolving relationships among deeply diverged species and compare its capabilities to those of conventional sequence alignment methods.

1.1.2 Tools for synteny detection

As mentioned, phylogenetic inference using a synteny-based methodology has been statistically and algorithmically challenging until recently. Early work focused on small datasets and restricted analysis to mitochondrial data, simple organisms and highly collinear genomes (Blanchette et al., 1999; Belda et al., 2005; Tang and Moret, 2003; Feng et al., 2017). Over the past five years, huge improvements have been made with a plethora of new synteny-block detection tools including MCScanX, i-ADHoRe, SynChro and Satsuma being published and used extensively in many genomics studies (Wang et al., 2012; Proost et al., 2012; Drillon et al., 2014; Grabherr et al., 2010). Synteny blocks are formally defined as regions of chromosomes between genomes that share a common order of orthologous genes derived from a common ancestor (Feng et al., 2017). A depiction of a synteny block, as defined in many of the synteny-based tools mentioned above, can be seen in Fig1.2. Features of these blocks include *anchors* which are stretches of orthologous genes shared between two or more species and breaks which are regions with no orthologs or a gap in synteny (Fig1.2).

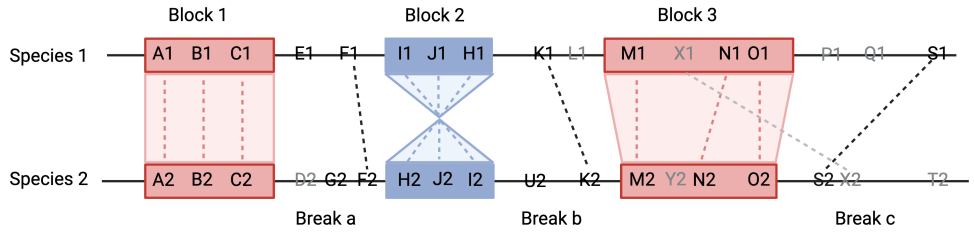

Figure 1.2 | Schematic of a synteny block Genes located on chromosomes of two species (species 1 and 2) are denoted as letters and the species they belong to e.g. A1 is gene A in species 1. Orthologous genes are connected by dashed lines and genes without an orthologous relationship are treated as gaps or breaks (light grey). Under the criteria of which you need at least three orthologous genes (anchors) to define a block, three scenarios are shown: Block 1, orthologs with the same order, with reversed order (Block 2), or as shown in Block 3, allowing some gaps. In contrast, synteny breaks are caused by a lack of orthologs (Break a and Break b) or gaps of synteny(Break c).

Synteny analysis, now common-place in most genome studies, can accompany alignment, be used to study evolutionary processes that have lead to diversity of chromosome number and identify conserved regions between genomes (Veltri et al., 2016; Lechner et al., 2014; Nakatani et al., 2021; Simakov et al., 2020; Walden and Schranz, 2023). Most of the tools available work in the most logical way, by filtering and organising regions of shared gene-order between genomes. This is achieved by identifying orthologs that can be used as *anchors* to position local alignments within and between genomes (see Fig.1.2 and Fig.1.3). Many methods use a gene homology matrix (Proost et al., 2012) or a reciprocal best hit algorithm (Wang et al., 2012; Drillon et al., 2014), both which achieve similar results that vary on resolution depending on the parameters used. A depiction of the general framework of synteny-based phylogenetic inference tools can be seen in Fig.1.3.

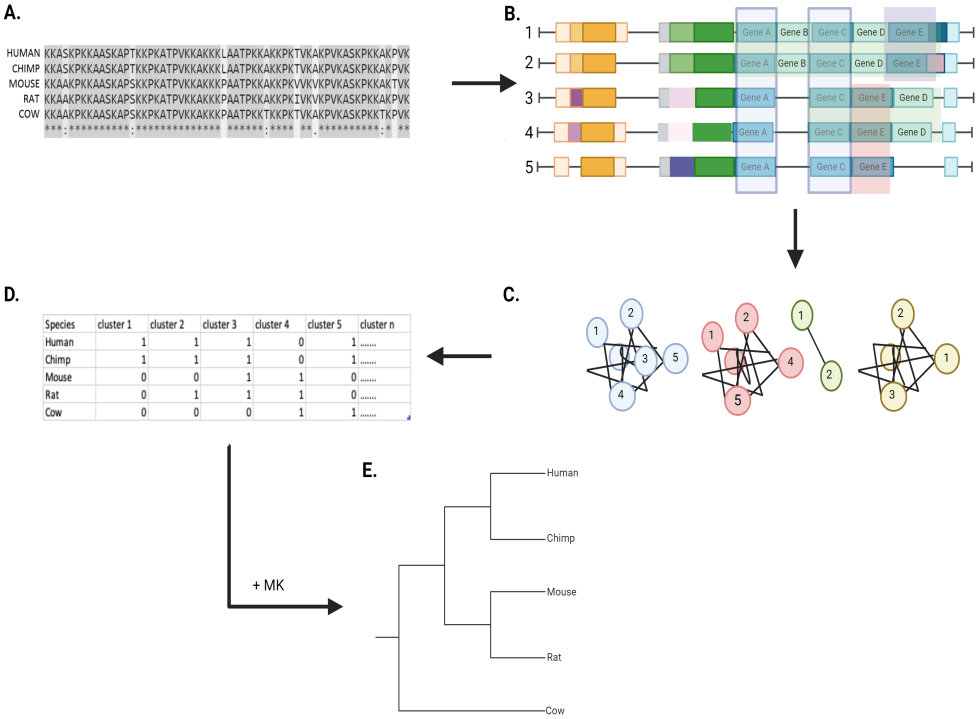

Figure 1.3 | Schematic of the general framework of a synteny-based phylogenetic inference tool (A) With your genomes of interest collected, the first step is orthology inference achieved using reciprocal best hits or tools like Orthofinder(Emms and Kelly, 2019) followed by alignment of orthologs. **(B)** Genomes are then stacked for detection of micro-syntenic blocks (Fig1.2) between and within them (only one block is shown here which produces one cluster in next step) **(C)** Block information is reshaped into clusters. **(D)** Cluster information is then transformed into a binary matrix representing presence and absence of a species in a cluster. **(E)** Phylogenetic trees are generated from this binary data using ML based phylogenetics tools and the MK morphological model (Lewis, 2001). This figure is adapted from a figure in Zhao and Schranz, 2017

Many of these synteny block detection tools have been adapted for phylogenetic inference (Drillon et al., 2020; Zhao et al., 2021). One of the first efficient methods using synteny blocks reconstructed a 20 species yeast phylogeny using a Double Cut and Join method (Feng et al., 2017). This method is a model for genome rearrangements and defines the distance between genomes based on gene order and orientation, rather than nucleotide sequence (Lin et al., 2010). *PhyChro*

was developed by scientists to reconstruct phylogenies based on chromosomal rearrangements and proved to be vastly more accurate and computationally efficient than previous attempts (Drillon et al., 2014; Drillon et al., 2020). The method is distance-based combined with a pairwise approach to synteny block detection which allowed for scaling and efficiency. For each breakpoint issued from each pairwise comparisons, the algorithm defines two disjoint sets of segments of genomes, called partial splits, which support the two block adjacency's defining the breakpoint between the genomes. When you have all the partial splits constructed from the pairwise analyses, *PhyChro* computes the distance between the two genomes based on the number of partial splits that separates them. This bottom up approach, by iteratively grouping sister genomes and minimizing genome distances, informs the tree reconstruction. The distance based method applied in *PhyCro* was applied to 13 vertebrate genomes and 21 yeast genomes and successfully reconstructed accurate phylogenies (Drillon et al., 2014).

Zhao et al., 2021, combined synteny detection with standard maximum likelihood based phylogenetics to reconstruct phylogenies for large whole-genome datasets of angiosperm and mammals (Zhao et al., 2021; Zhao et al., 2017; Zhao and Schranz, 2017). Similar to *PhyChro*, the pipeline uses a pairwise approach which has the potential to work on large numbers of genomes. The method follows a similar framework to that depicted in Fig.1.3 and includes four main steps, (1) synteny block detection using MCScanX (Wang et al., 2012), (2) network clustering (3) binary matrix representation of cluster information and (4) ML based tree inference. The framework for the approach is initially similar to standard approaches, beginning with orthology detection. This information is then networked to unveil a matrix with homology information in a syntenic context. The "microsynteny-based" trees built with the pipeline in Zhao et al., 2021, were highly congruent with phylogenies built using standard methodologies, with simulations confirm-

ing its accuracy and efficiency. Gene order has long been recognised as having a phylogenetic signal and with pipelines like *PhyChro* and this, *Syn-MRL*, it is now possible to exploit this information and reconstruct species relationships efficiently and accurately (Drillon et al., 2020; Zhao et al., 2021).

1.1.3 Gene order evolution versus molecular sequence evolution

The rate in which molecular sequence changes occur overtime differs from the frequency in which genes rearrange in the genome. Understanding if there is an interplay between these phenomena requires further investigation. Similar to the long standing question of whether or not there is a relationship between genomic and phenotypic evolution, here we ask if there is a relationship between gene order evolution and sequence evolution (Omland, 1997; Davies and Savolainen, 2006)? Until recently, it has been impossible to test this as micro-synteny-based phylogenetic methods were inaccurate and only reliable with simple data (Blanchette et al., 1999; Luo et al., 2008; Belda et al., 2005). As detailed above, many of these new tools have shown that the phylogenetic signal is strong enough to reconstruct phylogenies therefore making it possible to garner some rate and temporal information from the branches of these trees. Branch lengths in an ultrametric tree built using sequence alignments tell us how the rate of sequence evolution has changed through time. Following those same lines, branch length in trees built with micro-synteny information therefore tell us how the rate of gene order evolution has changed through time.

$$Branch\ Length\ = Time\ X\ Rate\ of\ Change \tag{1.1}$$

It is important when carrying out phylogenomic comparative analysis to be aware that different methodologies may introduce phylogenetic artefacts that could influence results. ML trees produced from the *Syn-MRL* pipeline are built with the MK morphological model(Lewis, 2001). The MK model is appropriate for discrete data, such as binary data and is analogous to the Jukes-Cantor model of sequence evolution (Lewis, 2001). The MK model assumes that transitions among these states follow a Markov process. This means that the probability of changing from one state to another depends only on the current state, and not on what has come before. The MK model has been shown to produce an inflation in the branch lengths due to ascertainment bias (Brown et al., 2017; Abadi et al., 2019). While more prevalent in morphological data, ascertainment bias may also affect micro-synteny data and previous work with morphological data excluded all terminal branches from correlation analysis (Liu et al., 2018; Bromham et al., 2002; Davies and Savolainen, 2006). Another crucial aspect to consider in this analysis is the potential presence of branch length auto-correlation within certain datasets Felsenstein, 1985. This phenomenon arises when branches in close proximity exhibit more similar values than those at greater distances causing them to group together in correlation investigations introducing unreliable results. Due to this clustering, the points are non-independent and so regardless of the relationship under scrutiny they may group together and erroneously suggest a false relationship. Felsenstein, 1985 formulated a method to account for these statistically non-independent data points, now common place when comparing biological trait-data across taxa in phylogenies (Felsenstein, 1985).

In this thesis, we examine the correlation between branch lengths within a fixed topology built through two distinct approaches: SA-based methods and innovative synteny methods for phylogenetic inference. Our aim is to uncover, if any, the interplay between these two molecular changes throughout the evolution of diverse phylogenies. Our analysis focuses on mammalian and angiosperm phylogenies, where we construct traditional sequence alignment trees, employ pre-computed synteny-derived phylogenies, and create presence/absence phylogenies. We compared branch lengths between these trees and investigated the correlation between the rate of change along the branches of the different data types. With this information we can shed light on the relationships between the evolution of molecular sequence changes, gene order and gene gain-and-loss respectively. We also look at the patterns of rate-change through time as it may be possible to pinpoint potential evolutionary, geographical, ecological or anomalous factors that may have influenced the rate of sequence change or gene order change at different time points. Factors that influence how the molecular sequence and other correlated factors have changed over time are important to understand. Events that influence the rate of fixed heritable genetic change, or substitution rate, and lead to adaption and ultimately speciation, are of a molecular evolutionists paramount interest and underpin the fundamental processes that drive organisms evolution.

1.2 Misassembly and collapsed repeat regions in genome assemblies

It is unclear to what extent published genome assemblies have been affected by the inherent error-prone nature of genome sequencing technologies. Most commonly, published genomes have been sequenced by second generation technologies, or

Next Generation Sequencing (NGS). While NGS has greatly enhanced our ability to sequence DNA at high throughput and for a low cost, sequencing errors and misassemblies are commonplace due to intrinsic errors in NGS methods. These errors can have substantial effects on research results and stunt the transformative capabilities genome sequencing could have in clinical settings (Salzberg and Yorke, 2005; Mardis, 2008; Kelley and Salzberg, 2010). A major challenge for sequencing and assembling diploid and polyploid non-model organisms is inaccurate resolution of duplicate genes and repetitive DNA (Salzberg and Yorke, 2005; Tørresen et al., 2019). While, Single Molecule Sequencing (SMS) (or Third Generation Sequencing (TGS)) offers relief from issues with repetitive DNA resolution, TGS is still not customary in most sequencing studies (Eid et al., 2009; Rothberg et al., 2011; Quail et al., 2012). Major progress has been made to overcome issues in genome assemblies but there is still a way to go before we can attest a perfect genome sequencing tool.

In chapter 4 we attempt to *re-duplicate* collapsed duplicates in the paddlefish genome following an adapted protocol by Du et al., 2020. The paddlefish is a complex case when it comes to assembling and annotating because the species experienced a WGD approximately 254.7-241.8 Myr that it shares with the sterlet sturgeon (Redmond et al., 2023). Following the WGD, the paddlefish transitioned from its ancestral tetraploid state to a functional diploid in an asynchronous process of rediploidization. Despite the long time that has elapsed since the event, both the sterlet and paddlefish still have signatures of polyploidy in their genomes. The current paddlefish assembly has fallen victim to the deficiency inherent to short-read assembly tools, with substantial evidence of collapsed duplicates and misassembled repeat regions.

1.2.1 A brief history of genome sequencing

From the very first assembly of the *Caenorhabditis elegans* in 1998, sequencing and assembling animal genomes has become a fundamental part of molecular biology (Consortium, 1998). The quantity of genomes available on databases like GenBank and UniProt has exponentially increased in the past twenty-three years and the technologies for sequencing, assembling and depositing data are being constantly updated and refined (Hotaling et al., 2021; Sayers et al., 2019). Sequencing data has revolutionised how we study all aspects of biology; from its use in clinical settings to the significant contribution it has made to research. With all this data, it is important that users are aware of the quality and the potential errors inherent in generating sequencing reads (Salzberg and Yorke, 2005; Tørresen et al., 2019; Kelley and Salzberg, 2010).

Nowadays, the two most common ways of sequencing are second-generation sequencing or short-read sequencing and third-generation sequencing, more commonly known as long-read sequencing. Second-generation sequencing or NGS is ubiquitous and has transformed the landscape of biology since its genesis at the start of the millennium. Despite it's profound effect on research and in clinical settings, it has its shortcomings. When NGS fragments are mapped to the reference genome, it can be challenging to determine the correct location of duplicated sequences due to multiple potential matching sites. This leads to repetitive DNA or duplicates incorrectly collapsing into a single locus resulting in misassembly (see Fig.1.5). TGS technologies, developed by PacBio and Oxford Nanopore Technologies in 2011 and 2014 respectively, have become more popular given their speed and ability to overcome these long-standing issues inherent to second generation methods (Eid et al., 2009; Rothberg et al., 2011). We will discuss these two methods in more depth below and explore the developments that have been made to address sequencing errors and improve technologies. In chapter 4 we expand

on an approach by Du et al., 2020 and attempt to resolve short-read errors in the paddlefish assembly (Cheng et al., 2020).

Duplicated sequences are notorious for introducing errors in assemblies and confounding results. The paddlefish genome has experienced a WGD which it shares with the only other extant lineage of the actinopterygi, the sturgeon (Redmond et al., 2023). This doubling of the genome results in many more duplicated regions, some containing ohnologs (gene duplicates originating from WGD), important for phylogenetic analysis. Due to a high sequence identity between duplicates, they can be mistakenly aligned to the same region , *a.k.a* collapsed, during the process of assembling individual short read sequences back to the reference genome (Fig.1.5). We show that this artefact impacted the Cheng et al., 2020 paddlefish genome and how *reduplicating* these reads improved the assembly, increased the gene count and facilitated the detection of more orthologs for enhanced phylogenetic analysis.

1.2.2 Genome sequencing and assembly errors

Since its inception, the velocity to which whole genome sequencing has revolutionised all aspects of the biological sciences - from research to clinical practices - has been prodigious (Schatz et al., 2010; Rothberg et al., 2011). Despite the slow and inefficient nature of early methods of DNA sequencing, pioneered by Sanger, they led to the first draft of the human genome sequence and marked the beginning of the genomics era (Sanger and Coulson, 1975; Lander et al., 2001). This slow process became dramatically degenerate following the establishment of NGS in the early 2000s which allowed for thousands of DNA molecules to be sequenced simultaneously (Margulies et al., 2005; Rothberg et al., 2011; Eid et al., 2009). A few years after NGS came TGS, or long-read sequencing technologies and over the twenty three years since then the number of whole genome sequences

available has exponentially grown (Mardis, 2008; Schadt et al., 2010; Quail et al., 2012).

NGS approaches were pioneered in the early 2000s but rumblings of the framework were initially brought to light in the late 90's through a sequencing technique called pyrosequencing (Ronaghi et al., 1996). Also known as sequencing-by-synthesis, this early technology is based on the measurement of luminescence generated as a result of pyrophosphate synthesis during sequencing and is generally regarded as an early-stage high-throughput sequencing technology. The key to NGS was the parallelisation of a large number of reactions which accelerated the pace at which it could sequence DNA molecules (Schatz et al., 2010).

By 2005, Jonathon Rothberg and colleagues pioneered a pyrosequencing technology in an automated system, the 454 system, that when acquired by Illumina became the worlds leading NGS platform, of which it is still to this day (Margulies et al., 2005; Shendure et al., 2005). All NGS approaches rely on a 'library' preparation of the extracted DNA. They begin with DNA fragmentation and fragment size selection followed by ligation of adapters to the ends of each fragment which can then be loaded onto the flow cell. These fragments are typically around 100-400 base pairs and need to be mapped back to the reference genome. This fragmentation and mapping step is the Achilles's heel of NGS technologies. Reconstructing complex or repetitive regions of the genome with these short, fragmented reads has the propensity to cause errors and thus, misassembly. While long-read approaches offer relief from this issue, most of the complete genomes of animals available on databases like NCBI and ensembl are short-read assemblies.

TGS was established by PacBio in 2010 with the introduction of the zero-mode waveguide (ZMW) technology (Rothberg et al., 2011; Eid et al., 2009). The technique uses "nanoholes" or *nanopores* that contain single DNA polymerases with a phosphate-labeled nucleotide attached to its nucleotide triphosphate (dNTP)

substrates. When a single DNA molecule is introduced into a ZMW, it binds to a DNA polymerase immobilized at the bottom of its flow cell. Labeled dNTPs are incorporated as the DNA polymerase synthesizes the complementary strand and they emit a specific fluorescent signal on meeting their complementary nucleotide. This signal is detected and recorded in real-time. Single-molecule sequencing was adapted by Oxford Nanopore Technologies who developed systems such as GridION and MinION (Ashton et al., 2015; Lu et al., 2016). These technologies use changes in electrical conductivity that occur when DNA strands pass through biological *nanopores* to identify the nucleotide sequence. Unlike NGS, long reads rather than short reads are generated and so complex and repetitive regions are easier to decipher (Schadt et al., 2010). However, third-generation technologies are not perfect and can often be more expensive per base compared to short-read technologies as well as requiring more DNA input, which can limit their uses when working with small samples (Amarasinghe et al., 2020). Another issue with SMS is its high error-rates during base calling, particularly when sequencing repetitive sequences of the same nucleotide. For example, the initial average error rate and fragment length of PacBio long reads were approximately 15% and 1.5 kb, respectively (Quail et al., 2012). A comparisom of TGS and NGS technologies can be seen in Fig1.4.

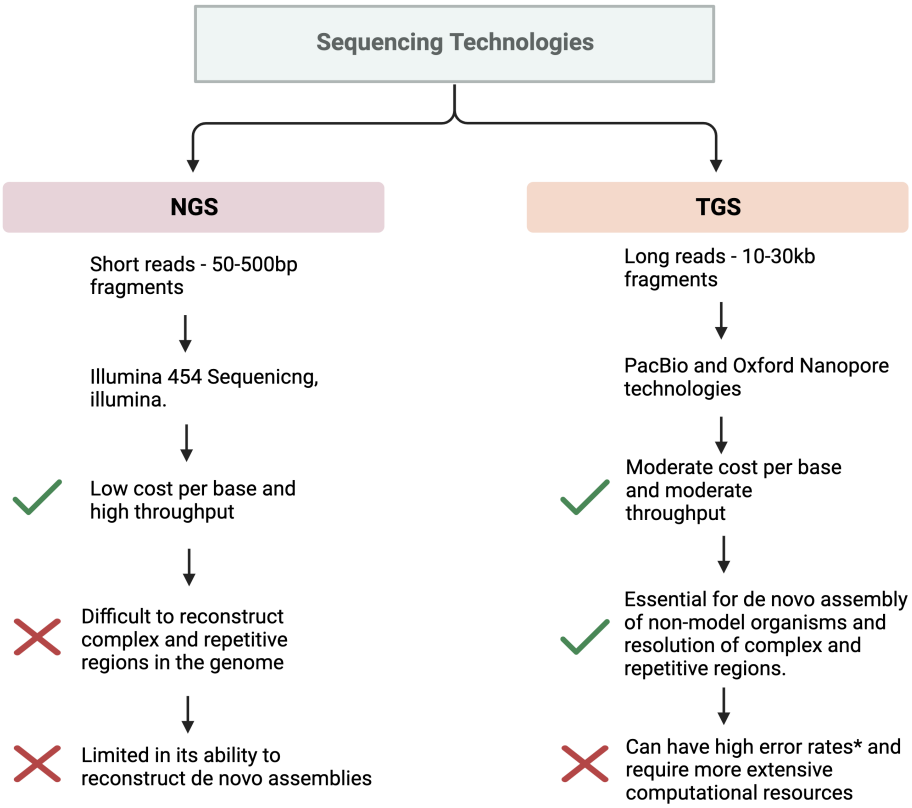

Figure 1.4 | Schematic comparing NGS and TGS technologies *Note: Novel techniques such as Circular Consensus Sequencing (CCS) and HiFi can produce exceptionally precise long reads, achieving a sequence accuracy of up to 99.8%. Nevertheless, it's essential to acknowledge that these methods are not widely adopted, and thus, the original point remains applicable in the majority of situations (Wenger et al., 2019; Nurk et al., 2020)

.

Despite these issues, SMS has revolutionised genomics since its genesis and work is continually being done to overcome many of these hurdles including novel methods for error correction like optimised circular consensus sequencing (CCS) or HiFi methods which have generated long reads of up to 13.5kb in length with sequence accuracy of up to 99.8% (Wenger et al., 2019; Cheng et al., 2021; Nurk et al., 2020). Short reads, however, remain ubiquitous in research due to their cost effectiveness, availability and familiarity but there are continuous strides to improve error prone short-read assemblies by upgrading short-read to long-read genomes as

well as ameliorating the second-generation assemblies with long-read sequencing information (Coombe et al., 2021; Cechova, 2020). Additionally, complementary methods, such as optical mapping or chromosome conformation capture techniques, may be used to aid in the characterization of repetitive regions and structural variations but in large error rates in assemblies of more complex genomes prevail in online databases (Schwartz et al., 1993; Dekker et al., 2002).

1.2.3 Genome assembly tools and repetitive DNA

It is unclear to what extent published genome assemblies have been impacted by the inherent error-prone nature of NGS and TGS. Second generation sequencing, or NGS, has greatly enhanced our ability to sequence DNA at high throughput and for a low cost. However, the method has intrinsic errors resulting in incomplete assemblies which can have major impacts on research results and impedes the transformative capabilities genome sequencing could have in clinical settings (Mardis, 2008; Schatz et al., 2010). While TGS offers relief from some of these issues it also has its flaws and so a perfect genome sequencing tool is yet to be realised.

Short and long read approaches use different assembly frameworks which have alternative strategies for how they approach repetitive regions. The two predominant approaches are De Bruijn Graph-Based Assemblers and Overlap-Layout-Consensus (OLC) Assemblers (Myers et al., 2000; De Bruijn, 1946; Simpson et al., 2009). The former uses k-mers which are short, contiguous sub-sequences of a fixed length, k, that are extracted from the short-reads produced by Illumina (Simpson et al., 2009; Zerbino and Birney, 2008). The De Bruijn method constructs a graph to represent overlaps between k-mers. The graph is parsed and depending on implementation, either scaffolds or contigs are generated. For repetitive regions,

this assembly method is flawed and inadequate as the repeat would have to be shorter than a k-mer to be properly resolved. While some assemblers using the De Bruijn method have been designed to deal with this issue by using multiple, overlapping k-mers, none are capable of overcoming some of the larger repeats found in vertebrate genomes(Mahadik et al., 2019). The other method, OLC, was used in the first assembly of the Drosophila genome in 2000 (Myers et al., 2000; Adams et al., 2000; Batzoglou et al., 2002; Mullikin and Ning, 2003; Miller et al., 2010). This method relies on finding overlaps between longer sequences to construct contigs and scaffolds. This method is more associated with third-generation technologies and is better at resolving repeat regions as it is not limited by k-mer size (Li et al., 2012; Miller et al., 2010). The overlap step works like a multiple sequence alignment and compares each read to all other reads. This step is computationally expensive if there are a lot of short-reads, like those produce by NGS, but for long-read sequencers the process is less taxing. There has been major work done on developing hybrid assemblers such as MaSuRCA, which can take Illumina short-read contigs or scaffolds and use the long reads from TGS technologies to disambiguate the regions of the assembly graph that have not been resolved (Zimin et al., 2013). The method combines the de Bruijin and OLC assembly approaches and has shown to yield high quality assemblies. MaSuRCA and other combinational approaches show promise for improved assemblies and better resolution of repeat regions in the future (Zimin et al., 2013; Coombe et al., 2021; McCoy et al., 2014).

Eukaryotic genomes can be as much as 80% repetitive, for example, the human genome is estimated to be 66-69% repetitive (Mkrtchyan et al., 2010; Biscotti et al., 2015a; Tóth et al., 2000). Ensuring that sequencing softwares and assembly tools can accurately map these repeat regions is therefore, paramount for there use in research. Many of the published "complete" genomes, including the human

genome, actually contain gaps which are more than often, long stretches of repeats that were uninterpretable by the sequencing and assembly software (Salzberg and Yorke, 2005; Schatz et al., 2010). Repetitive DNA occurs in all domains of life and comes in different forms that are either interspersed throughout the genome or occur as tandem repeats (Biscotti et al., 2015a; Mkrtchyan et al., 2010). The former includes transposable elements like LINEs (Long Interspersed Nuclear Elements), SINEs (Short Interspersed Nuclear Elements) and retrotransposons while tandem repeats are sequences that are contiguous in the genome and include satellite DNA or simple sequence repeats (Konkel et al., 2010; Kramerov and Vassetzky, 2011; Biscotti et al., 2015b). Other important elements that are ubiquitous in eukaryotic genomes and prone to collapse during assembly are segmental duplications and CNV's which play major roles in evolutionary studies and have been associated with diseases in humans (Rice and McLysaght, 2017; Zhang et al., 2009; Bailey and Eichler, 2006). Accurate detection and characterisation of repeat regions and the genes within repetitive regions is therefore important in clinical settings and more broadly in evolutionary studies and species comparisons.

1.2.4 Collapsed repeat regions in genome assemblies

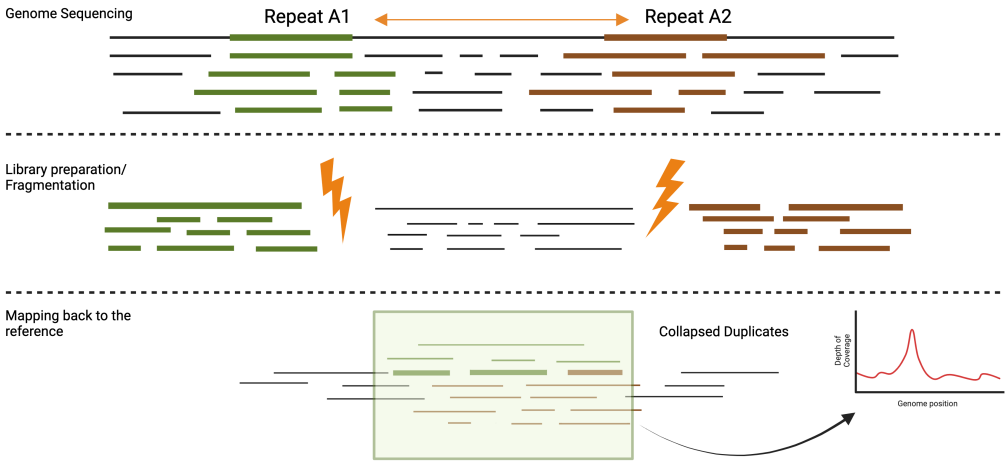

Figure 1.5 | Schematic of misassembled duplicates in a genome assembly An example of repeat genes, A1 and A2, separated by a unique sequence. Fragmentation of DNA for library preparation leads to errors when short-reads like this are mapped back to the reference. The duplicates are collapsed into a single read. A diagnostic for this can bee seen in a coverage plot as a peak of double the expected coverage.

We discussed how NGS works in the previous Section 1.2.2, detailing how library preparation involves fragmenting the genome and then assembling back to the reference and how this step can introduce errors. This process is depicted in Fig.1.5. In the figure you can see duplicates, A1 and A2, that are separated by a unique sequence (However, note that reads do not even need to be linked for collapsing to manifest). If these duplicates are long enough, during primary preparation there will be multiple fragments or short reads that span this region. When assembling the reads back to this region in the reference, it will confound the assembler leading to just one copy of the repeat being recognised and the unique region in the middle being mis-assembled. Given the repetitiveness of most eukaryotic genomes, there is a high prevalence of the situation described in Fig.1.5 occurring in sequencing attempts. While tools have been developed to overcome

these issues, there is no automated way of fixing this problem which becomes even more complex as genomes get larger and ploidy number increases.

Most assembly softwares aim to generate a haploid assembly and collapse allelic difference between chromosomes (Glusman et al., 2014; Chin et al., 2016). Alternatively, they use a reference genome to partition the reads by haplotype (reference-guided assembly)(Edge et al., 2017; Martin et al., 2016). They take a diploid or polypoid genome and because each chromosome is slightly diverged from the other, they recognise, for example in a diploid genome, two haplotypes and choose only one version of each heterozygous region for the final assembly thus creating a pseudo-haploid reference. This means that all diploid or polyploid assemblies are actually reduced representations of the actual complexity of that genome. In cases where there are tandem repeats (a sequence of two or more DNA bases that is repeated numerous times) it becomes difficult to distinguish a repeat from a true polypmorphism between haplotypes and so those repeats are collapsed and thus misassembled (Rhee et al., 2016; Cilibrasi et al., 2007). In certain instances, elevated levels of heterozygosity between alleles may mimic paralogy, which refers to gene duplicates created by a duplication event within a genome. This resemblance has the potential to confound algorithms, leading to the erroneous generation of false duplicates or gaps in genome assemblies. (Barrière et al., 2009). This becomes more complicated with more complex ploidy states like paleotetraploidy (functional diploidy) seen in salmon, paddlefish and sterlet genomes. In these scenarios, assemblers face the challenge of not only generating potential false duplicates, as previously described, but more critically, they may struggle to discern true ohnologs. The accurate identification of ohnologs is paramount for precise phylogenetic analysis (Romanenko et al., 2015; Robertson et al., 2017; Redmond et al., 2023). Rectifying these issues is of the utmost importance as much of the framework of comparative evolutionary studies starts with

identification of gene duplications and orthology analysis.

The emerging consensus suggests that addressing many of the challenges associated with current genome assemblies requires the generation of more high-coverage, long-read data from technologies such as PacBio and Nanopore. Complementing these with short-read data from NGS can contribute to effective error correction, enhancing the overall quality and accuracy of genomic assemblies (Zimin et al., 2013; McCoy et al., 2014; Coombe et al., 2021). Long-read assemblies are becoming more common but the the universality and accessibility of Illumina biases researchers choices and often, short-read data suffices for most research purposes. It is clear that a gold-standard framework for generating complete genome assemblies is at the cusp of being developed but economic issues and accessibility mean that collapsed duplicates and misassembly of repetitive regions prevails.

1.2.5 Fixing collapsed duplicates in genome assemblies

We are still a long way from having a "perfect" and complete genome assemblies for all of life. Sequencing assembly errors are omnipresent in most draft and finished genomes and this can have major impacts on scientific findings (Salzberg and Yorke, 2005; Mardis, 2008). Despite this, we are approaching a new frontier in genomics as genome sequencing methods become more accurate, more efficient and more affordable. The question arises: should we embark on a comprehensive re-sequencing effort or focus on refining existing data? The decision hinges on weighing the benefits of obtaining fresh, superior data against the value and feasibility of optimizing and correcting the wealth of genomic information already at our disposal.

Ensuring that the user is aware of the quality of the data that they are using is of the utmost importance. There are several standard tools and metrics that

we can use for quality assessment including Benchmarking Universal Single Copy Orthologs, BUSCO (Manni et al., 2021), N50 value and Recognition of Errors in Assemblies using Paired Reads, REAPR (Hunt et al., 2013). These metrics allow for evaluation of various elements of an assembly. N50, for example, is a length metric and provides a standard measure of the contiguity of an assembly, while BUSCO scores are a completeness score based off a set of genes that are universally distributed as orthologs across the specific clade you are interested in (Jauhal and Newcomb, 2021; Thrash et al., 2020). While all these metrics are broadly utilised, they really only capture limited aspects of an assembly and can miss subtle errors. Allowing these subtleties to persist can lead to incorrect conclusions and results in scientific literature.

As discussed in depth above, a major challenge for sequencing and assembling diploid and polyploid non-model organisms is inaccurate resolution of duplicates, repeats and haplotypes. In recent years, tools have been developed to improve assembly contiguity by unearthing collapsed duplicates and repeat regions from short read assemblies. The predominant method appears to be a hybrid approach in which you use both long and short read data during scaffolding (Coombe et al., 2021; Cechova, 2020; Du et al., 2020). First you map the long reads to the short read contigs. This alignment is then made into a graph which can be traversed to produce scaffolds in which gap sizes are estimated from the linking information (Kronenberg et al., 2021; Zimin and Salzberg, 2022). Most early tools of this kind only worked on small genomes but more recent developments, like Scaffolding Assemblies with Multiple Big Alignments or SAMBA can be used on larger genomes (Zimin and Salzberg, 2022). SAMBA is designed to work on assemblies with a 10-30X genome coverage and a set of existing contigs. As described, it uses long reads to re-scaffold contigs from an existing genome assembly in an effort to update misassembled short-read genomes.

In this thesis, we attempt to *re-duplicate* collapsed duplicates in the paddlefish genome following an adapted protocol by Du et al., 2020. The paddlefish is a complex case when it comes to assembling and annotating because the lineage experienced a WGD approximately 254.7-241.8 Myr that it shares with the sterlet sturgeon (Redmond et al., 2023). Following the WGD, the paddlefish transitioned from its ancestral tetraploid state to a functional diploid in an asynchronous process of rediploidization. Despite the long time that has elapsed since the event, both the sterlet and paddlefish genomes still have regions undergoing tetrasomic inheritance, that have not rediploidized (Redmond et al., 2023). This complex ploidy makes assembly and annotation difficult and can lead to collapsed repeat regions. This is evident from read-depth analysis where large parts of the genome have double the expected read coverage, a hallmark of collapsed duplicates (Fig.1.5). Like SAMBA (Zimin and Salzberg, 2022), we take a hybrid approach, using both Illumina short-reads and and PacBio long reads to improve the paddlefish assembly.

Short read assemblies will continue to be used despite their shortcomings and for that reason, utilising long reads in a hybrid approach like this will aid in fixing misassembled short-read genomes and update rather than totally discard the data. While we are approaching an era in which high quality TGS methods like HiFi (Cheng et al., 2021; Nurk et al., 2020) and CCS (Wenger et al., 2019)are becoming more accessible, it is important that the users are aware of the potential errors in most of the larger genome assemblies currently available on genome databases (NCBI and ensembl). Continued disregard for these errors will lead to a proliferation of incorrect conclusions throughout the literature.

1.3 Whole Genome Duplication and rediploidization

Whole genome duplication(WGD) has shaped the genomes of several major eukaryotic lineages. In vertebrates, it is widely accepted that there were two ancestral WGD (known as 1R and 2R) that occurred early in their evolution (Sidow, 1996; Dehal and Boore, 2005). A third WGD, shared by all teleosts, is thought to have occurred after fishes diverged from land vertebrates (see Fig.1.6). (Jaillon et al., 2004). In plants, polyploidy is widespread in monocots and core eudicots, with ancestral and recent WGD being an evolutionary hallmark of many lineages (Soltis et al., 2009; Wendel, 2000). The ancestral WGD that affected the ancestor of the baker's yeast, *Saccharomyces cerevisiae*, has been well researched and contributed heavily to our knowledge of WGD (Wolfe and Shields, 1997; Conant and Wolfe, 2008). These examples underscore the recurring nature of WGD throughout the evolutionary history of life on Earth. However, the fundamental questions persist: What are the mechanisms driving WGD events? Why do they happen, and what are the functional and evolutionary repercussions that ensue from these events?

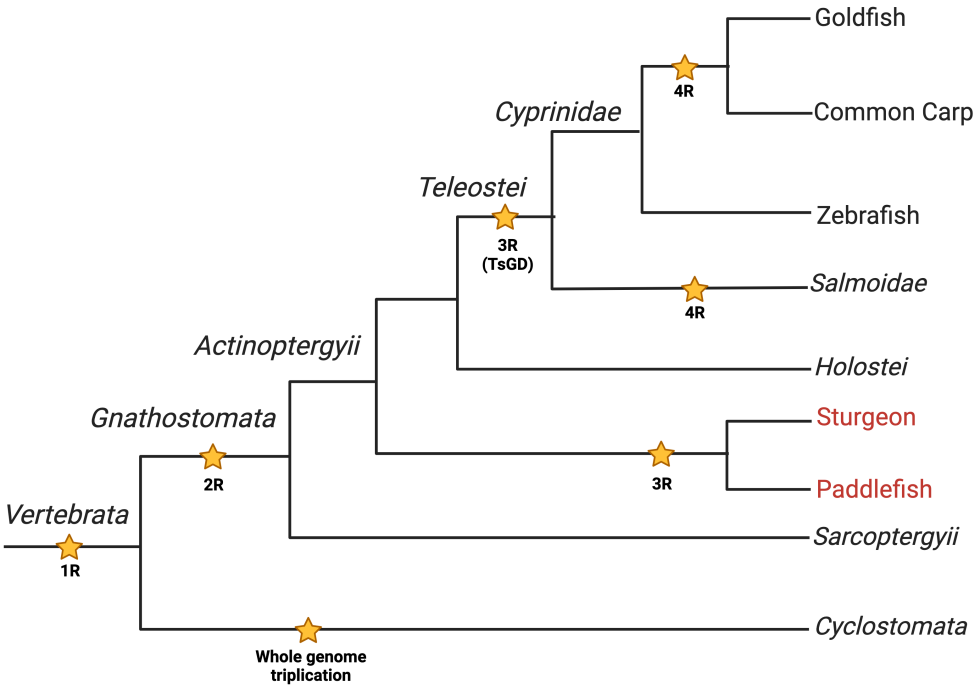

Figure 1.6 | Cladogram showing several important WGDs that have taken place in vertebrate evolution WGD events are represented by a yellow star. 1R-4R represent tetraploidy events while the whole genome triplication in *Cyclostomata* is a hexaploid event. For their importance in this thesis, sturgeon and paddlefish are highlighted in red. Branch lengths are not to scale.

Importantly, while it may seem from Fig.1.6, as well as the other examples listed above, it's important to recognize that WGDs are, in fact, infrequent. In a broader context, polyploidy is often considered an evolutionary "dead end" due to its adverse impacts on fertility and genomic stability (Comai, 2005). The polyploidy events that established themselves in the evolutionary history of the lineages listed above, more than likely had an adaptive advantage over their diploid progenitors (Peer et al., 2017; Conant and Wolfe, 2008). Early work on WGD, pioneered by Susumo Ohno, focused on a link between these doubling events and increased species diversity (Ohno, 1970). Given that gene duplication can provide evolutionary potentials for generating novel functions, an entire doubling of chromosomes provides a framework for immense innovation (Conant and Wolfe, 2008;

Peer et al., 2017; Cañestro et al., 2013). While it is accepted that compared with diploid, polyploid genomes have an increased mutational robustness, there is still a debate over whether or not WGD offers an increased environmental robustness and increased propensity for adaption (Dubcovsky and Dvorak, 2007; Cañestro et al., 2013). In vertebrates, it is believed that the early tetraploidisations, 1R and 2R, played a significant role in the evolution of vertebrate complexity and an ancient WGD occurring before the divergence of *Arabidopsis* and other dicots, potentially contributed to the radiation of eudicots (Otto, 2007; Dehal and Boore, 2005; Vanneste, Baele, et al., 2014). While it is hard to pin WGD as the sole reason for these advances, evidence suggests that it has played a big part.

In plants, there were many WGD occurring close to the Cretaceous–Paleogene (K–Pg) boundary, in which there was a meteor impact near Chicxulub (Mexico), appear to have aided in the survival of many lineages (Renne et al., 2015). While this catastrophe led to the extinction of 70% of all plant and animal life, including all non-avian dinosaurs, evidence in plants indicated that WGD provided protection from this major environmental instability (Vanneste, Maere, et al., 2014; Vanneste, Baele, et al., 2014). The paddlefish and sturgeon's shared WGD is also suspected to have supported survival during the Permian-Triassic (P-Tr) boundary mass extinction event (Redmond et al., 2023). There are also strong lines of evidence linking the two rounds of WGDs that occurred during early vertebrate evolution (500–550 Mya) and the immense species diversification associated with the Cambrian explosion (Wille et al., 2008). This is corroborated by studies showing that the acquisition of vertebrate defining features like neural crest cells and a complex tripartite brain occurred post 2R and probably facilitated the transition from filter-feeding, non-vertebrate chordates to complex vertebrate predators (Northcutt and Gans, 1983; Gans and Northcutt, 1983; Holland and Garcia-Fernàndez, 1996). Gene family diversification post-WGD has also been

well studied in many organisms with the most renowned gene family diversification happening to the Hox-gene cluster which went from a single cluster to four clusters during 1R and 2R events in early vertebrate evolution (Wagner et al., 2003; Málaga-Trillo and Meyer, 2001; Crow et al., 2012a; Holland and Garcia-Fernàndez, 1996). Hox genes are involved in morphogenesis and responsible for specifying regions of the body plan. The quadrupling of this gene-family contributed to morphological innovations in different vertebrate lineages including jaw formation, origin of limbs and also secondary loss of limbs in snakes (Holland and Garcia-Fernàndez, 1996; Wagner et al., 2003). However, like in all of biology, there are no rules and there are many examples of where WGD has not lead to diversification and complexity. Horseshoe crabs have experienced three rounds of WGD to no apparent enhancements or bursts of diversity, as have the sturgeon and paddlefish lineages (Nong et al., 2021; Cheng et al., 2020; Du et al., 2020). Consideration of extinct taxa and a rigorous inclusion of diverse fossil records have shown a tenuous association between a WGD and vertebrate body plan evolution (Donoghue and Purnell, 2005). Despite this, there are persistent streams of evidence that suggest WGDs have played an important role in certain developmental innovations and diversification of lineages. How do they happen, how are they maintained and if, in general they are "dead-ends", then what do the early stages of a prosperous duplicated genome look like?

In general, there are two types of polyploids: autopolylpoids and allopolyploids. The former are those who've experienced a self-doubling of their chromosomes, while allopolyploids are a product of a merger of chromosome-sets from different origins (Jackson, 1982; Ramsey and Schemske, 1998). On doubling of the chromosomes, or hybridisation of two sets of distinct chromosomes in the case of allopolyploids, one of the major challenges the genome now faces is the correct segregation of the chromosomes during meiosis. In the case of autopolyploids, the

homologs will begin randomly pairing with now, multiple related chromosomes creating complex structures called multivalents. These structures are unstable and recombination becomes a lot more difficult as the related chromosomes attempt to enter metaphase I simultaneously (Furlong and Holland, 2002; Ramsey and Schemske, 1998; Otto, 2007). For Allopolyploids, the path back to bivalent pairing is a lot less complicated as each chromosome can pair with the homolog from it's respective parental genome. In autopolyploids, the unstable pairing and homologous recombination of homologs can persist for millions of years in some parts of the genome, which as a result means those genes are not truly duplicated but rather, in the case of tetraploidy, a single locus with four alleles (Eric Schranz et al., 2012; Lien et al., 2016; Berthelot et al., 2014; Redmond et al., 2023).

1.3.1 Rediploidization

The transition of a polyploid genome back to a stable diploid state with bivalent pairing is known as redploidization (Wolfe, 2001; Furlong and Holland, 2002; Hokamp et al., 2003b; Lien et al., 2016; Robertson et al., 2017). We have mentioned how this process is different depending on the mode of polyploidy and for self-doubling, autopolyploids, multivalent pairing can persist in genomes for extended periods of time (Robertson et al., 2017; Redmond et al., 2023; Parey et al., 2022; Gundappa et al., 2021). Suppression of recombination in a tetraploid is suspected to be achieved through chromosomal rearrangements and successive mutations that allow loci to rediploidize into bivalent pairs (Allendorf et al., 2015; Lien et al., 2016). The loci can only be considered duplicated when this has occurred. These WGD-derived duplicate genes are known as ohnologs. The asynchronous resolution of ohnologs makes comparative and functional analysis of species with ancestral autopolyploids difficult and until recently very few studies considered

the delayed nature of rediploidization following autopolyploidisation (Berthelot et al., 2014; Lien et al., 2016; Robertson et al., 2017; Du et al., 2020; Gundappa et al., 2021; Redmond et al., 2023). Substantial evidence has been garnered from studies of the ancestral salmonid WGD which displays a rediploidization process that has been temporally protracted for over tens of millions of years (Lien et al., 2016; Robertson et al., 2017).

A delayed rediploidization can have major implications on the evolutionary trajectory of a lineage (Eric Schranz et al., 2012; Lien et al., 2016; Macqueen and Johnston, 2014). Duplicate genes generated following WGD have been well studied given their potential for functional divergence and innovation (Conant and Wolfe, 2008). Unlike small-scale duplications (SSD), doubling the entire genome allows for a balanced divergence of whole networks of genes allowing diversification of signalling pathways and ultimately promoting molecular and phenotypic enhancements (Ohno, 1970; Conant and Wolfe, 2008; Otto, 2007; Peer et al., 2017). For allopolyploidy events, we can expect that divergence of duplicate genes begins almost instantaneously and any apparent effects will follow promptly after the WGD. Expecting the same timescale for adaptions post-autopolyploidisation will not be constructive. As mentioned, most studies did not consider a delay and refuted a link between WGD and species radiations because of interludes between the two events (Donoghue and Purnell, 2005; Santini et al., 2009). For example, the Teleost-specific Genome Duplication (TsGD) is purported to have occurred 320-350 million years ago while the major teleost species radiation happened >200 mya after the doubling and because of the delay, studies ruled out a link between the major events (Santini et al., 2009). Now with better understanding of the rediploidsation process in autopolylpoids, work is being done to untangle these delayed species radiations and to better understand the effects of temporally protracted duplicate resolution and its effect on speciation events and species

diversification.

An early model that acknowledged the extensive lag time between some WGD and species diversification described how lineages that share an ancestral doubling event, millions of years before they diversified, may have acquired genetic material for novel traits at the WGD that became activated long-after when interplayed with lineage-specific ecological factors - like migration events and changing environmental conditions (Eric Schranz et al., 2012). This idea of a "lag-time" motivated researchers to delve deeper into the root of the delay from a molecular standpoint. As described above, autopolyploidisation involves a spontaneous doubling of the same genome within a cell. The four identical chromosome-sets will pair randomly with one another during meiosis and this perpetuates recombination between four alleles at a loci, preventing divergence of ohnologs (Furlong and Holland, 2002). The process of halting recombination at tetraploid loci and allowing bivalent paring to resume (a.k.a rediploidization), may be delayed for long periods after autpolyploidisation events which, when choosing the most parsimonious answer, may link back to the apparent delay in species diversifications post WGD (Lien et al., 2016; Macqueen and Johnston, 2014; Robertson et al., 2017). In salmonids, who experienced an ancestral WGD post-TsGD (Fig.1.6), the first studies probing a delayed rediploidization showed it to be mediated by large genomic rearrangements and bursts of transposon-mediated repeat expansions (Lien et al., 2016). While other work refuted a connection between genome reorganisation and rediploidization (Berthelot et al., 2014), it was clear from collinear block analysis in the Atlantic salmon, that there were defined duplicate regions with signatures of ancestral rearrangement events (Lien et al., 2016). They proposed that following WGD by autopolyploidisation, transposable element expansions followed by genomic rearrangements were driving forces in halting tetraploid inheritance and directing the genome toward a diploid state (Lien et al., 2016; Guillén and Ruiz,

2012). Similarly, in sturgeon and paddlefish, there is evidence of collinear blocks with probable origins in rearrangement events (Redmond et al., 2023). In chapter 5, we will investigate these collinear, or micro-syntenic, blocks and the role of rearrangement in rediploidization in these lineages. As explained above, a delay in rediploidization will delay ohnolog resolution and thus their functional divergence. This can have major effects on the evolutionary trajectory of the lineages, delaying species diversification and generating lineage-specific WGD histories.

1.3.2 Lineage-specific Ohnolog Resolution (LORe) model

Robertson et al., 2017, proposed the 'Lineage-specific Ohnologue Resolution' or 'LORe' model to describe the role of an asynchronous rediploidization process in the evolution of sister lineages that share an ancestral WGD. The framework focuses on a situation in which rediploidization happens in parallel with speciation i.e. speciation happens before all genes have completely resolved back to bivalent pairing. This leads to a unique situation in which some ohnologs have resolved in the ancestor of the sister lineages while other ohnologs rediploidized independently in each. Under LORe, the former ohnologs are described by 'Ancestral Ohnolog Resolution', or 'AORe' framework (Fig.1.7).

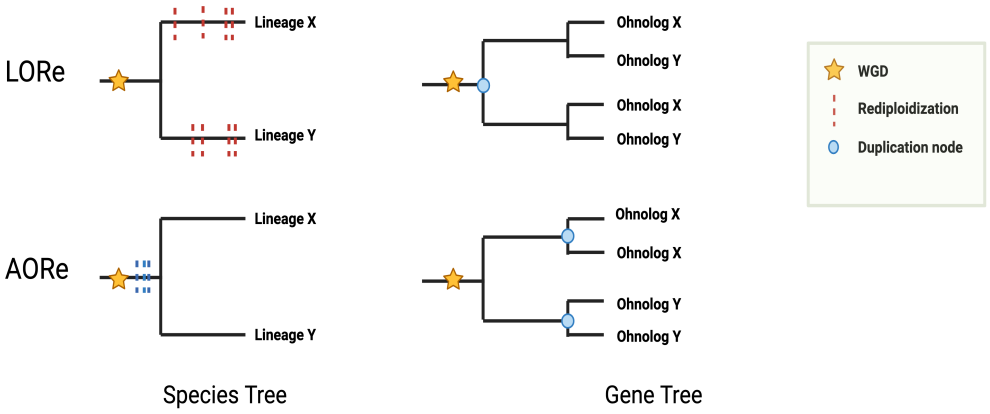

Figure 1.7 | The LORe model of post-WGD evolution following a delayed rediploidization The schematic illustrates the implications of the LORe model on ohnolog divergence in contrast to AORe. Blue and red dashed lines indicate rediploidisation events. On top, the LORe framework in which rediploidization takes place in a lineage-specific manner results in gene tree topologies (top-right) where ohnologs have a 2:2 orthology relationship. In contrast, AORe trees (bottom), where rediploidization happens before speciation and so ohnolog divergence happens in the ancestor. This leads to gene trees with 1:1 orthology assignments

These genes began functional divergence in the ancestor, succumbing to ancestral selective pressures. This increases the likelihood that the ancestral function of the gene will be conserved in the sister lineages. In contrast, genes that rediploidize independently in each lineage follow the "LORe" framework (Fig.1.7). Under this model, the functional divergence of ohnologs is independent in respective lineages and species-specific pressures may lead to stark differences in functions of these ohnologs. Another implication of this model, is that phylogenetically, lineage-specific ohnologs lack 1:1 orthology with the ohnolog pair from the sister lineage, setting a 2:2 homology relationship between the LORe genes. If LORe is not accounted for, this can confuse phylogenetic analysis and LORe ohnologs can look like paralogs or SSD. Testing for this in salmonids showed that 25% of the genome evolved under LORe, with evidence that LORe ohnologs developed lineage-specific functions and physiological adaptations that potentially facilitated

salmonid species radiation (Robertson et al., 2017).

Since the delineation of the model, studies using it have made some major revelations on post-WGD processes (Redmond et al., 2023; Parey et al., 2022; Du et al., 2020). By using the LORe model on sturgeon and paddlefish genomes it was found that what was previously believed to be two independent WGD events, was in fact, a shared event that was being masked by a delayed rediploidization process (Symonová et al., 2017; Crow et al., 2012b; Redmond et al., 2023). The incidence of gene trees with independent duplications histories (LORe gene trees) was higher than those with shared duplication histories and so a separate WGD seemed like the most parsimonious conclusion. As previously noted, the WGD, believed to have taken place more than 200 million years ago, could have played a pivotal role in enhancing survival amidst the environmental upheaval of the Permian-Triassic (P-TR) mass extinction. There is also some possibility of a connection between the protracted process of rediploidization and the survival of species during the Triassic-Jurassic mass extinction (Redmond et al., 2023). Using LORe, teleosts have been found to have regions in their genomes that have maintained tetraploidy for more than 60 million years after the TsGD, a time period interspersed with several speciation events of major teleost clades (Parey et al., 2022). Appreciation of LORe has significant effects on the outcome and understanding of the functional properties of genes and the evolutionary history of many lineages. The lag-time between a WGD and major events in the vertebrate lineage can be illustrated more clearly through the lens of this model. Rather than "bursts" of radiation and rapid functional divergence post WGD, the model predicts a gradient of effects, occurring over tens of millions of years in some cases. In salmonids, acipenseriformes and some teleosts it has been shown how LORe can explain delayed lineage-specific diversification episodes under prevailing ecological pressures (Redmond et al., 2023; Parey et al., 2022; Du et al., 2020; Lien et al.,

2016; Robertson et al., 2017). So far, the framework has been predominantly used in fish but there are hopes it will be expanded for probing the effects of WGD and rediploidization in other lineages such as plants, famed for their propensity for polyyploidy.

1.3.3 Mechanisms of rediploidization

Over the past decade we have gained a much better understanding of rediploidization following WGD. We discussed in detail how following auto-tetraploidisation, there can often be a delay in returning to stable bivalent pairing, in some cases taking tens of millions of years for the genome to fully rediploidize (Robertson et al., 2017; Redmond et al., 2023; Parey et al., 2022). While we are continuing to understand the functional effects of this process on the genome, we have not fully elucidated how the genome stops tetrasomic inheritance and begins reversion to disomic inheritance. The predominant theory is that it proceeds via temporally isolated chromosomal rearrangement events, which have the potential to suppress recombination and allow for resolution of ohnologs in that segment (Lien et al., 2016; Allendorf et al., 2015). Evidence for this has been shown in the salmonid lineage, which experienced an ancestral WGD, post-TsGD. A study recognised large collinear blocks with >87% sequence similarity in the Atlantic salmon genome, assumed to have originated via asynchronous rearrangements that occurred in the genome after the WGD (Lien et al., 2016) (see Fig.1.8 for a potential framework). Such blocks have also been found in sturgeon and paddlefish genomes, which, as discussed, experienced a shared WGD separate from the TsGD (Redmond et al., 2023). Possible mechanisms underlying the initiation of genomic rearrangements are not well understood. It has been suggested that transposable-element (TE) expansions are drivers of large genomic reorganisations. TE expansions, in turn,

are known to increase during times of genomic stress, for example post WGD (Slotkin and Martienssen, 2007). The interplay between WGD events, genomic stress, TE expansions and genome rearrangements have been postulated as possible pathways to a rediploidized genome (Lien et al., 2016). Loose evidence for this framework has been shown in salmonids but there is still more work to be done to unravel the universal mechanisms underlying termination of multivalent pairing and rediploidization (Lien et al., 2016).

The rearrangement mechanism described above, draws parallels to the suppression of recombination during the evolution of the mammalian sex chromosomes (Lahn and Page, 1999). The human X and Y chromosomes are thought to have evolved from an ordinary pair of autosomes whose evolutionary history was punctuated by at least four major rearrangement events. What are suspected to have been inversions, happened in a step-wise fashion along the length of the ancestral Y chromosome, each suppressing recombination one stratum at a time, without disturbing gene order on the X. This stepwise suppression of recombination between the pair of chromosomes may be analogous to the mechanisms which promote asynchronous rediploidization of genomes post WGD Lien et al., 2016; Redmond et al., 2023.

In the Acipenseriformes, analysis found that when visualising ohnolog pairs that had rediploidized before the speciation event and those that resolved after the speciation event, they were not randomly distributed in the genome but rather were found in distinct syntenic-blocks along uninterrupted sections of chromosomes (Redmond et al., 2023). For example, the "Pre-speciation" ohnolog blocks were conserved across 6 of the largest paddlefish and sturgeon chromosomes. This is evidence that these were not SSD but rather shared genome rearrangements that most likely happened in the ancestor of the sister lineages, after the WGD event. This follows that in a case where speciation follows a WGD and delayed

rediploidization, rearrangements that interrupt tetrasomic inheritance should be seen as large blocks of contiguous genes that share a common rediploidization history; be that pre-speciation or post-speciation ohnolog resolution (LORe or AORe genes). Like the strata in the mammalian sex chromosomes, the paddlefish and sturgeon chromosomes may also be stratified by divergence time of ohnologs within the synteny blocks (Redmond et al., 2023; Lien et al., 2016).

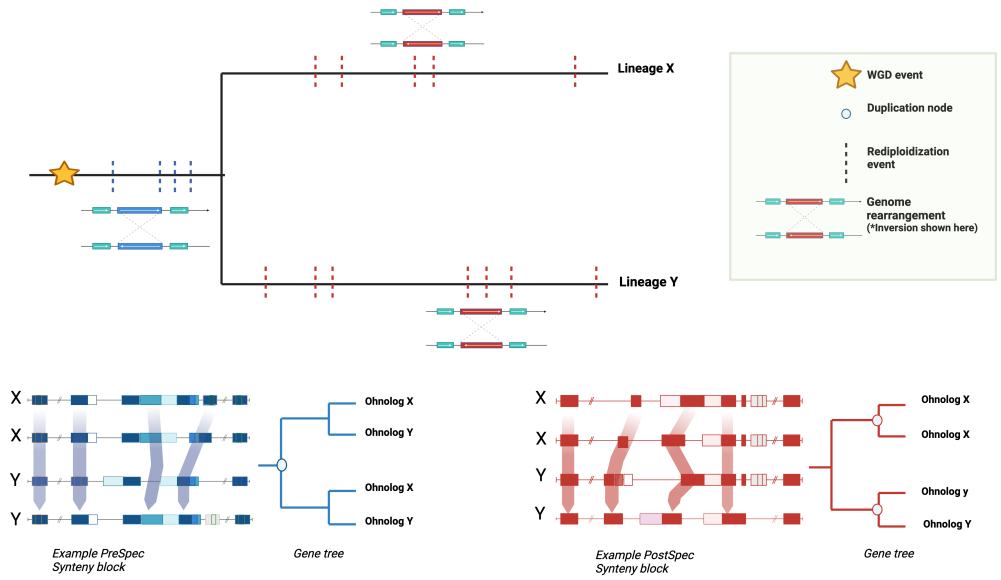

Figure 1.8 | Schematic of the mechanism for rediploidization through genome rearrangement and it's consequences The diagram illustrates a species tree of two lineages (top) that have undergone a shared whole-genome duplication (WGD), denoted by a star. Dashed lines in the species tree indicate rediploidization events. Blue dashed lines signify rediploidization events resulting from a genome rearrangement (depicted as an inversion here, but other rearrangements are possible) in the ancestor, while red represents events that occurred independently in the sister lineages. Below, blue synteny-blocks and gene trees represent the contiguous blocks of ohnologs that rediploidized in the ancestor of the lineages. Red synteny-blocks and gene trees represent blocks of ohnologs that have independently resolved in each lineage. Legend, top right.

Nearly 20% of the loci in the salmonid genome had ongoing tetrasomic inheritance well after the WGD (Robertson et al., 2017; Gundappa et al., 2021). In

the sturgeon and paddlefish, its estimated that more than half the genome had ongoing tetrasomic inheritance (Redmond et al., 2023).Considering that certain loci are still in the process of rediploidization, these well-studied instances of ancient tetraploidizations serve as valuable snapshots. They provide insights into the events leading to the suppression of recombination among tetrads of alleles, ultimately culminating in the resolution of ohnologs following WGD (Lien et al., 2016; Robertson et al., 2017; Redmond et al., 2023).

As mentioned, in salmonids, duplicated loci with almost identical alleles make up 20% of the genome. Known as isoloci, these alleles undergo homeologous recombination, perpetuating tetravalence. Isoloci have been found to be predominantly telomeric, which gives an insight into how they've prevailed in the genome (Allendorf et al., 2015; Lien et al., 2016). It's suggested that homelogous recombination can continue near the ends of chromosomes more readily than other parts of the chromosome and a model for this mechanism has been described in (Allendorf et al., 2015).

We still don't have a full picture of how ancient polylpoidisation events undergo this reversion process back to diploid inheritance. Despite this, a lot of progress using salmonid and non-teleost genomes has been made and there is increasing evidence to suggest that inversions and genome rearrangements play a big part in interrupting recombination and initiating rediploidization (Lien et al., 2016; Robertson et al., 2017; Redmond et al., 2023). In this work (Chapter 5) we take a deeper look at the mechanisms by identifying microsyntenic blocks of ohnologs in the paddlefish and sturgeon genomes. The ancestor of these lineages had an auto-tetraploidisation event that had not resolved to diploid when speciation took place resulting in complex patterns of shared and lineage-specific gene duplications dotted throughout the paddlefish and sturgeon genomes. With these lineages we can evaluate two scenarios: i) synteny-blocks that have a consistent

gene-tree topology showing a shared WGD, which we can assume rediploidized before the speciation event and have 1:1 orthology, and ii) blocks with a consistent gene tree topology defining a post-speciation WGD scenario, which would have rediploidized after speciation and would have a 2:2 orthology assignment. Testing whether blocks of ohnologs with conserved micro-synteny show a consistent topology, either the pre-speciation or post-speciation scenario, will allow us to test whether rearrangements halting recombination between duplicate homologous chromosomes is the primary mechanism of rediploidization in this lineage (Fig 1.8). The delay and asynchronicity of the rediploidization events we have discussed can lead to genomes with a mosaic of shared and lineage-specific gene duplications. With a clearer picture of the events that proceed polyploidy, we can better understand the evolutionary trajectories of these lineages and their complex histories. By re-examining known WGD events in the light of delayed rediploidization and mechanisms of this process, we can reinterpret the evolution of many lineages who've experienced WGD and probe difficult questions about the number and timings of WGD in early vertebrates.

1.4 Aim

This thesis will highlight the importance of gene-collinearity, its conservation and change, in aiding in determination of evolutionary relationships and the molecular histories of animals. In the third chapter I will examine the use of micro-synteny for phylogenetic inference. By comparing its abilities to sequence alignment methods we showcase its limitations while also revealing how through time, it unveils similar trends in mammalian and angiosperm evolution as standard sequence alignment frameworks. In chapter 4, we take a step away from synteny and show how to fix misassembled repetitive DNA in genomes, here working with the paddlefish

assembly. Coverage analysis revealed a large part of the genome had double the expected read-depth, suspected to be collapsed duplicates. By *reduplicating* these collapsed reads we "fix" the genome assembly for use in this and future work. In the final chapter before discussion, we investigate the mechanisms of rediploidization in the sturgeon and paddlefish genomes. We search for micro-syntenic blocks, shared and independent to the genomes of the two non-teleost ray-finned fishes, in order to exhibit the role of rearrangement in the rediploidiztaion process.

Chapter 2

Materials & Methods

This chapter provides an introduction and detailed explanation to some of the methods used throughout this thesis.

2.1 Phylogenetic inference

2.1.1 Phylogenetic inference with sequence alignments

In chapter 3 and chapter 5 we followed similar frameworks for phylogenetic inference with sequence alignments (SA) for our species trees and gene trees respectively. For the mammalian and angiosperm phylogenies in chapter 3 we followed a standard approach. 71 mammal and 99 angiosperm genomes were downloaded via the NCBI genome table (Table3.4, Table3.3). Orthologs for both groups were predicted using OrthoFinder (v2.5.4) (Emms and Kelly, 2019) and results were filtered and pruned using an in-house script which attempts to remove all hidden paralogs from the OrthoFinder output (github.com/Hidden_paralogs_script) Hidden paralogy is a major problem in molecular phylogenetics, referring to a situation where paralogous genes are mistakenly identified as orthologous due to their sequence similarity and functional properties. Hidden paralogy can have

significant effects on phylogenetic analyses and the accuracy of evolutionary reconstructions (Natsidis et al., 2021). Following these filtering steps, sequences were concatenated and then aligned using using MAFFT(v7.453) (Katoh et al., 2002) with standard parameters. Aligned sequences were ready for phylogenetic inference. We partitioned both protein datasets by gene and a mammalian nucleotide dataset by codon to achieve a tree with the greatest ML. Partitioning is important as it accounts for the heterogeneity in evolutionary processes that can occur across different regions of a SA. It recognizes that not all parts of a sequence may evolve in the same way, and by partitioning the data, you can apply distinct models and parameters to different segments of the SA. Model selection is the final step and is critical in phylogenetics because it influences the accuracy and reliability of phylogenetic analyses. The choice of an appropriate evolutionary model affects how well the data fits and, consequently, the quality of the inferred phylogenetic tree. The trees built with the partitioned protein alignments were inferred with the LG + G model[2.1.1] and the mammalian nucleotide data with the GTR + G model[2.1.1] in IQ-TREE (v1.6.12) (Minh et al., 2020). For the gene trees in chapter 5, orthology assignment is detailed below in Section 5.2.1. Ohnologs were aligned as before using MAFFT (v7.453) (Katoh et al., 2002). Phylogenetic inference by maximum likelihood was performed with IQ-TREE (v1.6.12) (Minh et al., 2020) with the JTT+G[2.1.1] model, -bb 1000 flag allowing for 1000 ultrafast bootstrap (UFBOOT) (Minh et al., 2013) replicates and -nt AUTO, which detected the optimal number of threads to be used for the analysis. UFBOOT

[1]**LG+G:** Le and Gascuel (LG) amino acid substitution model with a gamma distribution, G, to account for rate heterogeneity among sites

[2]**GTR+G:** General Time Reversible (GTR) model is a nucleotide substitution model that allows for different rates of nucleotide substitution between the four DNA or RNA bases while also considering the possibility of transitional substitutions and transversional substitutions. A Gamma distribution, G, was used, as before

[3]**JTT+G:** Jones-Taylor-Thornton (JTT) model is a general amino acid substitution model and assumes that the rates of amino acid substitutions are homogeneous across sites in the alignment. G, the gamma distribution, as before.

is used to compute the support of phylogenetic groups in ML based trees (Minh et al., 2013).

2.1.2 Phylogenetic inference with micro-synteny based alignments

The synteny-based phylogenetic trees in chapter 3 were generated using the pipeline described in the Zhao et al., 2021. The pipeline and data from this work is readily available at *github.com/SynNet-Pipeline*. The pipeline is loosely outlined in chapter 1, Fig.1.3 and comprises four main steps: (1) Synteny block detection using MCScanX (Wang et al., 2012), (2) network clustering (3) binary matrix representation of cluster information and (4) ML based tree inference in IQ-TREE (Minh et al., 2020). The framework for the approach is initially similar to standard approaches, beginning with orthology detection. In the Syn-MRL pipeline, orthology detection is done by pairwise genome comparison with Diamond (v0.9.30.131) (Buchfink et al., 2021). This information is then networked with MCScanX (Wang et al., 2012) to unveil a matrix with homology information in a syntenic context. Tree inference was performed using IQ-TREE (v1.6.12) (Minh et al., 2020) with the MK+R[2.1.2] morphological model. In IQ-TREE (Minh et al., 2020), the SA tree for each dataset, mammal and angiosperm, was input as the fixed topology using the *-te* flag. *-bb 1000* flag was used to allow for 1000 ultrfast bootstraps (UFBOOT)(Minh et al., 2013), the *-alrt 1000* flag specifies the number of replicates to perform SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010). UFBOOT and SH-aLRT are both used to measure the confidence of the branch placements i.e. branch support values. IQ-TREE

[4]**MK+R:** "M" stands for "Markov" and "k" refers to the number of states observed, usually binary, 0 or 1, for presence or absence. The model assumes that all transitions between character states are equal, and that all characters in the matrix have the same transition matrix. R (FreeRate) model was used to account for site-heterogeneity.

recommends using both measures so that each branch will then be assigned with SH-aLRT and UFBoot supports (Guindon et al., 2010; Minh et al., 2013; Minh et al., 2020). One would typically start to rely on the clade if its SH-aLRT >= 80% and UFboot >= 95%. *-st MORPH* flag was input to specify a binary sequence type.

2.1.3 Phylogenetic inference with gene presence/absence information

The presence/absence matrices were built using an in-house script available at *github.com/orthocounts2bin*. The script creates a binary gene presence/absence alignment from Orthogroups.GeneCounts.tsv file created following an OrthoFinder (v2.5.4) run (Emms and Kelly, 2019). This alignment can be used to construct phylogenies based on gene content. Tree inference for each dataset was preformed with IQ-TREE(v1.6.12) (Minh et al., 2020) with the MK+$R^{2.1.2}$ morphological model and as before, the -te flag was used to specify our fixed topologies, both mammal and angiosperm, built using SAs. *-bb 1000* flag was used to allow for 1000 ultrfast bootstraps (UFBOOT), the *-alrt 1000* flag specifies the number of replicates to perform SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010; Minh et al., 2013) and *-st MORPH* specifies a binary sequence type in IQ-TREE (Minh et al., 2020).

2.2 Genome assembly and annotation

2.2.1 *Re-duplicating* collapsed duplicates in the paddlefish genome

Cheng et al., 2020's American Paddlefish assembly has 60 pairs of chromosomes with a genome size of 1.54GB with 26,017 predicted protein-encoding genes. The assembly was sequenced to 30X coverage and short and long reads from this study were deposited in CNGB under project accession number CNP0000867. Cheng et al., 2020 note a smaller than expected genome size following a 17-mer analysis (Liu et al., 2020). In Redmond et al., 2023, we found that there were regions in their Paddlefish assembly that had double the expected genome coverage. This is an indication that duplicates may have collapsed during assembly to the reference. Following a similar method by Du et al., 2020, we identified these collapsed regions and attempted to *"reduplicate"* them (Ko et al., 2022; Kelley and Salzberg, 2010; Zhang et al., 2019).

PacBio long-reads were aligned with bwa (v0.7.17-r1198-dirty) (Li and Durbin, 2009) using standard parameters. This was sorted and indexed using SAMtools (v1.16.1) (Bonfield et al., 2021). The short reads were aligned with Bowtie (v2.4.2) (Langmead et al., 2009) using standard parameters and again, indexing and sorting was done with SAMtools (v1.16.1) (Bonfield et al., 2021). To assess the depth of coverage across the genome, the alignments were split into 10kb regions and mosdepth (v0.3.3) (Pedersen and Quinlan, 2018) was used to quantify read depth at each segment with parameter *–by*. As discussed in Section 1.2, regions of high sequence similarity have been shown to collapse or merge during assembly to the reference genome (Salzberg and Yorke, 2005; Kelley and Salzberg, 2010). The Paddlefish's WGD history, along with the fact much of the genome is still undergoing tetrasomic inheritance, means that large parts of the fish's genome may

be present in duplicate and some of these regions may have artificially collapsed into a single locus leading to gaps and misassembly, and as seen here, double the expected read-depth in parts of the assembly (Redmond et al., 2023).

The 10kb double coverage segments found using mosdepth (v0.3.3) (Pedersen and Quinlan, 2018) were separated from the rest of the genome for the next steps. Using FreeBayes (v1.3.6) (Garrison and Marth, 2012), a polymorphism VCF was generated from the short-read alignments. FreeBayes (Garrison and Marth, 2012) is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment. The output is a VCF (Variant Call Format) file, for storing gene sequence variations.

In the next steps, PacBio long-reads aligned with bwa (v0.7.17-r1198-dirty) (Li and Durbin, 2009) were used to decipher separate haplotypes in the double coverage regions. Paddlefish PacBio long reads can be found at CNGB (Cheng et al., 2020). HapCUT2 (Edge et al., 2017), a haplotype assembly tool, was used to reconstruct individual haplotypes in double-coverage mapped long-reads. The assembly tool allows regions with more than one haplotype to be identified, and is used here to decipher potential duplicates from multiple alternative alleles in the double-coverage long-reads BAM file. The inputs for HapCUT2 (Edge et al., 2017) are the mapped double coverage reads (BAM file) and a VCF file. The program only works with VCFs from diploid genomes as phasing is currently not possible for polyploid genomes. Given that parts of the Paddlefish genome appear to be tetraploid, the VCF needed to be filtered of any polyploid genotypes (e.g. 4/4, 3/4 etc.) to be used in HapCUT2 (Edge et al., 2017). Using a custom script, we forced the genotype (GT) fields that were for example, 4/4, to be 2/2.

This was not a perfect solution but rather than deleting the entry altogether, we removed two of the least common alternative alleles from a tetraploid entry, thus keeping most of the information (see Chapter 4). The HapCUT2 output for the double-coverage 10kb regions contained files in which there were multiple assembled haplotypic segments. These segments were split into individual files with information from the VCF file using a script by Du et al., 2020 (available at https://github.com/dukecomeback/sterletM_sch) that was modified for this study. These split files were then processed with fgbio's *HapCutToVcf* script to generate separate VCF's for each assembled haplotype. These VCF files and the reference were used to produce haplotypic contigs in fasta format using *vcf-consensus* from the bcftools package (v1.10.2) (Li, 2011a)[2.2.1]. The fasta files of the split regions were merged with the original average-coverage contigs using the Unix *cat* command to generate a "new" 1.66GB genome fasta for the Paddlefish.

2.2.2 Assembly and scaffolding of the paddlefish genome

Following haplotype splitting, the genome then needed to be reassembled and scaffolded. Cheng et al., 2020, described 60 pairs of chromosomes (n=120), finding 26 macro chromosomes and 34 smaller, micro-chromosomes, a number which aligns with previous karyotype studies (Symonová et al., 2017) and is equivalent to the sterlet genome (Du et al., 2020). Assembly and scaffolding were done using Juicer (v1.6) and 3d-DNA (v190716)(Durand et al., 2016; Dudchenko et al., 2017). The illumina short-reads were aligned to the "new" contigs with Juicer (v1.6)(Durand et al., 2016). 3d-DNA (v190716) (Dudchenko et al., 2017) was then used for assembling the genome with -r=0 flag to ensure no iterative rounds of mis-join correction were carried out. This flag was used to speed the process up and given that the assembly had already been scaffolded, iterative rounds of mis-

[5]**To note:** If not stated otherwise, default parameters were employed for all tools.

join correction were not necessary. Finally, the scaffolded assembly was manually reviewed using Juicebox assembly tools (v1.6)(Dudchenko et al., 2017).

2.2.3 Paddlefish genome annotation

Genome annotation was performed using three lines of evidence: homology annotation, de-novo annotation, and RNA-seq annotation. Firstly, assembly quality and completeness were assessed with BUSCO (v5.4.4) (Manni et al., 2021) under the *Actinopterygii odb9* database. The gene prediction flags, *-augustus*, and *-long* were implemented in the BUSCO run (Manni et al., 2021; Stanke and Morgenstern, 2005). AUGUSTUS (Stanke and Morgenstern, 2005) is a de novo gene prediction tool for eukaryotic genomes and can be run separately or as part of a BUSCO run (Manni et al., 2021). *-long* is used for optimization of AUGUSTUS self-training mode in BUSCO and while it adds considerably to the run time, it can improve gene prediction results for some non-model organisms. As repeat masking had already been carried out in the original assembly (Cheng et al., 2020), it was not necessary to do again here.

For homology annotation, we used a set of 11 diverse vertebrate proteomes from NCBI: American Paddlefish (*Polyodon spathula*; GCF_017654505.1) (Cheng et al., 2020), elephant shark (*Callorhinchus milii*; GCF_000165045.1) (Venkatesh et al., 2014), zebrafish (*Danio rerio*; GCF_000002035.6), medaka (Oryzias latipes; GCA_002234675.1), fugu (*Takifugu rubripes*; GCA_901000725.2), stickleback (*Gasterosteus aculeatus*; GCA_016920845.1), sea lamprey (*Petromyzon marinus*; GCA_010993605.1), spotted gar (*Lepisosteus oculatus*; GCF_000242695.1) (Braasch et al., 2016), human (*Homo sapiens*; GCF_000001405.39), mouse (Mus musculus; GCA_000001635.9) and sterlet (*Acipenser ruthenus*; GCA_902713425.2) (Du et al., 2020). The proteomes were run through CD-HIT to reduce redundancy

when aligning, which resulted in 229,665 proteins (Fu et al., 2012). These were aligned to the assembly using Exonerate (v2.2.0) (Slater and Birney, 2005) and GFF3 files were created for use in evidence based gene modelling in later steps. Exonerate (Slater and Birney, 2005) is a generic tool for pairwise sequence comparison. It allows you to align sequences using many alignment models and can be quick and general, producing either gapped or ungapped alignments.

For RNA-seq annotation, RNA-seq reads from the Brain, Kidney, Liver, Spleen, Skin, Skeletal Muscle, Eye, Gill filament, Gill raker, Rostrum, Spiral valve, Stomach, Heart, Gonad, Pyloric Caeca of five adult Paddlefish were aligned to the reference using hisat2 (v2.2.1) (Kim et al., 2015), trimmed using trimal (v1.4.1) (Capella-Gutiérrez et al., 2009) and assembled using stringtie (v2.2.1)(Kovaka et al., 2019). The resultant BAM files were then sorted and indexed with SAMtools (v1.16.1) (Bonfield et al., 2021) and merged using taco (v0.7.3) (Niknafs et al., 2017) to produce GFF3 files for use in gene modelling by EVidence Modeler(EVM) (v2.1.0) (Haas et al., 2008). TransDecoder (v5.7.0) was used in parallel to the hisat2/stringtie method for RNA-seq assembly. TransDecoder is used to identify likely coding regions within transcript sequences by identifying long open reading frames (ORFs) within transcripts. Transdecoder reports ORFs that encode sequences with compositional properties consistent with coding transcripts. All lines of gene evidence obtained from homology, RNA-seq (taco and transdecoder) and de-novo annotation (Augustus) were collected and transferred into EVM (v2.1.0)(Haas et al., 2008), where gene models conformed by all lines of evidence were extracted as high-quality gene models. General functional annotation was done using the AnnotaPipeline (v2.0) using SWISS-PROT and TrEMBL(uniprot) to annotate and validate predicted features in genomic sequences. A total of 35,930 protein coding genes were predicted by EVM of which 29,833 were functionally annotated.

2.3 WGD and Rediploidization analysis

2.3.1 Orthology Assignment

OrthoFinder (v2.5.4) (Emms and Kelly, 2019) was used for orthology inference. We included a diverse set of proteomes that spanned the jawed vertebrate phylogeny including ghost shark (*Callorhinchus milii*; GCF_000165045.1) (Venkatesh et al., 2014) from Chondrichthyes, human (*Homo sapiens*; GCF_000001405.39), African clawed frog (*Xenopus tropicalis*; GCF_000004195.4), and coelacanth (*Latimeria chalumnae*; GCF_000225785.1) from Sarcopterygii. From within Actinopterygii we selected zebrafish (*Danio rerio*; GCF_000002035.6) and used spotted gar (*Lepisosteus oculeatus*; GCF_000242695.1) (Braasch et al., 2016) as a representative of Neopterygii. Finally we used the Grey bichir *(Polypterus senegalus*; GCF_016835505.1) (Bi et al., 2021) as the combined sister group to the paddlefish and sturgeon. We included a species tree in our OrthoFinder (Emms and Kelly, 2019) run with flag *-s* in line with the accepted relationships in the jawed vertebrate phylogeny to augment orthology inference: *((Callorhinchus milii), ((Latimeria chalumnae,(Xenopus tropicalis,(Homo sapiens)),(Grey birchir,((Polyodon Spathula, Acipenser ruthenus),((Danio rerio), (Lepisosteus oculatus)))))));* Fig.5.3.

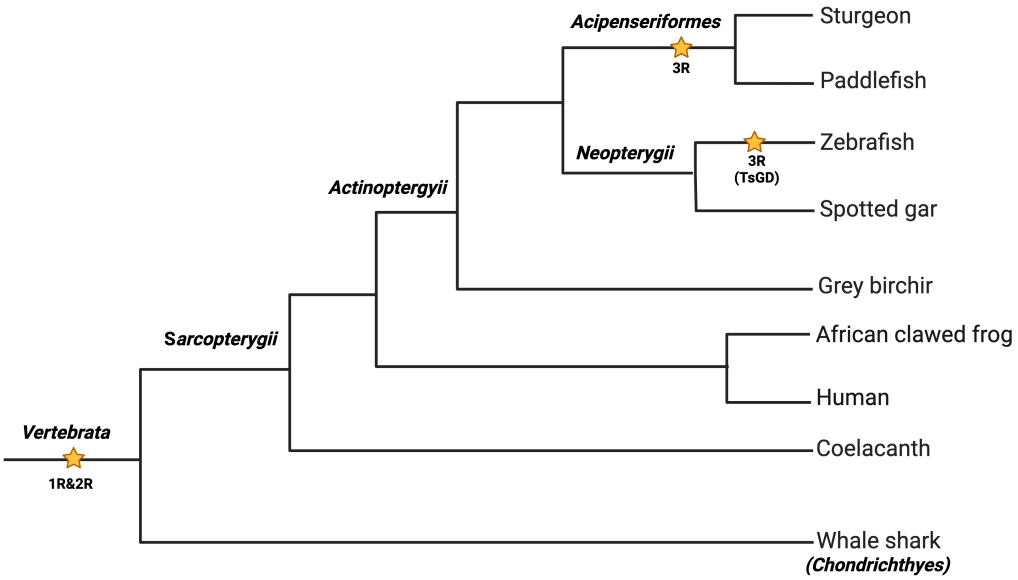

Figure 2.1 | Phylogeny of the species used for orthology inference in Chapter5 Species that span the jawed vertebrate phylogeny used in OrthoFinder (Emms and Kelly, 2019) for orthology inference. 1R-3R represent tetraploid events. Branch lengths are not to scale and so event placement is approximate.

We extracted the longest isoform from the proteomes using a custom protocol and then used CD-HIT (Fu et al., 2012) to reduce redundancy in the sequences and ran OrthoFinder (v2.5.4) (Emms and Kelly, 2019) with parameter *-y* and as mentioned *-s*, to include a rooted species tree. By using the *-y* flag, within the Phylogenetic Hierarchical Orthogroups (PHOGS) file in the OrthoFinder result, we split paralogous clades below root of a Hierarchical Orthogroup into separate Hierarchical Orthogroups. For example, we split clades that have genes that may have duplicated after the earliest diverging jawed vertebrate (here, the whale shark) resulting in separate PHOG files for each duplicate. Following a protocol from Redmond et al., 2023, further filtering of the PHOGs was done by extracting groups that included two sequences each from sturgeon and paddlefish and that

also had at least one outgroup for subsequent rooting of the sturgeon-paddlefish pair subtree. Sturgeon-paddlefish genes were only chosen from one of the 60 largest chromosomes therefore ohnologs from micro-chromosomes were not considered in this analysis. PHOG gene trees were built by first aliging with MAFFT (v7.453) (Katoh et al., 2002) and reconstructed with IQ-TREE (v1.6.12) (Minh et al., 2020), see Section 3.2.1 for more detail. We did not filter the PHOG set any further as it was not necessary for this analysis as further filtering was done in later steps using synteny. A strict high-confidence ohonolog set has also been published in Redmond et al., 2023. We verified that the sturgeon-paddlefish sequences formed a monophyletic group and ensured that the two paddlefish and sturgeon sequences (suspected ohnologs) diverged after the split from Neoptergyii. This was done by ensuring each gene tree had at least one sequence from Neopteregyii and a more distantly related outgroup for construction. We are aware that this set is not high-confidence and may include genes with complex histories that may look like ohnologs but may be paralogous. This is because gains and losses during the evolutionary history of theses lineages may be hard to disentangle from the information we can garner from the extant genomes used here. We cannot rule out introgression, incomplete lineage sorting or reciprocal gene gain/loss as an explanation for some of these ohnolog pairs. We also note that these ohnologs are conserved in pairs in both species and thus excludes genes that may have been lost in sturgeon, paddlefish or in the Acipenseriformes stem lineage.

2.3.2 Ohnolog duplication-time inference

The sturgeon-paddlefish ohnolog-pair described above, then underwent phylogenetic analysis to estimate the time of rediploidization relative to the speciation event. MAFFT (v7.453) was used for multiple SAs of the filtered PHOGs with

standard parameters. Phylogenetic inference by ML was performed with IQ-TREE (v1.6.12)(Minh et al., 2020) with the -m JTT+G2.1.1 flag, a general amino acid substitution model with four discrete rate categories, -bb 1000 flag allowwing 1000 ultrafast bootstrap (UFBOOT) (Minh et al., 2013) replicates Minh et al., 2020 and -nt AUT0, which detected the optimal number of threads to be used for the analysis. These ohnolog gene trees were pre-processed and analysed for duplication time inference (i.e rediploidization time). For this, we used the ETE(v3)toolkit python library (Huerta-Cepas et al., 2016) to check that the sturgeon and paddlefish sequences formed a monophyletic group in each of the filtered PHOG gene trees and then rooted each with the most distantly related sequence relative to the Acipenseriformes. Custom scripts adapted from Redmond et al., 2023 were used to perform strict gene-species tree reconciliation to infer speciation and duplication nodes/events. The resulting gene trees were summarised into different groups indicating their duplication time: PreSpec, PostSpec, Other[PreSpec-like, PostSpec-like].

2.3.3 Synteny analysis

To find the syntenic blocks between sturgeon and paddlefish genomes we used OrthoFinder (v2.5.4) (Emms and Kelly, 2019) and i-adhore (v3.0.01) (Proost et al., 2012). The sturgeon, *"reduplicated"* paddlefish and grey birchir proteomes were run through OrthoFinder(v2.5.4) (Emms and Kelly, 2019) using standard parameters. It was not necessary to do any post processing on the OrthoFinder output here as a refined ortholog dataset, as described in Section 5.2.1 has already been curated. Using the *Orthogroups.txt* and corressponding GFF3 files of the proteomes used in orthology inference, we prepared the genes for the i-adhore run. The *segments.txt* file in the i-adhore output was used to define the genomic

co-ordinates of the pairs of genes within a syntenic block (a.k.a *multiplicon* in i-adhore)(Proost et al., 2012). The gene names and co-ordinates were prepared for circos (v0.69-9) (Krzywinski et al., 2009) to be used as synteny links within and between the two genomes.

Chapter 3

Insights into the relationship between molecular sequence evolution and gene order evolution

3.1 Introduction

Phylogenetic reconstruction is accomplished by profiling molecular and morphological characters among a group of species, in an attempt to infer their evolutionary relationships. Most commonly, reconstruction is achieved by comparing changes in the genetic sequence between two or more organisms. Other characters such as morphological changes and gene order changes are, in most instances, overlooked, in part because they have been shown to be less accurate. Local conservation of gene order, or micro-synteny, describes the shared spatial ancestry of groups of genes in and between species. While incorporating micro-synteny information into phylogenetic analysis is not novel, many older methods were computationally expensive, relied on well developed models and worked on small, simple genomes (Belda et al., 2005; Tang and Moret, 2003; Luo et al., 2008; Feng et al., 2017). Over

the past six years, there have been significant advances in the use of micro-synteny for evolutionary analysis, introducing new tools that can be applied to orthology inference, the reconstruction of ancestral polyploid events, and, as explored in this study, phylogenetic profiling.

Firstly, we aim to reconstruct the deeply diverged bilaterian phylogeny using micro-synteny information alone. This is to test the framework's robustness against extended evolutionary time-frames and diverse sets of genomes with complex histories. We discuss the methods requirement for high quality genomes assemblies, the lack of those for many major groups in the bilateria and the current limitations in modelling binary data for phylogenetic inference. We then focus our analysis on mammalian and angiosperm phylogenies: This involved constructing standard sequence alignment trees, gene presence/absence phylogenies, and utilising pre-computed synteny-derived phylogenies to conduct a thorough comparative analysis of these various inference methods. Focusing on branch lengths, we explored the correlation between the rate of change along the branches across different data types. With this information we can shed light on the relationship between the evolution of molecular sequence changes, gene order changes and gene gain-and-loss respectively. Additionally, we construct timetrees to trace variations in gene order and molecular sequence rates across mammalian evolution. This aspect of the study can aid in identifying potential geographical, ecological, or anomalous factors that might have influenced the rate of change in these characteristics.

As mentioned, incorporating micro-synteny information into phylogenetics is not novel but older frameworks lacked complexity and worked only on small, simple genomes or mitochondrial genomes (Belda et al., 2005; Tang and Moret, 2003; Luo et al., 2008; Feng et al., 2017). Recently, there have been a wave of new techniques which use a distance-based approach based on breakpoints in syntenic blocks and combines synteny networks with ML based phylogenetics tools (Drillon et al.,

2020; Zhao et al., 2021). These approaches have succeeded in building phylogenies of large whole-genome datasets (Zhao et al., 2021; Zhao et al., 2017; Zhao and Schranz, 2017). In this work, we use phylogenies built with the *Syn-MRL* pipeline (Zhao et al., 2021). The pipeline uses a pairwise approach which unlike previous attempts, has the potential to work on a large numbers of genomes and do so efficiently. The method includes four main steps: (1) Synteny block detection using MCScanX (Wang et al., 2012), (2) network clustering (3) binary matrix representation of cluster information and (4) ML based tree inference with IQ-Tree (Minh et al., 2020) (See Fig.1.3). The framework for the approach is initially similar to standard methods, beginning with orthology detection achieved by a reciprocal best hit approach using Diamond (Buchfink et al., 2021). Using MCScanX (Wang et al., 2012), this information is then networked to unveil a matrix with homology information through a syntenic context. Angiosperm and mammalian phylogenies constructed with *Syn-MRL* have been shown to be highly congruent with phylogenies built using standard methodologies and simulations have also confirmed the pipelines accuracy and efficiency(Zhao et al., 2021; Zhao et al., 2017; Zhao and Schranz, 2017). With pipelines like *Syn-MRL*, it is now possible to exploit gene-order information and reconstruct species relationships efficiently and accurately (Belda et al., 2005; Tang and Moret, 2003; Luo et al., 2008; Feng et al., 2017; Drillon et al., 2020; Zhao et al., 2021).

As mentioned we first attempt to reconstruct deeply diverged phylogenies with the pipeline. It is assumed that synteny information will collapse as relationships become more diverged due to saturation, however this has not been formally tested. For success, *Syn-MRL* relies on high-quality genomes with adequate genome annotation as well as a comprehensive taxon-sampling of the groups you're interested in (Zhao et al., 2021; Liu et al., 2018). Here, we look to reconstruct the bilaterian phylogeny, estimated to be between ∼636.1-553.83 mya old (Reis et al., 2015;

Bengtson et al., 2012). By selecting the highest quality assemblies available from each of the major groups within Biliteria (Ecdysozoa and Lophotrochozoa making up Protostomia and Deuterostomia consisting of Xenambulcraria[3.1] and Chordata) we aim to construct an accurate phylogeny by relying solely on conserved micro-synteny among these lineages. In our analysis we attempt to show whether or not synteny information alone is adequate for reconstructing widely accepted clades. We delve into the challenges associated with poor-quality genome assemblies and the problems stemming from incomplete taxon sampling in many major bilaterian clades. We also discuss the lack of complex models for binary data and whether the assumptions of the MK[2.1.2] model are fitting for use with the micro-synteny data produced here.

With this novel method of species inference of large whole-genome datasets, it is now possible to carry out a comprehensive comparative analysis of micro-synteny based methods versus standard sequence alignment frameworks. Here, we ask how the rate of molecular sequence evolution correlates with the rate in which genes have rearranged themselves throughout the evolution of a phylogeny. Similar to the long standing question of whether or not there is a relationship between genomic and phenotypic evolution (Omland, 1997; Bromham et al., 2002; Davies and Savolainen, 2006), we ask is there a relationship between gene order evolution and sequence evolution? (Omland, 1997; Halliday et al., 2019) Until recently, it has been impossible to test this because micro-synteny based methods were incapable of producing phylogenies of the same caliber as those generated by sequence alignment methods. Now, with pipelines utilizing a ML reconstruction program, like the one described here, we have characters, like branch lengths, that can be compared to SA phylogenies. With these we can garner some rate and temporal

[1]**Xenambulcraria:** Here, Xenambulcraria defines a group contaning Xenoturbella and the acoelomorph worms (Xenacoelomorpha), echinoderms and hemichordates (Philippe et al., 2019). Other studies refute this grouping and place Xenacoelomorpha as the sister group of all other bilaterian animals and not sister to chordates (Cannon et al., 2016).

information that can tell us about the evolution of gene movement in these groups. Branch lengths in an ultrametric tree built using sequence alignment methods, tell us how the rate of sequence evolution has changed through time. Branch length in trees built with micro-synteny information therefore, tells us how the rate of gene order evolution has changed through time.

$$Branch\ Length\ = Time\ X\ Rate\ of\ Change \tag{3.1}$$

Comparing branch lengths of a fixed topology constructed using different inference methods will give us insight into how these molecular changes have interacted with one another over the evolutionary history of a lineage. Given the novelty of this investigation, the expectation for how molecular changes interact with each other through time are subjective. If the rate of gene order and sequence evolution do in fact correlated it would suggest that both are neutrally evolving under a similar underlying mechanism or influenced by a similar mechanism. If they do not correlate then different mechanisms influence their evolution and both are interetsing results. As discussed, the pipeline has been shown to generate mammal and angiosperm phylogenies that exhibit high congruence with phylogenies built with sequence alignment-based methods (Zhao et al., 2021). We use these pre-computed synteny phylogenies in this analysis. We also construct presence/absence phylogenies and compare the evolution of gene gain and loss through time in angiosperm and mammals to both synteny and sequence alignment methods. By also looking at the patters of rate change through time for each of these characters, it may be possible to pinpoint potential geographical, ecological or anomalous factors that may have influenced any fluctuations in rate.

3.2 Materials & Methods

3.2.1 Phylogenetic inference with sequence alignments

For the mammalian and angiosperm phylogenies we followed a standard approach. 71 mammal and 99 angiosperm genomes were downloaded via the NCBI genome table (Table 3.3, Table 3.4). Orthologs for both groups were predicted using Orthofinder (v2.5.4) (Emms and Kelly, 2019) and results were filtered and pruned using an in-house script which attempts to remove all hidden paralogs from the Orthofinder output (see Chapter 2 Methods, for more details on hidden paralogy). Following these filtering steps, sequences were assembled into a supermatrix and then aligned with MAFFT(v7.453) (Katoh et al., 2002). We partitioned both protein alignments by gene and the mammalian nucleotide alignments by codon to achieve a tree with the greatest ML. The trees built with the partitioned protein alignments were inferred using the site homogeneous LG + G model[2.1.1] and the mammalian nucleotide data with the GTR + G model[2.1.1] in IQ-TREE (v1.6.12) (Minh et al., 2020).

3.2.2 Phylogenetic inference with micro-synteny information

The synteny-based phylogenetic trees were generated using the pipeline described in Zhao et al., 2021. The pipeline and data from this work is readily available at *github.com/SynNet-Pipeline*. The pipeline is loosely outlined in chapter 1, Fig.1.3, and comprises four main steps: (1) Synteny block detection using MCScanX (Wang et al., 2012), (2) network clustering (3) binary matrix representation of cluster information and (4) ML based tree inference in IQ-TREE (Minh et al., 2020). In the *Syn-MRL* pipeline, orthology detection is done by pairwise genome comparison with Diamond (v0.9.30.131) (Buchfink et al., 2021). Tree inference was performed

using IQ-TREE (v1.6.12) (Minh et al., 2020) with the Mk+R$^{2.1.2}$ morphological model. For the bilaterian tree analysis, genomes were downloaded via the NCBI genome table (Table 3.1). They were formatted for use in the pipeline and the binary matrix output was prepared for inference by ML. For the angiosperm and mammal phylogenies, we used the pre-computed micro-synteny matrices available in supplementary of Zhao and Schranz, 2019, at *https://dataverse.harvard.edu* and adapted them for use in this work. The pipeline was then run from the network clustering checkpoint. In IQ-TREE (Minh et al., 2020), the SA trees for the mammal and angiosperm datasets, were input as the fixed topology using the *-te* flag. For both analysis', *-bb 1000* flag was used to allow for 1000 ultrfast bootstraps (UFBOOT)(Minh et al., 2013), the *-alrt 1000* flag specifies the number of replicates to perform SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010).*-st MORPH* flag was input to specify a binary sequence type.

3.2.3 Phylogenetic inference with gene presence/absence information

The presence/absence matrices were built using an in-house script available at *github.com/orthocounts2bin*. The script creates a binary gene presence/absence alignment from the *Orthogroups.GeneCounts.tsv* file created following orthology inference with Orthofinder (v2.5.4) (Emms and Kelly, 2019). This alignment can be used to construct phylogenies based on gene content. Tree inference for each dataset was preformed with IQ-tree(v1.6.12) (Minh et al., 2020) with the MK+R$^{2.1.2}$ morphological model and as before, the -te flag was used to specify our fixed topologies, both mammal and angiosperm. *-bb 1000* flag was used to allow for 1000 ultrfast bootstraps (UFBOOT), the *-alrt 1000* flag specifies the number of replicates to perform SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010; Minh et al., 2013) and *-st MORPH* specifies a binary

sequence type in IQ-TREE (Minh et al., 2020).

3.2.4 Timetree for the mammalian phylogeny

In this analysis, we used the high-confidence mammalian timetree published by Álvarez-Carretero et al., 2021. They applied a Bayesian molecular-clock dating method to construct a timetree encompassing 4,705 mammal species. We manually pruned this tree to retain only the species featured in our sequence alignment and micro-synteny trees. The upper-bound stem and crown family age estimates were used to date the nodes. For details on all mean stem and crown family ages, along with their 95% highest posterior density intervals, refer to Álvarez-Carretero et al., 2021. Following Formula.3.1 and now with knowledge of the time parameter, we could isolate for rate. Plots for rate against time were drawn with the Python package plotly (Inc., 2015).

3.2.5 Timetree for the angiosperm phylogeny

The angiosperm timetree published by Ramírez-Barahona et al., 2020 was used in this analysis. They constructed and dated a comprehensive family-level phylogeny of flowering plants, integrating 16 million geographic occurrence records for angiosperms globally. The tree was constructed using the Bayesian uncorrelated log-normal clock model implemented in BEAST version 2.5.1 (Bouckaert et al., 2014). We used the upper-bound stem and crown family age estimates for dating the nodes. For details on all mean stem and crown family ages, along with their 95% highest posterior density intervals, refer to the supplementary material of the publication (Ramírez-Barahona et al., 2020). We manually pruned this high-confidence tree to include only the species present in our sequence alignment and micro-synteny trees. After applying Formula 3.1 and obtaining the time parameter, we isolated the rate. Subsequently, plots depicting the correlation between

rate and time were generated using the Plotly Python package (Inc., 2015).

3.2.6 Statistical analysis

Statistical tests and correlation plots were all carried out and constructed in R (v4.2.1) (**R**). We excluded outgroup taxa from all correlations because they typically consist of long branches. This is primarily due to incomplete taxon sampling relative to the in-group, making them less reliable representatives of the true evolutionary rate within the ingroup taxa. To find the appropriate statistical test for quantifying strength of correlation between the two characters, we tested the distribution of branch lengths using a one-sample Kolmogorov-Smirnov(KS) test, and by graphical assessment. The data distributions were found to be non-normal, and so branch lengths for all trees were compared using the Spearman's correlation (Rho) test for non-parametric data. The Spearman's rank-order correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative relationship, 1 indicating a perfect positive relationship, and 0 indicating no relationship (independence) between the variables. The level of statistical significance for Spearman's rho was determined using a p-value cutoff of less than 0.05. We also looked for any instances auto-correlation in the datasets. Auto-correlation occurs in phylogenetic trees where branches at a closer proximity to one another display more similar values than those located at greater distances, as character differences accumulate between lineages in proportion to evolutionary time (Felsenstein, 1985). These data points would then be unreliable for correlation testing, as they may cluster in a manner that could spuriously suggest that molecular and gene order rates or branch lengths are correlated (statistical non-independence). We carried out a phylogenetic generalised least squares (PGLS) test to assess for auto-correlation artefacts in our datasets.

As a precautionary measure we removed all terminal branches when performing

correlation tests and removed micro-synteny terminal branches when looking at rates through time. When dealing with binary data reconstructed using the MK model[2.1.2], there is a potential risk of ascertainment bias. The Mk model represents the evolution of discrete-state morphological characters as a stochastic process. If there are any derived characters (autapomorphies) in extant taxa that are not considered, there is a possibility of underestimating rates at terminal branches (Lewis, 2001). While we are not using morphological data here, the *Syn-MRL* pipeline transforms the micro-synteny data into discrete 0 and 1's (absence of a synteny-block in that species/presence of a synteny-block in that species) and so any complex movements or evolutionary process happening in extant branches may not be accounted for leading to an underestimation of rates at terminal branch lengths relative to internal branches.

3.3 Results

3.3.1 Testing the *Syn-MRL* pipeline on deeply diverged taxa

For our first analysis we wanted to test how the pipeline would work with deeply diverged taxa, here the bilaterian phylogeny. Animal genomes were downloaded from NCBI genome table. In order to achieve a comprehensive representation of the entire bilaterian phylogeny, we selected a minimum of three species from each major bilaterian phylum. Species chosen had BUSCO scores >90 and were assembled at the scaffold level or higher (see list of species in Table 3.1). While previous work indicated that the pipeline would function sub-optimally with genomes with N50 <1mb (Liu et al., 2018; Zhao et al., 2021; Zhao and Schranz, 2019), we included some lower quality genomes as we believed sampling each major bilaterian

sub-phylum trumped the high-quality genome requirement. For example, *Stylea clava*, the only species available to represent Tunicates (Uorchordata), had a N50 of 0.74MB but BUSCO >90 and was assembled to scaffold level (Table 3.1). At the time of this analysis there was no Xenacoelomorpha genome assembled above contig level and so we did not include that sub-phylum in our analysis.

Running the pipeline effectively requires careful configuration of several parameters that significantly influence the size and quality of the micro-synteny blocks you're analyzing between genomes (see synteny block schematic in Fig.1.2). Identifying the most informative blocks is vital for successful inference, and selecting the right parameters plays a pivotal role in achieving this goal. The critical parameters you need to consider are k, s, and m:

1. **k (tophits):** This parameter determines the number of target sequences from the Diamond (Buchfink et al., 2021) results to retain for each query sequence. The default value is 6. Adjusting this parameter can impact the sensitivity and specificity of your results. A higher k value may capture more potential homologous sequences, but it may also introduce more noise into your analysis.

2. **s (anchors for synteny block):** The 's' parameter is the number of anchors required to define a synteny block when using MCScanX. The default value is 5. Increasing this number may lead to more stringent criteria for identifying synteny blocks, potentially reducing the number of detected blocks. Conversely, lowering the 's' value might result in the identification of smaller and possibly less reliable synteny blocks.

3. **'m'**: The number of genes you search for as anchors in both the upstream and

[2]**Doubt over deuterostome monophyly:** Recent molecular phylogenetic studies have not consistently supported deuterostome monophyly. Support for the deuterostome clade, widely accepted for over 100 years, may be the result of an artifact of tree reconstruction (Kapli et al., 2021).

downstream regions of a gene. The default value is 25. Lowering this value makes the anchor detection stricter and can help identify more conserved regions but may also reduce the number of detected blocks.

In Table 3.2, we adjusted and experimented with different parameter combinations to investigate their effects on the tree produced and to find the optimum parameters for this dataset. We employed the Robinson-Foulds (RF) distance as a metric to assess the accuracy of the topology in comparison to the true topology. The metric is a measure used in phylogenetics to quantify the dissimilarity or difference between two phylogenetic trees, here our synteny-based tree versus the generally accepted topology of these species (constructed by https://phylot.biobyte.de/ which generates phylogenetic trees based on the NCBI taxonomy). A lower RF distance indicates greater similarity between the two trees, implying that they share more common branching patterns. Conversely, a higher RF distance indicates more dissimilarity, with fewer shared bipartitions and less congruence in topology.

Species	**SuperPhylum**	**Acronynm**	**Scaf/Chr**	**N50(MB)**	**BUSC0**
Mus musculus	Chordata	mus	Chr	106.1	C:99.6%[S:51.1%,D:48.4%],F:0%,M:0.4%,n:137980
Homo sapiens	Chordata	hs	Chr	67.8	C:99.5%[S:39.1%,D:60.4%],F:0%,M:0.4%,n:13780
Pan paniscus	Chordata	ppn	Chr	141.8	C:95.2%[S:55.3%,D:39.9%],F:1.2%,M:3.6%,n:13780
Pan troglodytes	Chordata	pt	Chr	95.4	C:97.8%[S:50.3%,D:47.5%],F:0.5%,M:1.7%,n:13780
Sus scrofa	Chordata	ss	Chr	88.2	C:95.7%[S:39.1%,D:56.6%],F:0.7%,M:3.6%,n:13335
Branchiostoma belcheri	Chordata	br	Scaf	47.3	C:98.7%[S:75.3%,D:23.4%],F:0.8%,M:0.5%,n:954
Branchiostoma Floridae	Chordata	bf	Chr	50.4	C:95.5%[S:70.6%,D:24.9%],F:2.1%,M:2.4%,n:954
Branchiostoma lanceolatum	Chordata	bl	Scaf	1.2	C:48.7%[S:46.2%,D:2.5%],F:15.4%,M:35.9%,n:3354
Styela clava	Chordata	stc	Scaf	0.74	C:93.7%[S:76.9%,D:16.8%],F:0.6%,M:5.7%,n:954
Strongylocentrotus purpuratus	Xenambulacraria	sp	Chr	2.1	C:98.6%[S:55.0%,D:43.6%],F:0.6%,M:0.8%,n:954
Asterias rubens	Xenambulacraria	ar	Chr	20.6	C:98.7%[S:77.1%,D:21.6%],F:0.2%,M:1.1%,n:954
Acanthaster planci	Xenambulacraria	apl	Scaf	49.7	C:98.8%[S:62.6%,D:36.2%],F:0.5%,M:0.7%,n:954
Strongyloides ratti	Ecdysozoa	str	Chr	11.7	C:69.1%[S:66.1%,D:3.0%],F:2.6%,M:28.3%,n:954
Linepithema humile	Ecdysozoa	lh	Scaf	14.0	C:99.0%[S:73.0%,D:26.0%],F:0.5%,M:0.5%,n:954
Pollicipes pollicipes	Ecdysozoa	pp	Scaf	0.11	C:99.0%[S:68.0%,D:31.0%],F:0.1%,M:0.9%,n:1013
Lingula anatina	Lophotrochozoa	la	Scaf	0.46	C:98.6%[S:55.0%,D:43.6%],F:0.6%,M:0.8%,n:954
Crassotrea virginica	Lophotrochozoa	crv	Chr	75.9	C:98.2%[S:53.6%,D:44.6%],F:0.3%,M:1.5%,n:5295
Gigantopelta aegis	Lophotrochozoa	gga	Chr	461.8	C:95.9%[S:62.6%,D:33.3%],F:0.5%,M:3.6%,n:5295
Helobdella robusta	Lophotrochozoa	hel	Scaf	0.52	C:90.2%[S:89.2%,D:1.0%],F:4.6%,M:5.2%,n:954
Clytia hemisphaerica	Cnidaria*	cly	Scaf	1.2	C:81.1%[S:72.1%,D:9.0%],F:2.9%,M:16.0%,n:954

Table 3.1 | Species used in reconstruction with *Syn-MRL pipeline.* Table includes information of assembly level (either scaffold *(scaf)* or chromosome *(chr)*) and N50 and BUSCO scores as a measure of assembly quality. **Clytia hemisphericais* the outgroup species.

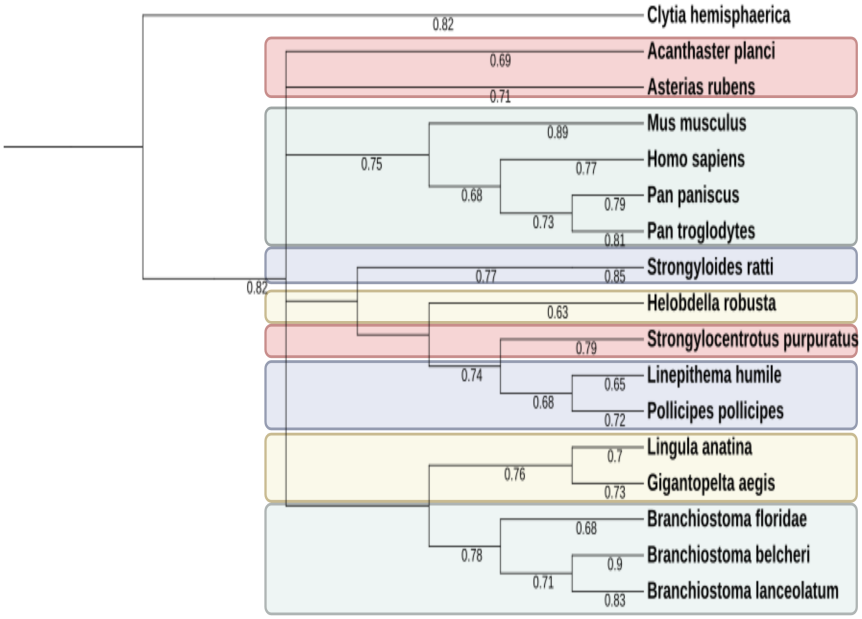

Figure 3.1 | Example of a phylogeny produced from the syn-MRL pipeline approach with an RF-distance of 64 to the true-tree. Cladogram depicting the best topology retrieved using the *Syn-MRL* pipeline with parameters k6s4m20, with RF distance of 64 to the true tree. The different bilaterian groups are highlighted to distinguish between them: Ecdyzosoa in blue, lophotrochozoa in yellow, xenambulacraria in red and chordata in green. *Clytia hemisphaerica* (cnidaria) was used as the outgroup species. Branch lengths are not to scale.

Referring to Table 3.2, which displays a list of RF distances calculated from trees reconstructed using changing *Syn-MRL* parameters, it becomes apparent that the pipeline struggled to generate an accurate representation of the species' evolutionary relationships. Very high RF distances (RF $>$ 60) indicate significant dissimilarities between synteny-based phylogenies and the expected or true species relationships. The lowest RF obtained was 62 and retrieved from a synteny-based tree constructed with parameters K=6, S=6 and M=25. Fig 3.1 is an example of a tree topology built with parameters k6s4m20. The tree has an RF distance of 64 to the true tree, indicating high discordance. By using distinct colours for each group. the figure highlights these errors and we can see xenambulacraria (red), a clade within Deuterostomia, branching alongside lophotrochozoa (blue).

K	S	M	Robinson Foulds Distance
6	4	15	72
6	4	20	64
6	4	25	72
6	4	30	74
6	4	35	70
6	5	15	68
6	5	20	74
6	5	25	68
6	5	30	70
6	5	35	70
6	6	15	64
6	6	20	74
6	6	25	62
6	6	30	64
6	6	35	68
6	7	15	78
6	7	20	74
6	7	25	78
6	7	30	74
6	7	35	78

Table 3.2 | Changing parameters from the *Syn-MRL* pipeline and it's effect on Robinson-Foulds distance to True Tree

These results indicate that the *Syn-MRL* pipeline struggles to accurately reconstruct the deeply diverged relationships of the bilaterian phylogeny. There are several factors that may have contributed to this failure. First and foremost, while we made efforts to encompass all major sub-phyla of protostomes and deuterostomes in our tree reconstruction, the lack of high-quality genome data for any Xenacoelomorpha species meant that this subphyla was not included. Furthermore, despite successfully finding at least one genome from all of the other major subphyla, it is evident that there is an insufficient availability of high-quality genome assemblies for certain less-studied metazoan groups making it impossible to achieve a comprehensive taxon sampling of all major groups. Other potential

issues may be lack of an appropriate model for such long divergence times. The MK model[2.1.2], used here, is the only model available on IQ-TREE (Minh et al., 2020) for binary data and lacks the complexity necessary for reconstructing such ancient relationships (this will be discussed further in Section 3.4).

3.3.2 Investigating the relationship between the rate of sequence evolution and gene order evolution

In this section we focus our attention on the branch lengths of these novel trees. In Equation 3.3.2, branch lengths in a tree constructed based on the molecular clock hypothesis, where the rate of molecular evolution is assumed to be relatively constant over time, is equal to the divergence time (in units of years or millions of years) multiplied by the rate at which genetic changes (e.g., nucleotide substitutions or gene order changes) accumulate along a branch. By concentrating on the branch lengths of trees constructed with either sequence data or micro-synteny information, and using fixed topologies, we are probing the possible link between the evolutionary patterns of molecular sequences and the dynamics of gene order or gene movement within the genomes of particular groups of species.

$$Branch\ Length\ Correlation\ =\ \frac{Evolutionary\ rate\ Synteny\ X\ Divergence\ time}{Evolutionary\ rate\ SA\ X\ Divergence\ time} \tag{3.2}$$

In Zhao and Schranz, 2019, they built highly accurate phylogenies of mammals and angiosperms using their micro-synteny based method of phylogenetic reconstruction, *Syn-MRL* which we have discussed in detail in this chapter. Rather than searching for alternative lineages, where we could not be sure the pipeline

would be capable of accurate inference, we chose to utilize these well-constructed phylogenies of mammals and angiosperm to explore the potential correlation between branch lengths in species trees based on synteny information and those built using sequence data.

To reconstruct the SA trees for angiosperm and mammals, we downloaded a similar dataset to that used in Zhao and Schranz, 2019 from NCBI genome table. A list of the angiosperm and mammal genomes used in this analysis can be found in Table 3.4 and Table 3.3. The trees built with *Syn-MRL* pipeline (Zhao et al., 2021) in this work were adapted versions of the phylogenies published in Zhao and Schranz, 2019 as there were some mammal and angiosperm genomes from their study that we could not locate on NCBI (Zhao and Schranz, 2019. They used 87 mammalian and 107 angiosperm genomes and in this work we used 82 mammalian (Table 3.3) and 99 angiosperm (Table 3.4) genomes). SA trees of these species, with widely accepted branching patterns were used as the fixed topology for all tree inferences in this phase of the study.

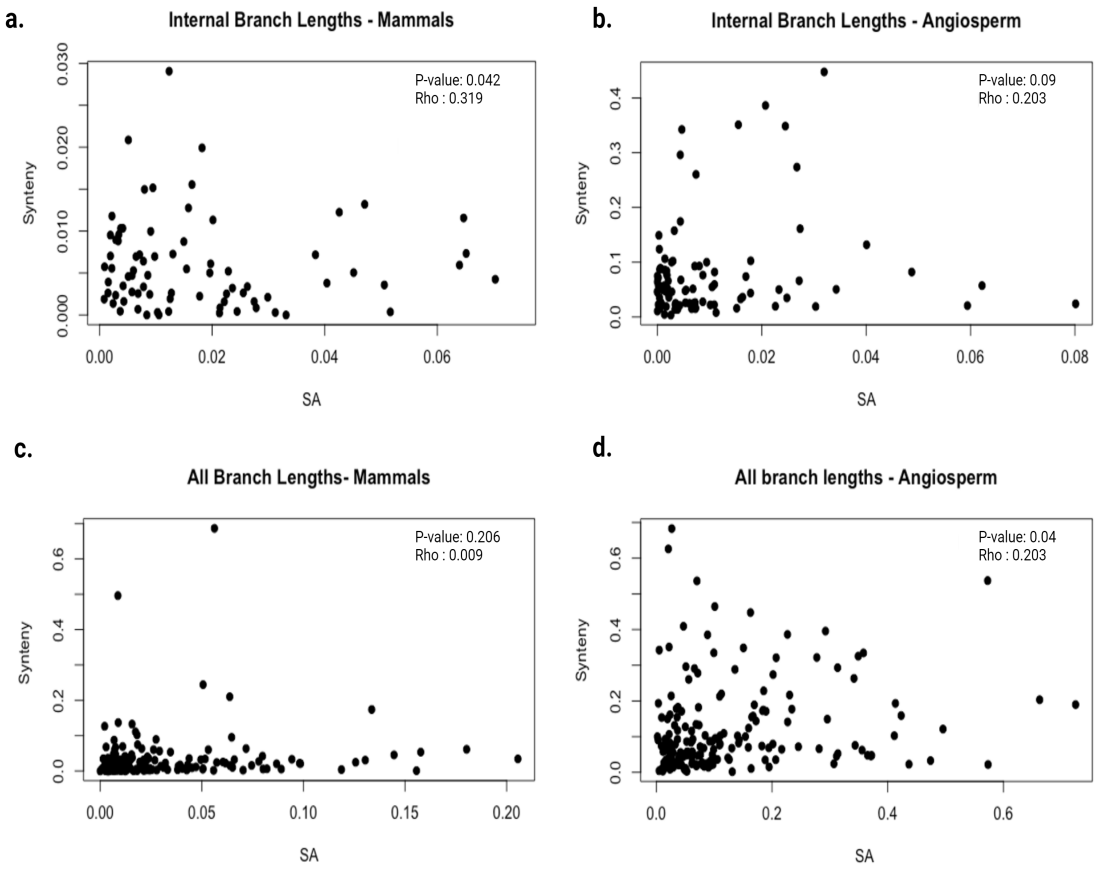

Figure 3.2 | Correlation analysis of branch lengths from SA trees and micro-synteny-based trees of Angiosperm and Mammal phylogenies. (A) and (B) show the correlation between the internal branch lengths in the angiosperm and mammals SA tree versus the micro-synteny-based tree. In (C) and (D), terminal branch lengths are included in the correlation analysis. We include Spearman's rank correlation coefficient (Rho) for each plot, which measures the strength and direction of the monotonic relationship between two variables. The level of statistical significance for Spearman's Rho was determined using a p-value cutoff of less than 0.05.

By employing a consistent species dataset and a fixed tree topology for all tree reconstructions, we effectively removed the divergence time component from Equation 3.3.2. This allowed us to focus exclusively on the rates of evolution of the characters. Correlation plots in Fig3.2, show direct comparisons of the evolutionary rates of these characters; gene order and molecular sequence. No-

tably, in Fig3.2 (A) and (B), where we plot the internal branch lengths of the micro-synteny and SA trees in angiosperms and mammals against one another, we observe a weak positive relationship between these rates in mammals and find no such evidence for a relationship in angiosperms. The level of statistical significance for Spearman's rank (Rho) was determined using a p-value threshold of less than 0.05. Similarly, in Figure 3.2 (C) and (D), we used the same dataset but included terminal branches from all phylogenies. While there is limited support for a relationship in angiosperms, this evidence is considered unreliable due to concerns about the accuracy of including terminal branches (discussed further in Section 3.4). Overall, our findings indicate little to no statistically significant relationships between the datasets for either lineage.

3.3.3 Gene presence/absence phylogenies

In this section, we look at another character that can be used for tree reconstruction; gene presence or absence. The phylogenies were built using an in-house script available at https://github.com/pnatsi/orthocounts2bin (Natsidis et al., 2021). The binary matrix generated by this script assumes that the 1's and 0's represent the presence and absence of genes within sets of genomes, allowing us to discern the acquisition and loss of genes throughout the course of evolution. Presence/absence phylogenies built with this method have been shown to be highly congruent with sequence-based phylogenies (Natsidis et al., 2021). Here, the accuracy to which a phylogeny can be built is less important as we are more interested in the branch lengths of the trees produced. The trees were built in IQ-TREE (Minh et al., 2020)with model $MK^{2.1.2}$ and as before, the SA tree for each group of species was set as the fixed topology.

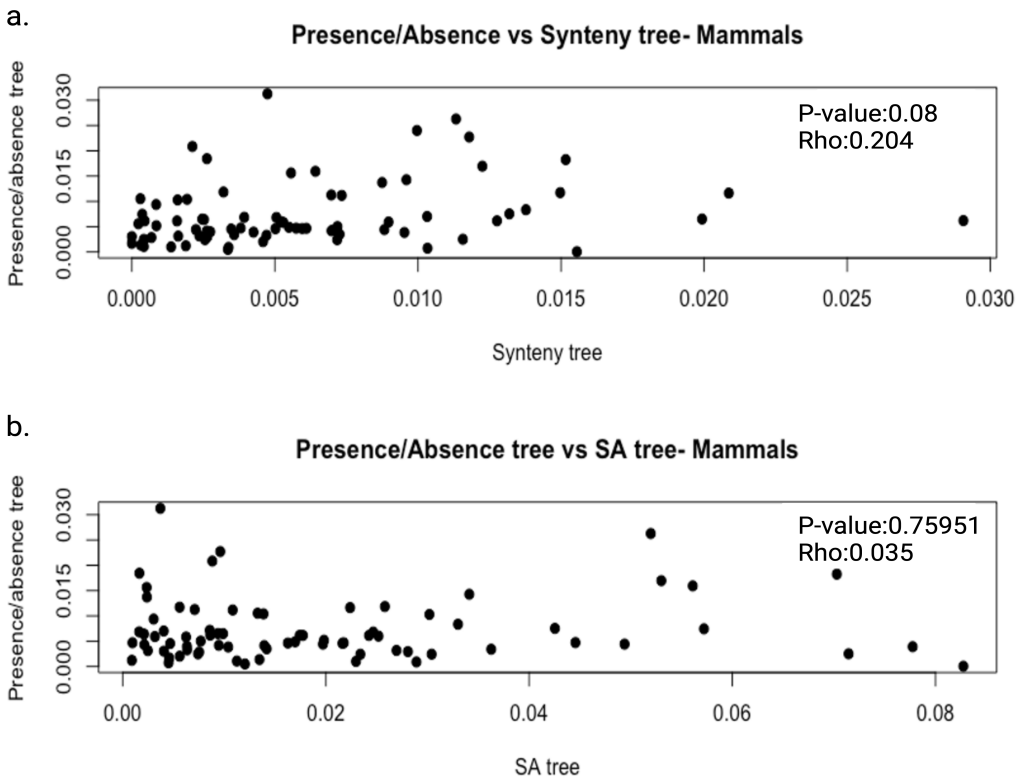

Figure 3.3 | Correlation analysis of branch lengths in gene presence/absence phylogenies compared to SA and micro-synteny-based trees. (A) illustrates the branch lengths in a mammalian phylogeny derived from gene presence/absence data versus the branch lengths in a synteny-based phylogeny. (B) shows presence/absence BL versus SA BL for a fixed mammal phylogeny. We include Spearman's rank correlation coefficient (Rho) for each plot, which measures the strength and direction of the monotonic relationship between two variables. The level of statistical significance for Spearman's Rho was determined using a p-value cutoff of <0.05.

In Fig3.3 we see the rate of gene gain and loss in mammals is not correlated with the rate of gene order evolution. There is also no apparent relationship found between SA branch lengths and presence/absence branch lengths (Fig3.3 (B), p-value = 0.681 for Spearman's Rho). It's worth noting that presence/absence phylogenies have previously shown strong congruence with SA trees, so the finding is unexpected (Natsidis et al., 2021). In Fig3.4 (A) and (B), once again, we show that there is not a statistically significant relationship between the branch lengths

in an angiosperm presence/absence phylogeny and the branch lengths of a micro-synteny-based phylogeny. However, we do identify a marginal level of significance in Fig3.4(B). Nevertheless, the lack of a relationship between the evolutionary dynamics of gene gain and loss and the rearrangement of genes within a genome in angiosperms and mammals represents a noteworthy discovery. This study marks the first time such a relationship has been explored.

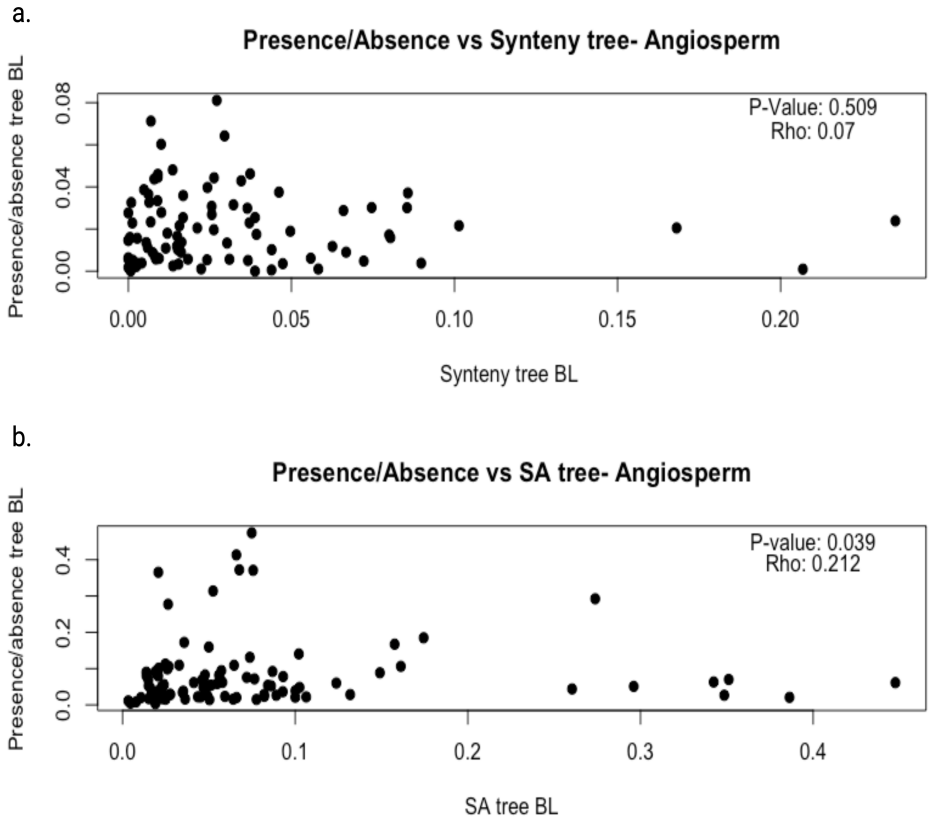

Figure 3.4 | Correlation analysis of branch lengths in a gene presence/absence phylogeny of angiosperm versus SA and micro-synteny-based phylogeny.(A) illustrates the branch lengths in an angiosperm phylogeny derived from gene presence/absence data versus the branch lengths in a synteny-based phylogeny. (B) shows presence/absence BL versus SA BL for a fixed angiosperm phylogeny. We include Spearman's rank correlation coefficient (Rho) for each plot, which measures the strength and direction of the monotonic relationship between two variables. The level of statistical significance for Spearman's Rho was determined using a p-value cutoff of <0.05.

3.3.4 Molecular sequence evolution and gene order evolution in mammals through time

Finally, we look at how these characters have evolved through time in mammals and angiosperm. In the previous section we used Equation 3.1, a formula for branch lengths in a tree based on the molecular clock hypothesis, to find the rate. We deliberately excluded the *divergence time* element by utilizing a fixed topology. In the present context, we reintroduce the concept of divergence time, which provides a temporal dimension to our analysis. This time component is then plotted alongside the rate, allowing us to gain deeper insights into the evolution of these phylogenetic characters. To incorporate divergence time, we require the use of timetrees.

Building a high-confidence timetree is a complex undertaking, presenting various challenges. Fortunately, there are well-constructed and appropriate timetrees available for mammals and angiosperms in the existing literature, simplifying this aspect of our work (Álvarez-Carretero et al., 2021; Ramírez-Barahona et al., 2020). In most of the analysis, we found that the rates of both molecular and gene order evolution appear to accelerate toward the present (Fig3.5, Fig3.6, Fig3.7).

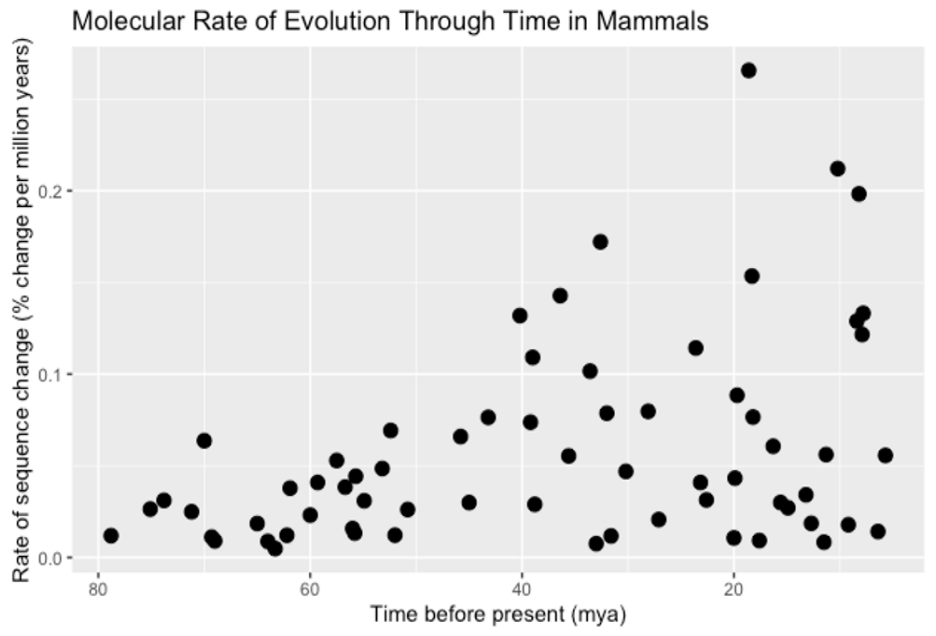

Figure 3.5 | Rates of molecular evolution through time in Mammals The X-axis is reversed showing past to present. Y-axis is the rate of evolution as percentage of character state changes occurring per million years (rate of 1.00 = 100% character state changes per million years)

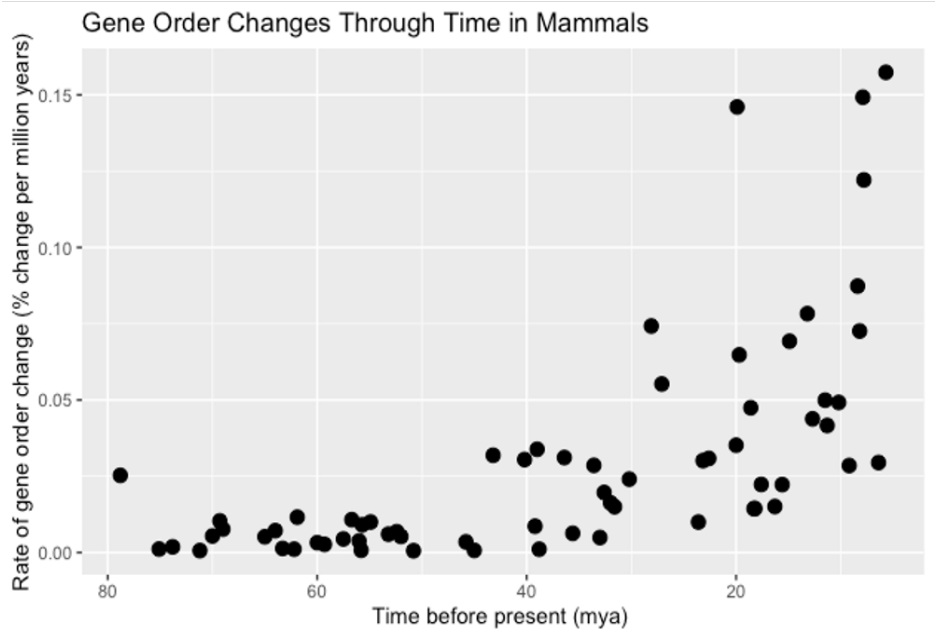

Figure 3.6 | Rates of gene order evolution through time in Mammals The X-axis is reversed showing past to present. Y-axis is the rate of evolution as percentage of character state changes occurring per million years (rate of 1.00 = 100% character state changes per million years)

For the mammalian phylogeny, as discussed in chapter 2 we used Álvarez-Carretero et al., 2021, a timetree of 4,705 mammal species built using a Bayesian molecular-clock dating approach. We pruned the tree to include only the species from our dataset and we illustrate the relationship between time and evolutionary rates using correlation plots which enable us to visually examine how these characters have evolved over time (see timetree Fig.3.8). In Figure 3.5, a distinct pattern emerges, revealing that rates of molecular evolution remained relatively constant during much of early mammalian evolution followed by a significant increase in rates on certain internal branches during the mid-late Eocene period (approximately 55-33.9 million years ago) and a continued sharp rise towards the present. A similar trend is observed in Figure 3.6, where we chart the rate of gene order changes through time. In this case, the increase in rates is less abrupt, with a noticeable gradual rise starting around 40 million years (Eocene epoch). These critical time points correspond with the Paleocene-Eocene radiation, which concluded around 34 million years ago(Jaramillo et al., 2010). This radiation event played a pivotal role in establishing all the major lineages of placental and marsupial mammals that exist today (Luo, 2007).

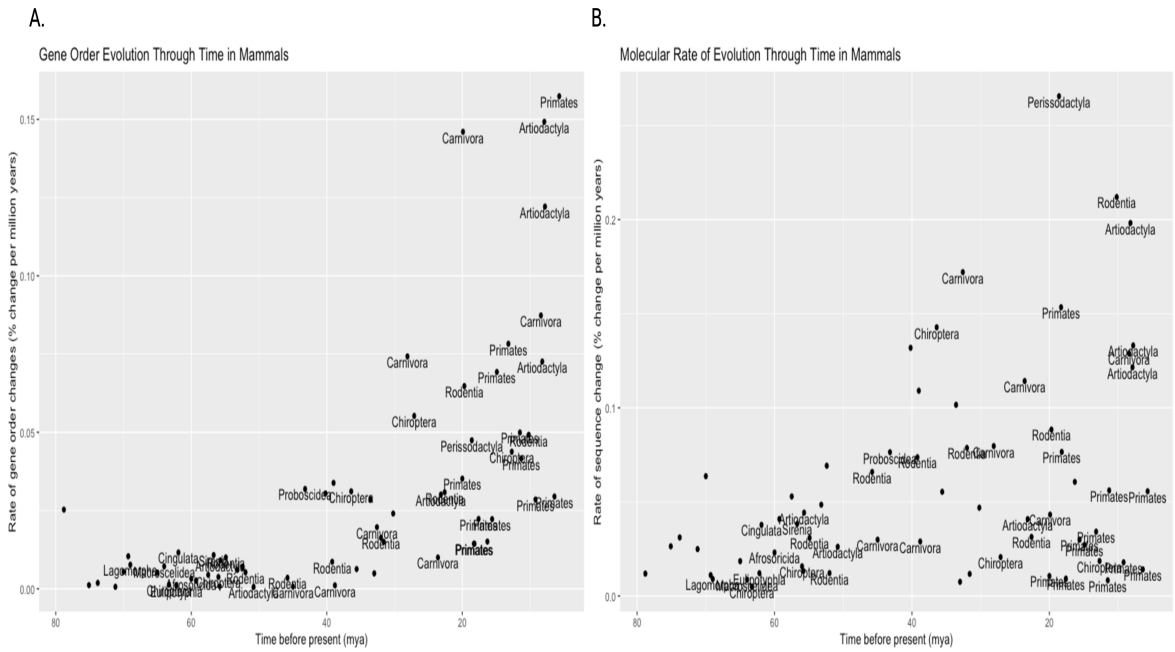

Figure 3.7 | Rates of character evolution through time with Mammalian order information. The X-axis in both plots is reversed (past to present). Y-axis is the rate of evolution as percentage of character state changes occurring per million years (rate of 1.00 = 100% character state changes per million years) (A) depicts molecular rate against time (mya), while (B) illustrates the rate of gene order evolution against time (mya). The orders presented in this plot are detailed alongside species information in Table 3.3.

In Fig3.7, we have labeled the orders associated with the branches. Since we have excluded the terminal branches from the gene order analysis, the labels on the plot (Fig3.7) are linked to the data points for the branches preceding the terminal branches but carry the order identities of the extant taxa represented by those terminal branches. The plot clearly demonstrates that no single order has experienced an isolated, abrupt increase in its rate over the past 40 million years. Instead, the acceleration in rates appears to be a universal phenomenon among all species. While we suspect that the gradual rise during the mid-late Eocene period (55-33.9 million years ago) may be linked to the species radiation known to have occurred during that epoch (Luo, 2007), we believe that the sharp rise observed closer to the present could be an artifact. This is in line with well-established

observations that rates of molecular evolution tend to scale with time (Harmon et al., 2021). Numerous analyses focusing on rates over time have unveiled a time-dependent bias in molecular rate estimates, likely resulting from various contributing factors (Gingerich, 1983; Henao Diaz et al., 2019). We will delve deeper into these factors in the discussion on this chapter.

3.3.5 Molecular sequence evolution and gene order evolution in angiosperm through time

In angiosperm, as in mammals, we see a similar acceleration in rates towards present. We used the time tree for flowering plants from Ramírez-Barahona et al., 2020 and pruned it to include only the species of interest in this study. The tree was built using a Bayesian uncorrelated log-normal clock model implemented in BEAST (Bouckaert et al., 2014). In this tree, a comprehensive set of 238 angiosperm fossils was used for calibration and here, we utilized the upper-bound estimate for dating the nodes. In Fig3.9, we see that while there is an acceleration in rates of gene order evolution towards the present, it is only on certain branches. This stands in contrast to the observed acceleration in molecular sequence evolution over time (see Fig3.10), with most orders showing universal increases starting around ∼50mya and sharply rising to the present day.

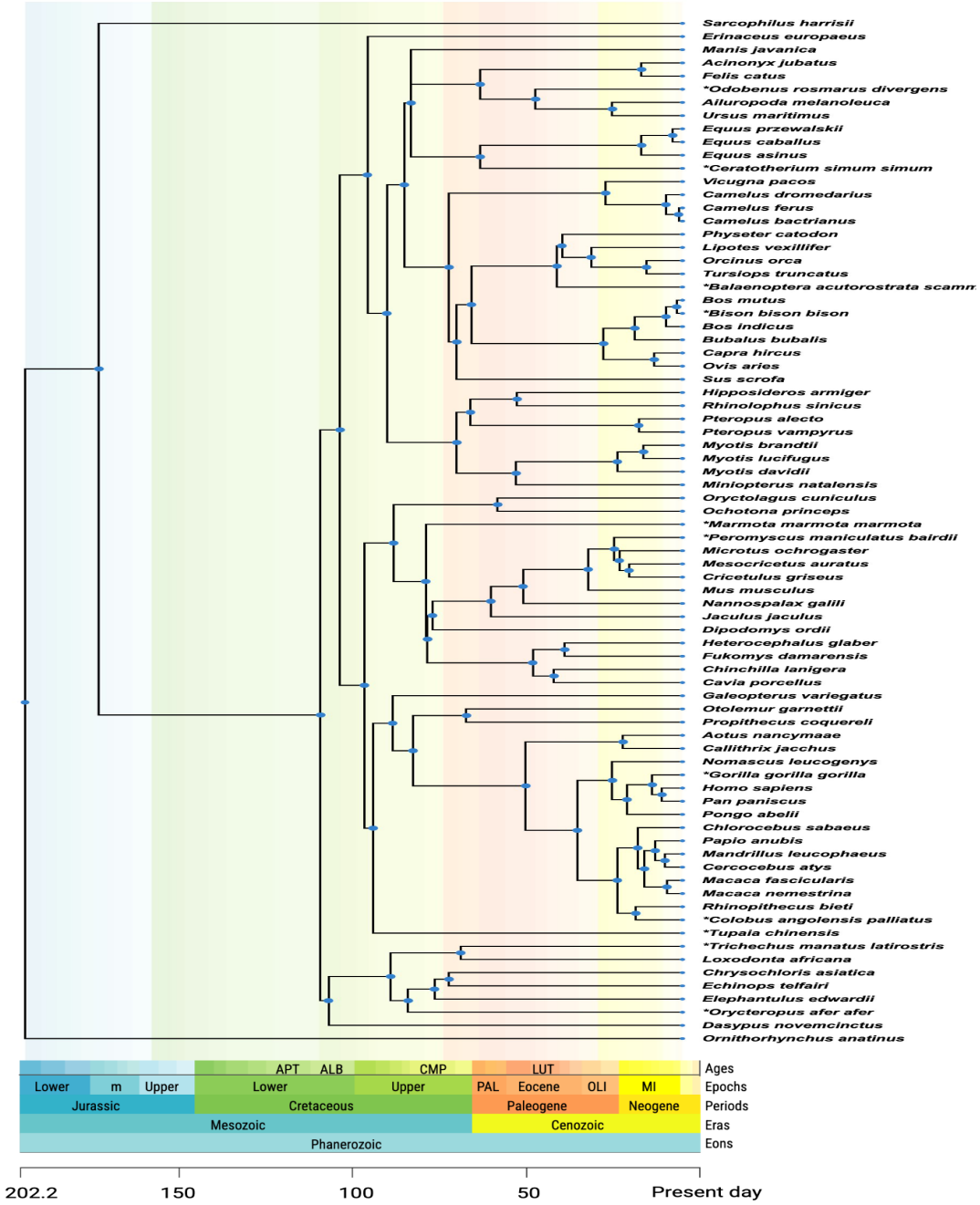

Figure 3.8 | Mammalian time tree used in analysis. Times were taken from Álvarez-Carretero et al., 2021. The branch lengths on the tree are proportional to time. We used the upper bound estimates to date the nodes but for information on mean stem and crown ages along with the 95% highest posterior density credibility interval see Álvarez-Carretero et al., 2021.

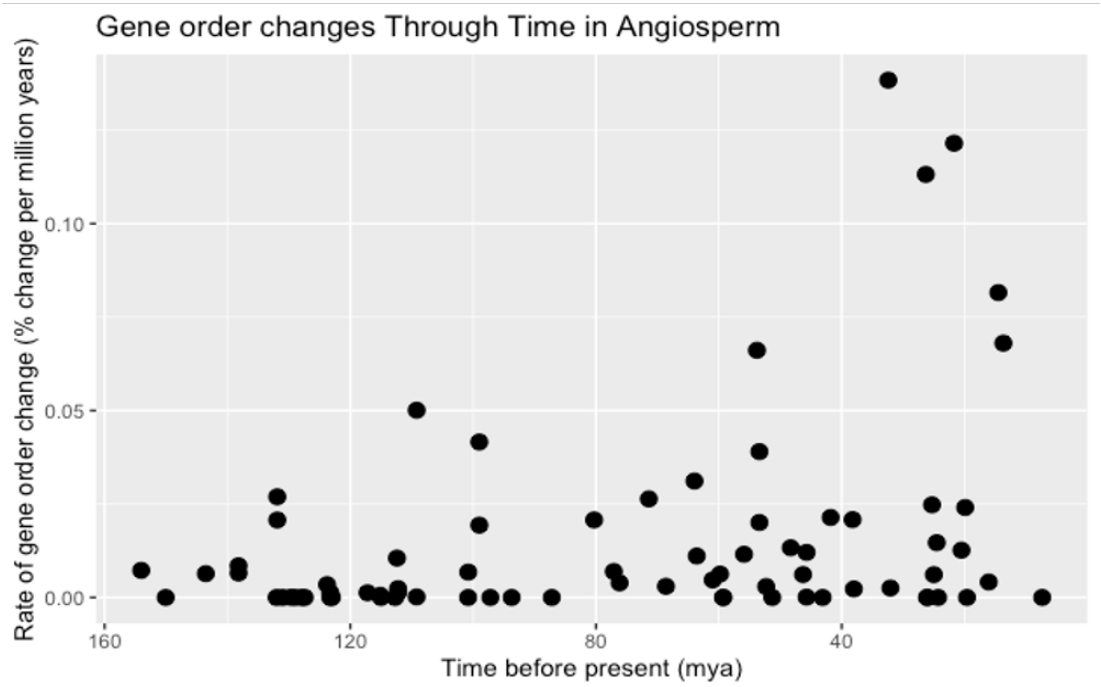

Figure 3.9 | Rates of gene order evolution through time in angiosperm The X-axis is reversed showing past to present. Y-axis is the rate of evolution as percentage of character state changes occurring per million years (rate of 1.00 = 100% character state changes per million years)

Angiosperms are known for the many ploidy events that have occurred throughout their evolution. Previous work on synteny in angiosperm has shown that their genomes are fractionated and reshuffled (Zhao and Schranz, 2019). Although it's reasonable to assume that ploidy events might result in gains, losses, and dynamic alterations within the genomes of the affected lineages, this does not appear to be mirrored in the rate of gene order evolution over time. (Fig.3.9)(Adams and Wendel, 2005; Vanneste, Baele, et al., 2014; Slotkin and Martienssen, 2007). The rise of flowering plants took place during the early Cretaceous (∼145–100mya) (Heimhofer et al., 2005). While it seems the rate of early angiosperm evolution proceeded at a steady pace, our results indicate a major increase in rate of, in particular, molecular evolution from ∼60mya to present day. This is in line with studies that indicate potential time-lags between lineage origination and species

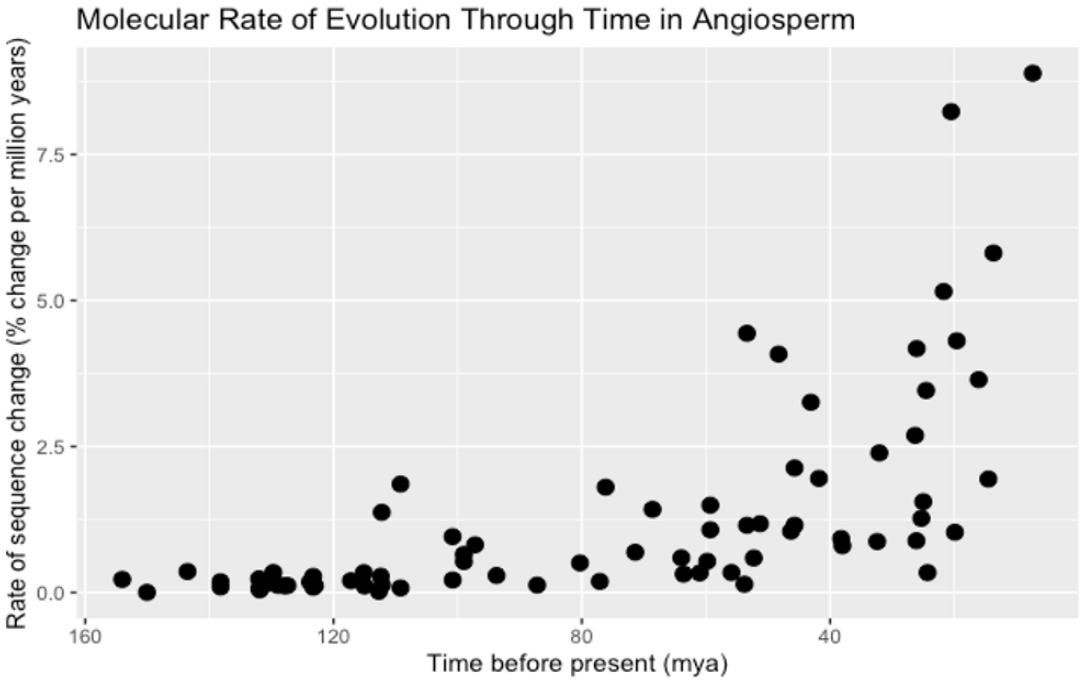

Figure 3.10 | Rates of molecular evolution through time in angiosperm The X-axis is reversed showing past to present. Y-axis is the rate of evolution as percentage of character state changes occurring per million years (rate of 1.00 = 100% character state changes per million years)

diversification that may be associated wth WGD events in these lineages(Vanneste, Maere, et al., 2014; Robertson et al., 2017; Eric Schranz et al., 2012; Lien et al., 2016). Some studies have shown that while most of the major lineages originated during the warmest phases of the Cretaceous(∼100-90Ma), many of them did not start to diversify until the start of the Cenozoic (∼66mya-*present*) and during the global warming of the Palaocene and Eocene(∼56mya) (Ramírez-Barahona et al., 2020; Jaramillo et al., 2010; Zachos et al., 2008; Heimhofer et al., 2005). This pattern is reflective in Fig3.9 and more noticeably in Fig3.10. However, we also consider the possibility that these results are an artifact, as in mammals, due to the time-dependence of rate estimates (see Section 3.4).

List of mammal genomes used in Chapter 3

Species	**Order**
Echinops telfairi	Afrosoricida
Chrysochloris asiatica	Afrosoricida
Vicugna pacos	Artiodactyla
Tursiops truncatus	Artiodactyla
Sus scrofa	Artiodactyla
Physeter catodon	Artiodactyla
Ovis aries	Artiodactyla
Orcinus orca	Artiodactyla
Lipotes vexillifer	Artiodactyla
Capra hircus	Artiodactyla
Camelus ferus	Artiodactyla
Camelus dromedarius	Artiodactyla
Camelus bactrianus	Artiodactyla
Bubalus bubalis	Artiodactyla
Bos mutus	Artiodactyla
Bos indicus	Artiodactyla
Bison bison bison	Artiodactyla
Balaenoptera acutorostrata scammoni	Artiodactyla
Ursus maritimus	Carnivora
Panthera tigris altaica	Carnivora
Odobenus rosmarus divergens	Carnivora
Felis catus	Carnivora

List of mammal genomes used in Chapter 3

Species	**Order**
Canis lupus familiaris	Carnivora
Ailuropoda melanoleuca	Carnivora
Acinonyx jubatus	Carnivora
Rhinolophus sinicus	Chiroptera
Pteropus vampyrus	Chiroptera
Pteropus alecto	Chiroptera
Myotis lucifugus	Chiroptera
Myotis davidii	Chiroptera
Myotis brandtii	Chiroptera
Miniopterus natalensis	Chiroptera
Hipposideros armiger	Chiroptera
Eptesicus fuscus	Chiroptera
Dasypus novemcinctus	Cingulata
Sarcophilus harrisii	Dasyuromorphia
Galeopterus variegatus	Dermoptera
Tupaia chinensis	Eulipotyphla
Erinaceus europaeus	Eulipotyphla
Oryctolagus cuniculus	Lagomorpha
Ochotona princeps	Lagomorpha
Elephantulus edwardii	Macroscelidea
Ornithorhynchus anatinus	Monotremata
Equus przewalskii	Perissodactyla

List of mammal genomes used in Chapter 3

Species	**Order**
Equus caballus	Perissodactyla
Equus asinus	Perissodactyla
Ceratotherium simum simum	Perissodactyla
Manis javanica	Pholidota
Saimiri boliviensis boliviensis	Primates
Rhinopithecus bieti	Primates
Propithecus coquereli	Primates
Pongo abelii	Primates
Papio anubis	Primates
Pan paniscus	Primates
Otolemur garnettii	Primates
Nomascus leucogenys	Primates
Mandrillus leucophaeus	Primates
Macaca nemestrina	Primates
Macaca fascicularis	Primates
Homo sapiens	Primates
Gorilla gorilla gorilla	Primates
Colobus angolensis palliatus	Primates
Chlorocebus sabaeus	Primates
Cercocebus atys	Primates
Callithrix jacchus	Primates
Aotus nancymaae	Primates

List of mammal genomes used in Chapter 3

Species	**Order**
Loxodonta africana	Proboscidea
Peromyscus maniculatus bairdii	Rodentia
Nannospalax galili	Rodentia
Mus musculus	Rodentia
Microtus ochrogaster	Rodentia
Mesocricetus auratus	Rodentia
Marmota marmota marmota	Rodentia
Jaculus jaculus	Rodentia
Heterocephalus glaber	Rodentia
Fukomys damarensis	Rodentia
Dipodomys ordii	Rodentia
Cricetulus griseus	Rodentia
Chinchilla lanigera	Rodentia
Cavia porcellus	Rodentia
Trichechus manatus latirostris	Sirenia
Orycteropus afer afer	Tubulidentata

List of angiosperm genomes used in Chapter 3

Species	**Order**
Lemna minor	Alismatales
Spirodela polyrhiza	Alismatales
Zostera marina	Alismatales

List of angiosperm genomes used in Chapter 3

Species	**Order**
Amborella trichopoda	Amborellales
Daucus carota	Apiales
Elaeis guineensis	Arecales
Phoenix dactylifera	Arecales
Asparagus officinalis	Asparagales
Dendrobium catenatum	Asparagales
Phalaenopsis equestris	Asparagales
Helianthus annuus	Asterales
Lactuca sativa	Asterales
Aethionema arabicum	Brassicales
Arabidopsis lyrata	Brassicales
Arabidopsis thaliana	Brassicales
Boechera stricta	Brassicales
Brassica napus	Brassicales
Brassica oleracea	Brassicales
Brassica rapa	Brassicales
Camelina sativa	Brassicales
Capsella grandiflora	Brassicales
Carica papaya	Brassicales
Cleome gynandra	Brassicales
Leavenworthia alabamica	Brassicales
Schrenkiella parvula	Brassicales

List of angiosperm genomes used in Chapter 3

Species	**Order**
Sisymbrium irio	Brassicales
Tarenaya hassleriana	Brassicales
Thellungiella halophila	Brassicales
Thellungiella salsuginea	Brassicales
Amaranthus hypochondriacus	Caryophyllales
Chenopodium quinoa	Caryophyllales
Citrullus lanatus	Cucurbitales
Cucumis sativus	Cucurbitales
Actinidia chinensis	Ericales
Arachis duranensis	Fabales
Cajanus cajan	Fabales
Cicer arietinum	Fabales
Glycine max	Fabales
Lotus japonicus	Fabales
Lupinus angustifolius	Fabales
Medicago truncatula	Fabales
Phaseolus vulgaris	Fabales
Trifolium pratense	Fabales
Vigna angularis	Fabales
Vigna radiata	Fabales
Castanea mollissima	Fagales
Juglans regia	Fagales

List of angiosperm genomes used in Chapter 3

Species	**Order**
Coffea canephora	Gentianales
Mimulus guttatus	Lamiales
Sesamum indicum	Lamiales
Utricularia gibba	Lamiales
Jatropha curcas	Malpighiales
Linum usitatissimum	Malpighiales
Manihot esculenta	Malpighiales
Populus trichocarpa	Malpighiales
Ricinus communis	Malpighiales
Gossypium raimondii	Malvales
Theobroma cacao	Malvales
Eucalyptus grandis	Myrtales
Xerophyta viscosa	Pandanales
Aegilops tauschii	Poales
Ananas comosus	Poales
Brachypodium distachyon	Poales
Hordeum vulgare	Poales
leersia perrieri	Poales
Oropetium thomaeum	Poales
Oryza glaberrima	Poales
Oryza rufipogon	Poales
Oryza sativa	Poales

List of angiosperm genomes used in Chapter 3

Species	**Order**
Phyllostachys heterocycla	Poales
Setaria italica	Poales
Sorghum bicolor	Poales
Triticum aestivum	Poales
Triticum urartu	Poales
Zea mays -V4	Poales
Nelumbo nucifera	Proteales
Fragaria vesca	Rosales
Humulus lupulus	Rosales
Malus domestica	Rosales
Morus notabilis	Rosales
Prunus mume	Rosales
Prunus persica	Rosales
Pyrus x bretschneideri	Rosales
Rubus occidentalis	Rosales
Ziziphus jujuba	Rosales
Citrus sinensis	Sapindales
Capsicum annuum	Solanales
Ipomoea nil	Solanales
Nicotiana attenuata	Solanales
Nicotiana benthamiana	Solanales
Nicotiana sylvestris	Solanales

List of angiosperm genomes used in Chapter 3

Species	**Order**
Nicotiana tomentosiformis	Solanales
Petunia axillaris	Solanales
Solanum lycopersicum	Solanales
Solanum melongena	Solanales
Solanum pennellii	Solanales
Solanum tuberosum	Solanales
Vitis vinifera	Vitales
Musa acuminata	Zingiberales

3.4 Discussion

In the first part of Chapter 3, we elucidated the limitations of the *Syn-MRL* pipeline for phylogenetic inference using micro-synteny information. The attempt to reconstruct the bilaterian phylogeny proved challenging, given the deep divergence times between the taxa involved, making it difficult to find accurate relationship information (Fig 3.1). The phylogeny depicted in Fig 3.1 appears to contain significant errors. Notably, there is an unexpected placement of xenambulacrarian species, *Strongyloocentrotus purpuratus*, which branches alongside protostome genomes. While the analysis successfully resolves closer relationships such as those within primates and Branchiostoma genomes, it fails to accurately depict the broader branching patterns of the entire phylogeny. We consider several factors which may contribute to this challenge. Firstly, the success of micro-synteny studies, as well as broader phylogenetic investigations, hinges on the availability of

high-quality genomes (Liu et al., 2018). While significant advancements have been made in genome sequencing for a wide range of animal groups, there remains a scarcity of high-quality genomes for various branches within the bilaterian tree, as evidenced by the absence of scaffold-level or higher quality genomes for xenacoelomorpha. This deficiency extends to other less explored phyla, including Annelida, Urochordata, and Hemichordata (see 3.1). The selection of taxa is critical for phylogenetic research, as incomplete, biased, or inadequately sampled taxa can lead to misleading results in the reconstruction of evolutionary relationships (Nabhan and Sarkar, 2012; Plazzi et al., 2010). Not having a representative assembly for an entire group or possessing low-quality genomes for significant phyla can hinder the accuracy and reliability of the phylogenies built with micro-synteny.

Secondly, the adoption of the Mk model (Lewis, 2001) in probabilistic analysis of phylogenetic characters has remained a subject of controversy. Concerns have been raised regarding whether such a simplistic framework can accurately account for the intricacies of evolution, whether it pertains to phenotypic changes or gene order evolution (Brown et al., 2017). The Mk model (Lewis, 2001) is a direct analog of the Jukes-Cantor (JC) model (Cantor and Jukes, 1966) for sequence evolution and applies to discrete characters with k unordered states. The model is symmetrical, and assumes that the rate of change from one character state to another is equal to the rate of reversal, meaning the probability of changing from 0 to 1 is the same as changing from 1 to 0. Unfortunately, this assumption might not hold for all gene order changes over time and introducing any form of asymmetrical rates of change is challenging. Adding complexity to models for nucleotide data is comparatively simpler because a nucleotide base, such as "T," shares the same properties across the alignment. This consistency from site-site to means that you can allow the nucleotide to have varying exchangeabilities across datasets as a function of the specific property of that nucleotide base. Micro-synteny data

lacks this consistency and the character definition is somewhat subjective. While we define a block based on a set of rules, these rules may vary from study to study, and converting this data to a binary matrix of 0s (indicating the absence of a micro-synteny block in the species) or 1s (indicating the presence of a micro-synteny block) lacks specificity. The Mk[2.1.2] model assumes that a 1 in one part of the binary alignment is equivalent to a 1 in another part, but this assumption can be unreliable. A change from state 1 to state 0 could represent the loss of a single syntenic ortholog in one block or could be a significant change, like multiple rearrangements in another and yet, it is treated the same in the model. While there are ways of adding complexity to the MK model[2.1.2] in programs like MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003 and IQ-TREE (Minh et al., 2020) by adding priors that allow for heterogeneity in character change symmetry this usage still provides an imperfect representation of the evolution of gene order in animals (Alekseyenko et al., 2008). Nevertheless, as shown here, the Mk model has proven to be a reliable tool for reconstructing relationships of well studied and closely related groups (Zhao and Schranz, 2019; Drillon et al., 2020). However, its efficacy in accurately resolving relationships may be limited for taxa that have undergone extensive divergence.

In the second part of this analysis, we explore the connection between gene order evolution and molecular sequence evolution. When comparing branch lengths in fixed-tree topologies, we observe that there is generally a weak or negligible relationship between these two traits in both angiosperm and mammals. In Figure 3.2 (C) and (D), where we include terminal branches, we do detect some signal in angiosperm datasets however, it's important to note that including terminal branches, especially when dealing with binary data and using the MK[2.1.2] model, can lead to ascertainment bias due to potential rate underestimations at terminal branches (Caldas and Schrago, 2019). As discussed, the primary use of the Mk

model[2.1.2] is in reconstructing the evolution of discrete-state morphological characters. Should there be any derived characters (autapomorphies)in extant taxa that are not accounted for, there is a risk of underestimating rates at terminal branches (Lewis, 2001). While we are not using morphological data here, the *Syn-MRL* pipeline transforms the micro-synteny data into discrete 0 and 1's so that if there are regions in the genome that experience more gene movement and rearrangement in extant taxa, these may not be accounted for. This leads to an underestimation of rates at terminal branch lengths relative to internal branches. For that reason, any major changes in the result's significance when terminal data points are introduced, will be treated dubiously.

Overall, the results imply that fluctuations in mutation rates do not significantly influence the movement of genes within the genome and vice-versa. While no significant relationship has been found here, perhaps over longer timescales and with more appropriate modelling this result could change. Future research that integrates morphological data into analysis looking at character relationships would add an extra layer of understanding. However, it's worth noting that there is currently a shortage of comprehensive and dependable databases containing morphological data for mammals, angiosperms, and other lineages. This limitation has posed challenges in prior studies examining the correlation between these traits (Bromham et al., 2002; Omland, 1997; Halliday et al., 2019).

In the final part of this study, we investigated the temporal evolution of these character rates. Our findings reveal a common trend in both angiosperms and mammals: a relatively steady rate during early lineage evolution, followed by a sharp increase in rates as we approach the present day. In the case of mammals, this initial rapid rise, occurring around 40 million years ago during the Eocene epoch (as indicated in Fig.3.5 and Fig.3.6), coincides with the Paleocene-Eocene radiation event that concluded approximately 34mya (Jaramillo et al., 2010; Luo,

2007). This radiation event played a crucial role in establishing the major lineages of placental and marsupial mammals existing today (Luo, 2007; Álvarez-Carretero et al., 2021). For angiosperms, our results (Fig.3.11) demonstrate a notable increase in the rate of molecular evolution from ∼60 million years ago to the present day. This increase might be associated with the diversification of flowering plants during the Cenozoic era (∼66 million- present) and the global warming of the Paleocene and Eocene epochs (∼56 million years ago) (Zachos et al., 2008; Jaramillo et al., 2010; Ramírez-Barahona et al., 2020; Heimhofer et al., 2005).

While we speculate about the potential alignment of these trends with the major geographical and environmental shifts noted above, we also acknowledge the possibility that this is an artifact. Evolutionary rate measures have consistently shown a pattern of scaling across various timescales, both micro- and macro- (Harmon et al., 2021). Over shorter timescales, rates tend to appear faster when compared to longer time intervals, as observed in numerous studies and in our own work here (Fig.3.5, Fig.3.6, Fig.3.11) (Gingerich, 1983; Ho et al., 2005; Ho et al., 2011).These studies suggest that rate estimates are accelerated for recent divergences however, it remains challenging to discern whether this pattern is driven by rate estimation errors or is a consequence of model misspecification or something else. While the concept of evolution perpetually accelerating is a possibility, it is also quite possible that this pattern is an artifact stemming from our limitations in accurately measuring the evolutionary process. An essential argument against the former explanation is that we observe this trend not only when considering extant taxa but it's also seen within the fossil record and on removal of extant taxa(Gingerich, 1983; Harmon et al., 2021; Henao Diaz et al., 2019).

Throughout Chapter 3, we have demonstrated that the use of micro-synteny in phylogenetic reconstruction has its inherent limitations. The effectiveness of synteny-based approaches relies on well-constructed genome assemblies, which remain

scarce, for many lesser-studied metazoan lineages. While we have not observed a statistically significant relationship between the rate of gene order evolution and molecular sequence evolution within lineages, an intriguing similarity emerges when we trace their evolutionary trajectories over time. More work needs to be done in order to unravel whether or not this trend is in fact real, mapping molecular and genomic changes with geographical and environmental shifts or if the the acceleration in rates witnessed for both characters — gene order and sequence evolution — may be attributable to model misspecification or due to inaccuracies in rate estimation.

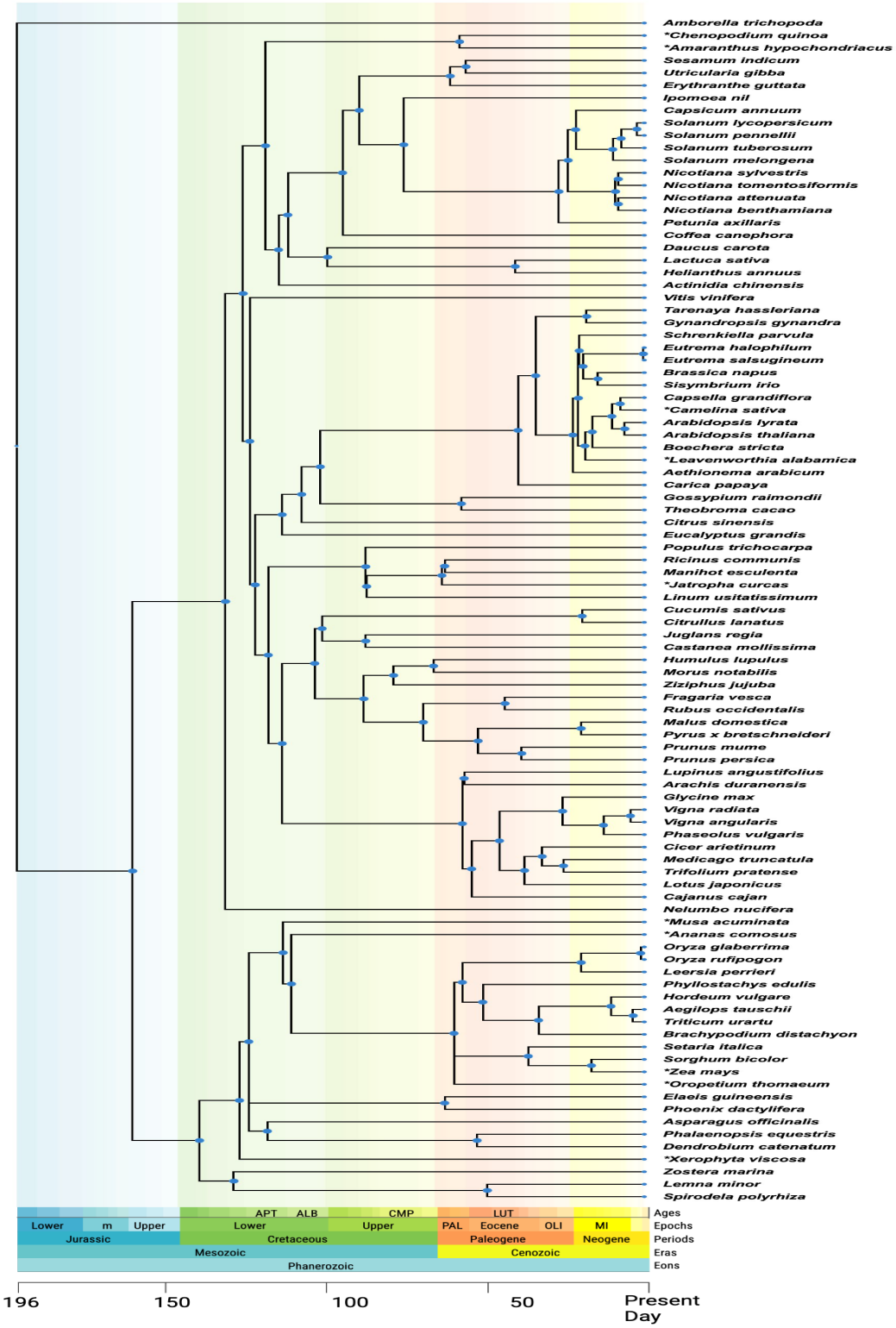

Figure 3.11 | The Angiosperm timetree utilised in these analyses is based on time estimates sourced from Ramírez-Barahona et al., 2020. The branch lengths on the tree are proportional to time. Times (in Ma) were estimated with MCMCtree(Rannala and Yang, 2007) using approximate likelihood. We use the upper-bound estimate from Ramírez-Barahona et al., 2020 as the age of the node but for information on all mean stem and crown family ages, along with their 95% highest posterior density intervals see Ramírez-Barahona et al., 2020.

Chapter 4

Reduplicating collapsed reads in the paddlefish assembly

4.1 Introduction

Significant progress has been made in genome sequencing technologies since the completion of the Human Genome Project in the early 2000s. Despite this, we are still a long way from achieving "perfect" and complete genome assemblies for all of life. Sequencing assembly errors are omnipresent in most draft and finished genomes and this can have major impacts on scientific findings (Salzberg and Yorke, 2005; Mardis, 2008). However, there is a sense of optimism on the horizon as genome sequencing methods are becoming more accurate, efficient, and cost-effective. This raises questions: Do we embark on a comprehensive re-sequencing effort of all current data or can we "fix" and enhance existing data?

In this chapter, we attempt to "un-collapse" or "*re-duplicate*" collapsed duplicates in the paddlefish genome following a protocol adapted from Du et al., 2020. The assembly and annotation of the paddlefish genome present a unique challenge due to its complex evolutionary history. The lineage underwent a WGD approximately

254.7-241.8 million years ago, occurring in the ancestor of the Acipenseriformes lineage. As a result, it shares this evolutionary event with sturgeons (Redmond et al., 2023). Post-WGD, the ancestor of these extant acipenseriformes embarked on a complex transition from its ancestral tetraploid state to a functional diploid, in an asynchronous process of rediploidization. Surprisingly, despite the considerable time that has elapsed since this evolutionary event, certain regions within the genomes of both extant species seem to still exhibit signatures of tetrasomic inheritance. This complex evolutionary history creates challenges for assembly and annotation, often resulting in collapsed repeat regions. Evidence of this issue is observed in read-depth analysis, where large segments of the genome exhibit double the expected read coverage, a telltale sign of collapsed duplicates and misassembly (see Fig.1.5). To address these faults, we adopt a hybrid approach that combines Illumina short-read data and PacBio long reads to 'uncollapse' these duplicated regions and ultimately enhances the paddlefish assembly.

It is unclear to what extent published genome assemblies have been affected by the inherent error-prone nature of second-generation technologies i.e. NGS. While, NGS has greatly enhanced our ability to sequence DNA at high throughput and for a low cost, sequencing errors and misassemblies are commonplace due to intrinsic errors in NGS methods (see Fig1.4). As discussed in detail in Section 1.2, a major challenge for sequencing and assembling diploid and polyploid non-model organisms is inaccurate resolution of duplicates, repeats and haplotypes.(Salzberg and Yorke, 2005; Tørresen et al., 2019). While, TGS methods offer some relief from issues with resolving repetitive DNA, it is still not customary to carry out long-read sequencing (see Fig1.4)(Eid et al., 2009; Rothberg et al., 2011; Quail et al., 2012). Major progress has been made to overcome issues in genome assemblies but there is still a way to go before we can attest a perfect genome sequencing tool.

In recent years, tools have been developed to improve assembly contiguity by unearthing collapsed duplicates and repeat regions from short read assemblies. The predominant method for this appears to be a hybrid approach in which you use both long and short read data during scaffolding (Coombe et al., 2021; Cechova, 2020; Du et al., 2020).Typically, the initial steps involve the alignment of long-reads to short-read contigs. This alignment is then made into a graph which can be traversed to produce scaffolds in which gap sizes are estimated from the linking information (Kronenberg et al., 2021; Zimin and Salzberg, 2022). Most early tools of this kind only worked on small genomes but more recent developments, like Scaffolding Assemblies with Multiple Big Alignments or SAMBA can be used on larger genomes (Zimin and Salzberg, 2022).

In this study, we implemented the approach described in Du et al., 2020. Initially, we identified double-coverage regions from the mapped Illumina short-reads. Subsequently, by employing Freebayes (Garrison and Marth, 2012) and HapCut2 (Edge et al., 2017), we identified genetic variants and reconstructed individual haplotypes within the PacBio long-reads. Regions exhibiting an excess of haplotypes were flagged as potential collapsed reads and split, these split regions were then reassembled back to the reference genome. In Du et al., 2020, this methodology resulted in a higher-quality genome assembly for sturgeon as well as enhanced gene discovery and a larger genome size more fitting with previous evidence (Du et al., 2020; Liu et al., 2013). Our efforts to find collapsed reads in the paddlefish genome not only increased the genome's size but, in combination with improved annotation procedures, revealed a significant number of genes that had not been previously documented.

Short read assemblies will continue to be used despite their shortcomings. Therefore, employing the method described here, which incorporates long reads, can help rectify misassembled short-read genomes and update the data, rather than

completely discarding it. While we are approaching an era in which high quality TGS methods like HiFi (Cheng et al., 2021; Nurk et al., 2020) and Circular Consensus Sequencing (CCS) (Wenger et al., 2019)are becoming more accessible, it is important that the users are aware of the potential errors in many genome assemblies currently available on genome databases (NCBI and ensembl). Continued disregard for these errors will lead to a proliferation of incorrect conclusions throughout the literature.

4.2 Materials & Methods

4.2.1 *Re-duplicating* collapsed duplicates in the paddlefish genome.

Cheng et al., 2020's American Paddlefish assembly has 60 pairs of chromosomes with a genome size of 1.54GB with 26,017 predicted protein-encoding genes. The assembly was sequenced to 30X coverage and short and long reads from this study were deposited in CNGB under project accession number CNP0000867. Cheng et al., 2020 note a smaller than expected genome size following a 17-mer analysis (Liu et al., 2020). Following a similar method by Du et al., 2020, we identified these collapsed regions and attempted to *"reduplicate"* them (Ko et al., 2022; Kelley and Salzberg, 2010; Zhang et al., 2019).

PacBio long-reads were aligned with bwa (v0.7.17-r1198-dirty) (Li and Durbin, 2009) using standard parameters. This was sorted and indexed using SAMtools (v1.16.1) (Bonfield et al., 2021). The short reads were aligned with Bowtie (v2.4.2) (Langmead et al., 2009) using standard parameters and again, indexing and sorting was done with SAMtools (v1.16.1) (Bonfield et al., 2021). To assess the depth of coverage across the genome, the alignments were split into 10kb regions and mosdepth (v0.3.3) (Pedersen and Quinlan, 2018) was used to quantify read depth

at each segment with parameter *–by*. As discussed in Section 1.2, regions of high sequence similarity have been shown to collapse or merge during assembly to the reference genome (Salzberg and Yorke, 2005; Kelley and Salzberg, 2010). The 10kb double coverage segments found using mosdepth (v0.3.3) (Pedersen and Quinlan, 2018) were separated from the rest of the genome for the next steps. Using FreeBayes[4.2.1] (v1.3.6) (Garrison and Marth, 2012), a polymorphism VCF was generated from the short-read alignments. In the next steps, we go back to the PacBio to decipher separate haplotypes in the double coverage regions. Paddlefish PacBio long-reads can be found at CNGB (Cheng et al., 2020). HapCUT2 (Edge et al., 2017), a haplotype assembly tool, was used to reconstruct individual haplotypes in double-coverage mapped long-reads. The assembly tool allows regions with more than one haplotype to be identified, and is used here to decipher potential duplicates from multiple alternative alleles in the double-coverage long-reads BAM file. The inputs for HapCUT2 (Edge et al., 2017) are the mapped double coverage reads (BAM file) and the VCF file. The program only works with VCFs from diploid genomes as phasing is currently not possible for polyploid genomes. Given that parts of the Paddlefish genome appear to be polyploid, according to the VCF, the VCF needed to be filtered of any polyploid or error genotypes (e.g. 4/4, 3/4 etc.) to be used in HapCUT2 (Edge et al., 2017). Using a custom script, we forced the genotype (GT) fields that were for example, 4/4, to be 2/2. This was not a perfect solution but rather than deleting the entry altogether, we removed two of the least common alternative alleles from a tetraploid entry, thus keeping most of the information (see Chapter 4.2). The HapCUT2 output for the double-coverage 10kb regions contained files in which there were multiple

[1]FreeBayes (Garrison and Marth, 2012) is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment. The output is a VCF (Variant Call Format) file, for storing gene sequence variations.

assembled haplotypic segments. These segments were split into individual files with information from the VCF file using a script by Du et al., 2020 (available at https://github.com/dukecomeback/sterletM_sch) that was modified for this study. These split files were then processed with fgbio's *HapCutToVcf* script to generate separate VCF's for each assembled haplotype. These VCF files and the reference were used to produce haplotypic contigs in fasta format using *vcf-consensus* from the bcftools package (v1.10.2) (Li, 2011a)[2.2.1]. The fasta files of the split regions were merged with the original contigs (minus the double coverage contigs) using the Unix *cat* command.

4.2.2 Assembly and scaffolding of the paddlefish genome

Following haplotype splitting, the genome then needed to be reassembled and scaffolded. Cheng et al., 2020, described 60 pairs of chromosomes (n=120), finding 26 macro chromosomes and 34 smaller, micro-chromosomes, a number which aligns with previous karyotype studies (Symonová et al., 2017) and is equivalent to the sterlet genome (Du et al., 2020). Assembly and scaffolding were done using Juicer (v1.6) and 3D-DNA (v190716)(Durand et al., 2016; Dudchenko et al., 2017). The Illumina short-reads were aligned to the "new" contigs with Juicer (v1.6)(Durand et al., 2016). 3D-DNA (v190716) (Dudchenko et al., 2017) was then used for assembling the genome with -r=0 flag to ensure no iterative rounds of mis-join correction were carried out. Finally, the scaffolded assembly was manually reviewed using Juicebox assembly tools (v1.6)(Dudchenko et al., 2017).

4.2.3 Paddlefish genome annotation

Genome annotation was performed using three lines of evidence: homology annotation, de-novo annotation, and RNA-seq annotation. Firstly, assembly quality and

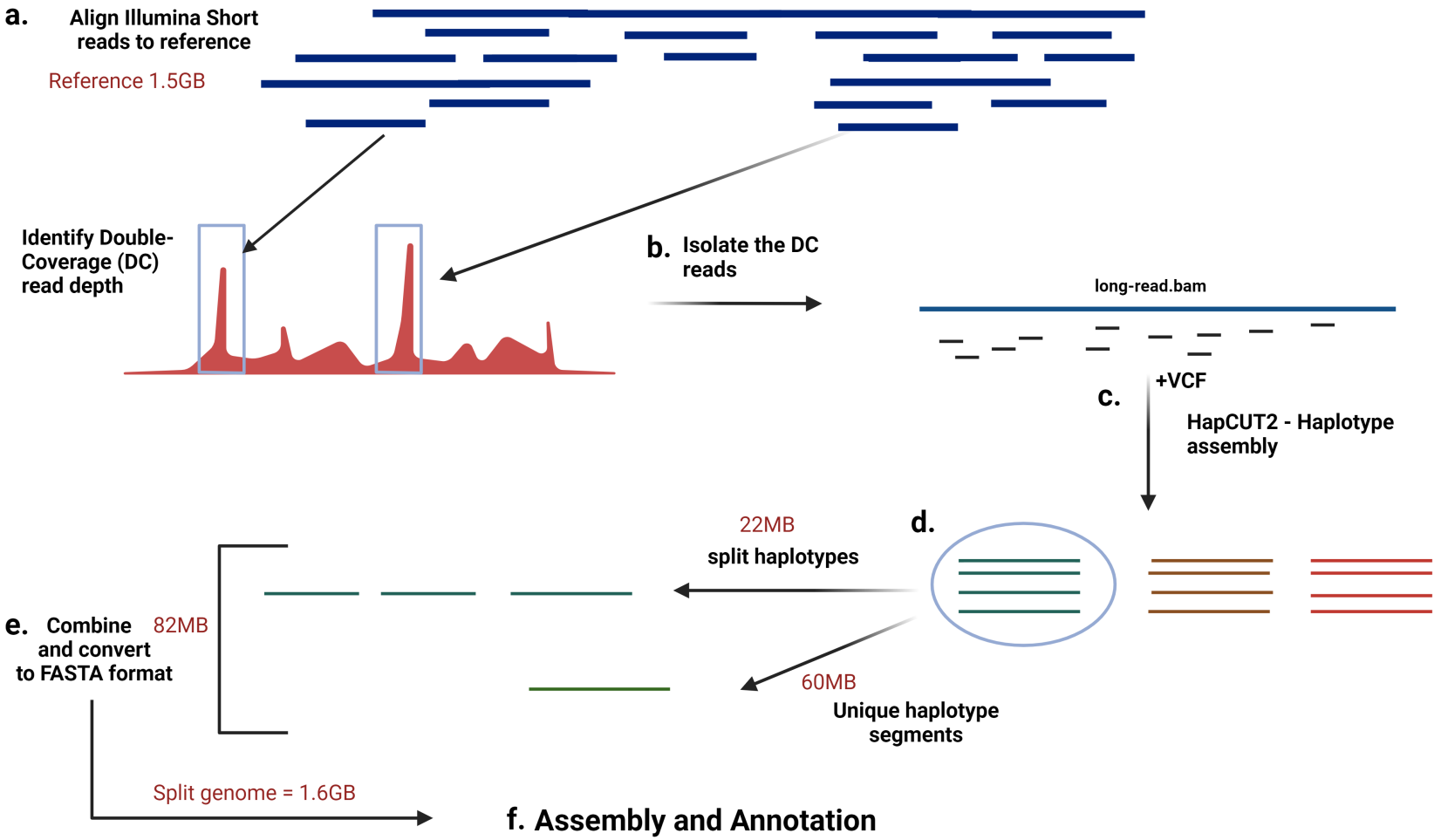

Figure 4.1 | Schematic of the methodology used to re-duplicating collapsed reads in the paddlefish genome (a) Illumina short reads are aligned to the reference, and regions with double coverage are identified using Mosdepth (v0.3.3) (Pedersen and Quinlan, 2018). (b) Double coverage (DC) regions are isolated from the mapped long reads (BAM file). (c) In conjunction with the DC BAM file, a VCF is utilized as input for HapCUT2 (Edge et al., 2017) to reconstruct individual haplotypes within the double-coverage mapped long reads.(d) For each double-coverage region, a haplotyped VCF file is generated. Some of these files contain more than one haplotypic segment. These segments are split based on the haplotypic information from the VCF file using an in-house script. (e) The split and unique regions are then integrated with the original fasta file (excluding the double-coverage regions).(f) Subsequently, the "new" fasta file for the genome is reassembled and annotated.

completeness were assessed with BUSCO (v5.4.4) (Manni et al., 2021) under the *Actinopterygii odb9* database. The gene prediction flags, *-augustus*, and *-long* were implemented in the BUSCO run (Manni et al., 2021; Stanke and Morgenstern, 2005). *-long* is used for optimization of AUGUSTUS self-training mode in BUSCO and while it adds considerably to the run time, it can improve gene prediction results for some non-model organisms.

For homology annotation, we used a set of 11 diverse vertebrate proteomes from NCBI: American Paddlefish (*Polyodon spathula*; GCF_017654505.1) (Cheng et al., 2020), elephant shark (*Callorhinchus milii*; GCF_000165045.1) (Venkatesh et al., 2014), zebrafish (*Danio rerio*; GCF_000002035.6), medaka *(Oryzias latipes*; GCA_002234675.1), fugu (*Takifugu rubripes*; GCA_901000725.2), stickleback (*Gasterosteus aculeatus*; GCA_016920845.1), sea lamprey (*Petromyzon marinus*; GCA_010993605.1), spotted gar (*Lepisosteus oculatus*; GCF_000242695.1) (Braasch et al., 2016), human (*Homo sapiens*; GCF_000001405.39), mouse (*Mus musculus*; GCA_000001635.9) and sterlet (*Acipenser ruthenus*; GCA_902713425.2) (Du et al., 2020). The proteomes were run through CD-HIT to reduce redundancy when aligning, which resulted in 229,665 proteins (Fu et al., 2012). These were aligned to the assembly using Exonerate (v2.2.0) (Slater and Birney, 2005) and GFF3 files were created for use in evidence based gene modelling in later steps. For RNA-seq annotation, RNA-seq reads from the Brain, Kidney, Liver, Spleen, Skin, Skeletal Muscle, Eye, Gill filament, Gill raker, Rostrum, Spiral valve, Stomach, Heart, Gonad, Pyloric Caeca of five adult Paddlefish were aligned to the reference using hisat2 (v2.2.1) (Kim et al., 2015), trimmed using trimal (v1.4.1) (Capella-Gutiérrez et al., 2009) and assembled using stringtie (v2.2.1)(Kovaka et al., 2019). The resultant BAM files were then sorted and indexed with SAMtools (v1.16.1) (Bonfield et al., 2021) and merged using taco (v0.7.3) (Niknafs et al., 2017) to produce GFF3 files for use in gene modelling by EVidence Modeler(EVM)

(v2.1.0) (Haas et al., 2008). TransDecoder (v5.7.0) was used in parallel to the hisat2/stringtie method for RNA-seq assembly. All lines of evidence obtained from homology, RNA-seq (taco and transdecoder) and de-novo annotation (Augustus) were collected and transferred into EVM (v2.1.0)(Haas et al., 2008), where gene models conformed by these evidences were extracted as high-quality gene models. General functional annotation was done using the AnnotaPipeline (v2.0) using SWISS-PROT and TrEMBL(uniprot) to annotate and validate predicted features in genomic sequences.

4.3 Results

4.3.1 Investigating double coverage regions in the paddlefish assembly

A thorough examination of the paddlefish assembly unveiled potential issues in specific genomic regions, suggesting the collapse of duplicates into single regions during the assembly process. Our investigation was initially prompted by suspicions regarding a lower gene count and a genome size smaller than anticipated (Liu et al., 2020; Cheng et al., 2020). It is noteworthy that the sterlet sturgeon genome, which shares a whole-genome duplication (WGD) event with the American paddlefish, displays a significantly higher number of genes compared to the paddlefish (Redmond et al., 2023). According to the NCBI, the sterlet is reported to have 38,193 protein-coding genes, while the American Paddlefish genome (Cheng et al., 2020) is documented with 30,722 protein coding genes (NCBI/Polyodon Spathula). While it is conceivable that the Paddlefish may have lost additional genes over the course of lineage divergence, the observed difference appeared to be larger than expected. Despite the published Paddlefish genome size being 1.54GB, 17-mer

depth frequency distribution analysis estimates a size range between 1.56GB to 1.59GB (Liu et al., 2020). The variations in gene count and genome size could be attributed to factors such as differences in assembly quality, annotation methods, or possibly actual biological distinctions between the species.

With this suspicion, our initial focus involved evaluating coverage in order to identify any reads exhibiting double, or greater than double, the read depth than expected. This approach aims to uncover potential irregularities in genome assembly, providing valuable insights into the reliability of the genomic data. We found that approximately 2.3% of the genome exceeded the average read depth. This strongly indicates that reads had collapsed during assembly to the reference. In Fig. 4.2, we use the ohnolog dataset published in Redmond et al., 2023. We categorized the ohnologs into two groups: genes present in double-copy in paddlefish (in red) and single-copy paddlefish orthologs where two copies exist in the sterlet, but one is presumed to be lost in the paddlefish (Du et al., 2020) (single-copy paddlefish ohnologs; blue). It is immediately evident that some of these genes may not be lost at all, but rather the single-copy ohnologs (here orthologs) have collapsed into a single locus during assembly to the reference. For our work in Chapter 5, it was intrinsically valuable to improve the genome in order to generate a thorough collection of ohnologs shared between these two extant acipenseriformes lineages for the synteny analysis.

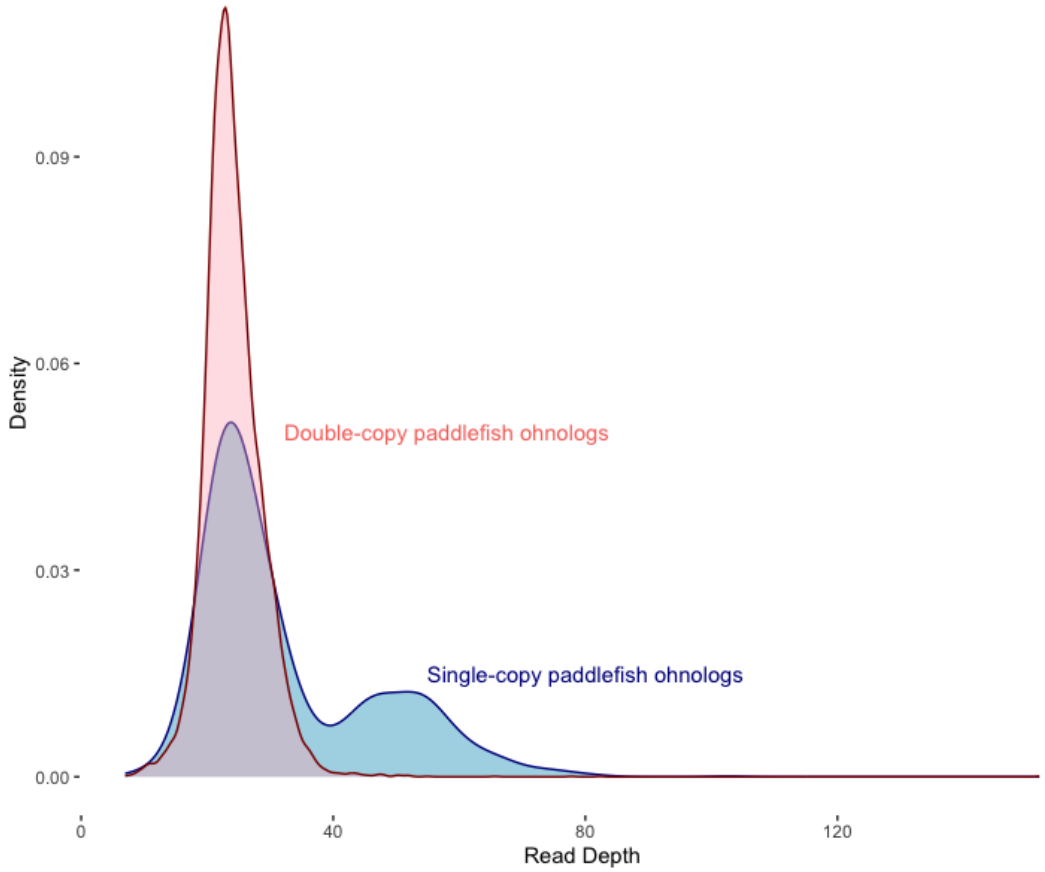

Figure 4.2 | Read-depth of paddlefish ohnologs present in two copies and paddlefish orthologs present in a single copy Plotted are the read-depths of the ohnolog pairs found in paddlefish (red) and the genes we suspect are single-copy ohnologs (here orthologs) in Paddlefish (blue). The peak at double the average coverage (blue) are potential duplicates that have collapsed into a single locus. Note: Read depth is measured in "X" or "fold coverage," where 1X means that each base in the genome has been sequenced once on average.

After pinpointing the regions with double coverage in the genome, our next challenge was to 'uncollapse' them. As illustrated in Fig4.1(C), the initial step in this process involves generating a Variant Call Format (VCF) file. A VCF is a standardized file format used for storing data related to genetic variants identified during DNA sequencing or genotyping. Most genome assemblers currently strive to produce haploid assemblies, wherein each genomic region is represented exactly once. However, for diploid or polyploid genomes, these haploid assemblies only

include one version of each heterozygous region. Consequently, they offer a simplified representation of the genome's actual complexity. To address this, haplotypes can be reconstructed from a reference assembly using long reads and tools such as HapCUT2 (Edge et al., 2017), as demonstrated in our study. By identifying all the haplotypes within the double coverage regions, we hope to separate the haplotypes from the duplicates.

While conducting the variant calling step (Fig 4.1 (C)), we observed unexpected entries in the VCF where the genotype field had more alleles than anticipated (Table 4.1). This aligns with the suspicion that the paddlefish assembly has collapsed duplicates and is potentially an indication of ongoing tetraploid inheritance in some parts of the genome, as has been observed in salmonids (Lien et al., 2016; Macqueen and Johnston, 2014). You can find a breakdown of the genotype counts found in the paddlefish genome in Table 4.1. Genotypes such as 0/0, 0/1, 1/2, 2/2, and 2/1 are standard entries for diploid genomes. However, the occurrence of a 3/3, 3/4 and 4/4 entry in the VCF of a diploid individual suggests there are 3 or 4 alternative alleles in the individual, which for a single diploid specimen is an error. Alternative reasons for such entries could be ascribed to data quality issues or misinterpretation of reads by Freebayes (Garrison and Marth, 2012). It has been recognised in several studies that artifacts in VCFs are the repercussion of errors accumulated in short-read alignment, which as discussed, often struggles to resolve complex structural variants and regions with high copy number (Li, 2011b). 1/3 and 3/3 entries require further probing given that we think duplicate reads have collapsed during short-read alignment. As well as this, such genotype entries may also be consistent with the hypothesis that parts of the paddlefish genome may have ongoing tetraploid inheritance. A 4/4 entry implies five alleles which we suspect is an errors during variant calling probably induced via a short-read alignment artefact (Tørresen et al., 2019).

Genotype (GT)	Count
0/0	9978282
0/1	2485561
0/2	29555
1/1	806637
1/2	15467
2/2	10532
2/1	5
4/4	1197
3/4	5
3/3	3136

Table 4.1 | Genotypes found in a VCF generated from paddlefish short-read data and their counts. Here, 0/0 represents a homozygous for the reference(ref) allele, 0/1 is heterozygous (one ref allele, one alternate(alt) allele) 1/1 is homozygous for the alt allele. 0/2, 2/2 etc. indicate that there are up to three alternate alleles at some positions in the genome e.g. if the reference is C, then perhaps you have an A, G and T(U) alternate allele. 3/3 indicates three alternate alleles perhaps representing collapsed reads or tetraploid inheritance and 4/4 genotypes, indicating 5 alleles, suggests a complex situation or error

4.3.2 Haplotyope assembly and splitting collapsed duplicates

After dividing the genome into smaller segments so that sequencing depth could be calculated and double coverage regions identified, these identified segments were extracted from the PacBio mapped long-reads for the next steps. Both the VCF and the BAM file of PacBio double coverage mapped long-reads underwent processing with HapCUT2 (Edge et al., 2017), a haplotype assembly tool. HapCUT2 produced a haplotyped VCF file for each double-coverage region. Some of these files contained more than one haplotypic segment and these were split according to the haplotypic segment information found in the VCF file, using an in-house script. Such segments were presumed to be collapsed duplicates rather

than representing alternative alleles. In total, 707 regions contained a higher than expected number of haplotypic segments and needed to be separated or split. These and the unique haplotyped VCF files (2108) were processed with fgbio's HapCutToVcf script to generate separate VCF's for each assembled haplotype. These were then transformed into fasta format file using vcf-consensus from the bcftools package (v1.10.2) (Li, 2011a). The unique and split files amounted to ~82MB and were merged with the rest of the reference and prepared for assembly and annotation (Fig.1.2).

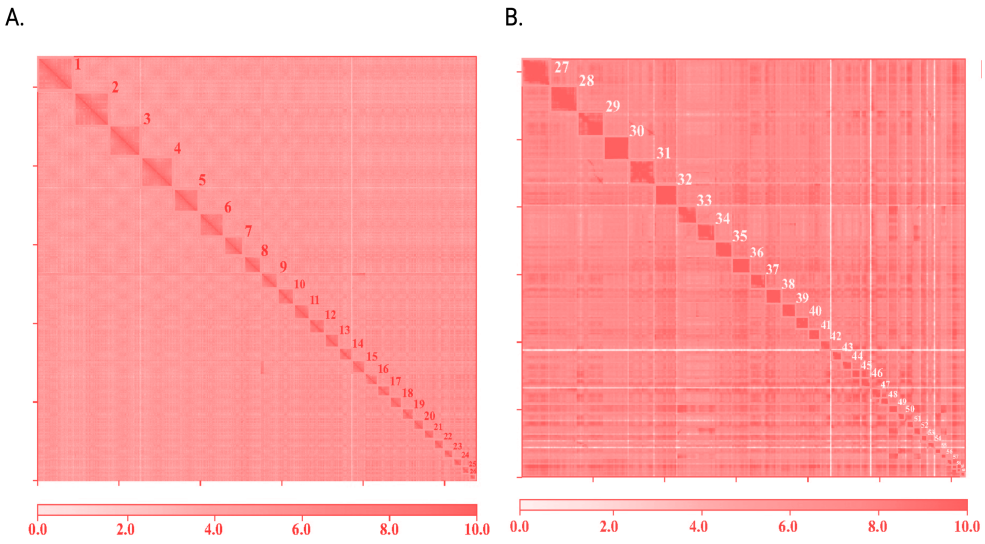

Figure 4.3 | A Hi-C-based chromosome-level genome assembly of the American paddlefish. (A) Contact map based on Hi-C data showing macro-chromosomes 1-26 of the American paddlefish. (B) Macro-chromosomes 27-60. The contact map is based on extracted Hi-C data after manual revisions and reruns with Juicebox assembly tools (v1.6)(Durand et al., 2016)

In Fig.4.4 the desity curve plot indicates that we have almost completely eradicated any signs of double coverage in the new assembly with our protocol. Comparing this figure to Fig.4.2 is striking, showing that the single-copy ohnologs found in the old paddlefish assembly have been "reduplicted" using the protocol presented here. In chapter 5 we attempt to recover and identify these ohnolog

pairs.

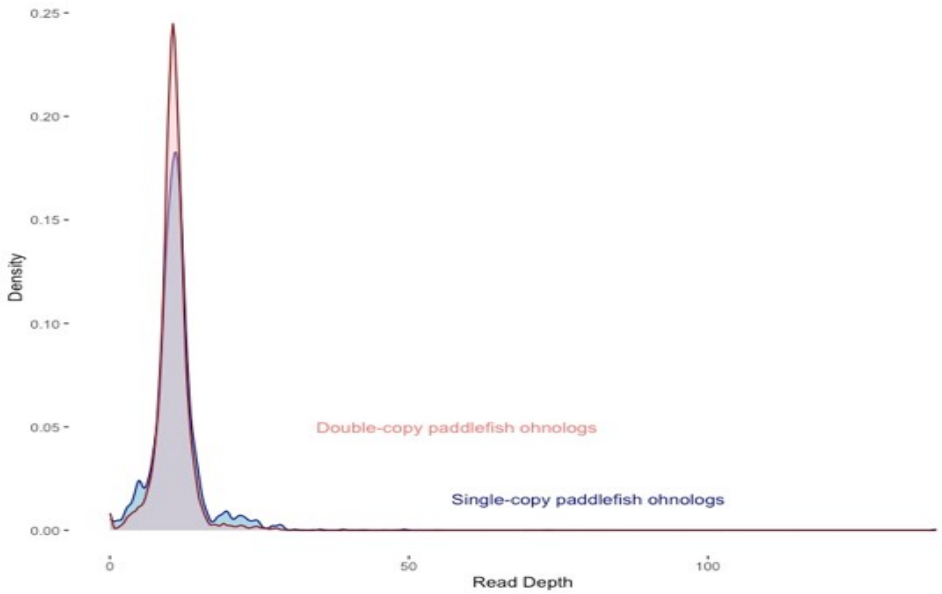

Figure 4.4 | Read-depth of paddlefish ohnologs present in two copies and paddlefish orthologs present in a single copy from old assembly Plotted are the read-depths of the ohnolog pairs found in paddlefish (red) and the genes we suspect are single-copy ohnologs (here orthologs) in Paddlefish (blue) in the new assembly. The peak shown previously at double the average coverage (blue) in Fig4.2 has almost disappeared. Note: Read depth is measured in "X" or "fold coverage," where 1X means that each base in the genome has been sequenced once on average.

4.3.3 Assembly and Annotation

As the reference had been split into 10KB regions, and with the addition of the split and unique haplotypic segments we generated, the genome needed to be reassembled. Following the same protocol as was done for Cheng et al., 2020 assembly, we used the programs Juicer and 3D-DNA(Dudchenko et al., 2017). In Fig4.3, we show a Hi-C contact map for the final assembly after editing with juciebox. We were able to identify 60 large chromosomes as was previously identified in Cheng et al., 2020.

For a comprehensive overview of annotation, please refer to Chapter 2. We firstly assessed the BUSC0 (Manni et al., 2021) quality. We found that the annotation

	Old P. Spathula	New P. Spathula
Size	1.54GB	1.63GB
Protein coding genes	26,017	35,930
Functionally annotated	25,886	29,833
Chromosome No.	60 pairs (2n = 120)	60 pairs (2n = 120)
BUSCO	C:92.0%[S:51.0%,D:41.0%],F:2.5%,M:5.5%,n:3640	C:92.0%[S:50.0%,D:42.0%],F:2.5%,M:5.5%,n:3640

Table 4.2 | Comparing the original Polyodon Spathula assembly from Cheng et al., 2020 with the improved version described in this study. BUSCO v5.1.2 (Manni et al., 2021) was used to assess assembly completeness with respect to the Actinopterygii odb10 dataset. BUSCO annotation; C= Complete, S=Complete and single-copy, D=Complete and duplicated, F=fragmented and M=Missing. n is the number of BUSCO genes in the Actinopterygii odb10 dataset

contained 92% out of 3640 conserved and complete Actinoptergyii genes (Table4.2). By integrating three strategies for annotation; homology, de novo, and transcriptome we predicted 35930 protein coding genes.

4.4 Discussion

We present a methodology aimed at enhancing the paddlefish short-read assembly through a process we refer to as *re-duplication*, which defines a method of "uncollapsing" duplicate reads that have collapsed during assembly to the reference genome. Our approach is adapted from the method outlined by Du et al., 2020. Here, we focus on ameliorating the chromosome-level genome of the American paddlefish, originally sequenced and assembled by Cheng et al., 2020. These improvements are expected to facilitate significant advancements in our comprehension of the evolutionary dynamics within this ancient fish lineage.

In Section 4.3, we identified regions with double the expected read-depth that we suspected were collapsed duplicates or collapsed repetitive regions in the paddlefish assembly (Fig4.2). This issue is not exclusive to this genome but is a common challenge when assembling short-read data. It underscores the imperative need for improved assembly methodologies (Wang et al., 2020; Tørresen et al., 2019;

Salzberg and Yorke, 2005).

Our protocol incorporates both long and short-read data collected during previous sequencing and assembly efforts (Cheng et al., 2020). Short-read sequencing is widely used despite being error-prone. This can be mitigated by employing long reads in a hybrid approach, it proves invaluable for rectifying misassembled short-read genomes; preserving and updating data rather than discarding it. As we transition into an era where high-quality TGS methods like HiFi (Cheng et al., 2021; Nurk et al., 2020) and CCS (Wenger et al., 2019) become more accessible, users must remain cognisant of potential errors in many larger genome assemblies available on databases such as NCBI and Ensembl. Though our approach enhanced the assembly by identification of more genes, a closer look at assembly statistics, such as BUSCO (Manni et al., 2021) scores, reveals that completeness levels remain comparable to the original assembly of the American paddlefish (Cheng et al., 2020). We observe a slight increase in the percentage of complete duplicate BUSCOs, as anticipated, aligning with our objective of identifying collapsed duplicates.

We acknowledge that not all collapsed reads were resolved, indicating that there is still improvement needed. We suspect that issues during variant calling may have hindered further enhancement. During the variant calling steps, complications can arise from poor data quality, misinterpretation of reads, and complex genetic variations. While we suggest in the text that 3/3 genotypes in the VCF may stem from collapsed reads or indicate ongoing tetraploid inheritance in parts of the genome, this explanation accounts for only part of the overall narrative in these regions. The most substantial challenges in variant calling emerge in hyper-variable and repetitive regions of the genome (Zverinova and Guryev, 2022). In fact, many of these challenges are attributable to the limitations of short-read sequencing methods. This creates a potential circular issue when attempting to

rectify an assembly with variant information, as demonstrated in our approach here (Zverinova and Guryev, 2022; Garrison and Marth, 2012; Du et al., 2020). Despite this, it is clear we have enhanced the assembly given the comparatively larger number of genes detected.

The HapCUT2 (Edge et al., 2017) program, used for haplotype assembly in this study, only works on diploid genomes and thus forced us to filter out any entries that indicated complexity or polyploidy from the VCF such as 4/4, 3/4, 3/3 etc. A custom script was written to force genotypes (GT) in the VCF to be diploid where the GT field could only be encoded by a 0 for the REF allele, 1 for the first ALT allele, 2 for the second ALT allele. For example, the script would transform a 4/4 allele into a 2/2 entry. The method preserved crucial information by excluding two of the least common alternative alleles, based on allele frequency, from these complex entries. It's important to note that this approach may result in some data loss and is not a flawless solution. As of now, in the absence of haplotype-assembly tools designed for polyploid genomes, this method provides a practical solution to address challenges posed by intricate genetic structures in the paddlefish genome and other complex genomes.

As previously mentioned, our annotation strategy embraced three distinct lines of evidence: homology annotation, de-novo annotation, and RNA-seq annotation. While Cheng et al., 2020 followed a similar approach, our study gained a significant advantage by incorporating high-quality RNA-seq data from five adult paddlefish. In contrast to the previous genome annotation, which relied on a single replicate and only used blood tissue, our RNA-seq analysis encompassed a diverse array of tissues, including Brain, Kidney, Liver, Spleen, Skin, Skeletal Muscle, Eye, Gill filament, Gill raker, Rostrum, Spiral valve, Stomach, Heart, Gonad, and Pyloric Caeca. This extensive tissue sampling, coupled with the use of five replicates, yielded substantially improved results. Through the amalgamation of all lines of

evidence into high-quality gene models, we identified a total of 35,930 genes. To unravel the potential functions embedded in this final gene set, we conducted functional annotation using public databases such as Pfam, TrEMBL, Swiss-Prot, and CDD (Maia et al., 2022). Our analysis successfully annotated 29,833 genes, providing valuable new insights into the genomic landscape of the American Paddlefish.

This improved genome and gene annotation is used in Chapter 5 in which we were able to delineate an improved ohnolog dataset for the paddlefish. These results also demonstrate that by utilizing both short and long-read data, we have developed an innovative protocol to enhance the genome. This method may serve as a valuable approach in future efforts to improve other assemblies facing similar challenges with collapsed reads. We underscore the limitations associated with short-read assemblies and emphasize the necessity for further advancements to transition into a new era of long-read sequencing. It is unclear to what extent published genome assemblies have been affected by the inherent error-prone nature of genome sequencing technologies. These errors can have substantial effects on research results and stunt the transformative capabilities genome sequencing could have in clinical settings (Salzberg and Yorke, 2005; Mardis, 2008; Kelley and Salzberg, 2010). While, long-read sequencing or Single Molecule Sequencing (SMS) offers relief from issues with repetitive DNA resolution, TGS is still not customary in most sequencing studies (Eid et al., 2009; Rothberg et al., 2011; Quail et al., 2012). Major progress has been made to overcome issues in genome assemblies but there is still a way to go before we can attest a perfect genome sequencing tool.

Chapter 5

Investigating the mechanisms of rediploidization in the paddlefish and sturgeon

5.1 Introduction

Ancient whole genome duplication's (WGD) have occurred several times throughout the vertebrate lineage (see Fig1.6). These doubling events are thought to offer raw genetic material which has the potential to facilitate evolutionary innovations in descendant species. In many cases, rediploidization follows a WGD in which, most conventionally, a tetraploid genome transitions back to a more stable, diploid state. WGD can occur by self doubling of the genome (autopolploidistaion) or through hybridisation of two different parent species (allopolyploidisation). Cytogenetically in the latter, rediploidization is perceived as happening instantaneously as the non-homologous chromosomes preferentially pair bivalently. For autoploylpoidisation, chromosomes most likely continue recombining multivalently during meiosis until suppressed. In the case of an auto-tetraploidisation event,

once bivalent loci have been uncoupled from tetraploidy they can then undergo functional divergence, marking the onset of their status as duplicated loci. The rediploidization process effectively serves to separate the genome duplication process from the gene duplication process and it is only after rediploidization takes place that the locus can be regarded as duplicated (Furlong and Holland, 2002; Redmond et al., 2023). Rediploidization is intrinsically linked to the suppression of recombination and is conceptually similar to the establishment of sex chromosomes (Lahn and Page, 1999). However, in both cases, the mechanism for how this happens is largely unknown. It is believed that the suppression of recombination is orchestrated by genomic rearrangements or the accumulation of mutations in specific regions of homologous chromosomes (Furlong and Holland, 2002; Lahn and Page, 1999).

In this chapter, we focus our analysis on the mechanisms of rediploidization in the sturgeon and paddlefish genomes, sister lineages, together representing extant Acipenseriformes. Previous studies had favoured independent WGD events in the sturgeon and paddlefish lineages, despite their close phylogenetic ties. However, recent work showed that the WGD was in fact shared but masked by a delayed rediploidization process (see Chapter 1.3.3)(Redmond et al., 2023). This shared auto-tetraploidisation is estimated to have occurred over 200 million years ago, potentially in proximity to the Permian-Triassic mass extinction. The extended transition to stable diploid inheritance likely conferred a survival advantage during the challenging environmental conditions of that era. This novel insight into the evolutionary history of these non-teleost, ray-finned fishes underscores the potential role of polyploidy and asynchronous rediploidization as an adaptive survival tactic. Understanding the processes involved in the reversion process of tetraploid to diploid inheritance, particularly its tendency to preserve concurrent tetraploid signatures, therefore holds invaluable importance.

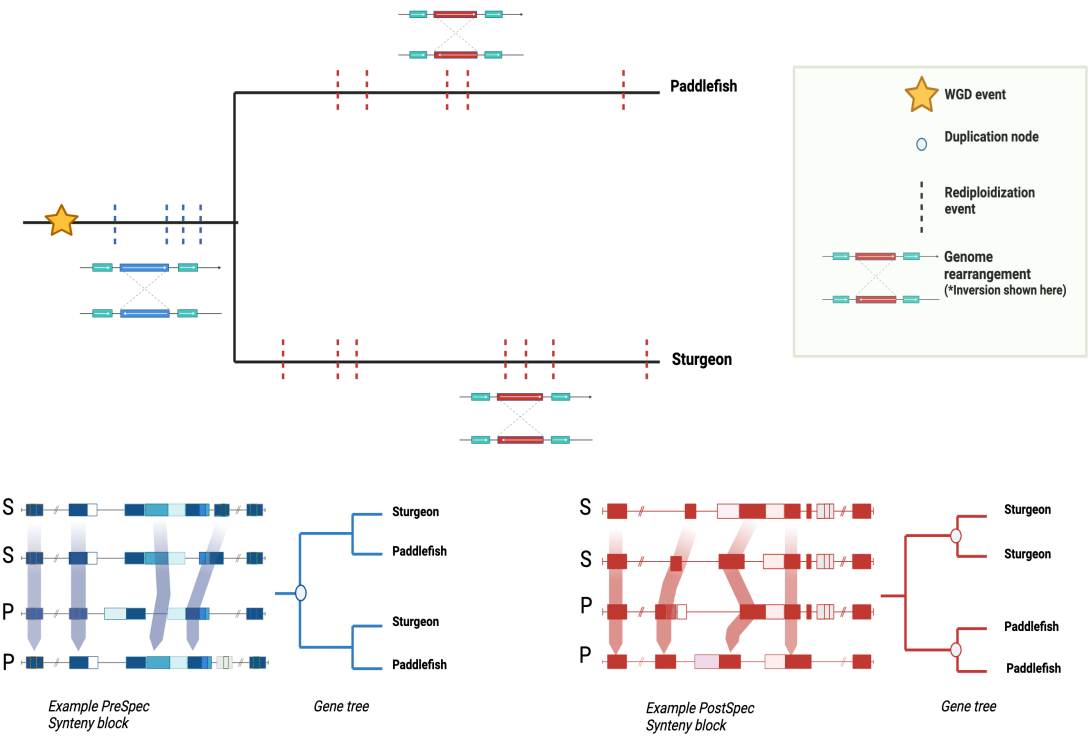

Figure 5.1 | Diagram illustrating the role of rearrangement in rediploidization in the sturgeon and paddlefish genomes Diagram illustrating rediploidization in sturgeon and paddlefish genomes. Blue dashed lines on the species tree (top) indicate rediploidization events stemming from ancestral genome rearrangement (shown here as inversions but we acknowledge other mechanisms are possible), while red lines represent lineage-specific events. Displayed below are example synteny blocks, with each row representing a stretch of genes from each chromosome of a species (S for sturgeon, P for paddlefish). The gene tree topologies indicate ancestral rediploidization events (blue) and independent occurrences in each lineage (red).

Delays in rediploidization can become increasingly difficult to unravel if the WGD in the ancestral lineage has not been resolved to diploid before the daughter species begins diverging, as happened with the sturgeon and paddlefish (Redmond et al., 2023). In cases like this, some ohnologs will resolve in a lineage-specific manner allowing for independent functional and regulatory divergence of the pairs of genes in response to lineage specific selective pressures. The LORe

(Lineage-specific Ohnolog Resolution) model describes this process (Robertson et al., 2017)(see Fig1.7). Such a process of asynchronous rediploidization has also been shown in salmonids and other teleosts (Parey et al., 2022; Lien et al., 2016; Robertson et al., 2017). In the Acipenseriformes, analysis found that that on visualisation of ohnolog pairs that had rediploidized before the speciation event and those that resolved after the speciation event, they were not randomly distributed in the genome but rather were found in distinct syntenic-blocks along uninterrupted sections of chromosomes (Redmond et al., 2023). This follows the hypothesis that in cases, like in the sturgeon and paddlefish, where a WGD and delayed rediploidization is followed by a speciation event, rearrangements that interrupt tetrasomic inheritance should be seen as large blocks of contiguous genes with common rediploidization history. This draws parallels to the strata in the mammalian sex chromosomes, the paddlefish and sturgeon chromosomes may also be stratified by the divergence time of ohnologs within the synteny blocks (Lahn and Page, 1999; Hokamp et al., 2003a; Redmond et al., 2023; Lien et al., 2016).

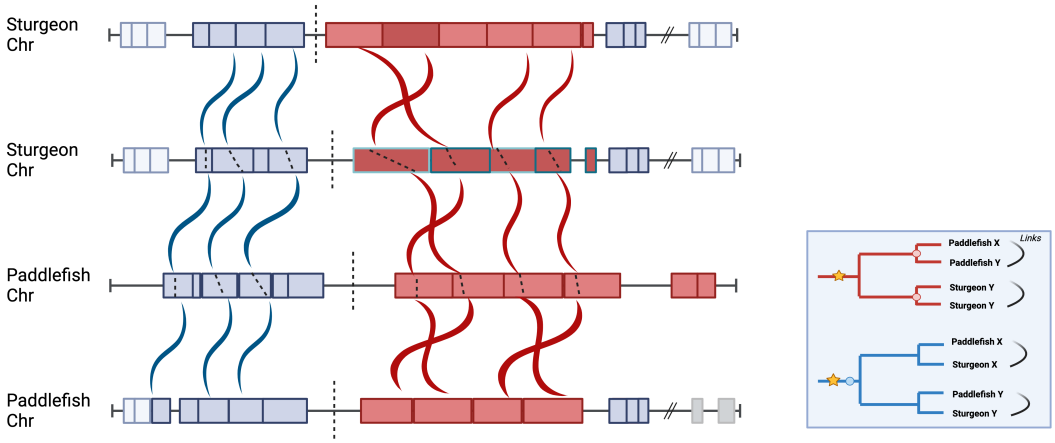

Figure 5.2 | Diagram of a micro-syntenic block demonstrating a genome rearrangement breakpoint and tree topology breakpoint Diagram illustrating regions on homologous chromosomes in sturgeon and paddlefish respectively that underwent a rearrangement event in the ancestor (left; blue genes, blue links), leading to a block of syntenic genes with PreSpec rediploidization histories and an adjacent block of genes with a lineage-specific rearrangement history (right; red genes, red links). Dashed lines between gene blocks denote rearrangement breakpoints. Dashed lines within genes link ohnologs between species. The legend showcases gene tree topologies of ohnologs, color-coded for PreSpec (blue) or PostSpec (red) rediploidization history.

While rearrangements have been proposed as a mechanism for rediploidization, formal testing of this hypothesis has been lacking. In this chapter, we carry out a comprehensive examination of this hypothesis, identifying micro-syntenic blocks of ohnologs within the paddlefish and sturgeon genomes. We evaluate two scenarios: i) synteny-blocks that have a consistent gene-tree topology showing a shared WGD, which we assume rediploidized before the speciation event and have 1:1 orthology, and ii) blocks with a consistent gene tree topology defining a post-speciation WGD scenario, which would have rediploidized after speciation and would have a 2:2 orthology assignment (Fig 5.1, (Fig5.2)). We show that there are blocks of ohnologs with conserved micro-synteny in paddlefish and sturgeon showing a consistent topology, either the pre-speciation or post-speciation scenario. Additionally, we identify blocks with distinct rearrangement breakpoints

characterised by a break in a consistent topology (see Fig 5.2). This provides clear evidence for temporally isolated rearrangement events in the species and suggests that rearrangements play a fundamental role as a primary mechanism for rediploidization (see Fig 5.2). Our findings also highlight the presence of some heterogeneity across blocks which suggest alternative mechanisms for how loci rediploidize in these lineages. We underscore the importance of understanding the methods contributing to this pivotal evolutionary process.

5.2 Materials & Methods

5.2.1 Orthology Assignment

Orthofinder (v2.5.4) (Emms and Kelly, 2019) was used for orthology inference. We included a diverse set of proteomes that spanned the jawed vertebrate phylogeny including from the Chondrichthyes; ghost shark (*Callorhinchus milii*; GCF_000165045.1) (Venkatesh et al., 2014). From the Sarcopterygii; human (*Homo sapiens*; GCF_000001405.39), African clawed frog (*Xenopus tropicalis*; GCF_000004195.4), and coelacanth (*Latimeria chalumnae*; GCF_000225785.1). From within Actinopterygii we selected zebrafish (*Danio rerio*; GCF_000002035.6) and used spotted gar (*Lepisosteus oculeatus*; GCF_000242695.1) (Braasch et al., 2016) as a representative of Neopterygii. Finally we used the Grey bichir (*Polypterus senegalus*; GCF_016835505.1) (Bi et al., 2021) as the combined sister group to the paddlefish and sturgeon (Fig5.3). We included a species tree in our Orthofinder (Emms and Kelly, 2019) run with flag *-s* in line with the accepted relationships in the jawed vertebrate phylogeny to augment orthology inference: *((Callorhinchus milii), ((Latimeria chalumnae,(Xenopus tropicalis,(Homo sapiens)),(Grey birchir,((Polyodon Spathula, Acipenser ruthenus),((Danio rerio),*

(Lepisosteus oculatus))))))); (Fig 5.3).

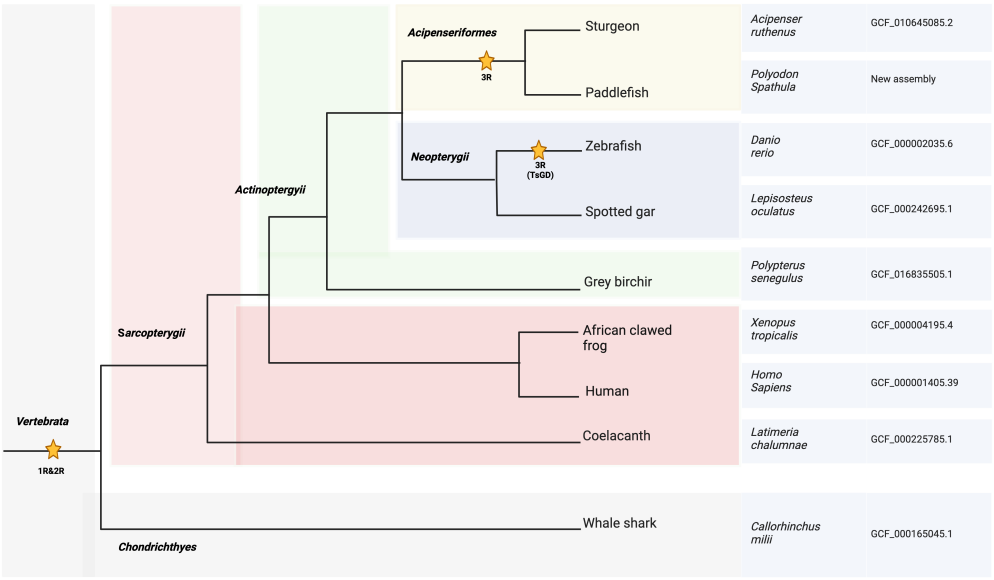

Figure 5.3 | Phylogeny of the species used for orthology inference. Species spanning the jawed vertebrate phylogeny used in Orthofinder (Emms and Kelly, 2019) for orthology inference. Common names, species names, and RefSeq assembly numbers are provided on the right. 1R-3R represent tetraploid events. Branch lengths are not to scale, and event placement is approximate.

We extracted the longest isoform from the proteomes using a custom protocol and then used CD-HIT(Fu et al., 2012) to reduce sequence redundancy and ran Orthofinder (v2.5.4) (Emms and Kelly, 2019) with parameter *-y* and as mentioned *-s*, to include a rooted species tree. By using the *-y* flag, within the Phylogenetic Hierarchical Orthogroups (PHOGS) file in the Orthofinder result, we split paralogous clades below the root of a Hierarchical Orthogroup into separate Hierarchical Orthogroups. For example, we split a group that had genes that duplicated after the earliest diverging jawed vertebrate (here, the whale shark) resulting in separate PHOG files for each duplicate. Following a protocol from Redmond et al., 2023, further filtering of the PHOGs was done by extracting groups that included two sequences each from sturgeon and paddlefish and that

also had at least one outgroup for subsequent rooting of the sturgeon-paddlefish pair subtree. Sturgeon-paddlefish genes were only chosen from one of the 60 largest chromosomes therefore ohnologs from micro-chromosomes were not considered in this analysis. PHOG gene trees were built by first aligning with MAFFT (v7.453) (Katoh et al., 2002) and reconstructed with IQ-TREE (v1.6.12) (Minh et al., 2020), see Section 3.2.1 for more detail. We did not filter the PHOG set any further as it was not necessary for this analysis and further filtering was done in later steps using synteny information. For a more strict, high-confidence ohnolog dataset see Redmond et al., 2023.

Next, we verified that the sturgeon-paddlefish sequences formed a monophyletic group and ensured that the two paddlefish and sturgeon sequences (candidate ohnologs) diverged after the split from Neoptergyii. This was done by ensuring each gene tree had at least one sequence from Neopteregyii and a more distantly related outgroup for construction. We are aware that this set is not truly "high-confidence" and may include genes with complex histories that may look like ohnologs but are paralogous. We also note that these ohnologs are conserved in pairs in both species and thus excludes genes that may have been lost in sturgeon, paddlefish or in the Acipenseriformes stem lineage but did in fact orginate from the WGD event.

5.2.2 Ohnolog duplication time inference

The sturgeon-paddlefish ohnolog pairs described above, were subjected to phylogenetic analysis to estimate the time of rediploidization relative to the speciation event. MAFFT (v7.453) was used for multiple SAs of the filtered PHOGs with standard parameters. Phylogenetic inference by ML was performed with IQ-TREE (v1.6.12)(Minh et al., 2020) with the -m JTT+G2.1.1 flag, a general

amino acid substitution model with four discrete rate categories, -bb 1000 flag allowing 1000 ultrafast bootstrap(UFBOOT) (Minh et al., 2013) replicates and -nt AUT0, which detected the optimal number of threads to be used for the analysis. These ohnolog gene trees were processed and analysed for duplication time inference (i.e rediploidization time). For this, we used the ETE(v3)toolkit python library (Huerta-Cepas et al., 2016) to check that the sturgeon and paddlefish sequences formed a monophyletic group in each of the filtered PHOG gene trees and then rooted each with the most distantly related sequence relative to the Acipenseriformes. Custom scripts adapted from Redmond et al., 2023 were used to perform strict gene-species tree reconciliation to infer speciation and duplication nodes/events. The resulting gene trees were summarised into different groups indicating their duplication time: PreSpec, PostSpec, Other[PreSpec-like, PostSpec-like].

5.2.3 Synteny analysis

To find the micro-syntenic blocks between sturgeon and paddlefish genomes we used Orthofinder (v2.5.4) (Emms and Kelly, 2019) and i-ADHoRe (v3.0.01) (Proost et al., 2012). The sturgeon, *"reduplicated"* paddlefish and grey birchir proteomes were run through Orthofinder(v2.5.4) (Emms and Kelly, 2019) using standard parameters. It was not necessary to do any post processing on the Orthofinder run for this analysis. Using the *Orthogroups.txt* and corresponding GFF3 files of the proteomes used before, we prepared the orthologs for the i-ADHoRe run. The *segments.txt* file in the i-ADHoRe output was used to define the genomic co-ordinates of the pairs of genes within a syntenic block and those that defined the boundaries of a block (first and last pairs of genes in a block) (Proost et al., 2012). These were prepared for use in circos (v0.69-9) (Krzywinski et al., 2009),

to be used as links within and between the two genomes.

We utilized a custom script to identify the previously curated ohnologs that intersected with the identified blocks. The quantification of gene-tree occurrences with specific topologies, namely PreSpec or PostSpec, within each block was carried out. A "PreSpec block" was defined if more than 98% of the ohnologs within the block exhibited this topology. Similarly, "PostSpec blocks" were characterized when over 98% of the ohnologs displayed PostSpec topologies. Instances deviating from these criteria are classified as "complex blocks". A "complex-PreSpec block" is all other blocks but where there are predominantly ohnologs with PreSpec histories. The same criteria apply to "complex-PostSpec blocks". Visualisation of these blocks was done using JCVI (Tang et al., 2015).This tool interfaces with MC-scanX (Wang et al., 2012), and thus outputs from i-ADHoRe were adapted using a custom script to make them compatible with JCVI.

5.3 Results

5.3.1 Ohnolog duplication time inference

Our initial step involved identifying ohnolog pairs between the paddlefish and sturgeon. In analyzing ohnolog gene trees, we considered two plausible topologies: (1) ohnolog gene trees featuring a single duplication node occurring before the sturgeon-paddlefish divergence, and (2) gene trees with independent duplication nodes arising after speciation. This analysis previously conducted inRedmond et al., 2023, was redone using the "re-duplicated" paddlefish genome as described in Chapter 4. To distinguish between these topologies, we built on a set of high confidence ohnologs previously identified in the sturgeon genome (Du et al., 2020), integrating extensive additional phylogenetic and syntenic evidence. Specifically, we incorporated a broad sampling of predicted proteomes from jawed vertebrate

Species	PreSpec trees	PostSpec trees	Other PreSpec	Other PostSpec	Total
Redmond et al. 2023	1448	2074	771	1138	5431
Casey et al. In prep	2054	1702	545	1392	5693

Table 5.1 | Gene tree topologies recovered in this study compared with Redmond et al., 2023

genomes (Fig.5.3) (see section 5.2).

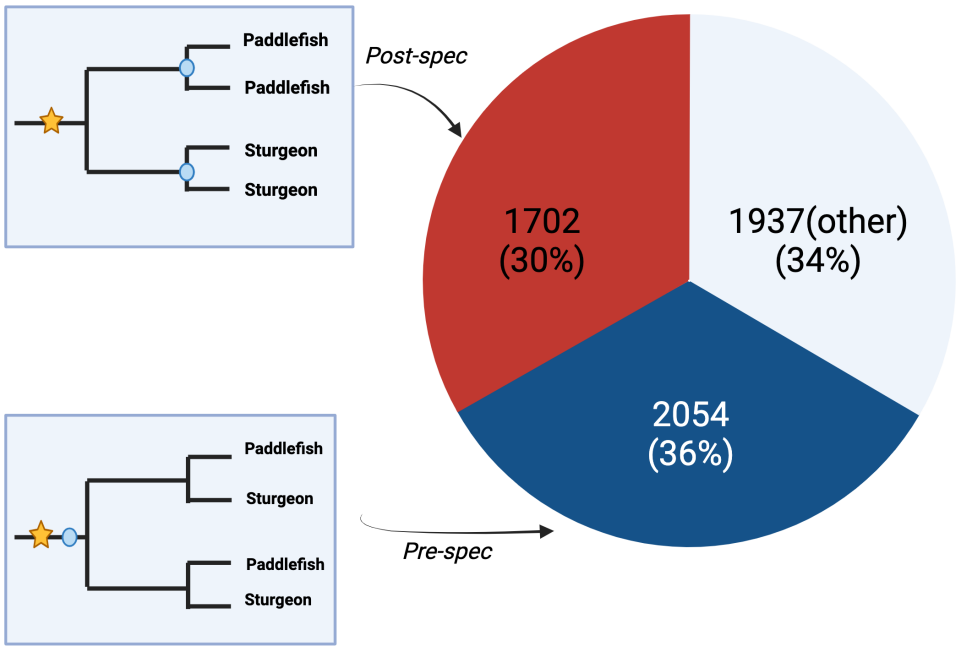

Figure 5.4 | Pie chart of the different gene tree topologies recovered from the ohnolog dataset The pie-chart illustrates both the frequency and relative occurrence of each topology, showcasing how often each topology was recovered. A PreSpec tree's duplication node is in the ancestor, while PostSpec trees have lineage-specific duplication nodes. "Other" trees are presumed error-trees that only partially match one of these scenarios (either, 'PreSpec-like', or 'PostSpec-like'; see Redmond et al., 2023 for more details).

Based on this data, we identified 5,693 gene families associated with protein coding, encompassing relatively high-confidence ohnolog pairs in both sturgeon

and paddlefish. Analysing maximum likelihood gene trees for each family we found that the gene tree with a shared duplication node was the most common topology (hereafter: 'PreSpec', for Pre-Speciation duplication node) being recovered 2054 times (36% of all trees; (Fig5.4)). The alternative ohnolog pair topology with two independent duplication nodes in each lineage ('PostSpec' for Post-Speciation) was recovered 1702 times (30% of all trees; (Fig5.4)). The remaining gene trees (1937, 34%; 'Other' for topologies other than PostSpec or PreSpec (Fig5.4)) failed to recover either of these topologies. The frequent recovery of the PreSpec topology differs from Redmond et al., 2023, where PostSpec gene trees were found to be more common, serving as an indication for the rationale behind previous studies inferring that sturgeons and paddlefish experienced independent WGDs (Table 5.1). Here, with our "re-duplicated" paddlefish genome, we find the PreSpec topology is more common. Variations in protocol may have had some influence on this result, but it is likely that the improved paddlefish assembly may have uncovered more collapsed duplicates from regions of the genome that rediploidized in the ancestor of the paddlefish. Further evidence of this is shown by the lower number of "Other PreSpec" trees recovered in our analysis than the previous study. This suggests that the *new* paddlefish assembly may have resolved trees, previously recognised as error-trees in (Redmond et al., 2023), as PreSpec in this analysis (see Discussion). The high prevalence of the PreSpec and PostSpec trees is consistent with a shared WGD followed by prolonged rediploidization extending past the sturgeon paddlefish speciation. The Ohnolog pairs are visualised in a circos plot, Fig5.5, where PreSpec ohnolog pairs (blue) are linked between the sturgeon and paddlefish and the PostSpec genes (red) are linked within the genomes of the two species respectively.

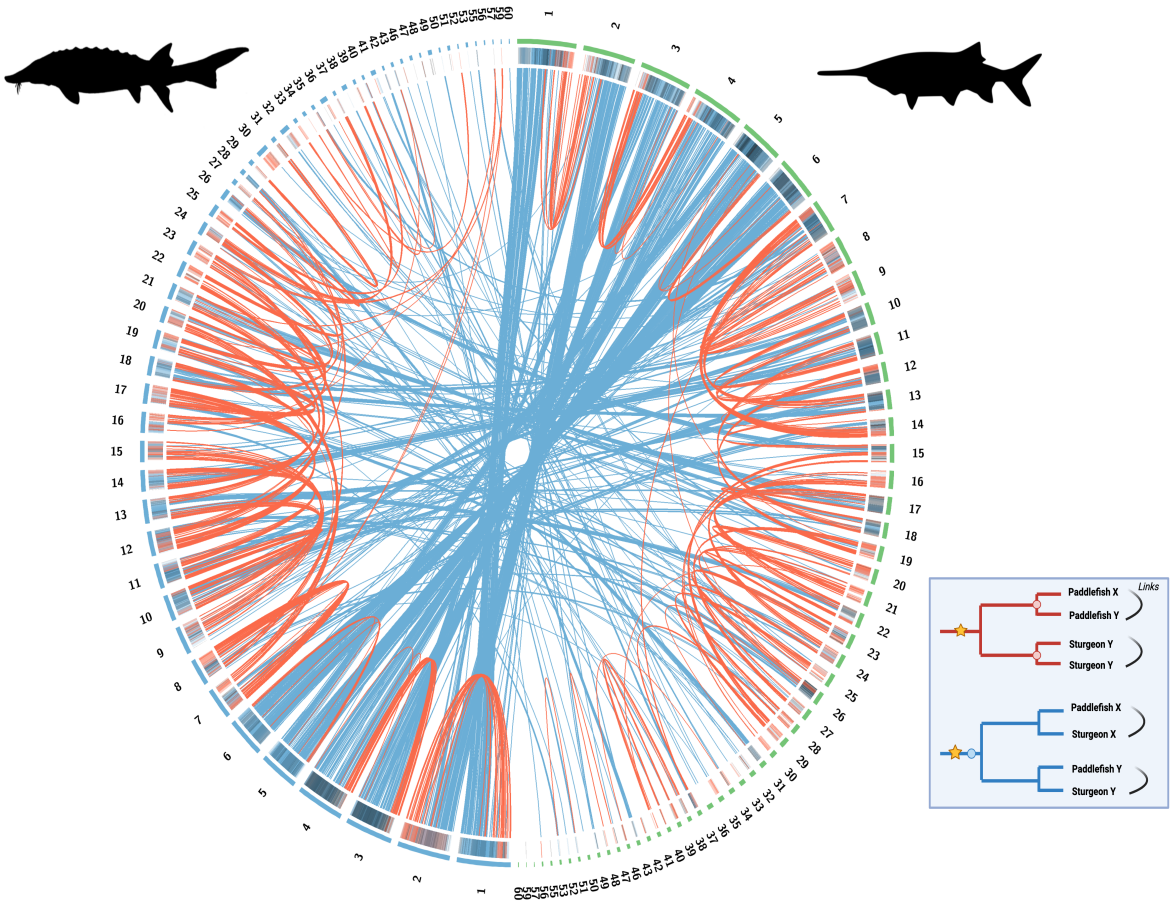

Figure 5.5 | Circos plot of the sterlet sturgeon genome and the American paddlefish genome and their ohnolog pairs Links showing the chromosomal locations of ohnolog pairs between and within their genomes. Links are coloured according to PreSpec (blue) or PostSpec (red) tree topologies. The outer ring illustrates the positions of micro-syntenic blocks identified through i-ADHoRe (Proost et al., 2012). PreSpec blocks (depicted in blue) are defined as blocks where over a sliding window of ohnologs, more than 98% of the genes have trees that exhibit PreSpec topologies. On the other hand, PostSpec blocks (depicted in red) have more than 98% of ohnologs exhibiting PostSpec topologies. Grey lines are complex blocks that don't fit this criteria (see Materials and Methods). Only the 60 macrochromosomes from each species are shown.

5.3.2 i-ADHoRe and syntenic block identification

Once we had identified the ohnolog pairs and their duplication times, we then identified syntenic blocks in the genomes of the species. We ran i-ADHoRe (Proost

Gap size	Cluster Gap	Anchor No.	Multiplicons	Segments
10	15	5	1338	2675
15	20	5	1050	2100
20	25	5	722	1443
25	30	5	713	1417
30	35	5	479	958

Table 5.2 | Adjusting i-ADHoRe Parameters and Observing Effects Gap size indicates the maximum distance that should exist between anchors. **Cluster gap** indicates the maximum distance that should exist between individual base clusters in a cluster. **Anchor number (no.)** indicates the minimum number of genes each segment in a multiplicon should have that are homologous to the other segments in that multiplicon. **Multiplicons** define the homologous segments with synteny found by i-ADHoRe with input gap size and anchor number. **Segments** are each part of a multiplicon (for two species you have 2 segments per multiplicon).[5.3.2]

et al., 2012) using several different parameter combinations (Table 5.2). While 5 was an agreed minimum amount of anchors, given it's use in other studies (Proost et al., 2012; Zhao et al., 2021; Liu et al., 2018) we found that varying the gap size had substantial effects on the number of syntenic blocks identified by i-ADHoRe(Proost et al., 2012). The authors recommend a cluster gap size to be approximately 5 units bigger than the gap size (Proost et al., 2012) and the standard parameters of the pipeline are a 15&20 combination. While our choice was somewhat arbitrary, we selected a gap size of 20 and a cluster gap size of 25. This configuration was found to yield a substantial number of syntenic blocks (Table 5.2). We considered this choice more reliable compared to a combination of 10&15, or 15&20, which might not be stringent enough, potentially resulting in small blocks lacking informative value.

In Fig5.6 we show the 20 largest chromosomes in the American paddlefish genome and the sturgeon sterlet genome and micro-syntenic blocks identified by i-ADHoRe between and within each species. There is a noticeable prevalence of chromosomal homology between the sturgeon and paddlefish genomes with notable robust homology between the six largest chromosomes (Fig5.6). This homology

[1]**iADHoRe:** for more information on the parameters used, see the iADHoRe manual.

becomes more evident in Fig5.5, where the conservation of gene order between the genomes is apparent. Here, we see compelling evidence for blocks of ohnolog pairs with shared duplication histories (PreSpec or PostSpec). This distinctive signature was also observed in Redmond et al., 2023, where it was proposed that the presence of these extensive gene blocks sharing consistent rediploidization histories suggests temporally isolated intrachromosomal rearrangement events, potentially playing a role in the evolutionary history of the species and facilitating the return to bivalent pairing.

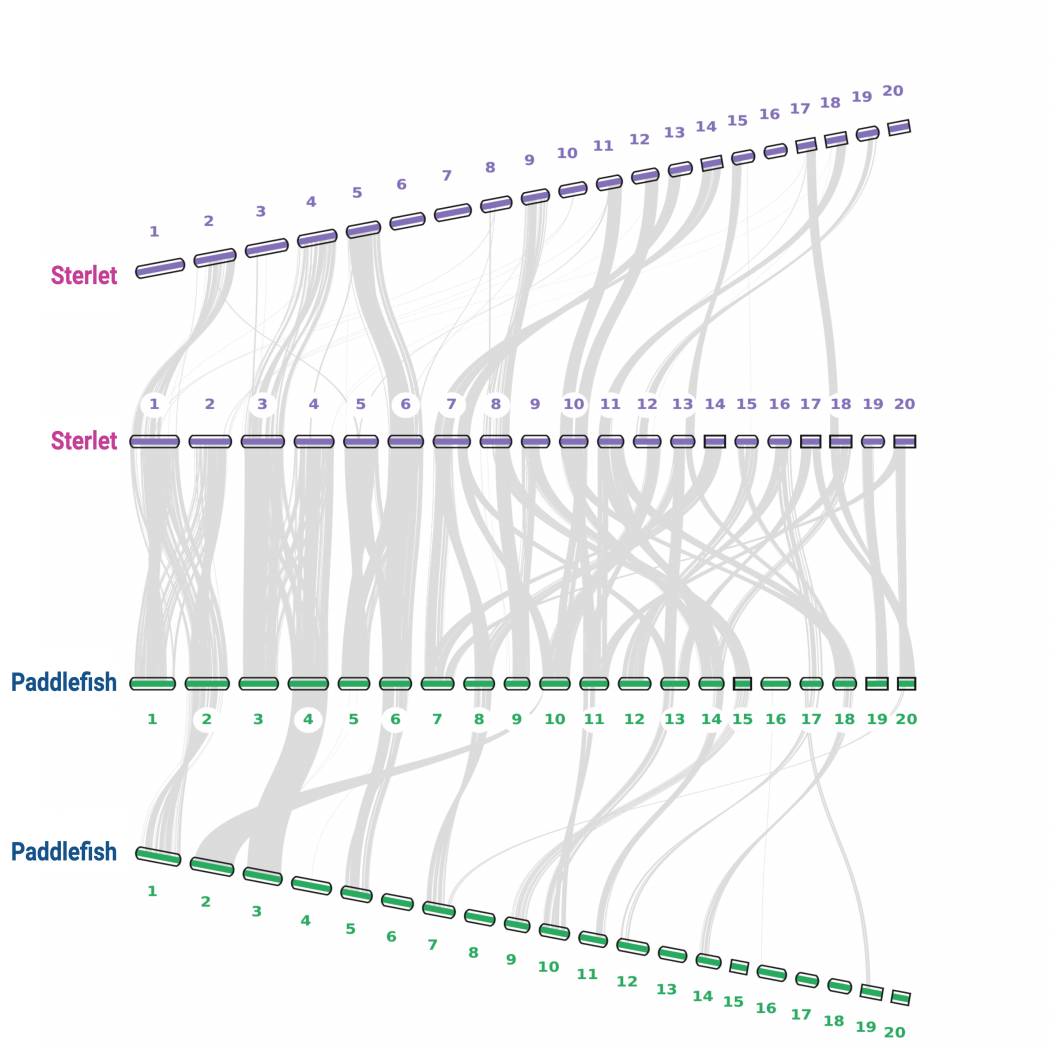

Figure 5.6 | Synteny blocks found using i-ADHoRe (Proost et al., 2012) within and between the 20 largest chromosomes of paddlefish (green) and sterlet (purple). Syntenic orthologs within each species are linked by grey ribbons, showcasing both intra- and inter-species synteny.)

5.3.3 Micro-syntenic blocks with common rediploidization histories

After identifying the syntenic blocks, a more in-depth analysis of select blocks was imperative. Table 5.3 shows the numbers and types of micro-syntenic blocks in the sturgeon and paddlefish genomes characterized by stretches of ohnologs with con-

sistent gene tree topologies, either PreSpec (699 PreSpec gene trees) or PostSpec (490 in paddlefish and 741 in sturgeon). Notably, there are substantially fewer genes within PostSpec blocks compared to PreSpec blocks. The former therefore, are smaller blocks with an average of ∼8 ohnologs per block, while the latter are larger with an average of ∼18 genes per block. This observation aligns with a narrative that PreSpec blocks, representing earlier rediploidization events, occur on larger chromosomes in extensive segments while more recent rediploidization events are characterized by smaller, localized rearrangement mechanisms(Lien et al., 2016; Redmond et al., 2023).

Sterlet and Paddlefish chromosomes 5 and 6 (Fig5.7) contain an examples of a PreSpec blocks, as described above. The PreSpec derived ohnologs show conservation of gene order across the chromosomes of the two lineages (Fig5.7)(A)). On chromosomes 1 and 2 of both species, there is increased gene movement. However, the ohnologs have consistent rediploidization histories in this blocks (Fig5.7 (B)). These visualisations are just two examples of many that we see following this structure throughout the two genomes.

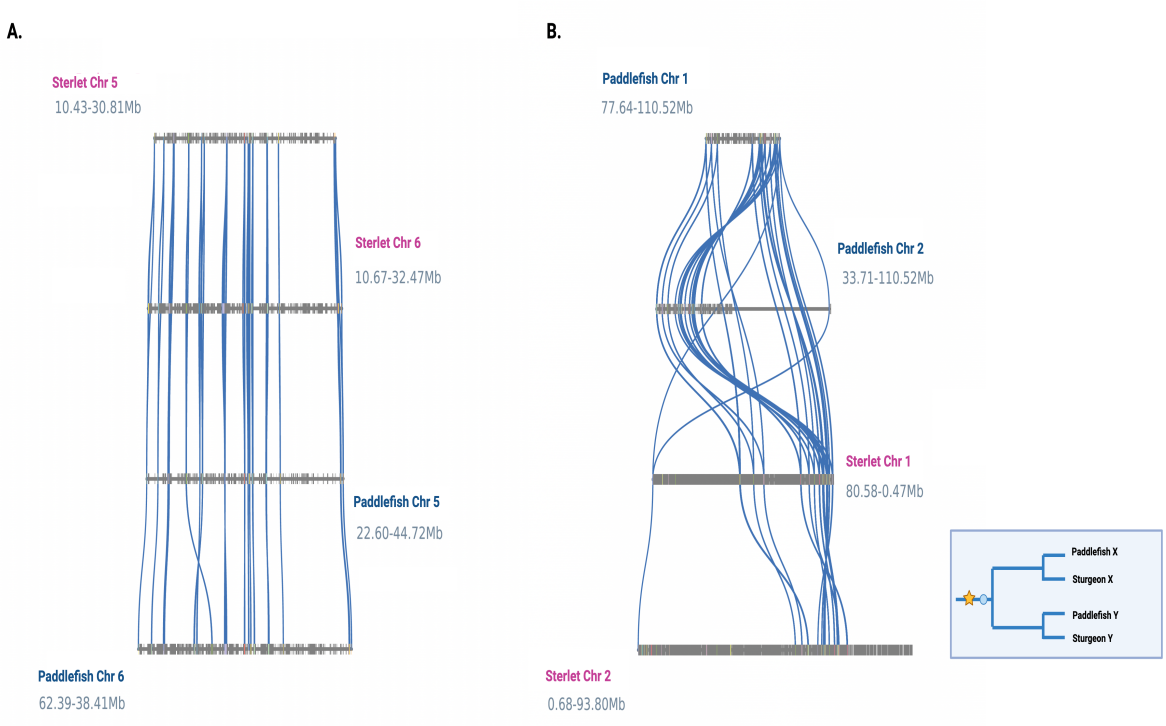

Figure 5.7 | Examples of PreSpec micro-syntenic blocks in the sterlet and American paddlefish genome Blue links connecting chromosomes are ohnolog pairs that rediploidized in the ancestor of the lineages. (A) Illustrates a syntenic block with highly conserved gene-order among PreSpec ohnologs in both species. (B) A micro-syntenic block found between chromosomes 1 and 2 of the paddlefish and sturgeon respectively, exhibiting some gene movement of the PreSpec ohnolog pairs (blue) within the block. The legend provides an illustration of the gene tree topologies of the ohnologs, coloured according to the PreSpec (blue) or PostSpec (red) tree topology.

A block of ohnologs with PostSpec rediploidization histories between paddlefish chromosomes 10 and 11 and Sterlet chromosomes 10 and 12 shows genes clustered in three smaller-blocks with intervening orthologs (links not shown). While its plausible their return to bivalent pairing may have happened all at once in each lineage for these genes, the clustering of the genes could also indicate six distinct rediploidzation events (3 in the sturgeon and 3 in the paddlefish). We visualise a more complex situation in In Fig5.8 (B). While majority of the block has a PostSpec rediploidization history, the end of the block has a stretch of

PreSpec ohnologs. This may be two adjacent blocks, that have had three distinct rearrangement events happening at different time points in the lineages histories - In the ancestor (blue ohnologs) and in a lineage-specific manner in the paddlefish and the sturgeon (red ohnologs) - followed by some gene movement within the block. Many other possibilities are plausible for this block.

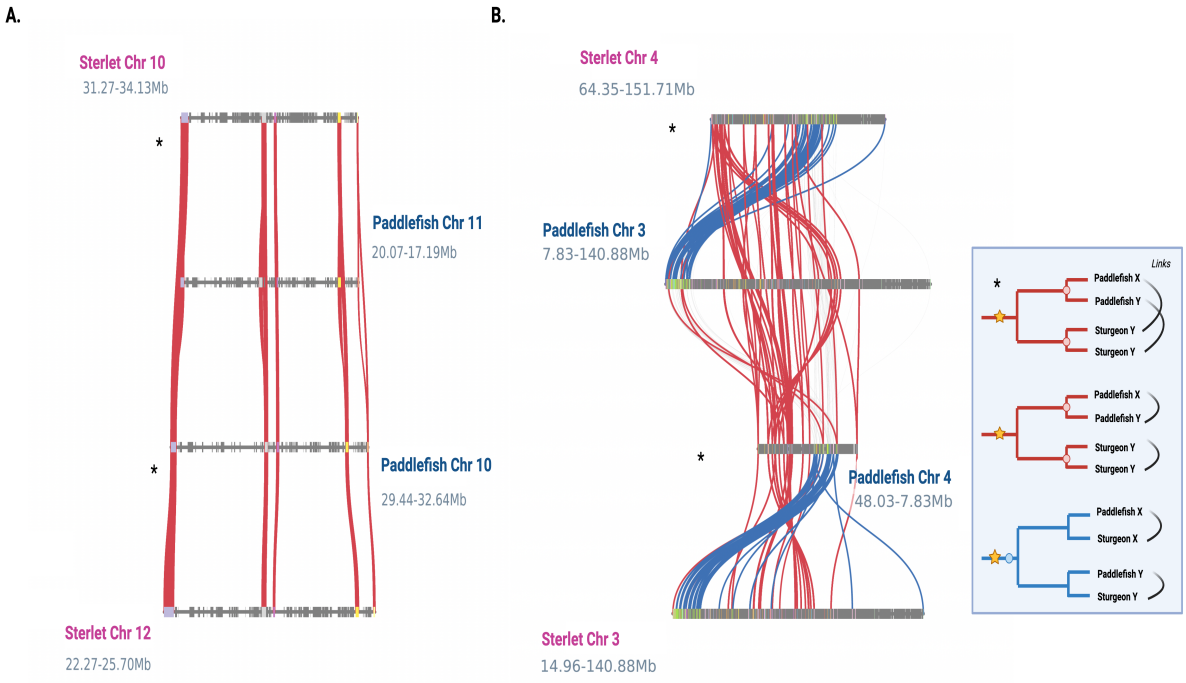

Figure 5.8 | Examples of PostSpec micro-syntenic blocks in the sterlet and American paddlefish genomes Links are coloured according to the PreSpec (blue) or PostSpec (red) rediploidization histories of the ohnologs. (A) Illustrates a syntenic block with highly conserved gene-order among PostSpec ohnologs (red) in both species (Interveining orthologs are not linked but are depicted as dashed lines on a chromosome). (B) A micro-syntenic block found between chromosomes 10 and 11 of paddlefish and 10 and 12 in sturgeon. This block has experienced both pre- and post-species rediploidization events. The legend illustrates gene tree topologies of ohnologs, color-coded for PreSpec (blue) or PostSpec (red) rediploidization history. Asterisks on gene trees in the legend correspond to ohnolog relationships in the figure with corresponding asterisks.

Finally, we observe complex blocks featuring both PreSpec and PostSpec ohnologs together (Fig. 5.9). In Fig. 5.9 (B), although the block is primarily

composed of PreSpec ohnologs, the positioning of PostSpec genes could suggest an explanation beyond rearrangement. An alternative interpretation is that an accumulation of mutations prevented homologous recombination between the tetravalent ohnologs, leading to the rediploidized loci we see scattered throughout the block here. Gene movement could also have played a part in this organisation of genes with different histories. Due to the length of time between the WGD event and the preceding speciation, it would not be uncommon to see a lot more movement within blocks.

Species	**PreSpec blocks**	**Complex PreSpec blocks**	**PostSpec blocks**	**Complex PostSpec blocks**	**Genes in PreSpec**	**Genes in PostSpec**
Paddlefish (60 chr)	699	20	490	15	6069	1793
Sturgeon (60 chr)	699	20	741	41	6247	2371

Table 5.3 | Micro-syntenic blocks found in and between the American paddlefish genome and the sterlet sturgeon genome and the ohnologs found within those blocks

In Table 5.3, following our criteria for defining PostSpec or PreSpec blocks, we identify 56 instances of complex situations, like these, where blocks consist of genes from both rediploidization histories. While our study provides clear evidence supporting rearrangement as a mechanism for rediploidization, it also must be noted that there are alternative methods that may contribute to this crucial evolutionary process.

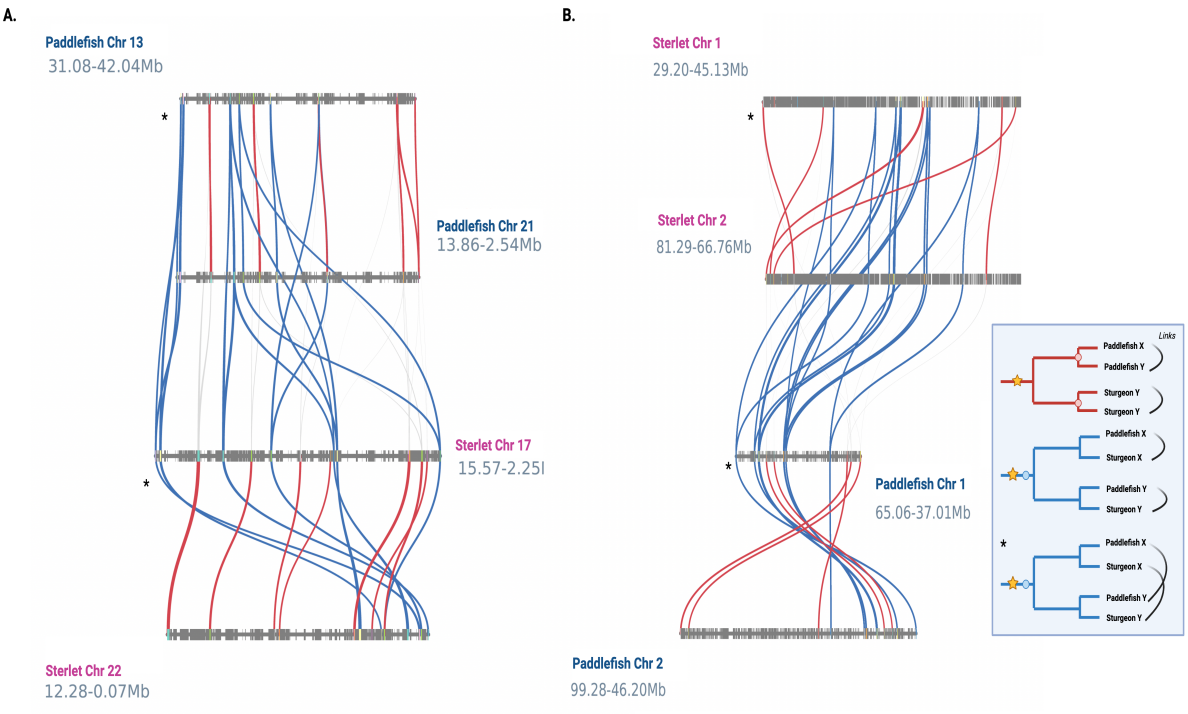

Figure 5.9 | Examples of micro-syntenic blocks in the sterlet and the American paddlefish genome with complex rediploidization histories (A) A micro-syntenic block identified between chromosomes 21 and 13 in paddlefish and 17 and 22 in sturgeon. This block has undergone both pre- and post-speciation rediploidization events, resulting in genes from different time-points scattered throughout the block. (B) A micro-syntenic block observed between chromosomes 1 and 2 in paddlefish and 1 and 2 in sturgeon. While this block predominantly contains PreSpec ohnologs, some ohnologs have independently undergone rediploidization in each species. Grey lines link PostSpec ohnologs and are not coloured to emphasis their species-specific rediploidization history. The legend illustrates gene tree topologies of ohnologs, color-coded for PreSpec (blue) or PostSpec (red) rediploidization history. Asterisks on gene trees in the legend correspond to ohnolog relationships in the figure with corresponding asterisks.

5.4 Discussion

In this final chapter, we take a closer look at the mechanisms of rediploidization following tetraploidisation in the American paddlefish and the sterlet sturgeon genomes. Using the newly assembled paddlefish genome described in Chapter 4,

we found 5693 high confidence ohnolog pairs in sturgeon and paddlefish. This is an improvement on the previous dataset published in Redmond et al., 2023. Notably, compared to previous work, we found much more gene-trees with shared duplication nodes, refereed to in this thesis as PreSpec. We recovered 2054 of these trees, over 500 more than the findings in Redmond et al., 2023, where more PostSpec ohnolog gene trees were reported. Our results imply that our enhanced paddlefish assembly may have exposed additional duplicates originating from genomic regions that underwent rediploidization in the common ancestor of the lineages. Interestingly, there is a noteworthy decrease in the number of PostSpec gene trees compared to Redmond et al., 2023. Over 300 ohnolog gene trees initially categorized as PostSpec in Redmond et al., 2023 have been reclassified here. 220 of these ohnologs are now placed in the Other-PostSpec category, contributing, therefore, to the increased number in this group. Additionally, 82 have been identified in the "Other PreSpec" group and the remaining missing PostSpec ohnologs were resolved as PreSpec gene trees in our current analysis. This was an unexpected result, as one might assume that PostSpec topologies would exhibit more similarity due to their more recent rediploidization history and, consequently, a more recent divergence time of the duplicates. We suspect that finding more PostSpec regions was not a result of increasing the size of the genome but rather an improved annotation of the genome contributed to finding more PostSpec genes. Also, while our methodology closely followed the previous analysis (Redmond et al., 2023), the use of different parameters for aligning the ML gene trees and some difference in animals used in orthology inference may have influenced the resulting topologies. These adjustments and the new paddlefish assembly likely contributed to the observed variation in the recovered gene trees and their categorization in the two studies.

As had been identified before, we found on visualisation of the ohnolog pairs from

our dataset that there are long-stretches of PreSpec and PostSpec genes found along the larger chromosomes of the paddlefish and sturgeon (Fig5.6). This follows the idea that if rearrangements are a mechanism for rediploidization, then large blocks of neighbouring genes sharing common rediploidization histories should be visible as largely non-overlapping syntenic blocks on different chromosomes, and present in both lineages; just as we see here. Examining the outcomes from i-ADHoRe (Proost et al., 2012), a clearer picture emerges. Micro-syntenic blocks, indicative of chromosomal homology, are conserved along the six largest chromosomes in both paddlefish and sturgeon. A closer inspection of these blocks, coupled with the identification of the ohnologs within the blocks and their duplication nodes, unveils, for the first time, compelling evidence supporting rearrangement as a primary mechanism driving rediploidization in both paddlefish and sturgeon (Fig5.7, Fig5.9, Fig5.8).

An important facet of this work is the timings of these events. Asynchronous rediploidization temporally separates ohnolog divergence from WGD, obscuring the dating of autopolyploidy events. While we can have some certainty using phylogenetics that an event happened before speciation or after speciation, the actual timings of these events is harder to unravel. For the pre-speciation events, the timings of some of these events may have occurred immediately post WGD and others closer to speciation. It appears evident that pre-speciation rearrangements predominantly occur on larger chromosomes and involve larger segments, as indicated in Table 5.3. This phenomenon might be ascribed to the overall larger *surface area* of these chromosomes. This increased size could potentially lead them to rediploidize sooner and cause the blocks to be more extensive. Future research should place emphasis on examining the genes within these early-recombining blocks. Investigating whether dosage or gene function influenced which regions returned to a diploid state first will be crucial for gaining insights into the dynamics

of rediploidization. For post-speciation blocks, rearrangements are lineage-specific. However, distinguishing between these timings is challenging, and most studies oversimplify the reality of these events and assume lineage-specific ohnologs undergo rediploidization simultaneously in each lineage (as we have done here). A more nuanced understanding of these temporal dynamics is essential for unraveling the mechanisms underlying the rediploidization process and the events following WGD.

While blocks with a consistent topology, either PreSpec or PostSpec, predominate the results, we also find that there are blocks where both topologies are evident in a stretch of orthologs (table 5.3). These findings shed light on the likelihood that, while rearrangements serve as the main mechanism, there may be alternative pathways for genes to revert back to bivalent pairing. In Fig. 5.9, instances are observed where distinct rediploidization histories are interspersed within a block. It is plausible that these genes accumulated mutations, rendering the alleles sufficiently dissimilar for homologous recombination between the four alleles to cease at this locus. A scenario also exists, where rearrangements were restricted to fewer or even a single gene or that subsequent rearrangements disrupted the contiguity of the blocks. While these animals exhibit considerable synteny across their genomes, the movement of genes is inevitable given the extended evolutionary history of these lineages (Redmond et al., 2023). Ohnologs within these complex blocks suggest that some loci initiated divergence in the ancestor, while others retained tetraploidy until after speciation. This complex scenario warrants further investigation into the mechanisms underlying these events, be it rearrangement or other processes, and the consequences of asynchronous divergence of duplicate genes and their roles in adaptive evolution. The rediploidization phenomenon explored in this study represents uncharted territory, highlighting the importance of further exploration in this field. It is apparent that more studies focusing on

the situations that emerge in the aftermath WGD are essential.

The asynchronicity of the rediploidization events we have discussed can lead to genomes with a mosaic of shared and lineage-specific gene duplications. With a clearer picture of the events that proceed polyploidy, we can better understand the evolutionary trajectories of these lineages and their complex histories. By re-examining known WGD events in the light of delayed rediploidization and mechanisms of this process, we can reinterpret the evolution of many lineages who've experienced WGD, including plants and animals, and probe difficult questions about the number and timings of WGD in early vertebrates.

Chapter 6

Conclusion

This work highlights the importance of synteny, its conservation and movement, in aiding in determination of evolutionary relationships. Throughout this thesis I have shown how synteny can be incorporated into various aspects of molecular evolution studies. This inclusion has proven instrumental in refining and, in some cases, offering novel perspectives on longstanding, as well as emerging, questions in the field.

In Chapter 3 we showed how micro-synteny can be used to build whole-genome phylogenies of diverse lineages. Gene order evolution follows an independent mechanism compared to sequence changes, providing new information to help resolve unanswered questions. While, our results suggest that it may not be optimal to use micro-synteny as a character for phylogenetic inference alone, it does highlights the potential for combining multiple methods to conduct a more comprehensive analysis of species relationships. Three pivotal aspects for understanding genome evolution are; nucleotide sequence changes, gene and genome duplication events and finally rearrangements or movement of genes. While most molecular evolution studies concentrate on sequence changes among lineages, this thesis underscores the equal significance of genome duplication events and genome rearrangements

in evolution. The integration of synteny information and WGD information, as demonstrated throughout this thesis, yields nuanced insights into the evolutionary histories of lineages.

Over the past few years, research adopting alternative methodologies beyond conventional sequence alignment-based phylogenetic inference has proved powerful. These approaches have been instrumental in resolving contentious species relationships and have introduced a fresh perspective on results from previous analyses (Parey et al., 2023; Simakov et al., 2020; Nakatani et al., 2021; Zhao et al., 2021). As we have shown in this thesis, the conservation of synteny across groups of genomes emerges as a potent character for examining relationships, offering a valuable complement to sequence-alignment approaches. Regrettably, the effective integration of genome organization information and sequence evolution is hindered by the lack of tools available and as we discussed in Chapter 3, inadequate evolutionary models for such endeavours (Parey et al., 2022; Zhao et al., 2021). In the coming years we expect that synteny-aware methods will become more commonplace which along with more high-quality reference genome assemblies and advanced phylogenetic methods, will allow more concrete answers for many of the unresolved questions in animal evolution.

Overall, the events that ensue after WGD have remained somewhat elusive, with many studies overlooking the intricate scenarios that unfold in the aftermath of these doubling events. In the final chapter, we look at a shared rediploidization event in the sturgeon and paddlefish. In the past few years there have been more publications on the topic of rediploidization following WGD, an event that until recently was assumed to happen instantaneously (Redmond et al., 2023; Lien et al., 2016; Robertson et al., 2017; Parey et al., 2022; Du et al., 2020; Macqueen and Johnston, 2014). In many cases, as we have seen here, rediploidization can be delayed and occur over tens of millions of year. Extensive lineage-specific

rediploidization has major implications for our understanding of genome evolution following polyploidy and our interpretation of duplicate gene evolution including their role in adaptive evolution. The framework for this analysis and the nuanced interpretation of evolution following WGD will prompt a re-examination of other autopolyploidy events including the 2R event at the base of the vertebrate lineage and the numerous events in plant lineages. In future work, it will be important to probe the effects of the asynchronous generation of duplicate genes on adaptive evolution. For example, what genes rediploidized first in these lineages, why? and what effects might that have had on the ancestral lineages. The complex histories of animals exhibiting delayed effects of WGD will offer valuable insight into the significance of the events in promoting the survival and prosperity of the lineages in which the event transpired. In the context of sturgeon and paddlefish, there is compelling evidence pointing to the potential role of polyploidy and asynchronous rediploidization in the lineage's resilience during the Permian-Triassic (P-Tr) and/or Triassic-Jurassic (Tr-J) mass extinction events (Redmond et al., 2023). Survival effects like this have also been proposed in plants where WGD may confer tolerance and adaptability to extreme environmental conditions, increasing fitness in the face of mass extinction events (Vanneste, Baele, et al., 2014; Jaramillo et al., 2010).

In this thesis, my aim is to demonstrate that the integration of synteny information and its evolutionary dynamics into phylogenetic analyses can enhance existing frameworks and contribute to resolving numerous unanswered questions in animal evolution. Furthermore, our exploration of rediploidization mechanisms calls for a reassessment and re-examination of evolution post WGD, probing events like 2R at the base of all vertebrates and the many WGD events within plant lineages.

Bibliography

Abadi, Shiran, Dana Azouri, Tal Pupko, and Itay Mayrose (2019). "Model selection may not be a mandatory step for phylogeny reconstruction". *Nature Communications* 10.1, p. 934. DOI: 10.1038/s41467-019-08822-w.

Adams, Keith L and Jonathan F Wendel (2005). "Polyploidy and genome evolution in plants". *Current Opinion in Plant Biology* 8.2. Genome studies and molecular genetics / Plant biotechnology, pp. 135–141. DOI: https://doi.org/10.1016/j.pbi.2005.01.001.

Adams, Mark D. et al. (2000). "The Genome Sequence of Drosophila melanogaster". *Science* 287.5461, pp. 2185–2195. DOI: 10.1126/science.287.5461.2185.

Alekseyenko, Alexander V, Christopher J Lee, and Marc A Suchard (2008). "Wagner and Dollo: a stochastic duet by composing two parsimonious solos". *Systematic biology* 57.5, pp. 772–784.

Allendorf, Fred W., Susan Bassham, William A. Cresko, Morten T. Limborg, Lisa W. Seeb, and James E. Seeb (2015). "Effects of Crossovers Between Homeologs on Inheritance and Population Genomics in Polyploid-Derived Salmonid Fishes". *Journal of Heredity* 106.3, pp. 217–227. DOI: 10.1093/jhered/esv015.

Álvarez-Carretero, Sandra, Asif U. Tamuri, Matteo Battini, Fabrícia F. Nascimento, Emily Carlisle, Robert J. Asher, Ziheng Yang, Philip C. J. Donoghue, and Mario dos Reis (2021). "A Species-Level Timeline of Mammal Evolution

Integrating Phylogenomic Data". *Nature*, pp. 1–8. DOI: 10.1038/s41586-021-04341-1.

Amarasinghe, Shanika L, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil (2020). "Opportunities and challenges in long-read sequencing data analysis". *Genome biology* 21.1, pp. 1–16.

Ashton, Philip M, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady (2015). "MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island". *Nature Biotechnology* 33.3, pp. 296–300. DOI: 10.1038/nbt.3103.

Bailey, Jeffrey A and Evan E Eichler (2006). "Primate segmental duplications: crucibles of evolution, diversity and disease". *Nature Reviews Genetics* 7.7, pp. 552–564.

Barrière, Antoine, Shiaw-Pyng Yang, Elizabeth Pekarek, Cristel G. Thomas, Eric S. Haag, and Ilya Ruvinsky (2009). "Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes". *Genome Research* 19.3, pp. 470–480. DOI: 10.1101/gr.081851.108.

Batzoglou, Serafim, David B Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P Mesirov, and Eric S Lander (2002). "ARACHNE: a whole-genome shotgun assembler". *Genome research* 12.1, pp. 177–189.

Belda, Eugeni, Andrés Moya, and Francisco J. Silva (Mar. 2005). "Genome Rearrangement Distances and Gene Order Phylogeny in 03b3-Proteobacteria". *Molecular Biology and Evolution* 22.6, pp. 1456–1467. DOI: 10.1093/molbev/msi134.

Bengtson, Stefan, John A. Cunningham, Chongyu Yin, and Philip C.J. Donoghue (2012). "A merciful death for the "earliest bilaterian," Vernanimalcula". *Evo-*

lution & Development 14.5, pp. 421–427. DOI: 10.1111/j.1525-142x.2012.00562.x.

Berthelot, Camille, Frédéric Brunet, Domitille Chalopin, Amélie Juanchich, Maria Bernard, Benjamin Noël, Pascal Bento, Corinne Da Silva, Karine Labadie, Adriana Alberti, Jean-Marc Aury, Alexandra Louis, Patrice Dehais, Philippe Bardou, Jérôme Montfort, Christophe Klopp, Cédric Cabau, Christine Gaspin, Gary H. Thorgaard, Mekki Boussaha, Edwige Quillet, René Guyomard, Delphine Galiana, Julien Bobe, Jean-Nicolas Volff, Carine Genêt, Patrick Wincker, Olivier Jaillon, Hugues Roest Crollius, and Yann Guiguen (2014). "The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates". *Nature Communications* 5.1, p. 3657. DOI: 10.1038/ncomms4657.

Bi, Xupeng, Kun Wang, Liandong Yang, Hailin Pan, Haifeng Jiang, Qiwei Wei, Miaoquan Fang, Hao Yu, Chenglong Zhu, Yiran Cai, Yuming He, Xiaoni Gan, Honghui Zeng, Daqi Yu, Youan Zhu, Huifeng Jiang, Qiang Qiu, Huanming Yang, Yong E. Zhang, Wen Wang, Min Zhu, Shunping He, and Guojie Zhang (2021). "Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes". *Cell* 184.5, 1377–1391.e14. DOI: 10.1016/j.cell.2021.01.046.

Biscotti, Maria Assunta, Ettore Olmo, and J. S. (Pat) Heslop-Harrison (2015a). "Repetitive DNA in eukaryotic genomes". *Chromosome Research* 23.3, pp. 415–420. DOI: 10.1007/s10577-015-9499-z.

— (2015b). "Repetitive DNA in eukaryotic genomes". *Chromosome Research* 23.3, pp. 415–420. DOI: 10.1007/s10577-015-9499-z.

Blanchette, Mathieu, Takashi Kunisawa, and David Sankoff (1999). "Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny". *Journal of Molecular Evolution* 49.2, pp. 193–203. DOI: 10.1007/pl00006542.

Bonfield, James K, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies (2021). "HTSlib: C library for reading/writing high-throughput sequencing data". *GigaScience* 10.2, giab007–. DOI: 10.1093/gigascience/giab007.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond (2014). "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis". *PLoS Computational Biology* 10.4, e1003537. DOI: 10.1371/journal.pcbi.1003537.

Braasch, Ingo, Andrew R Gehrke, Jeramiah J Smith, Kazuhiko Kawasaki, Tereza Manousaki, Jeremy Pasquier, Angel Amores, Thomas Desvignes, Peter Batzel, Julian Catchen, Aaron M Berlin, Michael S Campbell, Daniel Barrell, Kyle J Martin, John F Mulley, Vydianathan Ravi, Alison P Lee, Tetsuya Nakamura, Domitille Chalopin, Shaohua Fan, Dustin Wcisel, Cristian Cañestro, Jason Sydes, Felix E G Beaudry, Yi Sun, Jana Hertel, Michael J Beam, Mario Fasold, Mikio Ishiyama, Jeremy Johnson, Steffi Kehr, Marcia Lara, John H Letaw, Gary W Litman, Ronda T Litman, Masato Mikami, Tatsuya Ota, Nil Ratan Saha, Louise Williams, Peter F Stadler, Han Wang, John S Taylor, Quenton Fontenot, Allyse Ferrara, Stephen M J Searle, Bronwen Aken, Mark Yandell, Igor Schneider, Jeffrey A Yoder, Jean-Nicolas Volff, Axel Meyer, Chris T Amemiya, Byrappa Venkatesh, Peter W H Holland, Yann Guiguen, Julien Bobe, Neil H Shubin, Federica Di Palma, Jessica Alföldi, Kerstin Lindblad-Toh, and John H Postlethwait (2016). "The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons". *Nature Genetics* 48.4, pp. 427–437. DOI: 10.1038/ng.3526.

Bromham, Lindell, Megan Woolfit, Michael S. Y. Lee, and Andrew Rambaut (2002). "TESTING THE RELATIONSHIP BETWEEN MORPHOLOGICAL

AND MOLECULAR RATES OF CHANGE ALONG PHYLOGENIES". *Evolution* 56.10, pp. 1921–1930. DOI: 10.1111/j.0014-3820.2002.tb00118.x.

Brown, Joseph W, Caroline Parins-Fukuchi, Gregory W Stull, Oscar M Vargas, and Stephen A Smith (2017). "Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick et al." *Proceedings of the Royal Society B: Biological Sciences* 284.1864, p. 20170986.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost (2021). "Sensitive protein alignments at tree-of-life scale using DIAMOND". *Nature Methods* 18.4, pp. 366–368. DOI: 10.1038/s41592-021-01101-x.

Caldas, Ian V. and Carlos G. Schrago (2019). "Data partitioning and correction for ascertainment bias reduce the uncertainty of placental mammal divergence times inferred from the morphological clock". *Ecology and Evolution* 9.4, pp. 2255–2262. DOI: 10.1002/ece3.4921.

Cañestro, Cristian, Ricard Albalat, Manuel Irimia, and Jordi Garcia-Fernàndez (2013). "Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates". *Seminars in Cell Developmental Biology* 24.2. Gene Duplication in Vertebrate Development, pp. 83–94. DOI: https://doi.org/10.1016/j.semcdb.2012.12.008.

Cannon, Johanna Taylor, Bruno Cossermelli Vellutini, Julian Smith, Fredrik Ronquist, Ulf Jondelius, and Andreas Hejnol (2016). "Xenacoelomorpha is the sister group to Nephrozoa". *Nature* 530.7588, pp. 89–93.

Cantor, Charles R and Thomas H Jukes (1966). "The repetition of homologous sequences in the polypetide chains of certain cytochromes and globins." *Proceedings of the National Academy of Sciences* 56.1, pp. 177–184.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón (2009). "trimAl: a tool for automated alignment trimming in large-scale phylogenetic

analyses". *Bioinformatics* 25.15, pp. 1972–1973. DOI: 10.1093/bioinformatics/btp348.

Cechova, Monika (2020). "Probably Correct: Rescuing Repeats with Short and Long Reads". *Genes* 12.1, p. 48. DOI: 10.3390/genes12010048.

Cheng, Haoyu, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li (2021). "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm". *Nature methods* 18.2, pp. 170–175.

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 (2020). "The American Paddlefish Genome Provides Novel Insights into Chromosomal Evolution and Bone Mineralization in Early Vertebrates". *Molecular Biology and Evolution* 38.4, msaa326. DOI: 10.1093/molbev/msaa326.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, et al. (2016). "Phased diploid genome assembly with single-molecule real-time sequencing". *Nature methods* 13.12, pp. 1050–1054.

Cilibrasi, Rudi, Leo Van Iersel, Steven Kelk, and John Tromp (2007). "The complexity of the single individual SNP haplotyping problem". *Algorithmica* 49, pp. 13–36.

Comai, Luca (2005). "The advantages and disadvantages of being polyploid". *Nature Reviews Genetics* 6.11, pp. 836–846. DOI: 10.1038/nrg1711.

Conant, Gavin C. and Kenneth H. Wolfe (2008). "Turning a hobby into a job: How duplicated genes find new functions". *Nature Reviews Genetics* 9.12, pp. 938–950. DOI: 10.1038/nrg2482.

Consortium, C. elegans Sequencing (1998). "Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology". *Science* 282.5396, pp. 2012–2018. DOI: 10.1126/science.282.5396.2012.

Coombe, Lauren, Janet X. Li, Theodora Lo, Johnathan Wong, Vladimir Nikolic, René L. Warren, and Inanc Birol (2021). "LongStitch: high-quality genome assembly correction and scaffolding using long reads". *BMC Bioinformatics* 22.1, p. 534. DOI: 10.1186/s12859-021-04451-7.

Crow, Karen D., Christopher D. Smith, Jan-Fang Cheng, Günter P. Wagner, and Chris T. Amemiya (2012a). "An Independent Genome Duplication Inferred from Hox Paralogs in the American Paddlefish—A Representative Basal Ray-Finned Fish and Important Comparative Reference". *Genome Biology and Evolution* 4.9, pp. 937–953. DOI: 10.1093/gbe/evs067.

— (2012b). "An Independent Genome Duplication Inferred from Hox Paralogs in the American Paddlefish—A Representative Basal Ray-Finned Fish and Important Comparative Reference". *Genome Biology and Evolution* 4.9, pp. 937–953. DOI: 10.1093/gbe/evs067.

Davies, T. Jonathan and Vincent Savolainen (2006). "NEUTRAL THEORY, PHYLOGENIES, AND THE RELATIONSHIP BETWEEN PHENOTYPIC CHANGE AND EVOLUTIONARY RATES". *Evolution* 60.3, pp. 476–483. DOI: 10.1111/j.0014-3820.2006.tb01129.x.

De Bruijn, Nicolaas Govert (1946). "A combinatorial problem". *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam* 49.7, pp. 758–764.

Dehal, Paramvir and Jeffrey L Boore (2005). "Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate". *PLoS Biology* 3.10, e314. DOI: 10.1371/journal.pbio.0030314.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner (2002). "Capturing chromosome conformation". *science* 295.5558, pp. 1306–1311.

Donoghue, Philip C.J. and Mark A. Purnell (2005). "Genome duplication, extinction and vertebrate evolution". *Trends in Ecology Evolution* 20.6. SPECIAL ISSUE: BUMPER BOOK REVIEW, pp. 312–319. DOI: https://doi.org/10.1016/j.tree.2005.04.008.

Drillon, Guénola, Alessandra Carbone, and Gilles Fischer (2014). "SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes". *PLoS ONE* 9.3, e92621. DOI: 10.1371/journal.pone.0092621.

Drillon, Guénola, Raphaël Champeimont, Francesco Oteri, Gilles Fischer, and Alessandra Carbone (2020). "Phylogenetic Reconstruction Based on Synteny Block and Gene Adjacencies". *Molecular Biology and Evolution* 37.9, msaa114. DOI: 10.1093/molbev/msaa114.

Du, Kang, Matthias Stöck, Susanne Kneitz, Christophe Klopp, Joost M. Woltering, Mateus Contar Adolfi, Romain Feron, Dmitry Prokopov, Alexey Makunin, Ilya Kichigin, Cornelia Schmidt, Petra Fischer, Heiner Kuhl, Sven Wuertz, Jörn Gessner, Werner Kloas, Cédric Cabau, Carole Iampietro, Hugues Parrinello, Chad Tomlinson, Laurent Journot, John H. Postlethwait, Ingo Braasch, Vladimir Trifonov, Wesley C. Warren, Axel Meyer, Yann Guiguen, and Manfred Schartl (2020). "The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization". *Nature Ecology & Evolution*, pp. 1–12. DOI: 10.1038/s41559-020-1166-x.

Dubcovsky, Jorge and Jan Dvorak (2007). "Genome plasticity a key factor in the success of polyploid wheat under domestication". *Science* 316.5833, pp. 1862–1866.

Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, Ido Machol, Eric S. Lander, Aviva Presser Aiden, and Erez Lieberman Aiden (2017). "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds". *Science* 356.6333, pp. 92–95. DOI: 10.1126/science.aal3327.

Dunn, Casey W., Sally P. Leys, and Steven H.D. Haddock (2015). "The hidden biology of sponges and ctenophores". *Trends in Ecology & Evolution* 30.5, pp. 282–291. DOI: 10.1016/j.tree.2015.03.003.

Durand, Neva C, Muhammad S Shamim, Ido Machol, Suhas S P Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden (2016). "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments". *Cell Systems* 3.1, pp. 95–98. DOI: 10.1016/j.cels.2016.07.002.

Edge, Peter, Vineet Bafna, and Vikas Bansal (2017). "HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies". *Genome Research* 27.5, pp. 801–812. DOI: 10.1101/gr.213462.116.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner (2009). "Real-Time DNA Sequencing from Sin-

gle Polymerase Molecules". *Science* 323.5910, pp. 133–138. DOI: 10.1126/science.1162986.

Emms, David M. and Steven Kelly (2019). "OrthoFinder: phylogenetic orthology inference for comparative genomics". *Genome Biology* 20.1, p. 238. DOI: 10.1186/s13059-019-1832-y.

Eric Schranz, M, Setareh Mohammadin, and Patrick P Edger (2012). "Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model". *Current Opinion in Plant Biology* 15.2. Genome studies molecular genetics, pp. 147–153. DOI: https://doi.org/10.1016/j.pbi.2012.03.011.

Felsenstein, Joseph (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach". *Journal of molecular evolution* 17, pp. 368–376.

— (1985). "Phylogenies and the Comparative Method". *The American Naturalist* 125.1, pp. 1–15. (Visited on 08/06/2023).

Feng, Bing, Yu Lin, Lingxi Zhou, Yan Guo, Robert Friedman, Ruofan Xia, Fei Hu, Chao Liu, and Jijun Tang (2017). "Reconstructing Yeasts Phylogenies and Ancestors from Whole Genome Data". *Scientific Reports* 7.1, p. 15209. DOI: 10.1038/s41598-017-15484-5.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data". *Bioinformatics* 28.23, pp. 3150–3152. DOI: 10.1093/bioinformatics/bts565.

Furlong, Rebecca F. and Peter W. H. Holland (2002). "Were vertebrates octoploid?" *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357.1420, pp. 531–544. DOI: 10.1098/rstb.2001.1035.

Gans, Carl and R Glenn Northcutt (1983). "Neural crest and the origin of vertebrates: a new head". *Science* 220.4594, pp. 268–273.

Garrison, Erik and Gabor Marth (2012). *Haplotype-based variant detection from short-read sequencing.*

Gingerich, Philip D (1983). "Rates of evolution: effects of time and temporal scaling". *Science* 222.4620, pp. 159–161.

Glusman, G, HC Cox, and JC Roach (2014). *Whole-genome haplotyping approaches and genomic medicine. Genome Med 6: 1–16.*

Grabherr, Manfred G., Pamela Russell, Miriah Meyer, Evan Mauceli, Jessica Alföldi, Federica Di Palma, and Kerstin Lindblad-Toh (Mar. 2010). "Genome-wide synteny through highly sensitive sequence alignment: Satsuma". *Bioinformatics* 26.9, pp. 1145–1151. DOI: 10.1093/bioinformatics/btq102.

Guillén, Yolanda and Alfredo Ruiz (2012). "Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution". *BMC Genomics* 13.1, p. 53. DOI: 10.1186/1471-2164-13-53.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel (May 2010). "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0". *Systematic Biology* 59.3, pp. 307–321. DOI: 10.1093/sysbio/syq010.

Gundappa, Manu Kumar, Thu-Hien To, Lars Grønvold, Samuel A.M. Martin, Sigbjørn Lien, Juergen Geist, David Hazlerigg, Simen R. Sandve, and Daniel J. Macqueen (2021). "Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution". *bioRxiv*, p. 2021.06.05.447185. DOI: 10.1101/2021.06.05.447185.

Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and

the Program to Assemble Spliced Alignments". *Genome Biology* 9.1, R7. DOI: 10.1186/gb-2008-9-1-r7.

Halliday, Thomas J. D., Mario dos Reis, Asif U. Tamuri, Henry Ferguson-Gow, Ziheng Yang, and Anjali Goswami (2019). "Rapid morphological evolution in placental mammals post-dates the origin of the crown group". *Proceedings of the Royal Society B* 286.1898, p. 20182418. DOI: 10.1098/rspb.2018.2418.

Harmon, Luke J., Matthew W. Pennell, L. Francisco Henao-Diaz, Jonathan Rolland, Breanna N. Sipley, and Josef C. Uyeda (2021). "Causes and Consequences of Apparent Timescaling Across All Estimated Evolutionary Rates". *Annual Review of Ecology, Evolution, and Systematics* 52.1, pp. 1–23. DOI: 10.1146/annurev-ecolsys-011921-023644.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". *Journal of molecular evolution* 22, pp. 160–174.

Heimhofer, U., P.A. Hochuli, S. Burla, J.M.L. Dinis, and H. Weissert (2005). "Timing of Early Cretaceous angiosperm diversification and possible links to major paleoenvironmental change". *Geology* 33.2, pp. 141–144. DOI: 10.1130/g21053.1.

Henao Diaz, L Francisco, Luke J Harmon, Mauro TC Sugawara, Eliot T Miller, and Matthew W Pennell (2019). "Macroevolutionary diversification rates show time dependency". *Proceedings of the National Academy of Sciences* 116.15, pp. 7403–7408.

Higgins, Desmond G. and Paul M. Sharp (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer". *Gene* 73.1, pp. 237–244. DOI: https://doi.org/10.1016/0378-1119(88)90330-7.

Ho, Simon YW, Robert Lanfear, Lindell Bromham, Matthew J Phillips, Julien Soubrier, Allen G Rodrigo, and Alan Cooper (2011). "Time-dependent rates of molecular evolution". *Molecular ecology* 20.15, pp. 3087–3101.

Ho, Simon YW, Matthew J Phillips, Alan Cooper, and Alexei J Drummond (2005). "Time dependency of molecular rate estimates and systematic overestimation of recent divergence times". *Molecular biology and evolution* 22.7, pp. 1561–1568.

Hokamp, Karsten, Aoife McLysaght, and Kenneth H Wolfe (2003a). "The 2R hypothesis and the human genome sequence". *Genome Evolution: Gene and Genome Duplications and the Origin of Novel Gene Functions*, pp. 95–110.

— (2003b). "Genome Evolution, Gene and Genome Duplications and the Origin of Novel Gene Functions", pp. 95–110. DOI: 10.1007/978-94-010-0263-9_10.

Holland, Peter WH and Jordi Garcia-Fernàndez (1996). "HoxGenes and chordate evolution". *Developmental biology* 173.2, pp. 382–395.

Hotaling, Scott, Joanna L. Kelley, and Paul B. Frandsen (2021). "Toward a genome sequence for every animal: Where are we now?" *Proceedings of the National Academy of Sciences* 118.52, e2109019118. DOI: 10.1073/pnas.2109019118.

Huelsenbeck, John P and Fredrik Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees". *Bioinformatics* 17.8, pp. 754–755.

Huerta-Cepas, Jaime, François Serra, and Peer Bork (2016). "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". *Molecular Biology and Evolution* 33.6, pp. 1635–1638. DOI: 10.1093/molbev/msw046.

Hunt, Martin, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman, and Thomas D Otto (2013). "REAPR: a universal tool for genome assembly evaluation". *Genome Biology* 14.5, R47. DOI: 10.1186/gb-2013-14-5-r47.

Inc., Plotly Technologies (2015). *Collaborative data science*.

Jackson, R. C. (1982). "POLYPLOIDY AND DIPLOIDY: NEW PERSPECTIVES ON CHROMOSOME PAIRING AND ITS EVOLUTIONARY IMPLICATIONS". *American Journal of Botany* 69.9, pp. 1512–1523. DOI: https://doi.org/10.1002/j.1537-2197.1982.tb13400.x.

Jaillon, Olivier, Jean-Marc Aury, Frédéric Brunet, Jean-Louis Petit, Nicole Stange-Thomann, Evan Mauceli, Laurence Bouneau, Cécile Fischer, Catherine Ozouf-Costaz, Alain Bernot, Sophie Nicaud, David Jaffe, Sheila Fisher, Georges Lutfalla, Carole Dossat, Béatrice Segurens, Corinne Dasilva, Marcel Salanoubat, Michael Levy, Nathalie Boudet, Sergi Castellano, Véronique Anthouard, Claire Jubin, Vanina Castelli, Michael Katinka, Benoît Vacherie, Christian Biémont, Zineb Skalli, Laurence Cattolico, Julie Poulain, Véronique de Berardinis, Corinne Cruaud, Simone Duprat, Philippe Brottier, Jean-Pierre Coutanceau, Jérôme Gouzy, Genis Parra, Guillaume Lardier, Charles Chapple, Kevin J. McKernan, Paul McEwan, Stephanie Bosak, Manolis Kellis, Jean-Nicolas Volff, Roderic Guigó, Michael C. Zody, Jill Mesirov, Kerstin Lindblad-Toh, Bruce Birren, Chad Nusbaum, Daniel Kahn, Marc Robinson-Rechavi, Vincent Laudet, Vincent Schachter, Francis Quétier, William Saurin, Claude Scarpelli, Patrick Wincker, Eric S. Lander, Jean Weissenbach, and Hugues Roest Crollius (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype". *Nature* 431.7011, pp. 946–957. DOI: 10.1038/nature03025.

Jaramillo, Carlos, Diana Ochoa, Lineth Contreras, Mark Pagani, Humberto Carvajal-Ortiz, Lisa M Pratt, Srinath Krishnan, Agustin Cardona, Millerlandy Romero, Luis Quiroz, et al. (2010). "Effects of rapid global warming at the Paleocene-Eocene boundary on neotropical vegetation". *Science* 330.6006, pp. 957–961.

Jauhal, April A and Richard D Newcomb (2021). "Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO". *Molecular Ecology Resources* 21.5, pp. 1416–1421.

Kapli, Paschalia, Paschalis Natsidis, Daniel J. Leite, Maximilian Fursman, Nadia Jeffrie, Imran A. Rahman, Hervé Philippe, Richard R. Copley, and Maximilian J. Telford (2021). "Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria". *Science Advances* 7.12, eabe2741. DOI: 10.1126/sciadv.abe2741.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata (July 2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". *Nucleic Acids Research* 30.14, pp. 3059–3066. DOI: 10.1093/nar/gkf436.

Kelley, David R and Steven L Salzberg (2010). "Detection and correction of false segmental duplications caused by genome mis-assembly". *Genome Biology* 11.3, R28. DOI: 10.1186/gb-2010-11-3-r28.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg (2015). "HISAT: a fast spliced aligner with low memory requirements". *Nature Methods* 12.4, pp. 357–360. DOI: 10.1038/nmeth.3317.

Kimura, Motoo (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". *Journal of molecular evolution* 16, pp. 111–120.

Ko, Byung June, Chul Lee, Juwan Kim, Arang Rhie, Dong Ahn Yoo, Kerstin Howe, Jonathan Wood, Seoae Cho, Samara Brown, Giulio Formenti, Erich D. Jarvis, and Heebal Kim (2022). "Widespread false gene gains caused by duplication errors in genome assemblies". *Genome Biology* 23.1, p. 205. DOI: 10.1186/s13059-022-02764-1.

Konkel, Miriam K., Jerilyn A. Walker, and Mark A. Batzer (2010). "LINEs and SINEs of primate evolution". *Evolutionary Anthropology: Issues, News, and Reviews* 19.6, pp. 236–249. DOI: 10.1002/evan.20283.

Kovaka, Sam, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea (2019). "Transcriptome assembly from long-read RNA-seq alignments with StringTie2". *Genome Biology* 20.1, p. 278. DOI: 10.1186/s13059-019-1910-1.

Kramerov, D A and N S Vassetzky (2011). "Origin and evolution of SINEs in eukaryotic genomes". *Heredity* 107.6, pp. 487–495. DOI: 10.1038/hdy.2011.43.

Kronenberg, Zev N., Arang Rhie, Sergey Koren, Gregory T. Concepcion, Paul Peluso, Katherine M. Munson, David Porubsky, Kristen Kuhn, Kathryn A. Mueller, Wai Yee Low, Stefan Hiendleder, Olivier Fedrigo, Ivan Liachko, Richard J. Hall, Adam M. Phillippy, Evan E. Eichler, John L. Williams, Timothy P. L. Smith, Erich D. Jarvis, Shawn T. Sullivan, and Sarah B. Kingan (2021). "Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C". *Nature Communications* 12.1, p. 1935. DOI: 10.1038/s41467-020-20536-y.

Krzywinski, Martin, Jacqueline Schein, İnanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra (2009). "Circos: An information aesthetic for comparative genomics". *Genome Research* 19.9, pp. 1639–1645. DOI: 10.1101/gr.092759.109.

Lahn, Bruce T. and David C. Page (1999). "Four Evolutionary Strata on the Human X Chromosome". *Science* 286.5441, pp. 964–967. DOI: 10.1126/science.286.5441.964.

Lander, ES, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, et al. (2001). "many others. 2001. Initial sequencing and analysis of the human genome". *Nature* 409, pp. 860–921.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". *Genome Biology* 10.3, R25. DOI: 10.1186/gb-2009-10-3-r25.

Lechner, Marcus, Maribel Hernandez-Rosales, Daniel Doerr, Nicolas Wieseke, Annelyse Thévenin, Jens Stoye, Roland K. Hartmann, Sonja J. Prohaska, and Peter F. Stadler (2014). "Orthology Detection Combining Clustering and Synteny for Very Large Datasets". *PLoS ONE* 9.8, e105015. DOI: 10.1371/journal.pone.0105015.

Lewis, Paul O. (Nov. 2001). "A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data". *Systematic Biology* 50.6, pp. 913–925. DOI: 10.1080/106351501753462876.

Li, Heng (2011a). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". *Bioinformatics* 27.21, pp. 2987–2993. DOI: 10.1093/bioinformatics/btr509.

— (2011b). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". *Bioinformatics* 27.21, pp. 2987–2993. DOI: 10.1093/bioinformatics/btr509.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". *Bioinformatics* 25.14, pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li, Zhenyu, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, et al. (2012). "Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph". *Briefings in functional genomics* 11.1, pp. 25–37.

Lien, Sigbjørn, Ben F Koop, Simen R Sandve, Jason R Miller, Matthew P Kent, Torfinn Nome, Torgeir R Hvidsten, Jong S Leong, David R Minkley, Aleksey Zimin, Fabian Grammes, Harald Grove, Arne Gjuvsland, Brian Walenz, Russell A Hermansen, Kris von Schalburg, Eric B Rondeau, Alex Di Genova, Jeevan K A Samy, Jon Olav Vik, Magnus D Vigeland, Lis Caler, Unni Grimholt, Sissel Jentoft, Dag Inge Våge, Pieter de Jong, Thomas Moen, Matthew Baranski, Yniv Palti, Douglas R Smith, James A Yorke, Alexander J Nederbragt, Ave Tooming-Klunderud, Kjetill S Jakobsen, Xuanting Jiang, Dingding Fan, Yan Hu, David A Liberles, Rodrigo Vidal, Patricia Iturra, Steven J M Jones, Inge Jonassen, Alejandro Maass, Stig W Omholt, and William S Davidson (2016). "The Atlantic salmon genome provides insights into rediploidization". *Nature* 533.7602, pp. 200–205. DOI: 10.1038/nature17164.

Lin, Yu, Vaibhav Rajan, Krister M Swenson, and Bernard ME Moret (2010). "Estimating true evolutionary distances under rearrangements, duplications, and losses". *BMC Bioinformatics* 11.Suppl 1, S54. DOI: 10.1186/1471-2105-11-s1-s54.

Liu, Binghang, Yujian Shi, Jianying Yuan, Xuesong Hu, Hao Zhang, Nan Li, Zhenyu Li, Yanxiang Chen, Desheng Mu, and Wei Fan (2013). "Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects". *arXiv preprint arXiv:1308.2012*.

— (2020). *Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects*.

Liu, Dang, Martin Hunt, and Isheng J Tsai (2018). "Inferring synteny between genome assemblies: a systematic evaluation". *BMC Bioinformatics* 19.1. Sun, p. 26. DOI: 10.1186/s12859-018-2026-4.

Lu, Hengyun, Francesca Giordano, and Zemin Ning (2016). "Oxford Nanopore MinION Sequencing and Genome Assembly". *Genomics, Proteomics & Bioinformatics* 14.5, pp. 265–279. DOI: 10.1016/j.gpb.2016.05.004.

Luo, Haiwei, Jian Shi, William Arndt, Jijun Tang, and Robert Friedman (2008). "Gene Order Phylogeny of the Genus Prochlorococcus". *PLoS ONE* 3.12, e3837. DOI: 10.1371/journal.pone.0003837.

Luo, Zhe-Xi (2007). "Transformation and diversification in early mammal evolution". *Nature* 450.7172, pp. 1011–1019. DOI: 10.1038/nature06277.

Macqueen, Daniel J. and Ian A. Johnston (2014). "A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification". *Proceedings of the Royal Society B: Biological Sciences* 281.1778, p. 20132881. DOI: 10.1098/rspb.2013.2881.

Mahadik, Kanak, Christopher Wright, Milind Kulkarni, Saurabh Bagchi, and Somali Chaterji (2019). "Scalable Genome Assembly through Parallel de Bruijn Graph Construction for Multiple k-mers". *Scientific Reports* 9.1, p. 14882. DOI: 10.1038/s41598-019-51284-9.

Maia, Guilherme Augusto, Vilmar Benetti Filho, Eric Kazuo Kawagoe, Tatiany Aparecida Teixeira Soratto, Renato Simões Moreira, Edmundo Carlos Grisard, and Glauber Wagner (2022). "AnnotaPipeline: An integrated tool to annotate eukaryotic proteins using multi-omics data". *Frontiers in Genetics* 13, p. 1020100. DOI: 10.3389/fgene.2022.1020100.

Málaga-Trillo, Edward and Axel Meyer (2001). "Genome duplications and accelerated evolution of Hox genes and cluster architecture in teleost fishes". *American Zoologist* 41.3, pp. 676–686.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, and Evgeny M. Zdobnov (2021). "BUSCO: Assessing Genomic Data Quality and Beyond". *Current Protocols* 1.12, e323. DOI: https://doi.org/10.1002/cpz1.323.

Mardis, Elaine R. (2008). "Next-Generation DNA Sequencing Methods". *Annual Review of Genomics and Human Genetics* 9.1, pp. 387–402. DOI: 10.1146/annurev.genom.9.081307.164359.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors". *Nature* 437.7057, pp. 376–380. DOI: 10.1038/nature03959.

Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall (2016). "WhatsHap: fast and accurate read-based phasing". *BioRxiv*, p. 085050.

McCoy, Rajiv C., Ryan W. Taylor, Timothy A. Blauwkamp, Joanna L. Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A. Petrov, and Anna-Sophie Fiston-Lavier (2014). "Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements". *PLoS ONE* 9.9, e106689. DOI: 10.1371/journal.pone.0106689.

Miller, Jason R, Sergey Koren, and Granger Sutton (2010). "Assembly algorithms for next-generation sequencing data". *Genomics* 95.6, pp. 315–327.

Minh, Bui Quang, Minh Anh Thi Nguyen, and Arndt von Haeseler (2013). "Ultrafast Approximation for Phylogenetic Bootstrap". *Molecular Biology and Evolution* 30.5, pp. 1188–1195. DOI: 10.1093/molbev/mst024.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear (Feb. 2020). "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era". *Molecular Biology and Evolution* 37.5, pp. 1530–1534. DOI: 10.1093/molbev/msaa015.

Mkrtchyan, Hasmik, Madeleine Gross, Sophie Hinreiner, Anna Polytiko, Marina Manvelyan, Kristin Mrasek, Nadezda Kosyakova, Elisabeth Ewers, Heike Nelle, Thomas Liehr, et al. (2010). "The human genome puzzle—the role of copy number variation in somatic mosaicism". *Current genomics* 11.6, pp. 426–431.

Mullikin, James C and Zemin Ning (2003). "The phusion assembler". *Genome research* 13.1, pp. 81–90.

Myers, Eugene W, Granger G Sutton, Art L Delcher, Ian M Dew, Dan P Fasulo, Michael J Flanigan, Saul A Kravitz, Clark M Mobarry, Knut HJ Reinert, Karin A Remington, et al. (2000). "A whole-genome assembly of Drosophila". *Science* 287.5461, pp. 2196–2204.

Nabhan, Ahmed Ragab and Indra Neil Sarkar (2012). "The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy". *Briefings in Bioinformatics* 13.1, pp. 122–134. DOI: 10.1093/bib/bbr014.

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". *Nature Communications* 12.1, p. 4489. DOI: 10.1038/s41467-021-24573-z.

Natsidis, Paschalis, Paschalia Kapli, Philipp H. Schiffer, and Maximilian J. Telford (2021). "Systematic errors in orthology inference and their effects on evolutionary analyses". *iScience* 24.2, p. 102110. DOI: 10.1016/j.isci.2021.102110.

Needleman, Saul B. and Christian D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48.3, pp. 443–453. DOI: https://doi.org/10.1016/0022-2836(70)90057-4.

Niknafs, Yashar S, Balaji Pandian, Hariharan K Iyer, Arul M Chinnaiyan, and Matthew K Iyer (2017). "TACO produces robust multisample transcriptome assemblies from RNA-seq". *Nature Methods* 14.1, pp. 68–70. DOI: 10.1038/nmeth.4078.

Nong, Wenyan, Zhe Qu, Yiqian Li, Tom Barton-Owen, Annette Y. P. Wong, Ho Yin Yip, Hoi Ting Lee, Satya Narayana, Tobias Baril, Thomas Swale, Jianquan Cao, Ting Fung Chan, Hoi Shan Kwan, Sai Ming Ngai, Gianni Panagiotou, Pei-Yuan Qian, Jian-Wen Qiu, Kevin Y. Yip, Noraznawati Ismail, Siddhartha Pati, Akbar John, Stephen S. Tobe, William G. Bendena, Siu Gin Cheung, Alexander Hayward, and Jerome H. L. Hui (2021). "Horseshoe crab genomes reveal the evolution of genes and microRNAs after three rounds of whole genome duplication". *Communications Biology* 4.1, p. 83. DOI: 10.1038/s42003-020-01637-2.

Northcutt, R Glenn and Carl Gans (1983). "The genesis of neural crest and epidermal placodes: a reinterpretation of vertebrate origins". *The Quarterly review of biology* 58.1, pp. 1–28.

Nurk, Sergey, Brian P Walenz, Arang Rhie, Mitchell R Vollger, Glennis A Logsdon, Robert Grothe, Karen H Miga, Evan E Eichler, Adam M Phillippy, and

Sergey Koren (2020). “HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads”. *Genome research* 30.9, pp. 1291–1305.

Ohno, Susumu (1970). *Evolution by gene duplication*. Springer Science & Business Media.

Omland, Kevin E. (1997). “Correlated Rates of Molecular and Morphological Evolution”. *Evolution* 51.5, pp. 1381–1393. (Visited on 09/19/2023).

Otto, Sarah P. (2007). “The Evolutionary Consequences of Polyploidy”. *Cell* 131.3, pp. 452–462. DOI: 10.1016/j.cell.2007.10.022.

Parey, Elise, Alexandra Louis, Jerome Montfort, Olivier Bouchez, Céline Roques, Carole Iampietro, Jerome Lluch, Adrien Castinel, Cécile Donnadieu, Thomas Desvignes, Christabel Floi Bucao, Elodie Jouanno, Ming Wen, Sahar Mejri, Ron Dirks, Hans Jansen, Christiaan Henkel, Wei-Jen Chen, Margot Zahm, Cédric Cabau, Christophe Klopp, Andrew W. Thompson, Marc Robinson-Rechavi, Ingo Braasch, Guillaume Lecointre, Julien Bobe, John H. Postlethwait, Camille Berthelot, Hugues Roest Crollius, and Yann Guiguen (2023). “Genome structures resolve the early diversification of teleost fishes”. *Science* 379.6632, pp. 572–575. DOI: 10.1126/science.abq4257.

Parey, Elise, Alexandra Louis, Jérôme Montfort, Yann Guiguen, Hugues Roest Crollius, and Camille Berthelot (2022). “An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication”. *Genome Research* 32.9, pp. 1685–1697. DOI: 10.1101/gr.276953.122.

Pedersen, Brent S and Aaron R Quinlan (2018). “Mosdepth: quick coverage calculation for genomes and exomes”. *Bioinformatics* 34.5, pp. 867–868. DOI: 10.1093/bioinformatics/btx699.

Peer, Yves Van de, Eshchar Mizrachi, and Kathleen Marchal (2017). "The evolutionary significance of polyploidy". *Nature Reviews Genetics* 18.7, pp. 411–424. DOI: 10.1038/nrg.2017.26.

Philippe, Hervé, Albert J. Poustka, Marta Chiodin, Katharina J. Hoff, Christophe Dessimoz, Bartlomiej Tomiczek, Philipp H. Schiffer, Steven Müller, Daryl Domman, Matthias Horn, Heiner Kuhl, Bernd Timmermann, Noriyuki Satoh, Tomoe Hikosaka-Katayama, Hiroaki Nakano, Matthew L. Rowe, Maurice R. Elphick, Morgane Thomas-Chollier, Thomas Hankeln, Florian Mertes, Andreas Wallberg, Jonathan P. Rast, Richard R. Copley, Pedro Martinez, and Maximilian J. Telford (2019). "Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria". *Current Biology* 29.11, 1818–1826.e6. DOI: 10.1016/j.cub.2019.04.009.

Pisani, Davide, Walker Pett, Martin Dohrmann, Roberto Feuda, Omar Rota-Stabelli, Hervé Philippe, Nicolas Lartillot, and Gert Wörheide (2015). "Genomic data do not support comb jellies as the sister group to all other animals". *Proceedings of the National Academy of Sciences* 112.50, pp. 15402–15407. DOI: 10.1073/pnas.1518127112.

Plazzi, Federico, Ronald R Ferrucci, and Marco Passamonti (2010). "Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies". *BMC Bioinformatics* 11.1, p. 209. DOI: 10.1186/1471-2105-11-209.

Posada, David and Keith A Crandall (2001). "Selecting the best-fit model of nucleotide substitution". *Systematic biology* 50.4, pp. 580–601.

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". *Nucleic Acids Research* 40.2, e11–e11. DOI: 10.1093/nar/gkr955.

Quail, Michael A, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". *BMC Genomics* 13.1, p. 341. DOI: 10.1186/1471-2164-13-341.

Ramírez-Barahona, Santiago, Hervé Sauquet, and Susana Magallón (2020). "The delayed and geographically heterogeneous diversification of flowering plant families". *Nature Ecology & Evolution* 4.9, pp. 1232–1238. DOI: 10.1038/s41559-020-1241-3.

Ramsey, Justin and Douglas W Schemske (1998). "Pathways, mechanisms, and rates of polyploid formation in flowering plants". *Annual review of ecology and systematics* 29.1, pp. 467–501.

Rannala, Bruce and Ziheng Yang (2007). "Inferring speciation times under an episodic molecular clock". *Systematic biology* 56.3, pp. 453–466.

Redmond, Anthony K., Dearbhaile Casey, Manu Kumar Gundappa, Daniel J. Macqueen, and Aoife McLysaght (2023). "Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor". *Nature Communications* 14.1, p. 2879. DOI: 10.1038/s41467-023-38714-z.

Reis, Mario dos, Yuttapong Thawornwattana, Konstantinos Angelis, MaximilianJ. Telford, Philip C.J. Donoghue, and Ziheng Yang (2015). "Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales". *Current Biology* 25.22, pp. 2939–2950. DOI: 10.1016/j.cub.2015.09.066.

Renne, Paul R, Courtney J Sprain, Mark A Richards, Stephen Self, Loÿc Vanderkluysen, and Kanchan Pande (2015). "State shift in Deccan volcanism at the Cretaceous-Paleogene boundary, possibly induced by impact". *Science* 350.6256, pp. 76–78.

Rhee, Je-Keun, Honglan Li, Je-Gun Joung, Kyu-Baek Hwang, Byoung-Tak Zhang, and Soo-Yong Shin (2016). "Survey of computational haplotype determination methods for single individual". *Genes & Genomics* 38, pp. 1–12.

Rice, Alan M and Aoife McLysaght (2017). "Dosage sensitivity is a major determinant of human copy number variant pathogenicity". *Nature communications* 8.1, p. 14366.

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". *Genome Biology* 18.1, p. 111. DOI: 10.1186/s13059-017-1241-z.

Rokas, Antonis and Peter W.H. Holland (2000). "Rare genomic changes as a tool for phylogenetics". *Trends in Ecology Evolution* 15.11, pp. 454–459. DOI: https://doi.org/10.1016/S0169-5347(00)01967-4.

Romanenko, Svetlana A., Larisa S. Biltueva, Natalya A. Serdyukova, Anastasia I. Kulemzina, Violetta R. Beklemisheva, Olga L. Gladkikh, Natalia A. Lemskaya, Elena A. Interesova, Marina A. Korentovich, Nadezhda V. Vorobieva, Alexander S. Graphodatsky, and Vladimir A. Trifonov (2015). "Segmental paleotetraploidy revealed in sterlet (Acipenser ruthenus) genome by chromosome painting". *Molecular Cytogenetics* 8.1, p. 90. DOI: 10.1186/s13039-015-0194-8.

Ronaghi, Mostafa, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén (1996). "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release". *Analytical Biochemistry* 242.1, pp. 84–89. DOI: https://doi.org/10.1006/abio.1996.0432.

Ronquist, Fredrik and John P Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". *Bioinformatics* 19.12, pp. 1572–1574.

Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, Jeremy Hoon, Jan F. Simons, David Marran, Jason W. Myers, John F. Davidson, Annika Branting, John R. Nobile, Bernard P. Puc, David Light, Travis A. Clark, Martin Huber, Jeffrey T. Branciforte, Isaac B. Stoner, Simon E. Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline A. Fidanza, Eugeni Namsaraev, Kevin J. McKernan, Alan Williams, G. Thomas Roth, and James Bustillo (2011). "An integrated semiconductor device enabling non-optical genome sequencing". *Nature* 475.7356, pp. 348–352. DOI: 10.1038/nature10242.

Salzberg, Steven L. and James A. Yorke (2005). "Beware of mis-assembled genomes". *Bioinformatics* 21.24, pp. 4320–4321. DOI: 10.1093/bioinformatics/bti769.

Sanger, F. and A.R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". *Journal of Molecular Biology* 94.3, pp. 441–448. DOI: https://doi.org/10.1016/0022-2836(75)90213-2.

Santini, Francesco, Luke J Harmon, Giorgio Carnevale, and Michael E Alfaro (2009). "Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes". *BMC Evolutionary Biology* 9.1, p. 194. DOI: 10.1186/1471-2148-9-194.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi (Oct. 2019). "GenBank". *Nucleic Acids Research* 48.D1, pp. D84–D86. DOI: 10.1093/nar/gkz956.

Schadt, Eric E., Steve Turner, and Andrew Kasarskis (2010). "A window into third-generation sequencing". *Human Molecular Genetics* 19.R2, R227–R240. DOI: 10.1093/hmg/ddq416.

Schatz, Michael C., Arthur L. Delcher, and Steven L. Salzberg (2010). "Assembly of large genomes using second-generation sequencing". *Genome Research* 20.9, pp. 1165–1173. DOI: 10.1101/gr.101360.109.

Schwartz, David C., Xiaojun Li, Luis I. Hernandez, Satyadarshan P. Ramnarain, Edward J. Huff, and Yu-Ker Wang (1993). "Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping". *Science* 262.5130, pp. 110–114. DOI: 10.1126/science.8211116.

Shendure, Jay, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church (2005). "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome". *Science* 309.5741, pp. 1728–1732. DOI: 10.1126/science.1117389.

Sidow, Arend (1996). "Gen (om) e duplications in the evolution of early vertebrates". *Current opinion in genetics & development* 6.6, pp. 715–722.

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". *Nature Ecology & Evolution*, pp. 1–11. DOI: 10.1038/s41559-020-1156-z.

Simpson, Jared T, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven JM Jones, and Inanç Birol (2009). "ABySS: a parallel assembler for short read sequence data". *Genome research* 19.6, pp. 1117–1123.

Slater, Guy St C and Ewan Birney (2005). "Automated generation of heuristics for biological sequence comparison". *BMC Bioinformatics* 6.1, p. 31. DOI: 10.1186/1471-2105-6-31.

Slotkin, R. Keith and Robert Martienssen (2007). "Transposable elements and the epigenetic regulation of the genome". *Nature Reviews Genetics* 8.4, pp. 272–285. DOI: 10.1038/nrg2072.

Smith, T.F. and M.S. Waterman (1981). "Identification of common molecular subsequences". *Journal of Molecular Biology* 147.1, pp. 195–197. DOI: 10.1016/0022-2836(81)90087-5.

Soltis, Douglas E., Victor A. Albert, Jim Leebens-Mack, Charles D. Bell, Andrew H. Paterson, Chunfang Zheng, David Sankoff, Claude W. dePamphilis, P. Kerr Wall, and Pamela S. Soltis (2009). "Polyploidy and angiosperm diversification". *American Journal of Botany* 96.1, pp. 336–348. (Visited on 10/10/2023).

Stanke, Mario and Burkhard Morgenstern (2005). "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints". *Nucleic Acids Research* 33.suppl_2, W465–W467. DOI: 10.1093/nar/gki458.

Symonová, Radka, Miloš Havelka, Chris T. Amemiya, William Mike Howell, Tereza Kořínková, Martin Flajšhans, David Gela, and Petr Ráb (2017). "Molecular cytogenetic differentiation of paralogs of Hox paralogs in duplicated and rediploidized genome of the North American paddlefish (Polyodon spathula)". *BMC Genetics* 18.1, p. 19. DOI: 10.1186/s12863-017-0484-8.

Tamura, Koichiro and Masatoshi Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." *Molecular biology and evolution* 10.3, pp. 512–526.

Tang, Haibao, Vivek Krishnakumar, and Jingping Li (Oct. 2015). *jcvi: JCVI utility libraries*. Version v0.5.7. DOI: 10.5281/zenodo.31631.

Tang, Jijun and Bernard M.E. Moret (July 2003). "Scaling up accurate phylogenetic reconstruction from gene-order data". *Bioinformatics* 19.suppl$_1$, pp. i305–i312. DOI: 10.1093/bioinformatics/btg1042.

Thrash, Adam, Federico Hoffmann, and Andy Perkins (2020). "Toward a more holistic method of genome assembly assessment". *BMC bioinformatics* 21.4, pp. 1–8.

Tørresen, Ole K, Bastiaan Star, Pablo Mier, Miguel A Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, Marcin Grynberg, Andrey V Kajava, Vasilis J Promponas, Maria Anisimova, Kjetill S Jakobsen, and Dirk Linke (2019). "Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases". *Nucleic Acids Research* 47.21, pp. 10994–11006. DOI: 10.1093/nar/gkz841.

Tóth, Gábor, Zoltán Gáspári, and Jerzy Jurka (2000). "Microsatellites in Different Eukaryotic Genomes: Survey and Analysis". *Genome Research* 10.7, pp. 967–981. DOI: 10.1101/gr.10.7.967.

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". *Genome research* 24.8, pp. 1334–1347.

Vanneste, Kevin, Steven Maere, and Yves Van de Peer (2014). "Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution". *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1648, p. 20130353.

Veltri, Daniel, Martha Malapi Wight, and Jo Anne Crouch (2016). "SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species". *Nucleic Acids Research* 44.W1, W41–W45. DOI: 10.1093/nar/gkw330.

Venkatesh, Byrappa, Alison P. Lee, Vydianathan Ravi, Ashish K. Maurya, Michelle M. Lian, Jeremy B. Swann, Yuko Ohta, Martin F. Flajnik, Yoichi Sutoh, Masanori Kasahara, Shawn Hoon, Vamshidhar Gangu, Scott W. Roy, Manuel Irimia, Vladimir Korzh, Igor Kondrychyn, Zhi Wei Lim, Boon-Hui Tay, Sumanty Tohari, Kiat Whye Kong, Shufen Ho, Belen Lorente-Galdos, Javier Quilez, Tomas Marques-Bonet, Brian J. Raney, Philip W. Ingham, Alice Tay, LaDeana W. Hillier, Patrick Minx, Thomas Boehm, Richard K. Wilson, Sydney Brenner, and Wesley C. Warren (2014). "Elephant shark genome provides unique insights into gnathostome evolution". *Nature* 505.7482, pp. 174–179. DOI: 10.1038/nature12826.

Wagner, Gunte P., Chris Amemiya, and Frank Ruddle (2003). "Hox cluster duplications and the opportunity for evolutionary novelties". *Proceedings of the National Academy of Sciences* 100.25, pp. 14603–14606. DOI: 10.1073/pnas.2536656100.

Walden, Nora and Michael Eric Schranz (Feb. 2023). "Synteny Identifies Reliable Orthologs for Phylogenomics and Comparative Genomics of the Brassicaceae". *Genome Biology and Evolution* 15.3, evad034. DOI: 10.1093/gbe/evad034.

Wang, Peipei, Fanrui Meng, Bethany M. Moore, and Shin-Han Shiu (2020). "Read coverage as an indicator of misassembly in a short-read based genome assembly". *bioRxiv*, p. 790337. DOI: 10.1101/790337.

Wang, Yupeng, Haibao Tang, Jeremy D. DeBarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-ho Lee, Huizhe Jin, Barry Marler, Hui Guo, Jessica C. Kissinger, and Andrew H. Paterson (2012). "MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity". *Nucleic Acids Research* 40.7, e49–e49. DOI: 10.1093/nar/gkr1293.

Wendel, Jonathan F (2000). "Genome evolution in polyploids". *Plant molecular evolution*, pp. 225–249.

Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller (2019). "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome". *Nature Biotechnology* 37.10, pp. 1155–1162. DOI: 10.1038/s41587-019-0217-9.

Wille, Martin, Thomas F Nägler, Bernd Lehmann, Stefan Schröder, and Jan D Kramers (2008). "Hydrogen sulphide release to surface waters at the Precambrian/Cambrian boundary". *Nature* 453.7196, pp. 767–769.

Wolfe, Kenneth H and Denis C Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome". *Nature* 387.6634, pp. 708–713. DOI: 10.1038/42711.

Wolfe, Kenneth H. (2001). "Yesterday's polyploids and the mystery of diploidization". *Nature Reviews Genetics* 2.5, pp. 333–341. DOI: 10.1038/35072009.

Zachos, James C., Gerald R. Dickens, and Richard E. Zeebe (2008). "An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics". *Nature* 451.7176, pp. 279–283. DOI: 10.1038/nature06588.

Zerbino, Daniel R and Ewan Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". *Genome research* 18.5, pp. 821–829.

Zhang, Feng, Wenli Gu, Matthew E Hurles, and James R Lupski (2009). "Copy number variation in human health, disease, and evolution". *Annual review of genomics and human genetics* 10, pp. 451–481.

Zhang, Xingtan, Ruoxi Wu, Yibin Wang, Jiaxin Yu, and Haibao Tang (2019). "Unzipping haplotypes in diploid and polyploid genomes". *Computational and*

Structural Biotechnology Journal 18, pp. 66–72. DOI: 10.1016/j.csbj.2019.11.011.

Zhao, Tao, Rens Holmer, Suzanne de Bruijn, Gerco C. Angenent, Harrold A. van den Burg, and M. Eric Schranz (2017). "Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation". *The Plant Cell* 29.6, pp. 1278–1292. DOI: 10.1105/tpc.17.00312.

Zhao, Tao and M Eric Schranz (2017). "Network approaches for plant phylogenomic synteny analysis". *Current Opinion in Plant Biology* 36, pp. 129–134. DOI: 10.1016/j.pbi.2017.03.001.

— (2019). "Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes". *Proceedings of the National Academy of Sciences* 116.6, p. 201801757. DOI: 10.1073/pnas.1801757116.

Zhao, Tao, Arthur Zwaenepoel, Jia-Yu Xue, Shu-Min Kao, Zhen Li, M. Eric Schranz, and Yves Van de Peer (2021). "Whole-genome microsynteny-based phylogeny of angiosperms". *Nature Communications* 12.1, p. 3498. DOI: 10.1038/s41467-021-23665-0.

Zharkikh, Andrey (1994). "Estimation of evolutionary distances between nucleotide sequences". *Journal of molecular evolution* 39, pp. 315–329.

Zimin, Aleksey V and Steven L Salzberg (2022). "The SAMBA tool uses long reads to improve the contiguity of genome assemblies". *PLoS Computational Biology* 18.2, e1009860. DOI: 10.1371/journal.pcbi.1009860.

Zimin, Aleksey V., Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L. Salzberg, and James A. Yorke (Aug. 2013). "The MaSuRCA genome assembler". *Bioinformatics* 29.21, pp. 2669–2677. DOI: 10.1093/bioinformatics/btt476.

Zverinova, Stepanka and Victor Guryev (2022). "Variant calling: Considerations, practices, and developments". *Human Mutation* 43.8, pp. 976–985. DOI: 10.1002/humu.24311.