COMPARATIVE AND POPULATION GENOMICS OF SECONDARILY TEMPERATE *PARANOTOTHENIA ANGUSTATA*, NEW ZEALAND BLACK COD

BY

NIRAJ RAYAMAJHI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biology
with a concentration in Ecology, Ethology, and Evolution
in the Graduate College of the
University of Illinois Urbana-Champaign, 2024

Urbana, Illinois

Doctoral Committee:

Associate Professor Julian M. Catchen, Chair and Director of Research
Professor Chi-Hing Christina Cheng
Professor Rebecca Fuller
Professor Andrew Suarez

ABSTRACT

Notothenioids are a group of teleost fish that have undergone at least two thermal transitions in their evolutionary history. Due to paleo-geological, -climatic, and -oceanographic changes, the environment of Antarctica transitioned from temperate to cold. Antifreeze glycoproteins became a key evolutionary innovation that enabled a group of temperate, bottom-dwelling notothenioids to adapt to increasingly cold waters. With the availability of vacant ecological niches, the cold-resistant notothenioids diversified over evolutionary time. Most of these derived lineages became cold-specialized (e.g., *Trematomus borchgrevinki*). Remarkably, a few of them readapted to a warmer environment, becoming secondarily temperate (e.g., *Paranotothenia angustata*); however, the genetic architecture of readaptation for these organisms remains largely unknown.

In this dissertation, my first goal was to identify the optimal *de novo* genome assembly strategy for notothenioids, as robust assembly is required for genome-based projects (Chapter 2). I evaluated Illumina-, Nanopore-, and PacBio-based genome assembly strategies with *T. borchgrevinki*. My results suggest that the strategy based on long-reads only is the current best approach and can be optimized through a subsampling method. My results indicate that short-reads only and hybrid (short- and long-reads) based strategies produce low quality assemblies. My second goal was to identify genomic features associated with secondarily temperate adaptations of *P. angustata* (Chapter 3). My results suggest that I have produced high quality chromosome-level assemblies for *P. angustata* (a focal species) and *T. borchgrevinki* (an outgroup). They also indicate that the genome of *P. angustata* consists of lineage-specific DNA transposons, chromosomal fusion patterns, inversions (most of which co-localized with one to three protein-coding genes having signals of accelerated molecular evolution), and

translocations. This line of evidence calls for a detailed future investigation on the role of lineage-specific repeats and chromosomal rearrangements in non-polar adaptations of *P. angustata*. Based on results related to the *P. angustata*-specific signatures of positive selection, I propose that genes under selection, mainly associated with protein chaperoning, circadian rhythm, vision, erythrocyte differentiation and development, heme metabolism, mitochondria, and ribosomes, may have contributed to the adaptations of *P. angustata* in a temperate environment.

My third goal was to infer timing of origin of the *P. angustata*-specific adaptive loci (Chapter 4). I assessed genome-wide gene genealogical patterns from Restriction site-Associated DNA sequencing (RADseq)-based loci at homologous regions between *P. angustata* and *T. borchgrevinki*, as well as between McMurdo Station and Prydz Bay populations of *T. borchgrevinki*. Additionally, I estimated the time to the most recent common ancestor (T_{MRCA}) of alleles across RAD-loci within and between species and populations. I was unable to find distinct local signatures of positive selection because most of the gene trees had reciprocally monophyletic patterns (i.e., haplotypes from one species clustered to the exclusion of haplotypes from the other species, resulting in a monophyletic clade per species). However, some genealogical trees with reciprocally monophyletic patterns were also located within a) 92 candidates (from a group of 317 genes exhibiting accelerated molecular evolution in *P. angustata*) and b) structural variations (specific to *P. angustata*) which were presented in Chapter 3. Additionally, the average time to the most recent common ancestor (T_{MRCA}) of alleles between species appears to be lower than the time required for a genome-wide reciprocally monophyletic pattern to form under neutrality. These results are consistent with the idea that divergent selection contributed to the observed reciprocally monophyletic patterns.

Moreover, I did not find distinct local peaks of inter-species T_{MRCA}, suggesting that adaptations of *P. angustata* evolved after the divergence of the ancestral lineages of *P. angusta* and *T. borchgrevinki*. While one intra-species T_{MRCA} outlier was found within the *P. angustata*-specific inversion, none were within the candidate loci. Also, intra-species T_{MRCA} distributions within and outside of candidates (317 genes exhibiting accelerated molecular evolution) showed no significant difference, similar to those within and outside structural variations. These results further support a substantial contribution of *de novo* mutations in *P. angustata*'s temperate adaptations. Apart from these findings, I found incomplete lineage sorting between two populations of *T. borchgrevinki* (one from McMurdo Station and another from Prydz Bay). This result indicates high gene flow and no geography-specific selection between the populations. I found intra-species T_{MRCA} outliers within two translocations specific to *T. borchgrevinki* (mentioned in Chapter 3). These results call for future investigation into the role of structural changes in the continuing cold adaptation of *T. borchgrevinki*. Overall, my results provide an overview of how and when the secondarily temperate adaptations of *P. angustata* may have evolved and provide genomic resources for future comparative and population genomic analyses in non-polar and polar notothenioids.

ACKNOWLEDGMENTS

First and foremost, I extend my hearfelt gratitude to my parents, Tanka Raya and Indira Budhathoki, for their unwavering support and sacrifices that facilitated my academic journey. I also thank my wife (Pragya Adhikari) and sister (Shreejana), for their steadfast encouragement during both the highs and lows. I am thankful to Pragya and my son (Aarin), who sacrificed precious family time so that I could focus on my dissertation work. Additionally, I am grateful to Dr. Julian Catchen for granting me the opportunity to work on this dissertation and providing support. I was also supported by a research assistantship from funds provided by the University of Illinois Research Board grant to Drs. Rebecca Fuller and Julian Catchen for assembling and analyzing killifish genomes. These experiences allowed me to learn valuable bioinformatic skills and deepen my understanding of evolutionary genomics. I would also like to acknowledge the contributions of the past and present members of Catchen lab, including Dr. Nicholas C. Rochette, Dr. Angel G. Rivera-Colón, Bushra Fazal Minhas, and Giovanni Madrigal, for their assistance in the molecular and bioinformatic work required for the analyses conducted in this research. In addition, I appreciate Drs. Chi-Hing Christina Cheng, Rebecca Fuller, and Andrew Suarez for their valuable time in the meetings and constructive feedback on my work. Additionally, I express my thanks to the National Science Foundation for funding as well as to the Department of Evolution, Ecology, and Behavior (EEB) and School of Integrative Biology (SIB) for summer stipends and teaching assistantship. Lastly, I express my gratitude to Ashmita Khanal, Dr. Rabin Bhattrai, Urbashi Karki, Deep Phuyal, Sweta Adhikari, Dr. Avinash Gupta, Sweta Shah, Saurav Raj Kunwar, Devendra Tamrakar, and Ichhya RajKarnikar. Their approachability, helpfulness, and selfless nature not only enriched my experience but also

created cherished memories for my family and me during our stay in Urbana, Illinois as well as in Ankeny, Iowa.

TABLE OF CONTENTS

CHAPTER 1: GENERAL INTRODUCTION

Understanding genetic adaptation is one of the major goals of evolutionary biology (Bomblies and Peichel 2022). Three hypotheses that can explain genetic adaptation in organisms are a) mutations in the coding sequence (Zhang *et al.* 2002), b) mutations in non-coding (regulatory) regions (Chan *et al.* 2010), and c) variation in genome structure through changes in copy number, orientation, and chromosomal location of the functional elements (Tigano *et al.* 2018; Christmas *et al.* 2019; Wellenreuther *et al.* 2019; Dorant *et al.* 2020). The coding and non-coding sequence mutations can result from point mutations (e.g., single-nucleotide substitution, insertion, and deletion) as well as chromosomal rearrangements, both unbalanced (deletion, insertion, duplication) and balanced (inversion, translocation, and fusion/fission), while structural variation occurs only through chromosomal rearrangements (Futuyma and Kirkpatrick 2017). Point mutations in coding sequences can drive adaptation by altering the amino acid translation of pre-existing genes. For instance, consider the Baltic herring, where the replacement of a single amino acid within the light-sensing rhodopsin protein, due to a missense mutation in the rhodopsin gene, has been proposed to play a significant role in its adaptation. Specifically, this adaptation enables the herring to capture a greater number of photons from the red-shifted light prevalent in the Baltic Sea environment (Hill *et al.* 2019). Further, deletions and insertions in non-coding regulatory regions can alter the rate, timing, and/or location of expression of genes that may lead to adaptive phenotypes. For example, the repeated independent deletion of a Pitx1 enhancer in geographically isolated threespine sticklebacks has been associated with adaptation (i.e., loss of pelvic limb through loss of gene expression) of sticklebacks to a freshwater environment (Chan *et al.* 2010).

Moreover, duplication events can lead to adaptive phenotypes. For example, gene duplication can result in the gain of paralogs. The duplicate copy of a gene can diverge and become adaptive through neo-functionalization, in which the ancestral gene copy retains the original function while a new gene copy develops a new function. For example, in an Antarctic zoarcid fish, the type III Anti-Freeze Protein gene arose through the neofunctionalization of a duplicated sialic acid synthase gene (Deng *et al.* 2010). Structural variants, such as inversions, may facilitate adaptation through the clustering of co-adapted genes to form supergenes by suppressing recombination. For example, in the ruff, an inversion block containing 125 genes on chromosome 11 has been associated with alternative reproductive strategies in male morphs (Küpper *et al.* 2016). Mechanisms that modify gene order and orientation, such as inversions, translocations, and chromosomal fissions or fusions, have the potential to alter gene expression, to form new gene combinations, and to break linkage blocks that cross an inversion boundary (Vakirlis *et al.* 2016). For example, in the ruff, one break point of the inversion on chromosome 11 disrupted the CENP-N gene, which is essential for mitotic centromere assembly (Küpper *et al.* 2016).

Transposable elements (TEs) can directly cause insertions, but they may also facilitate genomic insertions, deletions, duplications, inversions, and translocations. TEs could modify gene regulation by integrating themselves into regulatory elements or impact protein function by inserting directly into genes. (Chuong *et al.* 2017). For instance, in the case of white females among Colias butterflies, the insertion of a TE into the regulatory region of the existing BarH-1 gene has been linked to an ecologically significant alternative life history strategy. In contrast to colored females, white females prioritize resource allocation toward reproduction over wing pigmentation. (Woronik *et al.* 2019). In domesticated silkworms, individuals with a TE insertion into the cis-regulatory region of the *ecdysone oxidase* (*EO*) gene were found to have more stable developmental phenotypes during food shortage compared to those without the TE insertion (Sun *et al.* 2014).

Multiple genome evolution mechanisms can act together to generate adaptive phenotypes. For example, gadids use anti-freeze glycoproteins (AFGP) to adapt to cold Arctic waters. The AFGP gene in gadids was generated *de novo* from a non-coding DNA region. Tandem duplication, translocation, single nucleotide substitution, and a one-nucleotide deletion contributed to the formation of this new gene. The single nucleotide deletion provided the frameshift that linked the signal peptide (an amino acid sequence that labels a protein for transportation), propeptide, and AFGP coding regions into a single open reading frame, which functionalized the emergent AFGP gene (Zhuang *et al.* 2019). In Douc langur primates, gene duplication, and non-synonymous substitutions in the duplicated gene contributed to its adaptation to a leafy diet (Zhang *et al.* 2002). Further, in the ruff, a 4.5 Mbp inversion, combined with subsequent structural changes maintained through balancing selection, is linked to alternative reproductive strategies among males (Lamichhaney *et al.* 2016).

Single-nucleotide variation is the most studied genetic variation, while structural variations are comparatively less studied (Rubenstein *et al.* 2019). Evidence shows that it is possible to find associations of structural variation to the environment without finding any association of single nucleotide polymorphisms (SNPs) to the same environment. For example, in lobsters, 48 copy number variants (deletions, insertions, and duplications) were found to be associated with the temperature of marine waters within the southern Gulf of St. Lawrence. However, SNPs did not show a genotype-temperature association (Dorant *et al.* 2020). Compared to SNPs, structural variants affect more bases and are abundant across populations and species (Wellenreuther *et al.* 2019). Additionally, adaptive genetic variation may exist within the population for a certain duration before it becomes beneficial following an environmental change. For example, the freshwater allele of the *Ectodysplasin* (*eda*) gene plays a crucial role in threespine sticklebacks by enabling adaptation to freshwater habitats through reductions in armor plating. The allele responsible for this change is present at a low frequency in marine populations (Colosimo *et al.* 2005).

Similarly, the functional allele of the *teosinte branched1* (*tb1*) gene in maize has evolved through the insertion of TEs, which existed in the teosinte ancestor of maize (Studer *et al.* 2011). These genetic changes are responsible for increased apical dominance, a trait selected by plant breeders for domestication. However, adaptive loci can also arise as *de novo* mutations after an environmental shift. For example, recent independent mutations in Arabidopsis plants located in the Cape Verde Islands have led to a simultaneous reduction in flowering time and increased fitness within distinct populations on different islands. This adaptation followed a shift in climate toward a more arid environment (Fulgione *et al.* 2022). These findings illustrate the efforts of biologists to not only understand the various genomic changes that have facilitated adaptations in organisms but also to pinpoint the timing of these adaptive loci. This is crucial for obtaining a more comprehensive understanding of the genetic underpinnings of adaptation (Bomblies and Peichel 2022). Furthermore, it emphasizes that the questions of which genomic changes are genuinely adaptive and when these adaptations initially occurred in organisms remain open in evolutionary biology.

In this dissertation, my primary focus revolves around the notothenioid teleost fish to contribute to the understanding of the genomic architecture involved in the adaptations that enabled the transition of these species from polar to temperate environments. Notothenioid fish are particularly compelling subjects due to their unique evolutionary history. They primarily inhabit the consistently cold regions of Antarctica and rarely species have relocated to relatively warmer, non-Antarctic areas. These distinct thermal histories position them as valuable models for investigating the genomic basis of cold adaptation and the subsequent re-adaptation to temperate conditions.

Contemporary Antarctica stands as the coldest and driest isolated continent on Earth, surrounded by the Southern Ocean with its perpetually cold and oxygen-rich waters, maintaining a temperature of approximately -2 degrees Celsius. However, in ancient times, Antarctica was physically connected to other continents, including South America, Australia, and New Zealand, approximately 110-90 million years ago (MYA) (Eastman 1993), and it exhibited a temperate climate (Zachos *et al.* 2001). Fossil evidence suggests that around 92 to 83 MYA, Antarctica was characterized by temperate rainforests (Klages *et al.* 2020). Over time, continental drift and tectonic forces gradually isolated Antarctica from the rest of the continents (Storey and Granot 2021).

Additionally, Antarctica's segregation from South America and Australia are characterized by the formation of the Drake Passage (~40 MYA) (Scher and Martin 2006) and the Tasmanian Gateway (between 33.5 and 35.5 MYA) (Stickley *et al.* 2004), respectively. These geological changes enabled marine waters to circumscribe Antarctica, leading to the complete separation of Antarctica from other continents. This also led to the development of oceanic features of the Southern Ocean, such as the Antarctic Circumpolar Current (ACC) (Beers and Jayasundara 2015). The periods of glaciation due to reduced carbon dioxide and the establishment of ACC have played a crucial role in freezing the environment of Antarctica (Kennett 1977; Clarke *et al.* 2004). As the ACC developed, the northern boundary of the Southern

Ocean was divided into temperate and Antarctic water masses (Eastman 1993). The ACC established a thermal barrier for water masses on the current's northern and southern sides (Kennett 1977) effectively trapping cooler water on its southern side, resulting in a frigid Antarctic environment. With Antarctica's cooling and the subsequent expansion of ice sheets, most of the temperate fish fauna disappeared, presumably due to their inability to tolerate cold (Eastman and DeVries 1986) and the destruction of their habitat by ice (Eastman 2005).

Nonetheless, a lineage of ancestral notothenioid fish belonging to the order Perciform and sub-order Notothenioidei managed to survive and thrive in these extreme conditions, at least partially due to the evolution of Anti-Freeze Glycoproteins (AFGPs) (Chen *et al.* 1997) that prevent ice crystal growth within fish (DeVries 1971). Remarkably, over a relatively short span of evolutionary time (i.e., 10.7 million years), this Antarctic lineage of notothenioids underwent rapid speciation (Bista *et al.* 2023), giving rise to numerous species collectively referred to as Antarctic notothenioids. They exploited the available ecological niches within the Southern Ocean, demonstrating remarkable adaptability. Interestingly, despite the absence of swim bladders, multiple lineages of Antarctic notothenioids independently colonized various water column habitats, including pelagic, semi-pelagic, and cryopelagic zones, all without facing significant competition. These colonization events were made possible by the acquisition of adaptive traits, such as reduced ossification and scale mineralization, as well as the accumulation of substantial lipid deposits (Eastman 1993).

Today, the sub-order Notothenioidei comprises eight families with 140 species (Eastman and Eakin 2021). Three of the families, including Bovichtidae, Pseudaphritidae, and Eleginopidae (**Figure 1.1.A**), are basal and their members are found in coastal regions of South America, Australia, and New Zealand (Hardy *et al.* 1988; Eastman 1993; Last *et al.* 2002;

Ceballos *et al.* 2012; Eastman and Eakin 2021). The family Eleginopidae with one species (*Eleginops maclovinus*) is a sister clade to the Antarctic notothenioid clade formed by the remaining five families. These families include Nototheniidae, Harpagiferidae, Artedidraconidae, Bathydraconidae, and Channicthyidae (**Figure 1.1.A**; Near *et al.* 2004) which dominate the fish fauna of the Southern Ocean constituting 95% of fish fauna biomass (La Mesa *et al.* 2004). Antarctic notothenioids are also an example of adaptive radiation in vertebrates. Most of the species from the Antarctic notothenioid clade are endemic to Antarctica and are cold-specialized (e.g., *Trematomus borchgrevinki* (**Figures 1.1.B & 1.2**)). In other words, they cannot tolerate elevated temperature. For example, cold-specialized *T. borchgrevinki* succumb at ~ 6^{o}C above their normal ambient temperature (Somero and DeVries 1967). AFGPs – derived from a pre-existing trypsinogen-like protease gene (Chen *et al.* 1997) – are a key adaptation of cold-specialized notothenioids (DeVries 1988) and arose only once in their evolutionary history (reviewed in Eastman and Clarke 1998).

Among other phenotypic changes, cold-specialized notothenioids also lost the ubiquitous inducible heat-shock response (Hofmann *et al.* 2000). They cannot upregulate heat shock proteins (Hsps, molecular chaperones), which are responsible for maintaining cellular protein homeostasis in response to heat or other stress. Given that the Hsp gene is intact and these notothenioids can produce Hsps constitutively, the function-altering mutations may have occurred in related regulatory regions (Place *et al.* 2004). Furthermore, within the most derived family of the Antarctic notothenioid clade (Channichthyidae), there are species with extreme phenotypes, such as a complete lack of hemoglobin expression and erythrocytes (red blood cells). These species can survive because they reside in the oxygen-rich waters of the Southern Ocean and have compensatory physiological mechanisms, including enlarged hearts with thickened myocardium, increased total blood volume, and excessive branching of blood vessels (Beers and Jayasundara 2015).

Remarkably, the Antarctic notothenioid clade consists of a few lineages known as secondarily temperate notothenioids (e.g., *Paranotothenia angustata* of the Nototheniidae family (**Figures 1.1.B** & **1.3**)), which diverged from an Antarctic ancestral lineage (Eastman and McCune 2000) and re-adapted to warmer waters of temperate regions, including the coastal waters of New Zealand (Beers and Jayasundara 2015). These secondarily temperate species either lack expression of AFGP or express severely reduced amounts of AFGP molecules (Cheng 2003). The most parsimonious explanation is the loss or severe mutation of the AFGP gene family due to relaxed selection for freeze avoidance (Coppes Petricorena and Somero 2007). These species vary in the evolutionary timing of their escape from Antarctica. For example, *Champsocephalus esox* (secondarily temperate icefish) diverged from *Champsocephalus gunnari* (cold-specialized icefish) about 1.6 MYA (Stankovic *et al.* 2002), whereas *P. angustata* diverged from an Antarctic lineage about 11 MYA (Cheng 2003). However, the genetic basis of secondarily temperate adaptations and the timing of their origins in notothenioids are largely unknown. Only one study on the genomic architecture of re-adaptation of secondarily temperate notothenioids has been conducted. This study focused on more recently evolved secondarily temperate notothenioid, *C. esox* (Rivera-Colón *et al.* 2023). Here, I focused on a more distant secondarily temperate notothenioid, *P. angustata*, because the genomic architecture of re-adaptation among secondarily temperate notothenioids can differ.

This dissertation comprises three core research chapters. In Chapter 2, my objective was to determine the optimal *de novo* assembly strategy for notothenioids. To this end, I evaluated Illumina-, Nanopore-, and PacBio-based *de novo* genome assembly strategies with *T.*

borchgrevinki. In Chapter 3, my objective was to determine the potential genetic basis of secondarily temperate adaptations in *P. angustata*. Additionally, my specific objectives were to: 1) create chromosome-level assemblies for *P. angustata* (a focal species) and the closely related, *T. borchgrevinki* (an outgroup), 2) identify and characterize chromosomal rearrangements specific to *P. angustata* using conserved gene synteny, 3) infer regions and genes under positive selection in *P. angustata* using Restriction site-Associated DNA sequencing (RADseq) and single-copy orthologs, respectively. In Chapter 4, I examined genealogical trees within and between *P. angustata* and *T. borchgrevinki*, specifically aiming to infer timing of the origin of *P. angustata*-specific adaptive loci.

FIGURES

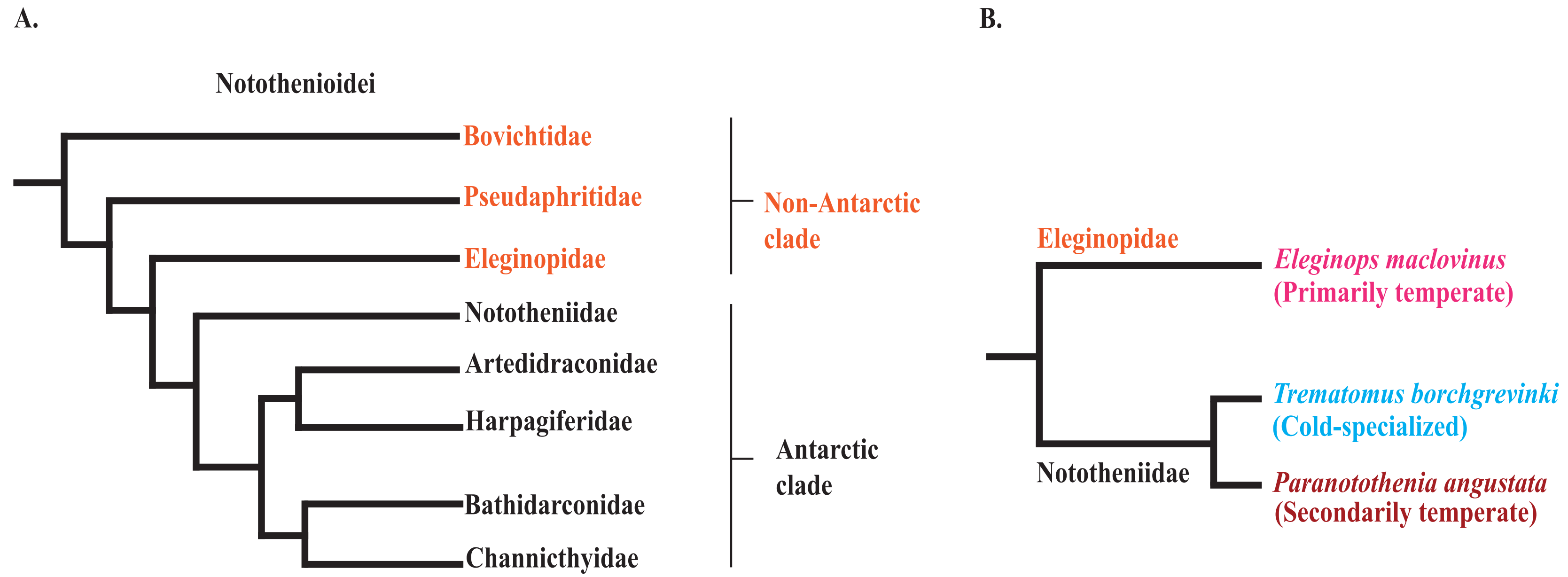


Figure 1.1 A) shows the phylogeny of eight families (three non-Antarctic and five Antarctic) of notothenioids within the order Perciform and sub-order notothenioidei. Non-Antarctic families are colored in orange, whereas Antarctic families are colored in black. **B)** shows a phylogenetic relationship among three notothenioids: *Eleginops maclovinus* (pink; primarily temperate), *Trematomus borchgrevinki* (blue; cold-specialized), and *Paranotothenia angustata* (red; secondarily temperate).

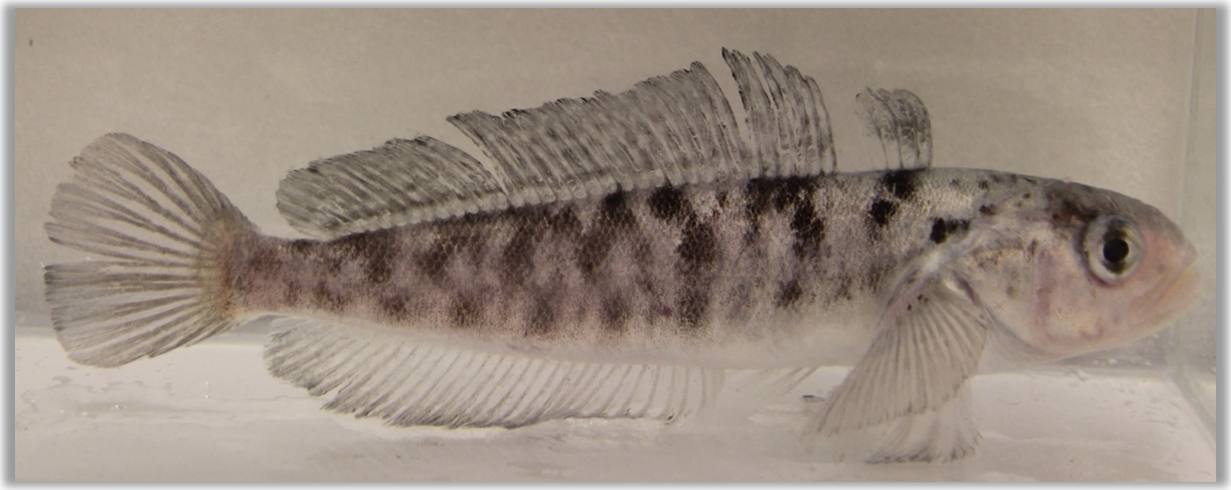


Figure 1.2 Image of one of the cold-specialized notothenioids, *Trematomus borchgrevinki* (commonly known as bald notothen). Image credit: Dr. Christina Cheng, Department of Evolution, Ecology, and Behavior, University of Illinois, Urbana-Champaign.

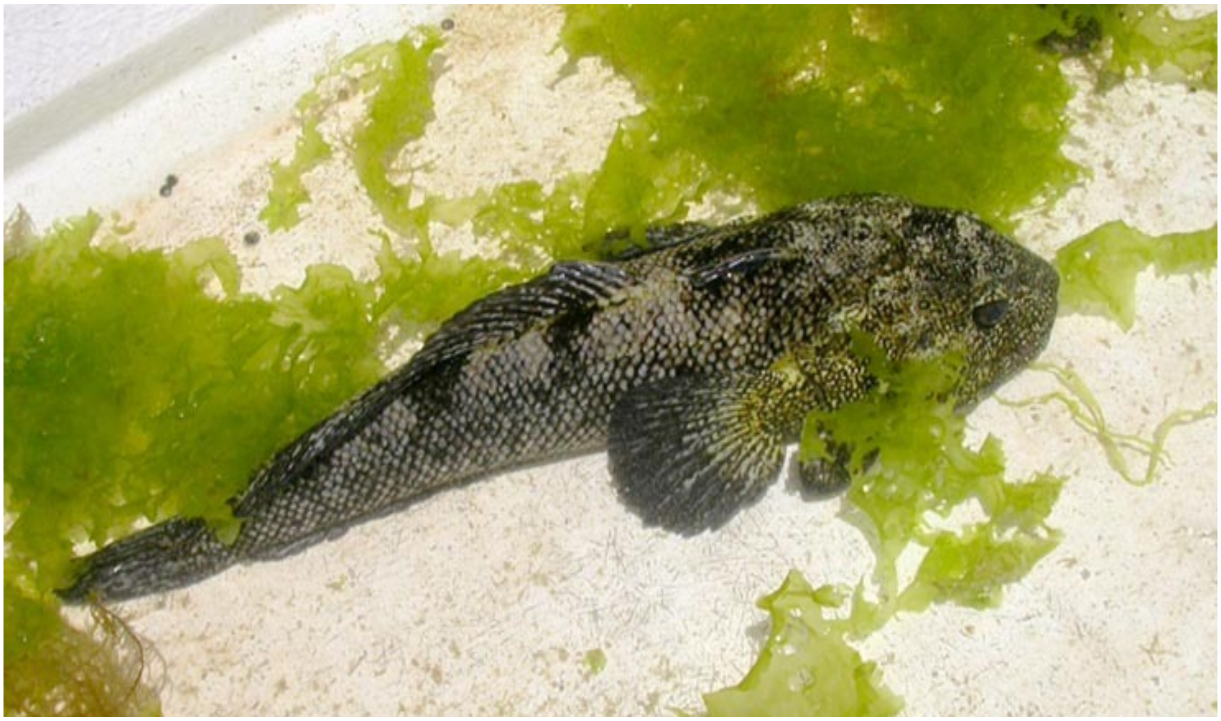


Figure 1.3 Image of one of the secondarily temperate notothenioids, *Paranotothenia angustata* (commonly known as New Zealand's black cod). Image credit: Dr. Christina Cheng, Department of Evolution, Ecology, and Behavior, University of Illinois, Urbana-Champaign.

CHAPTER 2: EVALUATING ILLUMINA-, NANOPORE-, PACBIO-BASED GENOME ASSEMBLY STRATEGIES WITH THE BALD NOTOTHEN, *TREMATOMOUS BORCHGREVINKI* [1]

ABSTRACT

For any genome-based research, a robust genome assembly is required. *De novo* assembly strategies have evolved with changes in DNA sequencing technologies and have been through at least three phases: i) short-read only, ii) short- and long-read hybrid, and iii) long-read only assemblies. Each of the phases has their own error model. We hypothesized that hidden short-read scaffolding errors and erroneous long-read contigs degrades the quality of short- and long-read hybrid assemblies. We assembled the genome of *T. borchgrevinki* from data generated during each of the three phases and assessed the quality problems we encountered. We developed strategies such as k-mer-assembled region replacement, parameter optimization, and long-read sampling to address the error models. We demonstrated that a k-mer based strategy improved short-read assemblies as measured by BUSCO while mate-pair libraries introduced hidden scaffolding errors and perturbed BUSCO scores. Further, we found that although hybrid assemblies can generate higher contiguity they tend to suffer from lower quality. In addition, we found long-read only assemblies can be optimized for contiguity by sub-sampling length-restricted raw reads. Our results indicate that long-read contig assembly is the current best choice and that assemblies from phase I and phase II were of lower quality.

[1] Chapter 2 has previously been published as Rayamajhi, N., C.-H. C. Cheng, and J. M. Catchen, 2022 Evaluating Illumina-, Nanopore-, and PacBio-based genome assembly strategies with the bald notothen, *Trematomus borchgrevinki*. G3 (Bethesda) 12: jkac192. It is reproduced here in adherence to copyright guidelines.

INTRODUCTION

The ultimate goal of genome sequencing is to connect the genome to phenotypes of interest. Genome sequencing can be used for the identification of rare variants associated with common human disease (Cirulli and Goldstein 2010), genes associated with agronomically important traits (Tao *et al.* 2019; Li *et al.* 2021), and structural variations potentially associated with adaptation to a novel environment (Kim *et al.* 2019). Sequencing technology has advanced enormously since its early implementation by the human genome project (HGP), launched in 1990 (Levy and Myers 2016). During the HGP, high-quality genome assemblies were generated by sequencing large insert size clones of human chromosomes using an automated Sanger sequencing approach, referred to as first-generation sequencing (Lander *et al.* 2001). However, while Sanger sequencing offered good read accuracy and approximately 1-kb read lengths, this method was expensive, laborious, and low throughput (Metzker 2005; Heather and Chain 2016).

With the advent of massively parallel, second-generation sequencing, the shortcomings of the Sanger strategy were bridged (Heather and Chain 2016), providing for the expansion and democratization of sequencing techniques (Rothberg and Leamon 2008) and a blooming of projects (Liao *et al.* 2019). However, second-generation sequencing reads were much shorter relative to Sanger sequencing (Schatz *et al.* 2010), which precluded resolving repeats longer than the insert size of the sequenced molecules (Alkan *et al.* 2011). Although certain molecular methods could extend the insert length (Berglund *et al.* 2011), they brought with them additional analysis challenges (Sahlin *et al.* 2016). And while the individual nucleotides of short reads have a very high fidelity, with an error rate of less than 1% (Bao and Lan 2017), the assemblies built with short-reads were highly fragmented, consisting of tens of thousands of scaffolds (Rhie *et al.* 2021).

In the recent decade, a third-generation of sequencing technology, long-read sequencing (LRS), including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) sequencing, are enabling researchers to generate high-quality, contig-level assemblies (Murigneux *et al.* 2020). LRS technologies can generate reads that are tens of kilobase pairs long. For example, continuous long reads (CLR) sequenced on a PacBio Sequel II machine can achieve a raw N50 length of 30–60 kb and an accuracy of 87–92%. The ONT MinIon/GridION sequencer can produce long and ultra-long reads with an N50 of 10–60 and 100–200 kb, respectively, with an accuracy of 87–98%. Using circular consensus sequencing, PacBio HiFi long-reads yield a reduced N50 of 10–20 kb, but with a significant improvement in accuracy (99%; Logsdon *et al.* 2020).

Furthermore, the long reads from PacBio and ONT can span repetitive regions (Rice and Green 2019), which second-generation short reads could not bridge, including most human genome repeats (Logsdon *et al.* 2020). Consequently, third-generation long reads have enabled genome assemblers to produce less-fragmented genome assemblies (Rice and Green 2019) with few or no gaps.

De novo genome assembly strategies have evolved along with changes in the underlying sequencing technologies resulting in 3 distinct phases: (Phase I) short-read-only, (Phase II) short- and long-read hybrid, and (Phase III) long-read-only assemblies. Phases I and II are now anachronistic strategies whereas the phase III assembly strategy is the current state-of-the-art. While phases I and II assemblies could not achieve chromosome-level results of high fidelity [at least, not without the aid of genomic resources such as very dense genetic maps (Fierst 2015)], phase III assemblies can yield full-length chromosomes in contig form, and scaffolding them—using chromosomal capture methods (Burton *et al.* 2013), optical maps (Leinonen and Salmela

2020), or genetic maps (Kim *et al.* 2019)—can reproduce a proper karyotype (Sedlazeck *et al.* 2018; Rice and Green 2019; Giani *et al.* 2020).

In phase I, short reads were generated primarily from Illumina sequencing platforms at large volume and low cost (with alternative technologies eventually outcompeted by Illumina). To generate contigs, short-read-only *de novo* genome assemblers used *de Bruijn* (Zerbino and Birney 2008; Compeau *et al.* 2011) or string graph structures (Myers 2005; Simpson and Durbin 2012) based on k-mers extracted from the reads. During the contig assembly process, when repetitive regions in the genome exceed the span of overlapping reads, the contiguity of the assembly breaks (Sullivan *et al.* 2015). While second-generation assemblies are highly accurate at a nucleotide level, they are usually highly fragmented because a significant number of repetitive regions are longer than the insert length of the sequenced molecule (Claros *et al.* 2012; Treangen and Salzberg 2012).

To resolve these repetitive regions, short-read-only assemblers typically used information from mate-pair reads (mapped onto assembled contigs) for ordering, orienting, and linking contigs, i.e. scaffolding. To obtain mate-pair reads, genomic DNA fragments sheared to several chosen lengths [from 2 to 20 kb (Ekblom and Wolf 2014)] are end-biotinylated and circularized to form separate libraries. The circular DNA is sheared again, and the small fragments, consisting of the biotin junction are captured and sequenced to obtain sequences from 2 opposite ends of the original, long DNA fragments. During the scaffolding process, an assembler would use the approximate mate-pair distance to estimate the size of gaps (Ns) within and between contigs (Simpson and Pop 2015). However, mate-pair reads are prone to introducing hidden scaffolding errors by joining distantly related contigs based on the presence of common repeats (Sohn and Nam 2018).

Phase II was marked with the advent of third-generation sequencing platforms, as produced by PacBio and ONT. LRS on early models and chemistries of these platforms was expensive, and data yield was low and laden with errors (10–15% error rate) such as spurious insertions, deletions, and mischaracterized homopolymer runs (Bao and Lan 2017; Salmela *et al.* 2017). In phase II, those long-reads were hybridized with short-read assemblies to increase contiguity (e.g. contig/scaffold N50), in at least 2 ways. The low-coverage, long-read contigs were either merged with high-coverage, short-read contigs with software like quickmerge (Chakraborty *et al.* 2016), or the gaps between and within scaffolds of short-read assemblies were filled with error-corrected long reads using software like PBJELLY (English *et al.* 2012).

Both the merging and gap-filling processes appear to improve contig and scaffold N50, however, the merging process could inflate genome size or duplicate genomic regions in the assembly, which becomes visible when examining the structure of single-copy ortholog genes, with software such as BUSCO (Benchmarking Universal Single-Copy Ortholog; Simão *et al.* 2015). For instance, when low-coverage contigs assembled with long reads are aligned and merged with short-read contigs, merging failure or hidden scaffolding errors can lead to generation of spurious duplicated BUSCO genes. When long reads are aligned to a short-read assembly to fill gaps between contigs, misjoins from mate-pair reads can result in spurious genome size expansion.

Phase III commenced when new iterations of long-read sequencer technology and improved molecular protocols led to less expensive and higher-throughput sequencing runs—for example, PacBio has reduced costs by 2-fold and increased throughput 10-fold (van Dijk *et al.* 2018). In phase III, the large volume of long reads can be used to directly assemble contigs with assemblers such as Falcon (Chin *et al.* 2016), Canu (Koren *et al.* 2017), WTDBG2 (Ruan and Li 2020), or Flye (Kolmogorov *et al.* 2019). In general, phase III has dramatically increased the contiguity of assembly components (Amarasinghe *et al.* 2020). Errors in long reads can be corrected through a nonhybrid approach in which instead of using short reads to correct long reads or contigs, the information from overlapping long reads alone is used (Chen *et al.* 2021)—although such self-error correction processes need higher sequencing coverage (Salmela *et al.* 2017; Zhang *et al.* 2020). However, reads of extreme length (tens of thousands of kilobases) or excessive coverage can still degrade the quality of long-read contig assemblies, potentially due to the presence of chimeric reads (Fichot and Norman 2013; White *et al.* 2017). Tools such as yacrd (Marijon *et al.* 2020) have been developed to identify and filter such chimeric reads to improve assembly contiguity.

For any *de novo* genome-based research, the challenge is not only to assemble a genome of high contiguity but also with high accuracy and completeness. Critical data analysis is required to obtain such accuracy. It is a common practice to use high values of completeness of BUSCO annotations and contiguity metrics (e.g. N50) as a proxy for quality; however, there is a general lack of critical evaluation of these results in the literature. Furthermore, genomes built using a phase II strategy have been widely reported (Das *et al.* 2020; Moran *et al.* 2020) and practitioners new to genome-scale research may assume such assemblies are of high quality solely based on the apparent high contiguity reported in the study. Thus, a critical retrospection of the accuracy of those assemblies, as well as the technical underpinnings of such results, will be a useful resource for the broader research community.

We hypothesize that when short-read-only assemblies have hidden scaffolding error and when low-coverage long-read contigs are erroneous, the quality of short- and long-read hybrid assemblies degrades. In this study, we assembled the genome of *Trematomus borchgrevinki*, a cold specialized Antarctic notothenioid fish with an estimated genome size of 1.28 Gb (Chen *et al.* 2008), for which we had all 3 phases of assembly data to investigate assembly quality problems. We show what a more in-depth analysis of BUSCO scores can reveal about assembly quality, and we developed strategies such as k-mer-assembled region replacement and parameter optimization to address phases I and II error models, while demonstrating that long-read sampling can be used to optimize phase III assemblies.

MATERIALS AND METHODS

Specimens, blood sampling and agarose embedding of red blood cells

Specimens of the Antarctic notothenioid fish *Trematomus borchgrevinki* were caught from McMurdo Sound (78ºS), Antarctica by hook and line through holes drilled through annual sea ice, and transported back to the aquarium facility at McMurdo Station. Fish were anesthetized using MS222 (Sigma) and heparinized blood was drawn from the caudal vein using needle and syringe. All fish handling complied with the University of Illinois, Urbana-Champaign (UIUC), IACUC approved protocol. The red blood cells (RBCs) were gently spun down and washed with notothenioid PBS (phosphate buffered saline, 500 mOsm, pH 8.4). Aliquots of buffer-washed RBCs of known concentration (determined with a hemocytometer) from a single male *T. borchgrevinki* were embedded in 1% low melting point agarose plugs using BioRad plug molds (1 cm×0.5 cm×0.75 cm) to prevent shearing of high molecular weight (HMW) genomic DNA, following Miyake and Amemiya (2004). Each plug contained an appropriate number of RBCs to provide about 20 μg of DNA, based on an estimated 1C genome size of 1.1 pg. The agarose embedded RBCs were then lysed exhaustively *in situ* using a 1% LDS lysis buffer (1% lithium dodecyl sulfate, 10mM Tris-HCl pH 8.0, 100mM EDTA, pH 8.0), and preserved in a 20% NDS solution (0.2% N-laurylsarcosyl, 2mM Tris-HCl, 100mM EDTA, pH 9.0). The preserved agarose plugs were returned to the University of Illinois, Urbana-Champaign (UIUC) for DNA extraction.

High molecular weight (HMW) genomic DNA preparation

The agarose plugs were first thoroughly desalted by equilibration with 0.5x TE (5mM Tris-HCl, 0.5mM EDTA, pH 8.0) at 4°C, followed by equilibration with 1x β-agarase buffer (10 mM Bis-Tris, 1 mM EDTA, pH 6.5). Individual plugs were then heated at 65°C for 15 minutes to melt the agarose, then cooled to 42°C. Two units of β-Agarase I (New England BioLabs) per plug of molten agarose was added and gently stirred in, and the sample was incubated at 42°C for 1-2 hour. The digested (liquified) plug was then incubated with proteinase K (final concentration of 2 mg/mL) at 55°C for one hour.

HMW DNA was prepared for sequencing on three different platforms – Illumina, Oxford Nanopore, and Pacific Biosciences (PacBio) Sequel II. For Illumina sequencing, the DNA was recovered from the digested plug by one gentle extraction with phenol:chloroform (1:1), transferred into Spectra/Por 3 dialysis tubing (MW cutoff 3500 Da), and dialyzed exhaustively against 0.5x TE. For Nanopore long read and PacBio CLR (continuous long read) sequencing, the DNA was recovered from the digest using the SPRI paramagnetic bead-based GenFind V3 kit (Beckman Coulter) following vendor instructions, but with two additional DNA elutions (for a total of three). The concentrations of recovered HMW DNA were determined using Qubit dsDNA Broad Range Assays and Qubit v.3 fluorometer (Invitrogen). The integrity and size range of the DNA were assessed by pulsed-field electrophoresis using a BioRad CHEF Mapper XA system. The DNA were of high purity and integrity, and achieved MW of 35kp to ≥150Kbp with the phenol:chloroform extraction method, and 48Kbp to ≥190Kbp with GenFind v.3, with insignificant fraction below the lower bound.

Sequencing

High molecular weight (HMW) DNA was extracted from red blood cells of a male and a female specimen of *T.borchgrevinki*, caught from McMurdo Sound (78°S), Antarctica. For the male, sequencing libraries were constructed for sequencing on 3 different platforms, Illumina, Oxford Nanopore, and PacBio Sequel II (see Supplementary text for details). For the female sample, sequencing was performed only on PacBio Sequel II.

For Illumina sequencing, 5 libraries (2 whole-genome shotgun libraries and 3 mate-pair libraries) were constructed. Two shotgun libraries were prepared using the Hyper Library construction kit (Kapa Biosystems) with no PCR amplification. For the first and the second libraries, insert size ranges of 400–500 and 700–800 bp fragments, respectively, were selected and sequenced on a single lane of HiSeq2500 to generate 250 and 160 bp paired-end reads, respectively. Three mate-pair libraries with insert size ranges of 2–5, 5–7, and 8–12 kb fragments, were constructed using the Nextera Mate Pair Library Sample prep kit (Illumina) followed by the TrueSeq DNA Sample Prep kit (we will refer to them as the 5, 7, and 12 kb mate-pair libraries subsequently). Each mate-pair library was sequenced on one lane of HiSeq2500 for 160 bp paired-end reads, which we refer to as mate-pair reads when paired-end reads are generated from mate-pair libraries.

For Oxford Nanopore sequencing, 12 libraries were made using the SQK-LSK109 ligation sequencing kit (Oxford Nanopore) to produce 1D reads, and each library was sequenced on one SpotON R9.4.1 FLO-MIN106 flowcell using a GridIONx5 sequencer. For PacBio CLR sequencing, 1 library for the female and 2 libraries for the male were constructed with unsheared HMW DNA based on PacBio recommendations, selecting for final library fragments ≥45 kb in length. The library was sequenced on Sequel II SMRT cells with 40 h of data collection. Illumina and Nanopore sequencing were carried out at the Roy J. Carver Biotechnology Center, University of Illinois Urbana-Champaign, and PacBio CLR sequencing was performed at the Genomics and Cell Characterization Core Facility, University of Oregon.

Construction and comparison of de novo short-read-only genome assemblies with different k-mer sizes

For each sequenced mate-pair library, the adaptors were removed with NxTrim v0.4.1 (O'Connell *et al.* 2015) and reads with a proper mate-pair orientation were separated from those with unknown orientation using the `–justmp` and `–separate` parameters. These mate-pair and paired-end reads were assembled with Meraculous (v2.2.2.5, Chapman *et al.* 2011), which employs a Hamiltonian *de Bruijn* graph framework based on k-mers to produce a *de novo* genome assembly. The assembly process was independently repeated 5 times, each time employing a different k-mer size (i.e. 51, 61, 71, 81, and 91 bp; **Figure 2.1**).

These 5 phase I assemblies were named after their respective k-mer sizes, as k51, k61, k71, k81, and k91 respectively. For each assembly, we executed QUAST v4.6.2 (Gurevich *et al.* 2013) to estimate contiguity metrics, and we assessed the completeness of 4,584 single-copy orthologs from Actinopterygii-specific OrthoDB v9 using BUSCO v3.0.2 with the default parameters. BUSCO classifies orthologs as (1) single copy and complete (hereafter complete), (2) complete but duplicated (hereafter duplicated), (3) fragmented, or (4) missing. At its core, BUSCO is a wrapper of 3 bioinformatic tools: TBLASTN (Camacho *et al*. 2009), AUGUSTUS (Keller *et al.* 2011), and HMMER (Eddy 2011).

Reverse complementation and reassembly of k71 as well as AUGUSTUS parameter changes

During the comparative assessment of completeness among the k51, k61, k71, k81, and k91 assemblies, we observed that a subset of k71 scaffolds containing fragmented BUSCO genes was assembled in the opposite orientation in alternative assemblies and contained complete versions of the same BUSCO genes. To test whether changing the orientation of a scaffold can convert a fragmented BUSCO gene to a complete one, we reverse complemented the k71 scaffolds (revcom-k71) and repeated the BUSCO analysis.

We next tested whether the inclusion of mate-pair data can affect an assembly and influence BUSCO scores by reassembling k71 while varying the number of mate-pair libraries in the assembly. First, only paired-end reads were used for reassembly. Next, 3 mate-pair libraries with insert sizes of 5, 7, and 12 kb were added separately to the paired-end data to produce 3 independent assemblies. In addition, the combination of 2 mate-pair libraries having 5 and 7 kb insert size as well as that of all 3 mate-pair libraries with paired-end data was employed separately for reassembling k71. We also reverse complemented scaffolds of the assemblies generated from paired-end reads and (1) one mate-pair library or (2) 2 mate-pair libraries.

We further re-executed BUSCO on the k71 assembly by changing the internal default BUSCO parameter `–singlestrand` from false to true. This allows one to find overlapping gene models, i.e. alternative transcripts producing different protein-coding sequences, located on opposite strands (by default BUSCO does not permit overlapping gene models). To validate these findings, we ran BUSCO v5.2.0 on the reference genome assembly of zebrafish, GRCz11 (Ensembl v106) as well as on k71 assembly using OrthoDB v10 in 3 ways. In the first and the second round, `–singlestrand` parameter was toggled false and then true, respectively. Third, we reverse complemented chromosomes or scaffolds with BUSCO genes that were fragmented in the first round but became complete in the second round.

A k-mer based strategy to improve the completeness of BUSCO genes in a short-read assembly

We developed and optimized a k-mer-based strategy to improve the completeness of k71 by writing 2 custom Python scripts, INFO and CONTEX. INFO enumerates the following elements of the BUSCO evaluations: (1) the names of fragmented genes in k71, (2) the enclosing scaffolds for those genes, (3) the start and the end basepair positions of each gene, (4) scaffold names in alternative assemblies (k51, k61, k81, and k91) with a complete gene, (5) the start and end basepair positions of those complete alternative genes, and (6) scaffold sequences from k71 and alternative assemblies.

CONTEX imports the data generated by INFO to improve k71 by translocating complete genes from alternative assemblies using a k-mer-based strategy (**Figure 2.2**). For each fragmented gene, CONTEX retrieves the k71 scaffold as well as the scaffold with a complete gene from an alternative assembly and syncs their orientation. It then k-merizes the whole k71 scaffold and the flanking sequences of the complete gene from the alternative assembly. Whenever k-mers of the flanking sequences and the whole scaffold match, CONTEX replaces the enclosing contig(s) (**Figure 2.2**). The improved k71 assembly generated by CONTEX was named *cork71*.

The additional details on algorithm are CONTEX as follows. CONTEX parses the csv file generated by INFO in a way that information related to each fragmented BUSCO gene is extracted one at a time. It applies to filter any fragmented BUSCO gene from downstream analysis, if the gene is found as complete in the reverse complemented scaffold. Then, the direction of each scaffold with fragmented BUSCO gene relative to that with complete BUSCO gene is determined. The comparison is performed by using gene(s) flanking both fragmented as well as complete genes only one or both side(s). If the comparison shows that order of flanking gene(s) along the scaffold is consistently same or opposite relative to complete as well as fragmented genes, then the directions of scaffolds are considered same or opposite to each other, respectively. If the relative order of adjacent gene(s) is inconsistent or if there is overlap between either between complete or fragmented gene and neighboring gene(s), then direction of the scaffold with fragmented BUSCO gene is not determined and the gene is filtered out from the downstream analysis.

The direction of the scaffold with only one gene is also determined by CONTEX based on two step process. However, the second step is only performed when the first step is unsuccessful. In the first step, k-mers of flanking sequences from one or both sides of complete BUSCO gene, depending on start and end positions of the gene, are searched against unique sets of k-mers generated independently from non-reverse and reverse completed scaffold with fragmented BUSCO gene. The directions of scaffolds with complete and fragmented gene are considered same or opposite, if the flanking sequences matches only to the k-mers from non-reverse or reverse complemented scaffolds, respectively. If the flanking sequences map to k-mers either from both non-reverse and reverse complemented scaffold or from none of them, CONTEX maps the k-mers of whole scaffold having complete BUSCO to the k-mers of whole scaffold having fragmented BUSCO gene. CONTEX implements user defined percentage of shared k-mers between the scaffolds to define the relative direction of scaffolds.

After determining the direction of scaffolds, CONTEX grabs each scaffold with fragmented BUSCO genes as well as kmerizes it and retains non-repetitive and non-palindromic k-mers. CONTEX also kmerize the flanking sequences complete BUSCO gene versions. The k-mers of flanking sequences are search against k-mers of the scaffold with fragmented BUSCO gene. Once the match between the k-mers from the two different sources are found, the contig(s) with fragmented BUSCO genes are replaced with contig(s) containing complete BUSCO gene.

Construction of de novo short- and long-read hybrid genome assemblies

As the *cork71* assembly of *T. borchgrevinki* was still highly fragmented, we employed 2 phase II hybrid genome assembly strategies to increase contiguity. The first strategy involved merging low-coverage, long-read-based contigs with *k71*. In detail, first, the raw Nanopore reads were independently assembled with Canu (v1.8, Koren *et al.* 2017) and WTDBG2 (v2.3, Ruan and Li 2020) assemblers and assessed with QUAST. Since the assembly from WTDBG2 had a higher contig N50 it was chosen for further analysis. However, the error-corrected Nanopore reads that Canu generated were reserved. Next, 2 rounds of polishing were executed on the WTDBG2 assembly with Pilon (v1.23, Walker, *et al.* 2014). In the first round, we only corrected small indels and SNPs using the Illumina 2 × 250 bp reads, whereas in the second round, we also included the 2 × 160 bp mate-pair reads and allowed for local reassembly. Since the second polishing strategy resulted in a higher N50, we proceeded only with this data set, which we named as *corNpor*. The assemblies *corNpor* and *k71* were aligned to each other using the nucmer program from the MUMMER package (v3.1, Kurtz *et al.* 2004). For the alignments, *corNpor* was used as the "reference" whereas *k71* as the "query." The alignments generated due to repeats and duplicates were filtered out with the MUMMER delta-filter program by manipulating the minimum alignment identity (`-i`) and minimum length of alignment (`-l`) parameters, including (1) `-i 95 -l 0` (default), (2) `-i 95 -l 1,000`, (3) `-i 95 -l 5,000`, and (4) `-i 95 -l 10,000`. After filtering alignments, finally, we merged the reference *corNpor* and the query *cork71* using quickmerge (v0.3, Chakraborty *et* *al.* 2016) with parameters `-hco 5.0 -c 1.5 -l 803500 -ml 5,000` and 5 independent hybrid assemblies were obtained.

These quickmerge-based hybrid assemblies were named, *mergedA*, *mergedB*, *mergedC*, and *mergedD*, after their respective delta-filter values. The overlapping (OVL) to non-overlapping (n-OVL) sequence ratio between two contigs determines the merging of two contigs in quickmerge. By default, any alignment with an OVL/n-OVL ratio less than 1.5 is not considered for merging. The hybrid assemblies were assessed with BUSCO and QUAST and a comparative analysis was performed to determine the factor(s) contributing additional duplicated BUSCO genes.

Filling gaps within and between scaffolds of a phase I assembly with long-reads

In a second strategy to obtain a phase II assembly, the gaps between and within scaffolds of *k71* were filled using PBJELLY (PBSUITE v15.4; English *et al.* 2012) with the error-corrected long reads. Default parameters were used except in the mapping (`--mpqv 40`) and assembly stages (changed -1, which means never timeout during local reassembly, to 2, which means timeout in 2 seconds). This gap-filled, *de novo* hybrid genome assembly was referred to as *filk71*.

Construction and optimization of a phase III assembly

To further improve our *T. borchgrevinki* assembly, we generated a phase III assembly using PacBio CLR reads with WTDBG2. A subsampling strategy was developed to improve the contiguity of the long-read-only assembly, through different permutations of minimum and maximum raw read length and total raw read coverage to generate different subsets of CLR reads.

We developed a custom Python program, `sample_reads.py`, to perform the subsampling: the user supplies an estimate of the genome size, a minimum and maximum read length, a target coverage, and given those parameters, the program will randomly sample reads from the input files until the coverage limit is reached. If the user wishes to reconstruct a sampled set of reads, they may specify the same "random" seed to subsequent executions of the script. Each set of sampled reads were then assembled with WTDBG2 and analyzed with BUSCO and QUAST. One round of polishing was performed in the final assembly with the arrow module in GCpp (v2.0.0 Pacific Biosciences) and analyzed with BUSCO. Ten random reads with length greater than 45 kb was chosen and aligned to the WTDBG2 assembly using minimap (v2.1; Li 2018) and alignments were analyzed with samtools (v1.12; Li *et al.* 2009) to test if a read was chimeric.

DATA AVAILABILITY

Raw Illumina and Nanopore reads are available from NCBI under BioProject PRJNA861284. The phase I and II assemblies are hosted on Dryad under DOI 10.5061/dryad.ghx3ffbs3. The custom Python scripts for methods are available in https://bitbucket.org/CatchenLab/scripts_contig_replacement_repo/src/master/.

RESULTS

Short- and long-read sequence data

The sequencing of Illumina libraries selected for 400–500 and 700–800 bp insert lengths separately generated 344,314,404 (83.57× coverage) and 95,269,368 (14.79×) reads, respectively. Three mate-pair libraries with insert sizes 2–5, 5–7, and 8–12 kb generated 115,968,758 (18.01× coverage), 116,808,220 (18.14×), and 133,442,224 (20.72×) reads, respectively. In addition, Nanopore sequencing generated 3,872,632 reads with a mean and average N50 length of 6.6 and 10.5 kb, respectively, for 24.29 Gb total length (23.58× coverage). The PacBio CLR sequencing from a single SMRT cell generated 118.42 Gb (114.97× coverage) in 7,651,558 reads with a mean and N50 length of 23.7 and 33.4 kb, respectively.

The k71 assembly showed high scaffold N50 but low completeness of BUSCO genes

Among 5 *de novo* short-read-only assemblies (k51, k61, k71, k81, and k91) generated with Meraculous, k71 had the highest scaffold N50 (746 kb, **Table 2.1; Figure 2.3**). However, results from BUSCO analyses showed that the number of single-copy, complete genes was the highest in k51 (4,221), with k71 (4,177) in third place (**Table 2.2**). In addition, a fraction of BUSCO genes that were fragmented in k71 were complete in other assemblies, specifically 62, 46, 30, and 35 fragmented genes in k71 were found complete in k51, k61, k81, and k91, respectively.

Reverse complementation, reassembly, and AUGUSTUS parameter modification reclassified BUSCO genes

When all the scaffolds of k71 were reverse complemented, a total of 29 fragmented BUSCO genes were reclassified as complete (**Tables 2.3 & 2.4**). These 29 cases of gene reclassification were almost always accompanied by changes in gene lengths; however, the underlying candidate genomic regions (i.e. potential gene locations outlined by the TBLASTN component of BUSCO) remained the same or highly similar. For the 29 reclassified genes, typically, the complete gene versions were shorter in length compared to their fragmented versions, while the start and the end positions of these complete versions were mapped within the boundaries of the originally fragmented version. In rare cases, when the complete version was longer than its fragmented version, the start and the end positions of the candidate gene model mapped to 2 different gene models, which were identified as candidates for the fragmented version (**Figure 2.4**).

The effect of mate-pair libraries on assembly metrics and BUSCO scores was observed through reassembling k71 and the reverse complemented versions. In general, when one or more mate-pair libraries were added to the paired-end reads of k71, the scaffold N50 increased and the number of scaffolds decreased (**Table 2.5**). In addition, the number of complete and duplicated BUSCO genes increased whereas the number of fragmented and missing BUSCO genes decreased (**Table 2.6**). Also, the assembly contiguity and BUSCO score were better when 3 mate-pair libraries were added to paired-end data rather than 1 or 2 mate-pair libraries (**Tables 2.5 & 2.6**). However, with further investigation, we found inconsistencies in the status of BUSCO genes across reassembled genomes. For example, when the same set of 29 reclassified BUSCO genes in k71 were scanned across the reassembled genomes, the genes that were complete in one reassembled genome were not always complete across other reassembled genomes (**Tables 2.7 & 2.8**). In addition, with the replacement of one mate-pair library of a given insert size with another, or the addition of more mate-pair libraries, when a BUSCO gene converted from fragmented to complete and vice-versa (**Table 2.7**), the corresponding scaffolds with different complete/fragmented gene status were typically found to be oriented in the opposite direction. Also, for some genes, when these scaffolds with different orientations were manually set to the same direction, the status of the same BUSCO gene in the scaffolds across assemblies became the same (**Table 2.9**).

Instead of reverse complementing all scaffolds in the k71 assembly or reassembled genomes, when we simply enabled the AUGUSTUS “singlestrand” parameter (see *Materials and Methods*), 26 fragmented versions of the 29 reclassified genes converted into their complete versions. In these 26 cases, 22 and 4 complete BUSCO genes became shorter (**Figure 2.5.A**) and longer (**Figure 2.5.B**), respectively. These 26 complete versions had the exact same gene length and corresponding protein sequence as those we obtained by reverse complementing the scaffolds.

To ensure our results were not anomalous to our *T. borchgrevinki* genome or the specific set of BUSCO annotations, we repeated the analysis using the model zebrafish genome as well as k71 with BUSCO v5.2.0. We found that 6 and 12 fragmented BUSCO genes in zebrafish and k71, respectively, became complete and their length changed, when "singlestrand" was set as true as well as when chromosomes or scaffolds containing them were manually reverse complemented.

Contig replacement lowered the number of fragmented BUSCO genes in k71

The CONTEX program identified 79 of 130 BUSCO genes that were fragmented in k71 but complete in at least one of the other assemblies (k51, k61, k71, k81, and k91). Using a k-mer size of 31, CONTEX corrected 39 of the 79 fragmented BUSCO genes resulting in the *cork71* assembly (**Table 2.10**). Of the remaining 40 genes, 39 genes were not corrected because they could not be translocated between assemblies without causing problems with neighboring genes, or the directionality of scaffolds could not be reliably determined between assemblies, or genes showed inconsistent fragmentation status with a change in scaffold direction (i.e. genes were fragmented in one direction but not in another).

Phase II assemblies increased contiguity and the number of BUSCO gene duplicates

When comparing the *corNpor* assembly at the nucleotide level using Pilon, the total number of bases confirmed against the Illumina short reads was 84.24%. Compared to the phase I *cork71* assembly, all phase II merged assemblies (*A*, *B*, *C*, and *D*) not only had higher scaffold

N50 and fewer gaps (Ns per 100 kb, **Table 2.11)** but also a higher number of duplicated BUSCO genes. As a reminder (see *Materials and Methods*), we increased the required minimum alignment length between *cork71* and *corNpor* contigs in each assembly from *mergedA to mergedD*. The duplicates decreased from 172 in *mergedA* to 143 in *mergedB* but increased further in *mergedC* (181) and *mergedD* (212, **Table 2.11; Figure 2.6**).

By comparing many-to-one alignments between scaffolds of *cork71* (*query*) to contigs in *corNpor* (*reference*), we observed many cases in which erroneous BUSCO gene duplication occurred when at least 2 conditions were met. First, at least one query (e.g. Illumina scaffold-1) was merged with the reference (e.g. Nanopore contig-1) to form a hybrid sequence. Second, at least one other distinct query (e.g. Illumina scaffold-2) failed to merge with the same reference (Nanopore contig-1), but both of them contained the same or similar set of BUSCO genes. When only the first condition was met, gene duplications did not occur. However, when the second condition was satisfied (i.e. when merging failure occurred), the set of BUSCO genes became duplicated as the hybrid sequence—generated from the alignments between the reference (Nanopore contig-1) and the query (Illumina scaffold-1) that merged—and the unmerged query (Illumina scaffold-2) were placed together in the merged assembly. Such failures can occur when the OVL portion of the reference and the query sequences was either low or absent **(Figure 2.7**).

In addition, we observed numerous cases in which an increase in the stringency of the minimum alignment length parameter reduced or even removed the overlapping portion of the alignment. Moreover, the overall number of alignments with a high alignment percentage decreased with the increase in parameter stringency **(Figure 2.8)**. When the stringency was low, we found a case in which the linear order of alignment fragments was disrupted by the inclusion of small, nonhomologous regions of the query and reference sequence. That, in turn, spuriously changed the start position of the query causing quickmerge to calculate a false high value of n-OVL portion of the alignment. This drastically lowered the OVL/n-OVL ratio (see *Materials and Methods*) to a value less than the merging threshold and resulted in merging failure and duplication of BUSCO genes (**Figure 2.9**). This error, however, was not observed, when the stringency was high as more small alignments were filtered out.

Comparing many-to-one alignments from *corNpor* back to *cork71*, we identified a case in which each merged assembly (*A*, *B*, *C*, and *D*) had 2 sets of 23 genes (46 in total) that were duplicates of each other—the highest we found. These gene sets were in 2 distinct hybrid sequences clustered in a row. These 2 hybrid sequences had one common corresponding query sequence (a scaffold in *cork71*; **Figure 2.10**) that contained the 23 complete genes. This common query scaffold mapped to regions in 4 distinct reference sequences (contigs of *corNpor*), one mapped to the distal portion of the common query, a second mapped to the proximal portion, and regions from the remaining 2 references mapped in between. While some of these mappings could be eliminated by changing the alignment stringency parameter, the duplication could not be fully prevented. However, when the common query was manually split into 2 parts by breaking it at a gap located upstream of its portion overlapping to the second reference, the duplicated 23 BUSCO genes converted to single-copy, complete genes, confirming the source of the duplication.

Gap-filling the short-read assembly with long-reads inflated genome size

As an alternative to creating a phase II assembly using quickmerge, we filled gaps in the *k71* assembly using error-corrected Nanopore reads with PBJELLY, generating the assembly *filk71*. Compared to *k71*, the *filk71* had a higher contig N50 (14 kb) and fewer gaps (Ns per 100 kb; 5.6 kb) as well as a longer total length (187 Mb larger; **Table 2.11**). However, we found 28,377 gaps in *filk71* were overfilled by PBJELLY. A gap is overfilled when long reads from either side of a gap extend into the gap from its flanking regions expanding the size of the original gap without closing it (**Figure 2.11**). From BUSCO, we observed that the number of duplicated genes was higher in *filk71* (2.3%, or 105 genes) than in k71 (2.1%, 95 genes; **Table 2.11**) and that 37 complete BUSCO genes in *k71* became duplicated in *filk71*.

Creating and optimizing a phase III assembly

We found that all assemblies built by subsampling raw PacBio long-reads improved the contiguity metrics compared to those obtained from assembling all raw long reads (**Table 2.11**; **Table 2.10; Figure 2.12**). For example, generating 70× coverage (based on a 1 Gb genome size estimate) using read lengths that ranged from 10–40, 15–40, and 15–45 kb, and assembling each subset of reads increased contig N50 more than 3 times, decreased number of contigs by half, and increased the largest contig length by more than 3.5 Mb compared to assembling all raw reads. We also observed variation in contiguity statistics for genome assemblies built with different sets of subsampled reads that represented the same amount of data. For example, shifting the minimum read length from 10 to 15 kb and the maximum read length from 40 to 45 kb, the amount of coverage was the same (70 Gb); however, the number of contigs increased by 370 and the contig N50 decreased by 0.16 Mb (**Table 2.12**). Also, we found evidence for chimeras among the longest reads, with one read of length 99,920 bp that aligned to 2 contigs of the WTDBG2 assembly with mapping quality of 60.

DISCUSSION

Here, we aim to elucidate the common sources of error in 3 distinct phases of genome assembly to yield some useful insights. First, for phase I assembly, although mate-pair reads increase contiguity (e.g. N50), they can inflate or deflate the BUSCO score of gene completeness. Mate-pair libraries of different insert sizes can interfere with each other, and a single best combination of mate-pair library types does not appear to exist in our data. A phase I assembly can be improved using a k-mer-based contig replacement strategy, though inconsistencies in alternative assemblies place limits on its efficacy. Second, for phase II assembly, when merging contigs created from low volume long reads with phase I contigs, the presence of sequence errors or small repeat alignments can quickly degrade the quality of the hybrid assembly. This problem grows as more assemblies are merged and in general, it is essential to optimize the alignment parameters used for the merging process. Furthermore, hidden scaffolding error generated from mate-pair libraries in the phase I assembly will further degrade the quality of hybrid assemblies. A critical analysis of BUSCO scores is necessary to evaluate the quality of any hybrid assembly that appears to have high contiguity. Finally, for phase III assembly, long reads generate highly contiguous assemblies; however, chimeric long reads or excessive coverage can lower the contiguity of the assembly. Sampling long reads can improve the contiguity of the long-read-only contig-level assembly.

Phase I

A single k-mer size cannot produce an optimal assembly, as measured by BUSCO

For our phase I assemblies, the short-read assembly with the highest N50 did not have the highest number of complete BUSCO genes while the number of fragmented BUSCO genes varied among assemblies using different k-mer lengths. These patterns are consistent with what was reported by Moran *et al.* (2020) for 4 phase I assemblies of orange throat darter fish. The authors reported that 4 assemblies built with k-mer sizes 49, 59, 69, and 79 had (1) 4,247, 4,241, 4,233, and 4,219 complete BUSCO genes, respectively, (b) 2.4, 2.2, 2.5, and 2.3 Mb of scaffold N50, and (3) 86, 93, 86, and 91 fragmented BUSCO genes. These results suggest that different regions of the genome would assemble better with different k-mer sizes, due to the interaction of k-mer length, the commonality of those k-mers in the genome, and sequencing coverage.

It is well recognized that having nonoptimal k-mer size affects the contiguity of short-read assemblies. Having a k-mer size that is too large can increase assembly fragmentation as large k-mers tend to have difficulty in finding overlapping, adjacent k-mers resulting in gaps. However, having a small k-mer size can increase misassembly as it favors collapsing repeats (Chikhi and Medvedev 2014), which can result in chimeric joins (while additionally, mate-pair reads can spuriously join genomic regions that are far apart; Treangen and Salzberg 2012). In both cases, the intron/exon structures of genes can be prevented from being properly assembled, as reflected in BUSCO results. While some *de novo* assemblers attempt to apply different k-mer sizes (e.g. Spades, Bankevich *et al.* 2012), it is in practice a difficult problem and one that has been superseded by newer, phase III approaches.

Mate-pairs can inflate or deflate BUSCO scores by generating aberrations in phase I assemblies

We found reverse complementing scaffolds can convert some fragmented BUSCO genes to complete versions and vice-versa, although TBLASTN searches, used by BUSCO to outline genomic regions to annotate, yielded the same candidate gene regions in the forward and reverse complemented scaffolds. This evidence suggests that some complete/fragmented BUSCO genes are aberrations that are only counted when contigs end up being in one particular orientation. Since mate-pair reads determine the orientation of a contig within a wider scaffold, they may be the primary culprit for these types of errors.

Swapping mate-pair libraries in our k71 assembly, we observed that corresponding scaffolds in alternative assemblies that had complete or fragmented versions of the same BUSCO gene typically had different orientations. The same pattern occurred when we increased the number of mate-pair libraries for reassembled genomes, and we found some cases in which manually forcing the scaffold orientation to be in the same direction generated the same gene version in all of them. This means that when mate-pair libraries with different insert sizes are mixed together, they can interfere with each other, and in turn, the completeness of a BUSCO gene can change. As mate-pair reads often lead to misjoins in the scaffolding process due to repeats, we think it is a fundamental nature of genomic repeats—and the inability of short reads to bridge them—that is responsible for the errors. Finally, our comparative analyses indicate that potentially the default "singlestrand" parameter in AUGUSTUS can trigger the misannotation of BUSCO genes, depending upon how mate-pair reads orient the underlying contigs, and consequently can contribute to the generation of annotation aberrations. Researchers involved in the application of BUSCO may benefit from varying this parameter in their own assemblies.

Importantly, with BUSCO, when the underlying assembly changes, the genomic lengths of the corresponding single-copy orthologs can change as well. Our comparative analyses suggest that these changes in the BUSCO gene lengths occur through at least 3 processes. First, the length can decrease due to the splitting of a long gene model in one direction into smaller gene models in the alternative direction (**Figure 2.5.A**). Second, the shift in the start or end position of the gene model can decrease **(Figure 2.5.A**) or increase (**Figure 2.5.A**) length. Third, BUSCO gene length can increase through the combination of smaller gene models (**Figure 2.5.B**). Here we refer to gene models as alternative transcripts resulting in different protein products from the same underlying gene.

No combination of mate-pair libraries can be considered better than another for assembly optimization

When we observed 29 BUSCO genes that were fragmented in *k71* but complete in the reverse complemented *k71*, their fate differed among k71 assemblies containing different complements of mate-pair libraries. Whether increasing the number of mate-pair libraries or swapping out mate-pair libraries with different insert sizes, inconsistent patterns in the completeness of BUSCO genes appeared. These results suggest that different mate-pair library combinations create different scaffolding errors and therefore some BUSCO genes will only be complete with a specific mate-pair or combination of mate-pair libraries. Changes in the BUSCO classification of genes most commonly appeared when mate-pair libraries changed the orientation of the underlying scaffold confirming the effect of mate-pairs on the assembly process and further highlighting the susceptibility of BUSCO classifications to errors due to underlying contig orientation.

Conitg-based gene replacement can improve fragmented BUSCO genes in phase I assemblies

We hypothesized that short-read assemblies could be improved by incorporating successful components of different assemblies. Our k-mer-based gene replacement strategy successfully improved 39 of the 79 fragmented BUSCO genes to produce our *cork71* assembly. However, the underlying genomic architecture of the focal genome limits the success of this strategy, as we were unable to fix the 30 additional gene models. While translocating a contig from one assembly to another may fix an assembly error, it also may create additional, new assembly errors highlighting the difficulty of integrating different regions of a genome assembled with different k-mer lengths (whether such an integration is done algorithmically or manually).

Phase II

Erroneous sequence, repeats, and misjoins of contigs can increase duplicated BUSCO genes in hybrid assemblies

We generated hybrid assemblies using quickmerge and compared them to our improved k71 assembly (*cork71*). Our phase II assemblies had higher N50 than *cork71*, however, they also contained a higher number of duplicated BUSCO genes. We found that merging failures between the reference (contigs of the long-read-based *corNpor*) and the query (scaffolds of the short-read-based *cork71*) with same or similar set of BUSCO genes contributed to the inflation of duplicates in our phase II merged assemblies. We observed that setting alignment parameters nonoptimally can halt the merging of a set of phases I and II contigs by reducing or even removing the overlapping portions of an alignment between them.

When a specific query contig is aligned to the reference by nucmer, the matching sequence segments of the query are aligned in a linear fashion if the sequences of the query and reference share high nucleotide sequence identity. However, if the query sequence is repetitive, then the alignment order of the query sequence blocks can be disrupted. Regardless, the summation of all the lengths of all aligned and overlapped blocks of the specific query contig to specific reference contig provides the total length of the alignment (i.e., overlapped and aligned (OVL) portion of the query) for that query sequence. Apart from OVL portion, the query contig may contain sequence that overlaps the reference but does not align (n-OVL) as well as sequence that neither overlaps nor aligns (overhang). When delta-filter is employed, it removes alignment blocks below a minimum identity and length. Quickmerge takes the alignment information to calculate the ratio of OVL to n-OVL and to determine any overhangs of the alignment. It considers merging the reference and query contigs based on the OVL/n-OVL ratio: any alignment with a ratio less than 1.5 is not considered for merging. If it merges the contigs, any overhang of the reference and/or query are included in the final product. The OVL of the reference sequence gets priority over the OVL of query while merging.

Large alignment blocks may fail to form if either the reference or query are highly erroneous. We observed that overall number of alignments with a high alignment percentage decreased when the parameter was increased. Moreover, approximately 16% of the nucleotides of the *corNpor* assembly were unconfirmed against Illumina short reads. As contigs of *cork71* (query) are highly accurate at a nucleotide level, the results suggest that contigs of *corNpor* (reference) still possessed sequence errors that favored the formation of many small alignment blocks between the query and the reference. The nonlinear alignment blocks, which we observed when the stringency of alignment length parameter was low, can be explained by genomic repeats because (1) such blocks were filtered out at high stringency and (2) the alignments of small length are more likely to be formed by repeats than due to true homologous regions. Moreover, when merging failure occurs due to any of these conditions, remnants of the unaligned reference sequences can still get dragged into the final merged assembly resulting in additional, duplicated BUSCO genes. This can happen when a single reference sequence overlaps with 2 or more queries at different portions and at least one of the overlaps surpasses the threshold for merging which we observed in our data (**Figure 2.6; Figure 2.9**).

We also observed a case in which the erroneous duplication of 23 BUSCO genes occurred when portions of multiple contigs in *corNpor* were present in a single scaffold of *cork71.* And, we found that when the scaffold was manually broken, the duplicated BUSCO genes were converted to single-copy complete genes. These results suggest that the scaffold consisted of misjoined contigs. This also means that the presence of hidden scaffolding error in the short-read-only assembly can also lead to generation of spurious duplicates (**Figure. 2.10**).

All in all, our results have shown that while merging 2 assemblies, optimization of the alignment filtration parameter is vital. Thus, it should be set in a way that minimizes the number of duplicated BUSCO genes in the hybrid assembly. The limitation of this parameter optimization is that it may not improve the number of duplicated genes if these duplicates are due to the presence of hidden scaffolding error from mate-pair libraries used in the original, phase I short-read assembly. In our results, some BUSCO duplicates generated due to mate-pair error persisted in all hybrid assemblies.

We find the pattern of increased duplicated BUSCO genes in phase II assemblies in our study was consistent with the pattern found in the genomes assembled by Xu *et al.* (2021). The authors built a chromosome-level assembly for a diploid, Canadian 2-row malting barley cultivar using Illumina, PacBio, 10X Genomics Chromium linked reads, and Hi-C data following 6 steps. One of the intermediate steps involved the merging of Illumina and PacBio contigs (built with corrected reads and polished with Illumina reads) using quickmerge. In this hybrid assembly, the number of duplicated BUSCO genes (107) was higher than those in genomes of 6-row malting barley cultivar, morex (36) and European 2-row malting barley cultivar, Golden Promise (42) built with Illumina data only.

However, the authors did not interpret their BUSCO scores for any step. We argue that the duplicated BUSCO genes could have increased when generating the phase II assembly due to merging failures since the minimum alignment length was 10 kb, which is potentially high because the long-read contigs were assembled with low coverage data (22X). This coverage is too low to for self-correction (Watson and Warr 2019; Zhang *et al.* 2020) and despite further correcting them with Illumina reads, the contigs will still possess errors (such as insertions and deletions) due to the difficulty in mapping the Illumina reads because of repeats (Watson and Warr 2019) but also due to errors in the underlying contigs. Consequently, not all errors disappear.

Similarly, Das *et al.* (2020) assembled the genome of a diploid snapping turtle, *Chelydra serpentine*. In their study, a phase II assembly was generated by filling gaps in the short-read-only assembly with PacBio long reads (average coverage of 11.4×). This gap-filled assembly was further merged with contigs, independently assembled from Nanopore reads (average coverage of 9.6×), employing quickmerge. The number of duplicated BUSCO genes in *C. serpentine* (70) was higher than in the genomes of related reptiles, including *Chelonia mydas* (21; Illumina-based genome), *Chrysemys picta* (17; Illumina and Sanger-based genome), and *Pelodiscus sinensis* (14; Illumina-based genome), and lower than in *Terrapenemexicana* (253; Illumina and 10X Genomics-based but the protocol is unknown). The "minimum alignment length" of 5 kb was set to merge Illumina scaffolds and Nanopore contigs, which, in our data sets, was large enough to result in merging failures and increased duplicated BUSCO genes. Since mate-pair libraries are also used in their phase I assembly, hidden scaffolding errors could have also contributed to the increased number of duplicated BUSCO genes.

Our results are also useful to interpret an increase in duplicated BUSCO genes found in more complex phase II assemblies generated by the hybridization of assemblies produced by 2 or more assemblers from the same, underlying long-read libraries. For example, Ou *et al.* (2019) generated an assembly of pear tree ("Zhongai 1") using PacBio CLR reads and an Hi-C library for scaffolding. However, in an intermediate stage, they merged contigs generated by the Canu and WTDBG2 assemblers that were built from the same sequencing libraries. They report that the number of duplicated BUSCO genes from this hybrid assembly was 28% (407) without interpretation. Such a result may indicate that errors in the long-read contigs could have increased the duplicated BUSCO score through merging failure. Based on our results, we argue that such assemblies need to be reanalyzed for their accuracy. Our results suggest that it is useful to keep track of both N50 and BUSCO scores from different stages of the assembly process and interpreting them to evaluate the results of each stage.

Underlying scaffolding errors can inflate genome size in phase II assemblies

Our phase II assembly, *filk71*, was created by the hybridization of our phase I, Illumina-based Meraculous assembly with Canu-corrected Nanopore reads, using PBJELLY. This resulted in an increased contig N50 size and drastically lowered the number of assembly gaps. However, the number of duplicated BUSCO genes increased and some genes that were complete in *cork71* became duplicated in *filk71*, which suggests that increase in genome length of *filk71* may be of low fidelity. PBJELLY maps the long reads onto the short-read contigs and fills the gaps in 3 ways. First, a long read may cleanly span a gap within or between scaffolds (**Figure 2.11.A**). Second, a long read extends into a gap without spanning the gap (**Figure 2.11.B**). Third, long reads overfill the gap (**Figure 2.11.C**). In *filk71*, we found numerous cases in which gaps were overfilled. This suggests that scaffolds of Illumina assembly possess hidden scaffolding error. When contigs are misjoined, long reads can align to opposite flanking sequences of a gap between 2 contigs, but those reads cannot align to each other and spuriously expand the genome size.

The problem of overfilling is usually unaccounted by researchers. In the literature, we can find examples that potentially indicate spurious genome size expansion but without any explanation. For example, the gap-filled genome of the snapping turtle assembled by Das *et al.* (2020) had an estimated size of 2.20 Gb. They assembled a phase I genome using Illumina paired-end and mate-pair read libraries with ALLPATHS-LG and subsequently filled the gaps with PBJELLY using error corrected PacBio reads. The size of the genome increased by 186 Mb (from 2.13 to 2.31 Gb), which indicates the gaps are potentially overfilled and this increase in genome size could be a spurious expansion. However, the authors did not quantify the number of overfilled gaps.

All the evidences generated from phase II genome assembly strategies suggest that higher N50 does not necessarily mean higher genome quality, and indicate that BUSCO scores may be informative for genome quality. Researchers typically simply report N50 values and BUSCO scores, without interpretation, and place their analytical emphasis on maximizing N50. Furthermore, they then report high BUSCO "completeness" scores, even if the remaining incomplete BUSCO genes offer a wealth of assembly information that is not being examined or interpreted. A step-wise interpretation of BUSCO scores, along with assembly statistics such as N50 and gap length, can provide researchers with significant information relative to the success of their assembly, and indicate sequencing libraries or analysis algorithms that may be degrading the assembly process. In particular, this type of analysis would make clear when to stop hybridizing different assemblies or assembly components (e.g. specific mate-pair libraries) together.

Phase III

<u>Long-read contig assembly can be tuned for higher contiguity through random sampling of reads</u>

For pure long-read assemblies, we observed that filtering by read length and coverage improves the contiguity of the genome compared to using the maximal number of raw PacBio reads. Generally, researchers use all of the CLR reads that pass a minimum read length threshold for *de novo* genome assembly. However, CLR reads of extreme length may be of low accuracy due to polymerase errors occurring within the SMRT cell, for example, the polymerase may not loop around the DNA molecule more than once. While the inclusion of reads of extreme length seems desirable for achieving high assembly contiguity, error rate seems to correlate with read length and, consequently, such reads could actually reduce contiguity.

In addition, PacBio reads may be chimeric, i.e. reads from distant parts of the genome joined together. In our analysis, we found a read of long length (>90 kb) that mapped to 2 distinct regions, and the supplementary alignment matched more than 2 kb of the reference with high quality. Excluding these reads is an easy approach to ameliorate this problem. Furthermore, chimeric reads will be rare in the data (Tvedte *et al.* 2021) and regions of an assembly graph that are linked by such reads will contain low coverage. By randomly sampling all reads down to a base, sufficient level of coverage, these regions of the assembly graph are likely to be excluded, improving the overall assembly. Our result shows that optimizing assembly by subsampling different read sets can help to improve the contiguity of contig-level assemblies. While we provide a program to do the sampling, alternatives, such as seqtk (https://github.com/lh3/seqtk; accessed 2022 Aug 17) are available. Furthermore, tools, such as yacrd (Marijon *et al.* 2020), present an alternative available for reducing chimeric reads in long-read data. Yacrd searches for reads with poor-quality segments based on an all-vs-all alignment of raw reads and selectively filters chimeras. However, it can take a great deal of time and space to process such a set of reads. The subsampling strategy reduces the large data processing time and space consumption for the users. In summary, based on our results, the phase III assembly strategy is the current best state-of-the-art for genome assembly and the resulting contiguity can be tuned by subsampling reads and limiting read lengths.

TABLES

Table 2.1 Genome statistics for five different short-read-only genome assemblies built with five different k-mer sizes ranging from 51 to 91.

Assembly	K-mer size	Number of scaffolds	Scaffold N50	Scaffold L50	Total scaffold length	Max. scaffold length	GC %	N's per 100 Kbp	Contig N50	Contig L50	Total contig length
k51	51	8,018	686,912	294	744,619,395	4,513,143	40.12	25,658.16	5,102	28,468	553,563,782
k61	61	8,561	695,226	285	744,343,530	5,414,719	40.13	24,808.03	5,323	27,463	559,686,532
k71	71	9,399	726,105	271	746,021,077	4,901,101	40.16	23,813.61	5,374	27,669	568,366,538
k81	81	10,579	721,191	257	745,435,592	5,652,300	40.19	22,871.49	5,404	27,739	574,943.361
k91	91	13,160	689,102	272	741,361,822	5,102,311	40.2	22,243.97	5,183	28,922	576,453,487

Table 2.2 Summary of Benchmarking Universal Single-Copy Orthologs (BUSCOs) specific to Actinopterygii clade in the five different short-read-only genome assemblies built with five different k-mer sizes ranging from 51 to 91.

Assembly	K-mer size	Complete	Complete and single-copy	Complete and duplicated	Fragmented	Missing	Total BUSCO groups searched
k51	51	4318 (94.2%)	4221 (92.1%)	97 (2.1%)	94 (2.1%)	172 (3.8%)	4584
k61	61	4288 (93.5%)	4186 (91.3%)	102 (2.2%)	118 (2.6%)	178 (3.9%)	4584
k71	71	4272 (93.2%)	4177 (91.1%)	95 (2.1%)	130 (2.8%)	182 (4.0%)	4584
k81	81	4242 (92.5%)	4146 (90.4%)	96 (2.1%)	150 (3.3%)	192 (4.2%)	4584
k91	91	4213 (92.0%)	4110 (89.7%)	103 (2.2%)	148 (3.2%)	223 (4.9%)	4584

Table 2.3 The number of BUSCO genes in k71 that converted the status from one version of the gene to another in reverse complemented k71 (revcomp-k71).

	Complete in revcomp-k71	**Duplicated in revcomp-k71**	**Fragmented in revcomp-k71**	**Missing in revcomp-k71**
Complete in k71	4111	19	20	27
Duplicated in k71	19	76	0	0
Fragmented in k71	29	0	98	3
Missing in k71	12	0	3	167

Table 2.4 The status of twenty-nine fragmented BUSCO genes from k71 in reverse complemented k71 (revcomp-k71).

Gene	revcomp-k71
EOG090C031F	Complete
EOG090C06A3	Complete
EOG090C09GB	Complete
EOG090C0BB3	Complete
EOG090C0CPN	Complete
EOG090C0CYM	Complete
EOG090C0E4A	Complete
EOG090C0FHB	Complete
EOG090C0FKI	Complete
EOG090C0GDD	Complete
EOG090C01H0	Complete
EOG090C04AG	Complete
EOG090C04VT	Complete
EOG090C08YF	Complete
EOG090C0AN8	Complete
EOG090C0B2Q	Complete
EOG090C0DUI	Complete
EOG090C0FHE	Complete
EOG090C03HW	Complete
EOG090C04JV	Complete
EOG090C05LL	Complete
EOG090C0FY1	Complete
EOG090C03FY	Complete
EOG090C0ARU	Complete
EOG090C0E9K	Complete
EOG090C01VQ	Complete
EOG090C03H9	Complete
EOG090C07FU	Complete
EOG090C0AHG	Complete

Table 2.5 Genome statistics for six different reassembled genomes built with k-mer size of k71.

Reassembled k71	Nuber of Scaffolds	Scaffold N50	Total scaffold length	N's per 100kbp	Number of contigs	Contig N50	Total contig length
PE	92,837	11,287	587,324,760	4,792.79	104,967	7,668	507,922,453
PE+5Kbp	22,019	133,778	696,585,992	18,074.86	108,006	7,240	506,616,768
PE+7Kbp	23,130	164,338	736,792,643	23,032.04	108,599	7,304	504,701,234
PE+12Kbp	31,969	152,151	775,606,865	27,490.87	111,141	7,078	502,856,587
PE+5+7Kbp	12,786	375,608	723,717,036	21,332.22	112,871	6,741	502,880,470
PE+5+7+12Kbp	9,397	718,560	745,779,012	23,786.12	116,549	6,381	500,201,938

PE indicates k71 reassembled with paired-end data only
PE+5Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kbp insert size
PE+7Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 7Kbp insert size
PE+12Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 12Kbp insert size
PE+5+7Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kp and 7Kbp insert sizes
PE+5+7+12Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kp, 7Kbp, and 12Kbp insert sizes

Table 2.6 Summary of Benchmarking Universal Single-Copy Orthologs (BUSCOs) specific to Actinopterygii clade in the six different reassembled genomes built with k-mer size of k71.

Reassembled k71	Complete	Complete and single-copy	Complete and duplicated	Fragmented	Missing	Total BUSCO groups searched
PE	2918 (63.7%)	2860 (62.4%)	58 (1.3%)	962 (21.0%)	704 (15.3%)	4584
PE+5Kbp	4093 (89.3%)	4001 (87.3%)	92 (2.0%)	278 (6.1%)	213 (4.6%)	4584
PE+7Kbp	4149 (90.5%)	4058 (88.5%)	91 (2.0%)	210 (4.6%)	225 (4.9%)	4584
PE+12Kbp	4051 (88.4%)	3957 (86.3%)	94 (2.1%)	261 (5.7%)	272 (5.9%)	4584
PE+5+7Kbp	4257 (92.8%)	4169 (90.9%)	88 (1.9%)	136 (3.0%)	191 (4.2%)	4584
PE+5+7+12Kbp	4267 (93.1%)	4182 (91.2%)	85 (1.9%)	129 (2.8%)	188 (4.1%)	4584

Table 2.7 The status of twenty-nine BUSCO genes (fragmented in k71 but complete in reverse complemented k71) across six different k71 reassembled genomes

Gene	PE	PE+5kbp	PE+7Kbp	PE+12Kbp	PE+5+7Kbp	PE+5+7+12Kbp
EOG090C031F	Complete	Complete	Complete	Complete	Complete	Complete
EOG090C06A3	Complete	Complete	Complete	Complete	Fragmented	Fragmented
EOG090C09GB	Complete	Fragmented	Complete	Fragmented	Complete	Complete
EOG090C0BB3	Complete	Fragmented	Fragmented	Fragmented	Complete	Fragmented
EOG090C0CPN	Complete	Fragmented	Fragmented	Complete	Fragmented	Fragmented
EOG090C0CYM	Complete	Complete	Complete	Complete	Complete	Fragmented
EOG090C0E4A	Complete	Complete	Fragmented	Fragmented	Complete	Fragmented
EOG090C0FHB	Complete	Fragmented	Fragmented	Complete	Complete	Fragmented
EOG090C0FKI	Complete	Complete	Complete	Fragmented	Complete	Fragmented
EOG090C0GDD	Complete	Complete	Complete	Complete	Complete	Fragmented
EOG090C01H0	Fragmented	Complete	Complete	Complete	Complete	Complete
EOG090C04AG	Missing	Complete	Fragmented	Complete	Fragmented	Complete
EOG090C04VT	Fragmented	Complete	Fragmented	Fragmented	Complete	Fragmented
EOG090C08YF	Fragmented	Complete	Fragmented	Complete	Complete	Complete
EOG090C0AN8	Missing	Complete	Complete	Missing	Complete	Complete
EOG090C0B2Q	Fragmented	Complete	Fragmented	Complete	Complete	Complete
EOG090C0DUI	Fragmented	Complete	Complete	Complete	Fragmented	Complete
EOG090C0FHE	Fragmented	Complete	Fragmented	Complete	Fragmented	Complete
EOG090C03HW	Fragmented	Missing	Complete	Fragmented	Fragmented	Fragmented
EOG090C04JV	Fragmented	Fragmented	Complete	Complete	Complete	Fragmented
EOG090C05LL	Fragmented	Fragmented	Complete	Fragmented	Complete	Fragmented
EOG090C0FY1	Fragmented	Fragmented	Complete	Complete	Fragmented	Complete
EOG090C03FY	Fragmented	Fragmented	Fragmented	Complete	Complete	Complete
EOG090C0ARU	Fragmented	Fragmented	Fragmented	Complete	Complete	Fragmented
EOG090C0E9K	Fragmented	Fragmented	Fragmented	Fragmented	Fragmented	Complete
EOG090C01VQ	Missing	Fragmented	Fragmented	Missing	Fragmented	Fragmented
EOG090C03H9	Missing	Fragmented	Missing	Missing	Fragmented	Fragmented
EOG090C07FU	Missing	Fragmented	Fragmented	Missing	Fragmented	Complete

Table 2.7 – Continued

EOG090C0AHG	Fragmented	Fragmented	Fragmented	Fragmented	Fragmented	Fragmented

PE indicates k71 reassembled with paired-end data only
PE+5Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kbp insert size
PE+7Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 7Kbp insert size
PE+12Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 12Kbp insert size
PE+5+7Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kp and 7Kbp insert sizes
PE+5+7+12Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kp, 7Kbp, and 12Kbp insert sizes

Table 2.8 The number of complete versions for twenty-nine BUSCO genes (fragmented in k71 but complete in reverse complemented k71) across six different k71 reassembled genomes

revcomp-k71	PE	PE+5Kbp	PE+7Kbp	PE+12Kbp	PE+5+7Kbp	PE+5+7+12Kbp
29/29	10/29	14/29	13/29	16/29	17/29	13/29

Table 2.9 The status of twenty-nine BUSCO genes (fragmented in k71 but complete in reverse complemented k71) across six different k71 reassembled genomes and their reverse complemented versions

Gene	PE+5 Kbp	revcomp-PE+5Kbp	PE+7 Kbp	revcomp-PE+7Kbp	PE+12 Kbp	revcomp-PE+12Kbp	PE +5+7 Kbp	revcomp-PE+5+7Kp	PE +5+7+12 Kbp
EOG090C031F	Comp	Comp	Comp	Frag	Comp	Frag	Comp	Comp	Comp
EOG090C06A3	Comp	Comp	Comp	Comp	Comp	Comp	Frag	Comp	Frag
EOG090C09GB	Frag	Comp	Comp	Comp	Frag	Comp	Comp	Comp	Comp
EOG090C0BB3	Frag	Comp	Frag	Comp	Frag	Comp	Comp	Frag	Frag
EOG090C0CPN	Frag	Frag	Frag	Comp	Comp	Frag	Frag	Comp	Frag
EOG090C0CYM	Comp	Comp	Comp	Comp	Comp	Comp	Comp	Frag	Frag
EOG090C0E4A	Comp	Frag	Frag	Comp	Frag	Comp	Comp	Comp	Frag
EOG090C0FHB	Frag	Comp	Frag	Comp	Comp	Frag	Comp	Frag	Frag
EOG090C0FKI	Comp	Comp	Comp	Comp	Frag	Comp	Comp	Comp	Frag
EOG090C0GDD	Comp	Comp	Comp	Comp	Comp	Comp	Comp	Frag	Frag
EOG090C01H0	Comp	Frag	Comp	Frag	Comp	Frag	Comp	Comp	Comp
EOG090C04AG	Comp	Comp	Frag	Comp	Comp	Frag	Frag	Frag	Comp
EOG090C04VT	Comp	Frag	Frag	Comp	Frag	Comp	Comp	Frag	Frag
EOG090C08YF	Comp	Comp	Frag	Comp	Comp	Frag	Comp	Comp	Comp

Table 2.9 - Continued

EOG090C0AN8	Comp	Comp	Comp	Frag	Miss	Miss	Comp	Comp	Comp
EOG090C0B2Q	Comp	Frag	Frag	Comp	Comp	Frag	Comp	Comp	Comp
EOG090C0DUI	Comp	Frag	Comp	Frag	Comp	Miss	Frag	Frag	Comp
EOG090C0FHE	Comp	Frag	Frag	Comp	Comp	Frag	Frag	Comp	Comp
EOG090C03HW	Miss	Frag	Comp	Comp	Frag	Frag	Frag	Comp	Frag
EOG090C04JV	Frag	Comp	Comp	Comp	Comp	Frag	Comp	Frag	Frag
EOG090C05LL	Frag	Frag	Comp	Frag	Frag	Comp	Comp	Frag	Frag
EOG090C0FY1	Frag	Comp	Comp	Frag	Comp	Frag	Frag	Comp	Comp
EOG090C03FY	Frag	Comp	Frag	Comp	Comp	Frag	Comp	Frag	Comp
EOG090C0ARU	Frag	Comp	Frag	Comp	Comp	Frag	Comp	Comp	Frag
EOG090C0E9K	Frag	Frag	Frag	Comp	Frag	Comp	Frag	Frag	Comp
EOG090C01VQ	Frag	Frag	Frag	Frag	Miss	Miss	Frag	Frag	Frag
EOG090C03H9	Frag	Frag	Miss	Miss	Miss	Miss	Frag	Comp	Frag
EOG090C07FU	Frag	Frag	Frag	Frag	Miss	Frag	Frag	Frag	Comp
EOG090C0AHG	Frag	Frag	Frag	Frag	Frag	Frag	Frag	Comp	Frag

Table 2.9 - Continued

PE indicates k71 reassembled with paired-end data only
PE+5Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kbp insert size
revcomp- PE+5Kbp indicates reverse complemented PE+5Kbp
PE+7Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 7Kbp insert size
revcomp- PE+7Kbp indicates reverse complemented PE+7Kbp
PE+12Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 12Kbp insert size
revcomp- PE+12Kbp indicates reverse complemented PE+12Kbp
PE+5+7Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kp and 7Kbp insert sizes
revcomp- PE+5+7Kbp indicates reverse complemented PE+5+7Kbp
PE+5+7+12Kbp indicates k71 reassembled with paired-end data plus mate-pair reads with 5Kp, 7Kbp, and 12Kbp insert sizes
Comp, Frag, and Miss indicate complete, fragmented, and missing respectively.

Table 2.10 Thirty-nine BUSCO genes fixed (i.e. convert from fragmented to complete versions) using CONTEX

Fragmented BUSCO gene in k71	Source assembly used to fix the fragmented BUSCO gene	Status (after editing)
EOG090C00H3	K51	Complete
EOG090C01CE	K51	Complete
EOG090C01JC	K51	Complete
EOG090C01QA	K51	Complete
EOG090C01QT	K51	Complete
EOG090C01T5	K51	Complete
EOG090C01T6	K51	Complete
EOG090C02EI	K51	Complete
EOG090C02LX	K51	Complete
EOG090C02NA	K51	Complete
EOG090C02NK	K51	Complete
EOG090C02ZZ	K51	Complete
EOG090C03AV	K51	Complete
EOG090C03P2	K51	Complete
EOG090C03TB	K51	Complete
EOG090C04IH	K51	Complete
EOG090C04LE	K51	Complete
EOG090C04O0	K51	Complete
EOG090C04U0	K51	Complete
EOG090C0502	K51	Complete
EOG090C0563	K61	Complete
EOG090C05AY	K51	Complete
EOG090C05M1	K51	Complete
EOG090C06C9	K51	Complete
EOG090C06X2	K51	Complete

Table 2.10 – Continued

EOG090C0879	K51	Complete
EOG090C09IE	K51	Complete
EOG090C09LR	K51	Complete
EOG090C09XA	K51	Complete
EOG090C0AEQ	K51	Complete
EOG090C0AX8	K51	Complete
EOG090C0BAB	K51	Complete
EOG090C0CJD	K51	Complete
EOG090C0DGW	K51	Complete
EOG090C0A7K	K61	Complete
EOG090C0BZH	K61	Complete
EOG090C04CH	K81	Complete
EOG090C02PR	K91	Complete
EOG090C0C11	K91	Complete

Table 2.11 Summary of genome statatiscs and Benchmarking Universal Single-Copy Orthologs (BUSCOs) specific to Actinopterygii clade for phase I, phase II, and phase III assemblies we assembled.

Assembly	**#Scaf**	**Scaf N50 (Mbp)**	**Scaf total length (Mbp)**	**N's per 100Kbp**	**# Contigs**	**Contig N50 (Kbp)**	**Total contig length (Mbp)**	**C**	**CS**	**CD**	**F**	**M**	**Total Genes searched**
k71	9,399	0.72	746.02	23,813.61	116,693	5.37	568.36	4,272 (93.2%)	4,177 (91.1%)	95 (2.1%)	130 (2.8%)	182 (4.0%)	4584
cork71	9,399	0.72	746.13	23,818.37	116,706	5.37	568.41	4,312 (94.1%)	4,217 (92.0%)	95 (2.1%)	91 (2.0%)	181 (3.9%)	4584
corNpor	N/A	N/A	N/A	N/A	5,394	807.66	843.87	4,435 (96.8%)	4,322 (94.3%)	113 (2.5%)	43 (0.9%)	106 (2.3%)	4584
mergedA	8,426	1.47	751.63	15,018.08	56,003	1,024.86	638.75	4,298 (93.8%)	4,126 (90.0%)	172 (3.8%)	76 (1.7%)	210 (4.5%)	4584
mergedB	8,654	1.40	752.05	15,351.44	57,113	1,001.96	636.60	4,299 (93.8%)	4,156 (90.7%)	143 (3.1%)	75 (1.6%)	210 (4.6%)	4584
mergedC	9,145	1.22	759.96	17,734.96	70,158	470.71	625.18	4,303 (93.8%)	4,122 (89.9%)	181 (3.9%)	78 (1.7%)	203 (4.5%)	4584

Table 2.11 - Continued

mergedD	9,269	0.94	764.50	20,155.11	86,994	9.76	610.41	4,302 (93.8%)	4,090 (89.2%)	212 (4.6%)	83 (1.8%)	199 (4.4%)	4584
filk71	8,055	0.9	933.94	5,639.23	95,999	14.57	881.28	4,372 (95.4%)	4,267 (93.1%)	105 (2.3%	81 (1.8%)	131 (2.8%)	4584
WTDBG2^{r*}	N/A	N/A	N/A	N/A	10,848	758.71	1098.31	N/A	N/A	N/A	N/A	N/A	4584
WTDBG2^{Sr*}	N/A	N/A	N/A	N/A	4,409	2,962.48	924.00	4205 (91.7%)	4085 (89.1%)	120 (2.6%)	134 (2.9%)	245 (5.4%)	4584
WTDBG2^{Sra}	N/A	N/A	N/A	N/A	4,409	2,964.76	924.72	4426 (96.6%)	4317 (94.2%)	109 (2.4%)	37 (0.8%)	121 (2.6%)	4584

k71 indicates original, uncorrected *de novo* short-read only assembly; *cork71* indicates k71 assembly corrected at BUSCO gene level; *corNpor* indicates contig level assembly built with corrected Nanopore reads with low coverage; *mergedA*, *mergedB*, *mergedC*, and *merged* indicates four independent quickmerge-based hybrid assemblies; *filk71* indicates gap-filled k71 with corrected Nanopore-reads

*indicates uncorrected assembly

C: complete; CS: complete and single-copy; CD: complete and duplicated; F: fragmented; M: missing

WTDBG2^{r*} indicates uncorrected long-read only assembly built with raw PacBio data using WTDBG2 assemble

WTDBG2^{Sr*} indicates uncorrected long-read only assembly built with 70Gbp subsampled PacBio data (generated by sampling minimum and maximum read lengths of 10Kbp and 40 Kbp, respectively) using WTDBG2 assembler

WTDBG2^{Sra} indicates polished long-read only assembly built with 70Gbp subsampled PacBio data (generated by sampling minimum and maximum read lengths of 10Kbp and 40 Kbp, respectively) using WTDBG2 assembler

Table 2.12 Genome statistics for assemblies built with raw PacBio data as well as subsampled data

Data	Data amount (Gbp)	Min (Kbp)	Max (Kbp)	N50 (Mb)	# contigs	largest contig (Mb)	Total length (Mb)	L50	Estimated genome size-s	Estimated genome size-a
Raw	181.4	N/A	N/A	0.76	10848	13. 76	1098	279	N/A	1000
Subsampled	80.00	10	40	2.18	6472	19.54	974	103	1000	780
Subsampled	80.00	10	40	2.04	7127	16.05	989	112	1000	900
Subsampled	80.00	10	40	2.88	4491	17.33	926	80	1000	1000
Subsampled	72.00	10	40	1.92	7057	17.42	983	108	900	900
Subsampled	70.00	10	40	2.96	4409	20.24	924	80	1000	1000
Subsampled	70.00	15	40	2.96	4449	17.70	932	74	1000	1000
Subsampled	70.00	15	45	2.80	4779	21.76	939	73	1000	1000
Subsampled	70.00	10	40	1.87	7102	17.44	983	123	1000	900
Subsampled	63.00	10	40	2.78	4416	21.27	921	76	900	1000
Subsampled	63.00	10	40	1.92	7087	16.20	984	114	900	900
Subsampled	54.60	10	40	1.74	7045	19.41	977	126	780	900
Subsampled	54.60	10	40	2.02	6398	14.52	963	110	780	780

Estimate genome size-s indicates the value of genome size used as parameter for subsampling PacBio reads from the raw data
Estimated genome size-a indicates the value of genome size used as parameter in the WTDBG2 assemb

FIGURES

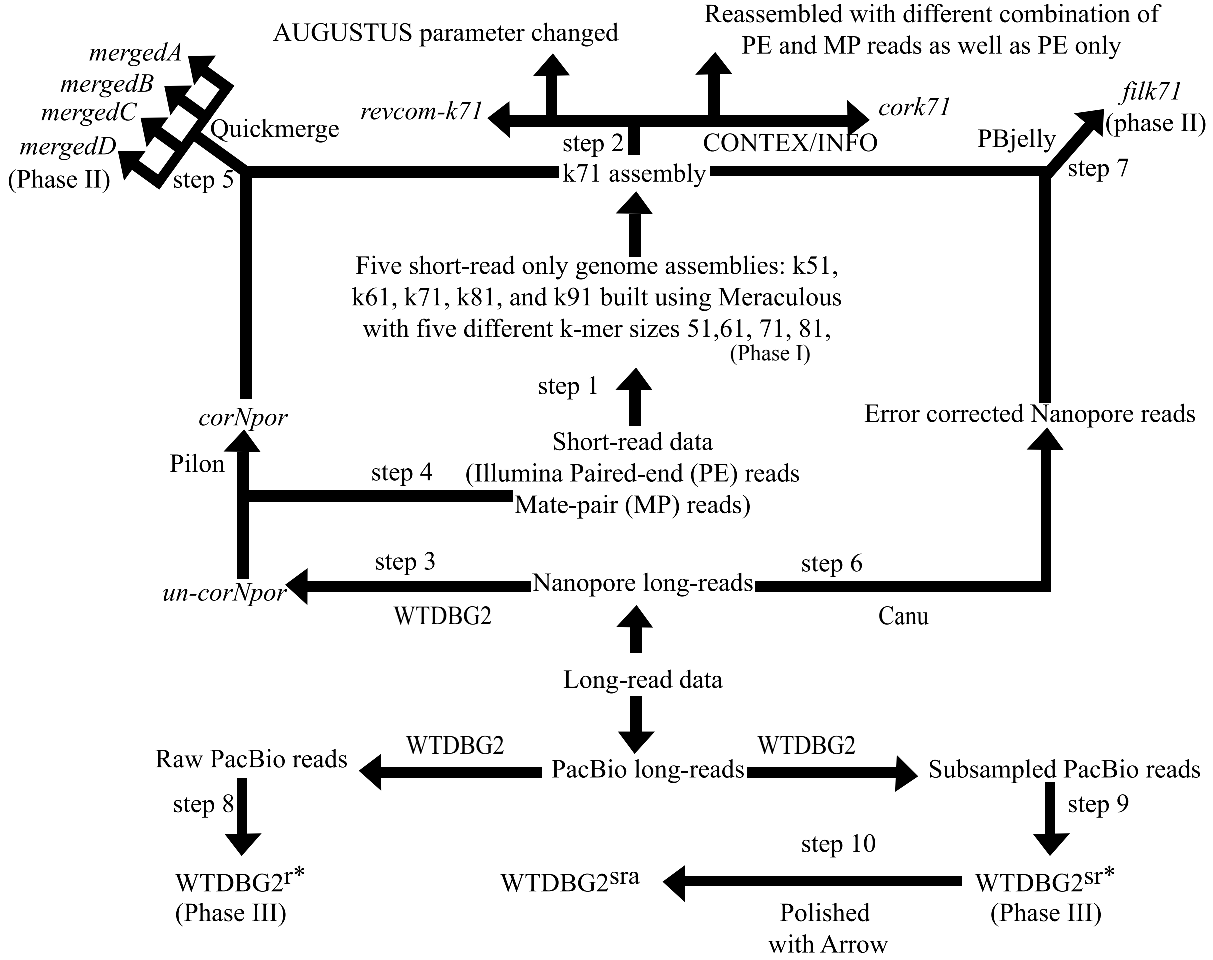


Figure 2.1 Flow chart showing ten steps employed to assemble *de novo* genomes with phase I, II, and III strategies by using Illumina short-reads, Oxford Nanopore long-reads, and Pacific Biosciences long-reads. Step 1: Five short-read assemblies were built with different k-mer sizes of 51 to 91bp using paired-end and mate-pair short-reads. Step 2a: for the k71 assembly, scaffolds were reverse complemented (*revcom-k71*); Step 2b: BUSCO analysis was performed while changing the AUGUSTUS parameter; Step 2c: fragmented BUSCO genes replaced with their complete version using CONTEX/INFO scripts; Stedp 2d: reassemblies were completed with different combinations of mate-pair and paired-end data. Step 3: Nanopore long-reads were assembled with WTDBG2 to produce low coverage, contig-level assembly (*un-corNpor*). Step 4: The *un-corNpor* was polished with short-reads using Pilon to create an error-corrected assembly (*corNpor*). Step 5: The k71 and *corNpor* assemblies were merged as *query* and *reference*, respectively, using Quickmerge by changing the minimum length of alignment in 4 different ways (0, 1000, 5000, 10000) at a minimum alignment identity of 95% to produce hybrid assemblies *mergedA, B, C,* and *D*. Step 6: Nanopore long-reads were corrected with Canu. Step 7: Gaps were filled using the error-corrected Nanopore reads with PBjelly. Step 8: Raw PacBio long-reads were assembled natively using WTDBG2^{r*}. Step 9: Raw PacBio reads were subsampled and assembled to generate contig-level assembly, WTDBG2^{sr*}. Step 10: Error correction was performed on the assembly from step 9 to generate a polished assembly, WTDBG2^{sra}.

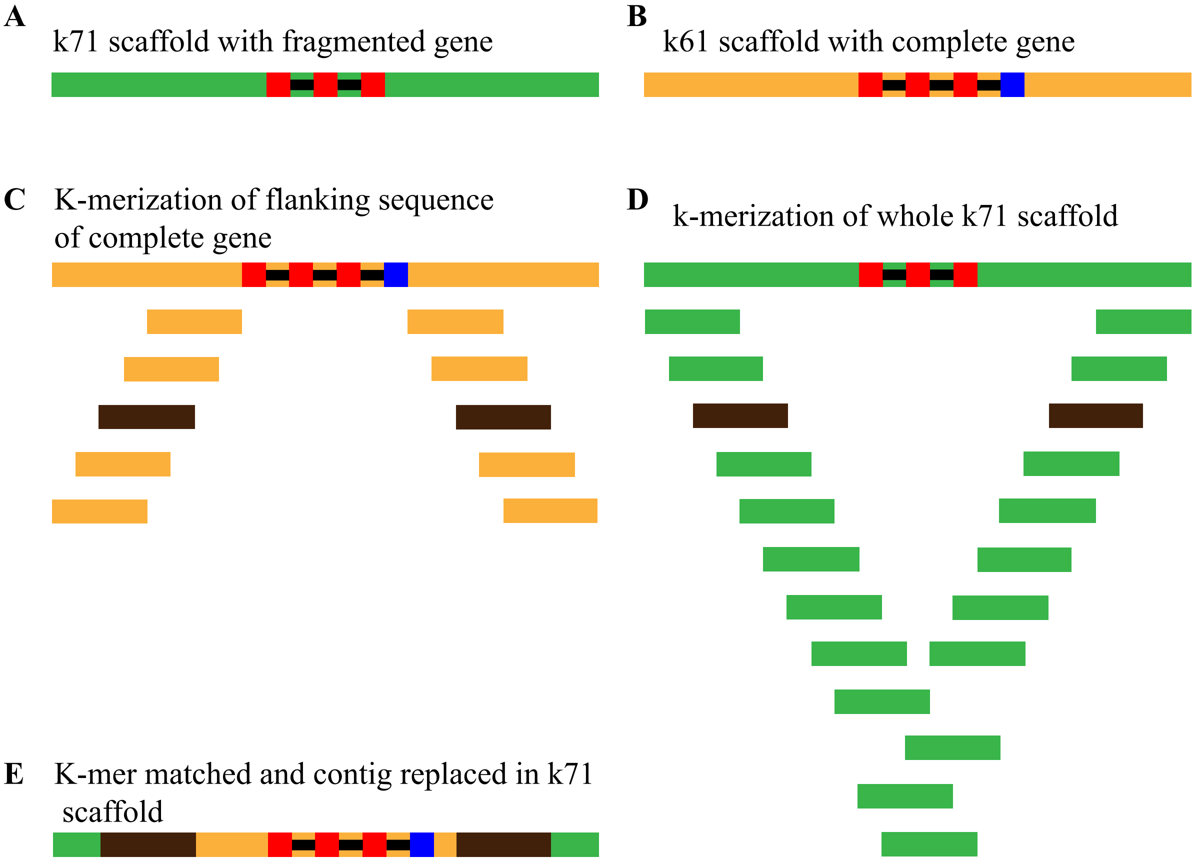

A
k71 scaffold with fragmented gene
B
k61 scaffold with complete gene
C
K-merization of flanking sequence
of complete gene
D
k-merization of whole k71 scaffold
E
K-mer matched and contig replaced in k71
scaffold

Figure 2.2 The five core steps of the CONTEX algorithm. A) Identify the k71 scaffold that contains a fragmented BUSCO gene. B) Identify a scaffold in an alternative assembly (e.g., k61) containing a complete version of the same BUSCO gene. C) K-merize the flanking sequences of the complete BUSCO gene. D) K-merize the whole k71 scaffold and search for matching k-mers in the alternative flanking sequence. E) If the k-mers match, replace the contig within the k71 scaffold with the contig from the alternative assembly.

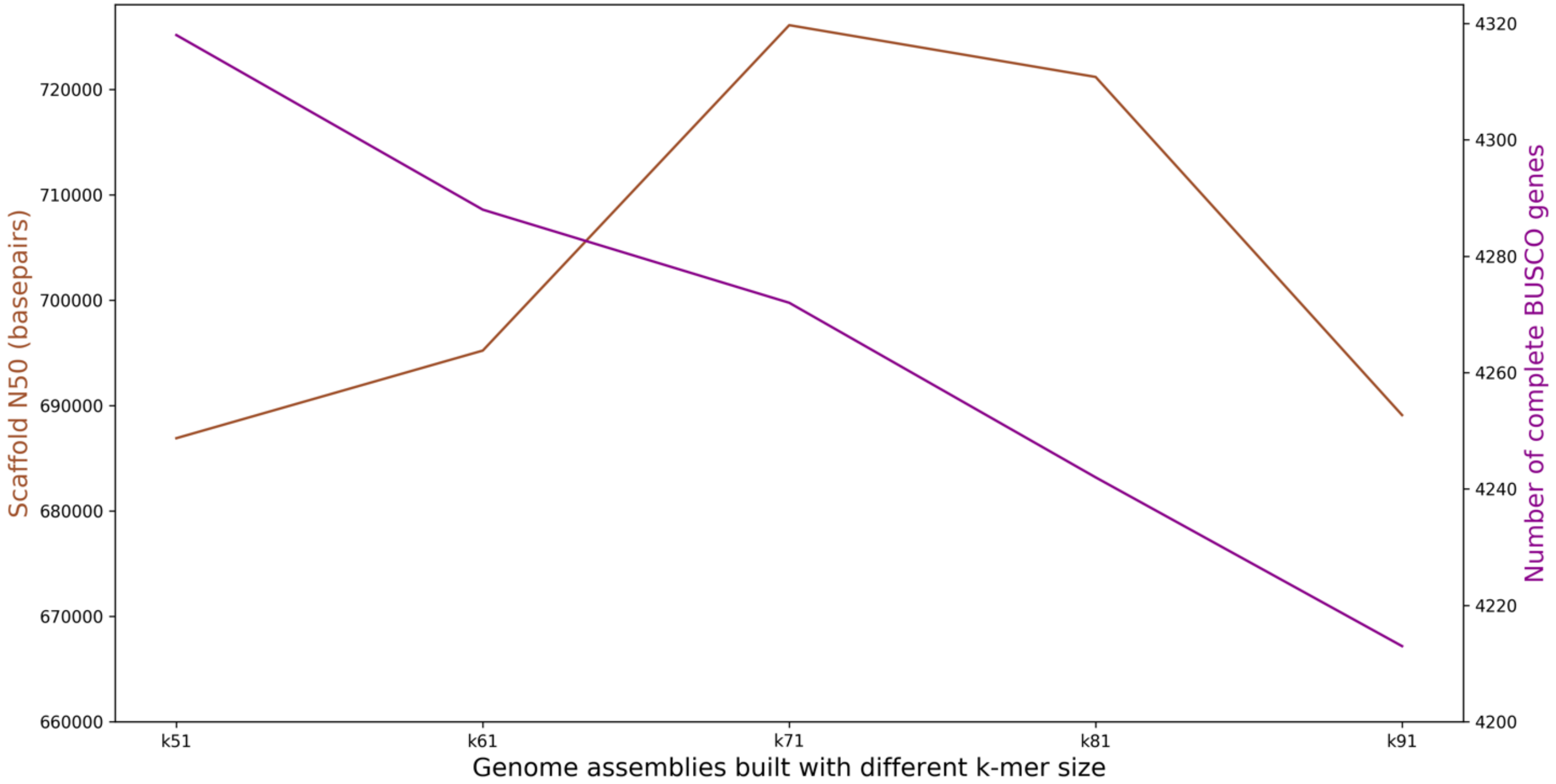


Figure 2.3 Assembly with high contiguity showed low BUSCO gene completeness. This figure shows the contiguity and the completeness of BUSCO genes (specific to Actinopterygii clade), for the short-read only assemblies of *Trematomus borchgrevinki* built with five different k-mer sizes ranging from 51 to 91.

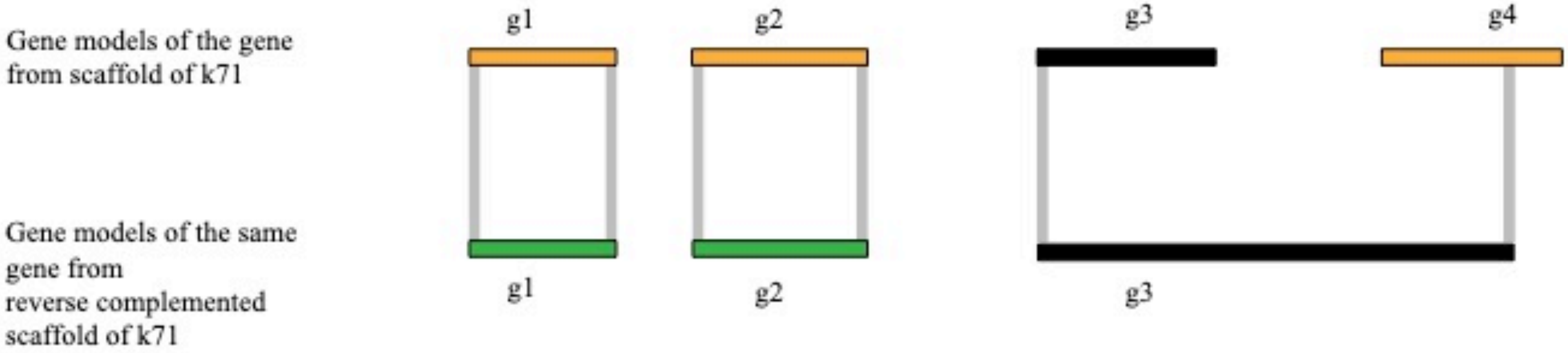

Gene models of the gene from scaffold of k71
g1
g2
g3
g4
Gene models of the same gene from reverse complemented scaffold of k71
g1
g2
g3

Figure 2.4 Reverse complementing a scaffold reduced the number of gene models and increased the length of one of those gene models (g3, black color). Genes g1-g4 are transcripts (gene models) of the same underlying BUSCO gene in k71. After reverse complementing the scaffold containing these gene models, g3 and g4 are merged, resulting in a longer version of g3.

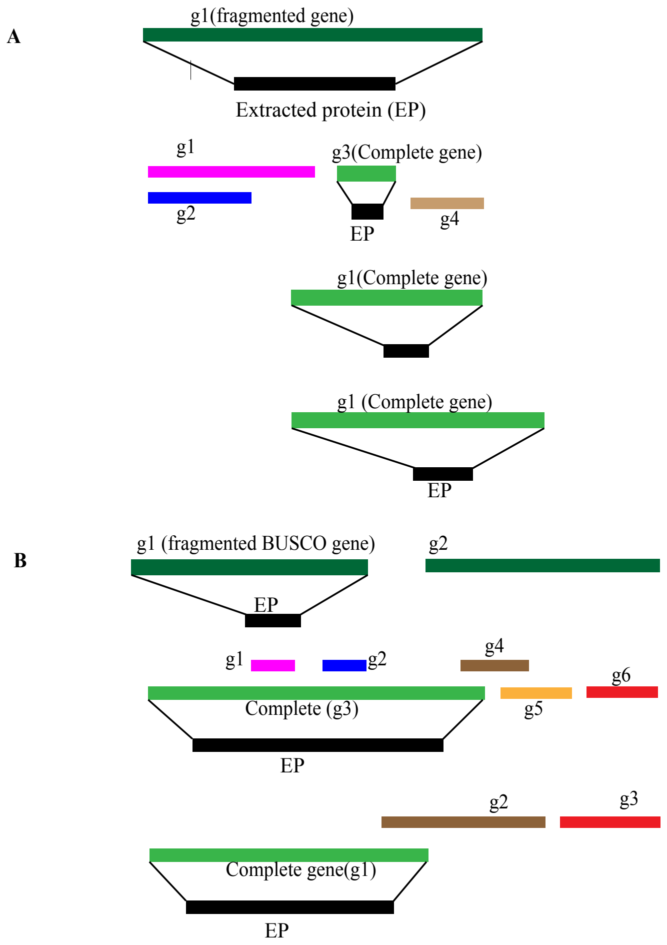


Figure 2.5 Change in gene length with --singlestrand=true parameter in AUGUSTUS. Dark green gene models are predicted by AUGUSTUS with --singlestrand=false whereas green models and all other colors are predicted by AUGUSTUS with --singlestrand=true. A) The fragmented gene model (g1) became complete (light green g3 and g1) through a reduction of gene size. The gene coordinates of the complete versions fell within those of the fragmented version or one of its coordinates shifted outside the boundary of fragmented version. B) The fragmented gene (g1) became complete (light green g1) through an increase in size when the parameter was true. The complete versions overlapped other gene models of the same gene.

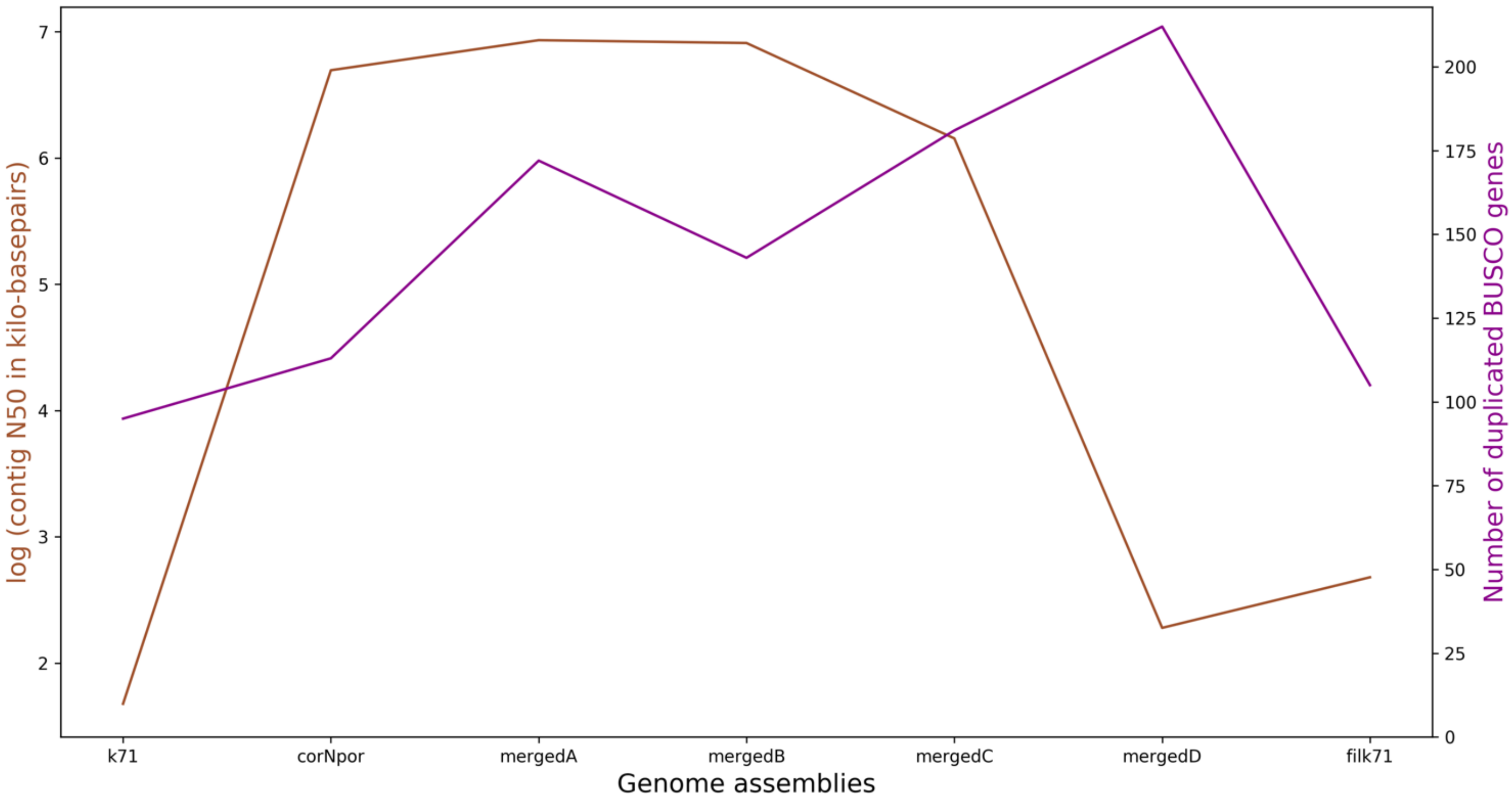


Figure 2.6 The number of duplicated BUSCO genes and contig N50 increased in Quickmerge-based hybrid assemblies (mergedA, B, C, and D) compared to their query (k71) and reference (corNpor) assemblies as well as in gap-filled, PBjelly-based hybrid assembly (filk71) compared to k71 assembly with unfilled

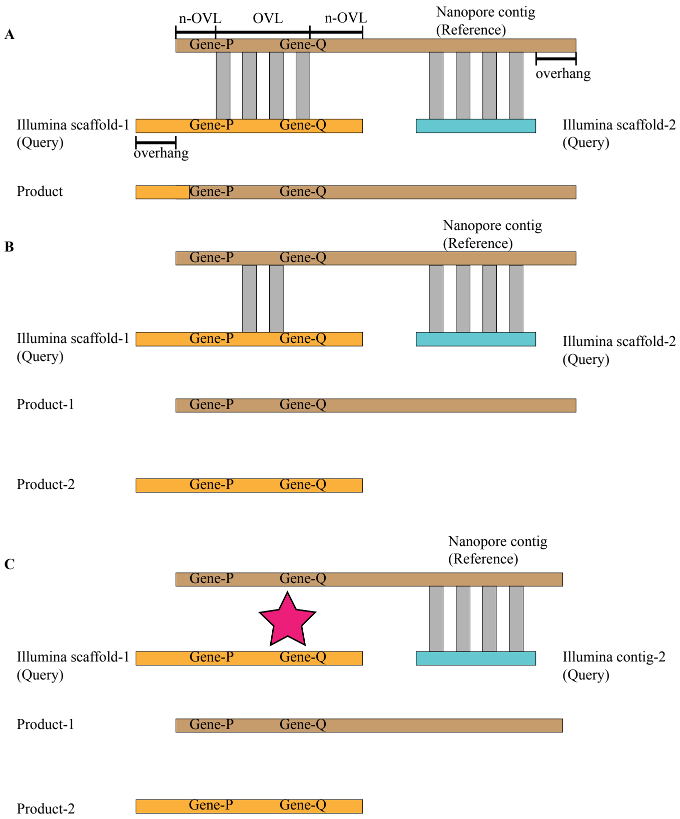


Figure 2.7 Duplication of BUSCO genes via the Quickmerge algorithm. **A**) Successful alignment and merging of one nanopore and two Illumina scaffolds without generating duplicated BUSCO genes. When a query contig (e.g., Illumina scaffold-1) is aligned to a reference contig (Nanopore contig), the alignment has three components: overhang, overlapped but unaligned (n-OVL), and overlapped and aligned (OVL) sequence (Vertical grey bars represent aligned regions). Quickmerge uses the ratio of OVL/n-OVL to determine if the query and reference contigs should be merged (product). Note if overhangs are present in the reference and/or the query, they are retained in the product. **B**) One nanopore contig aligned to two Illumina scaffolds in which Illumina scaffold-2 merged with the nanopore contig; however, Illumina scaffold-1 failed to merge. Consequently, two products were produced with the same BUSCO genes. **C**) Alignment between Illumina scaffold-1 failed with nanopore contig. Consequently, two products are produced with the same BUSCO gene set. Star sign indicates alignment failure.

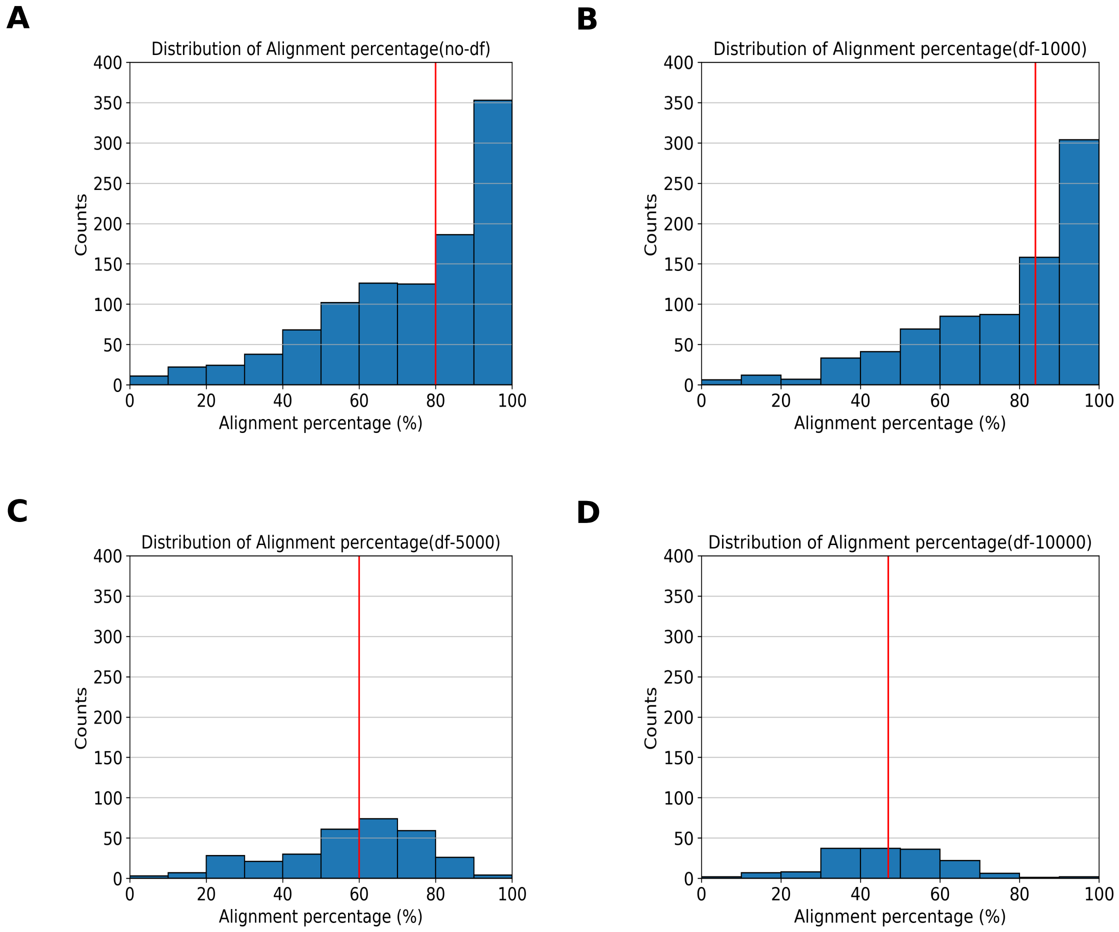


Figure 2.8 Distribution of alignments generated with different parameter settings when merging assemblies to create a phase II hybrid assembly. The plot shows the distribution of alignment percentage for different minimum length of alignment (L) parameters employed in the Mummer delta-filter (df) program when setting a 95% minimum alignment identity. For A) L equals to 0, B) L equals to 1000, C) L = 5000, and D) L equals to 10000. X-axis represents alignment percentage whereas y-axis represents counts of those alignment percentages. Red line represents median of the alignment percentage. The number of high percentage alignments decreased with an increase in the stringency of minimum length of the alignment (shown in figure A to D

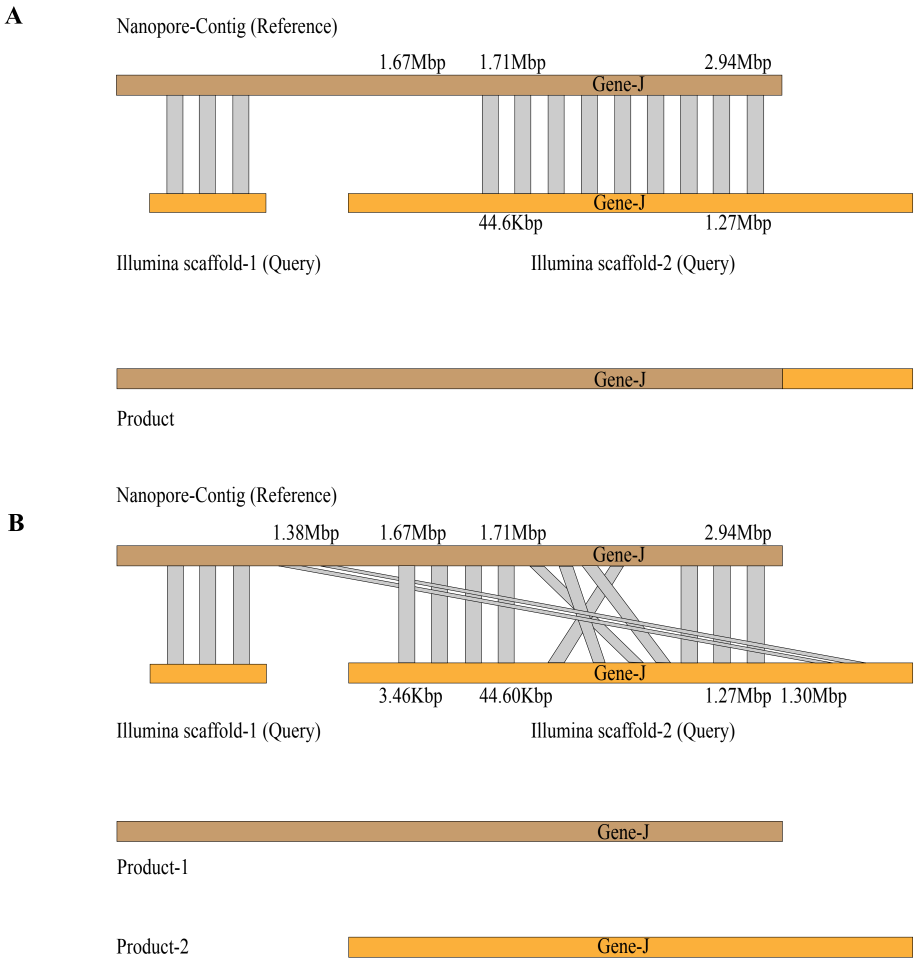

A
Nanopore-Contig (Reference)
1.67Mbp
1.71Mbp
2.94Mbp
Gene-J
Gene-J
44.6Kbp
1.27Mbp
Illumina scaffold-1 (Query)
Illumina scaffold-2 (Query)
Gene-J
Product
Nanopore-Contig (Reference)
B
1.38Mbp
1.67Mbp
1.71Mbp
2.94Mbp
Gene-J
Gene-J
3.46Kbp
44.60Kbp
1.27Mbp
1.30Mbp
Illumina scaffold-1 (Query)
Illumina scaffold-2 (Query)
Gene-J
Product-1
Product-2
Gene-J

Figure 2.9 The disruption of the linear order of nucmer alignments between the query (Illumina scaffold) and reference (Nanopore) contigs, as implemented by Quickmerge (using the set of Mummer alignment tools), resulting in duplicated BUSCO gene (e.g. Gene-J). A) Successful, linearly ordered alignment and merged product. B) The disrupted order of alignments between query and reference contigs due to small and spurious alignments leading to merging failure and a duplicated BUSCO ge

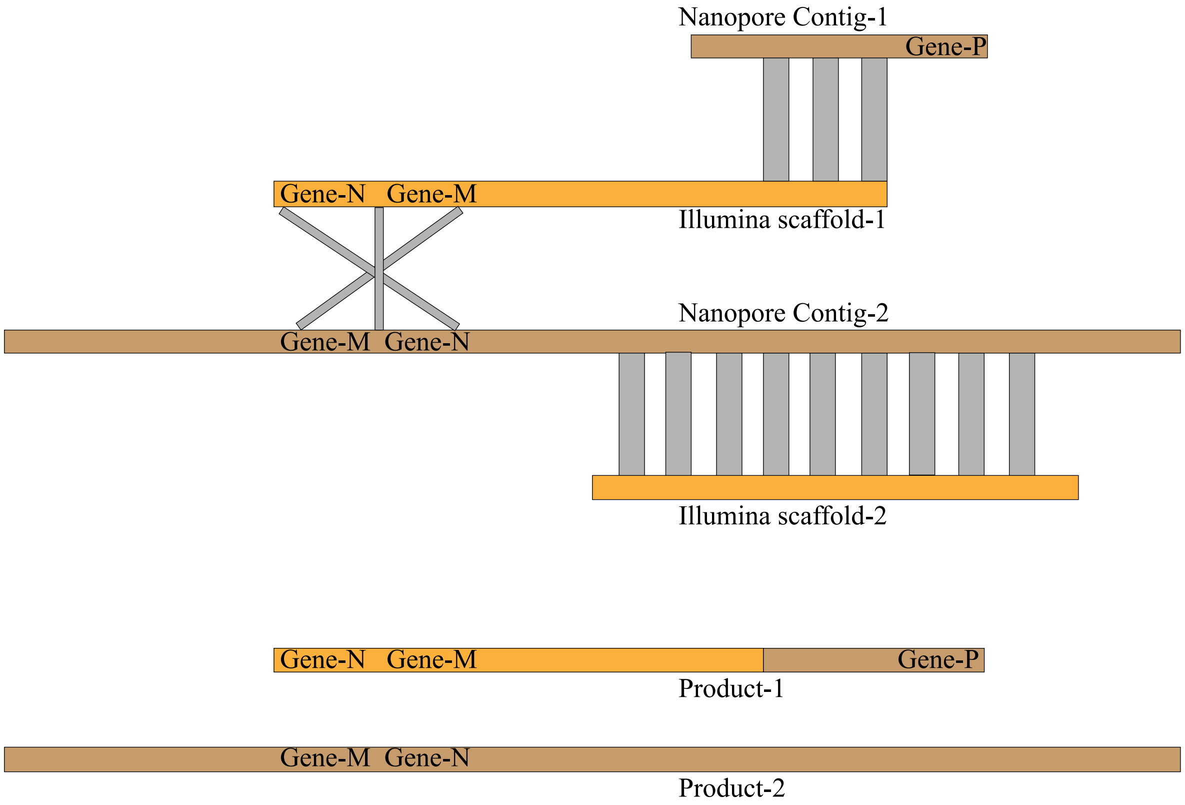


Figure 2.10 Duplication of BUSCO genes (e.g. Gene-M, Gene-N) in the merged, phase II assembly due to the effect of mis-joined contigs in Illumina scaffold-1 (composed of portions of different Nanopore contigs). A) Illumina scaffold-1 composed of distantly related contigs aligned and merged to at-least two Nanopore contigs-1 and -2. B) The merged assembly contained duplicated BUSCO genes.

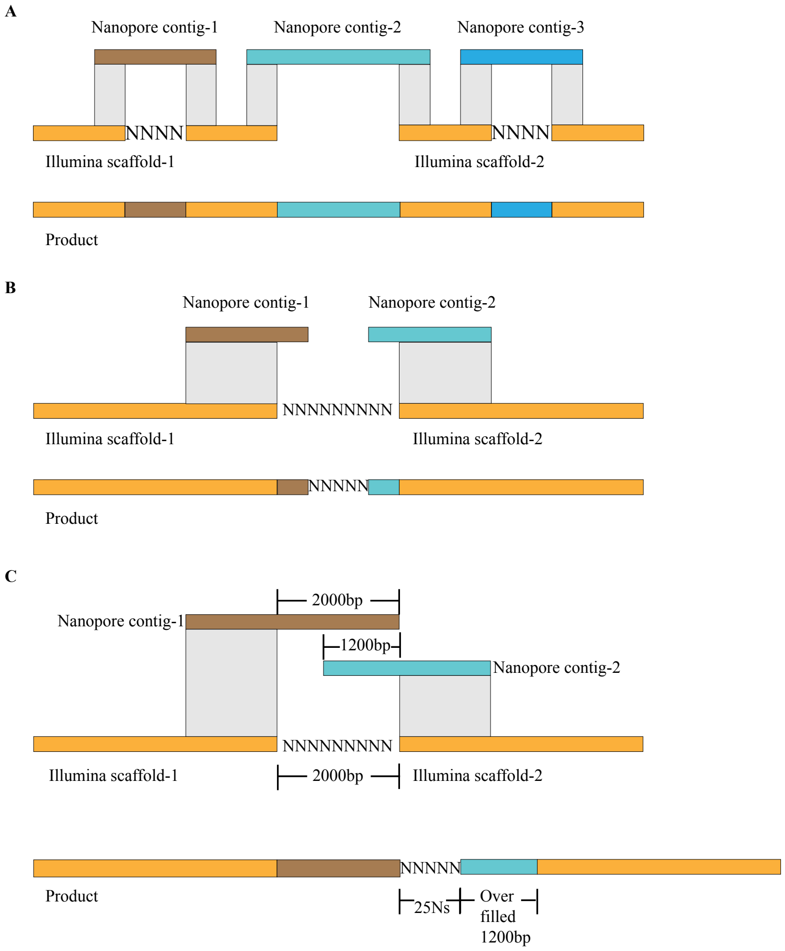


Figure 2.11 The gap-filling process of PBJELLY. A) When Nanopore long-reads span across gaps (Ns), the gap is filled. B) When a long-read extends into the gap, the gap is partially filled. C) When distinct nanopore long-reads extend into the gap from either side, but do not align with one another, the gap is extended according to the lengths of the individual reads, potentially overfilling the gap (and an additional gap of 25 Ns is added by PBJELLY). For example, if the total length of a gap is 2000bp prior to merging, and nanopore reads extend into the gap 2000bp on one side and 1200bp on the other, then the total gap is extended to 3200bp (plus 25bp of Ns)

A

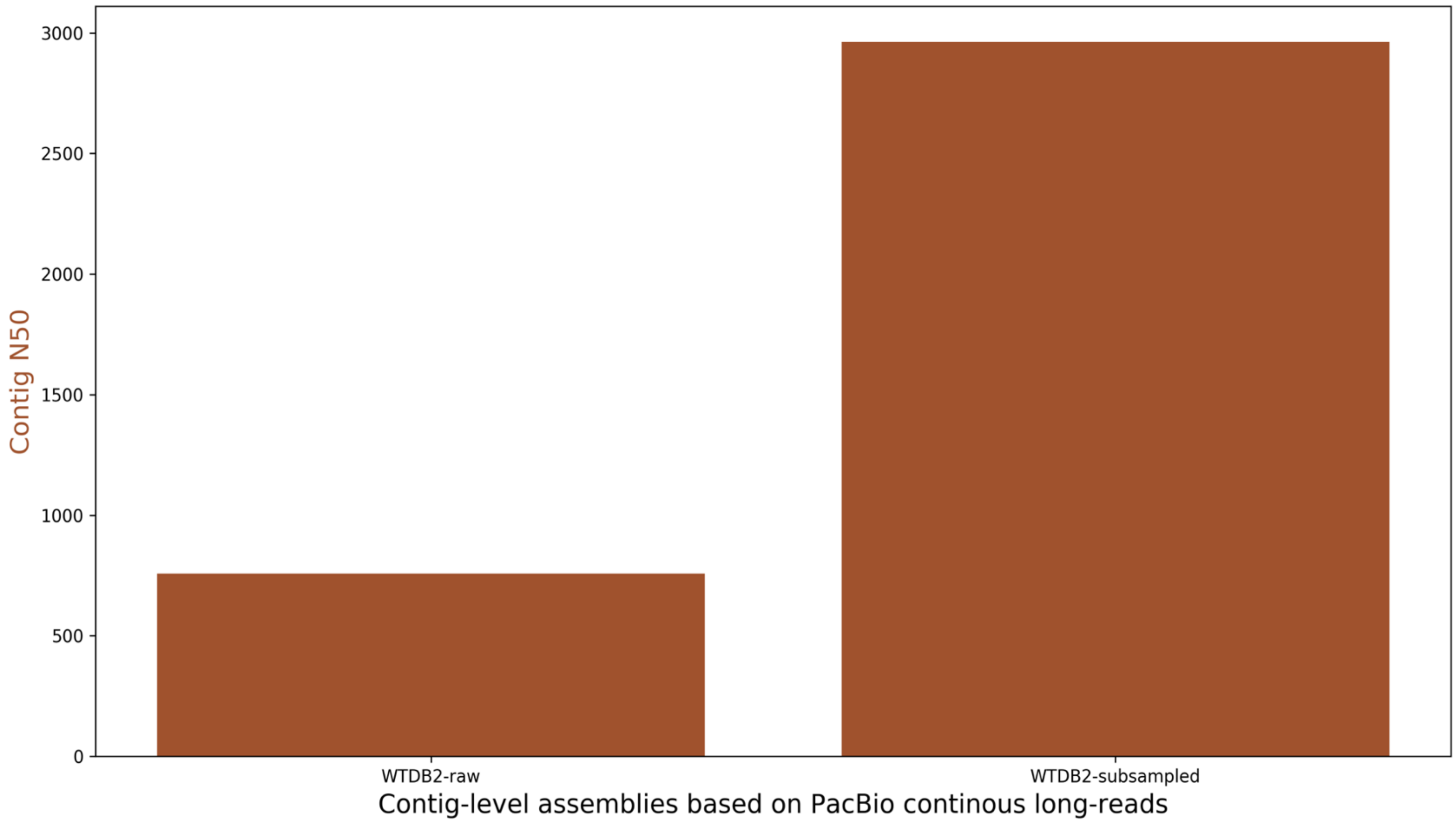


Figure 2.12 Subsampling of PacBio contiguous long-reads can increase contiguity for contig-level assembly and such assemblies' BUSCO gene completeness can be increased by polishing through self-error correction protocol. A) This figure shows that assembly WTDBG2-subsampled assembly (WTDBG2^{Sr*} in Table 1) built by subsampling PacBio reads has high contiguity metric N50 than WTDBG2-raw (WTDBG2^{r*} in Table 1)
built with raw reads.

B

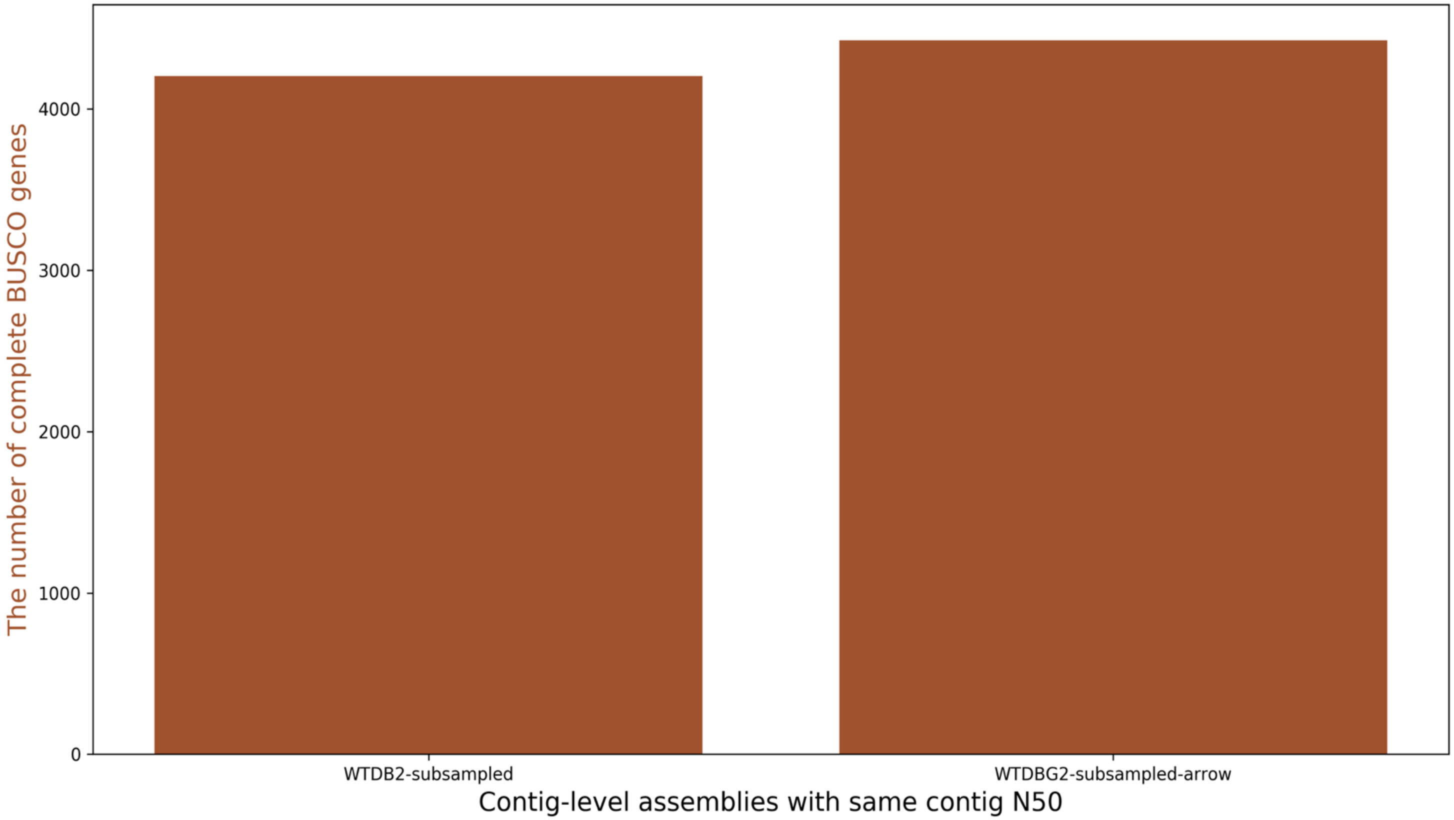


Figure 2.12 – Continued. B) This figure shows that error corrected, subsampled assembly WTDBG2-subsampled-arrow (WTDBG2^{Sra} in Table 1) has more BUSCO gene completeness than uncorrected, WTDBG2-subsample

CHAPTER 3: GENOMIC INSIGHTS INTO SECONDARILY TEMPERATE ADAPTATIONS OF NEW ZEALAND'S BLACK COD, *PARANOTOTHENIA ANGUSTATA* (MAORI CHIEF)

ABSTRACT

Most species within the Antarctic clade of notothenioid fish are endemic to Antarctica, cold-specialized, and stenothermal (e.g., *Trematomus borchgrevinki*). However, a few have secondarily adapted to temperate conditions, including *Paranotothenia angustata*. The specific genetic changes underlying the adaptations of this notothenioid have remained largely unknown. To shed light on the genetic adaptation of secondarily temperate notothenioids, I generated high quality chromosome-level assemblies, annotations, and Restriction site-Associated DNA Sequencing (RADseq)-based population-level, genetic variation data on *P. angustata* (secondarily temperate) and *T. borchgrevinki* (Antarctic). I focused on genetic changes specific to *P. angustata* and used the related *T. borchgrevinki* as an outgroup. I found high repeat content in both species and lineage-specific expansion of DNA transposons in *P. angustata*. I observed evidence of chromosomal rearrangements such as fusions, inversions, and translocations potentially specific to *P. angustata*. I found that the orientations of chromosomes that formed the fusions are predominantly unique to *P. angustat*a. I identified inversions with one to three genes exhibiting a significant non-synonymous to synonymous nucleotide substitution ratio, indicating directional selection. Genes related to protein chaperoning, circadian rhythm, vision, erythrocyte differentiation and development, heme metabolism, vision, mitochondria, and ribosomes appear to be under positive selection in *P. angustata*. Overall, my results provide insight into genomic adaptations that may have enabled the ancestor of *P. angustata* to adapt to more temperate environments.

INTRODUCTION

Notothenioids are a group of teleost fish that evolved about 47 million years ago (MYA) (Bista *et al.* 2023). Their evolutionary history reflects transitions between temperate and cold Antarctic environments. Currently, Antarctica is characterized by a polar environment and is encircled by the Southern Ocean. However, historically, Antarctica had a temperate climate (Zachos *et al.* 2001; Klages *et al.* 2020) and was part of the Gondwana supercontinent (reviewed in Faure and Mensing 2010). Over millions of years, temperate Antarctica progressively separated from other land masses due to continental drift and tectonic forces (Storey and Granot 2021). It shifted to its current location at the south polar position and became surrounded by marine waters, forming the Southern Ocean. With the onset of the Antarctic Circumpolar Current (ACC), the northern boundary of the Southern Ocean segregated into temperate and Antarctic water masses. As a result, the temperate notothenioid stocks divided into non-Antarctic and Antarctic components (reviewed in Eastman 1993). Glaciation, induced by the establishment of the ACC, and reduced carbon dioxide concentration in the atmosphere contributed to the cooling of Antarctica (Kennett 1977; Clarke *et al.* 2004). As temperatures decreased, most of the temperate fauna from the Southern Ocean disappeared (Daane and Detrich 2022).

Persistence of an ancestral stock of originally temperate notothenioids in these low-temperature waters was facilitated by the presence of Anti-Freeze Glycoproteins (AFGPs), which originated between 10.7 and 26.3 MYA (Bista *et al.* 2023). The AFGPs prevent ice growth in Antarctic fish (DeVries 1971). Around 10 MYA, the cold-adapted Antarctic notothenioid lineage diversified (Bista *et al.* 2023), likely because of vacated ecological niches. Over time, most of these derived species became cold-specialized and endemic to Antarctica (e.g., *Trematomus borchgrevinki*) (Eastman 1993). However, a few lineages escaped Antarctica and re-adapted to warmer waters of temperate regions (Coppes Petricorena and Somero 2007; Daane and Detrich 2022). Today, notothenioids consist of non-Antarctic and Antarctic clades. The non-Antarctic clade consists of three families (Bovichtidae, Pseudaphritidae, and Eleginopidae), which have never experienced freezing temperatures (Patarnello *et al.* 2011); the Eleginopidae family has a single species, *Eleginops maclovinus,* which is sister to the Antarctic clade. The Antarctic clade itself consists of five families (Nototheniidae, Harpagiferidae, Artedidraconidae, Bathydraconidae, and Channicthyidae) (Near *et al.* 2004), includes both cold-specialized and secondarily temperate members, and is more speciose than the non-Antarctic clade due to its adaptive radiation within the frigid Southern Ocean (Eastman 2013; Beers and Jayasundara 2015). Most of the secondarily temperate species belong to Nototheniidae, the most speciose family (Eastman and Eakin 2021). While the evolution of AFGPs enabled notothenioids to adapt to chronically cold environments (DeVries 1988), the underlying genetic adaptations of secondarily temperate notothenioids remained understudied.

Here, I focus on secondarily temperate notothenioid, *Paranotothenia angustata* (family Nototheniidae; commonly known as New Zealand black cod or Maori Chief), which is endemic to the coastal waters of Southern New Zealand, with a temperature of 6-18 degrees centigrade (Lau *et al.* 2001). It diverged from the Antarctic nototheniod lineage about 11 million years ago (Cheng 2003), and some of its traits reflect its polar ancestry. For example, *P. angustata* still has a few small AFGP coding genes and can produce minuscule levels of protein (Cheng 2003). Additionally, the number of hemoglobin isoforms and their structure and function in *P. angustata* is highly similar to that of closely related cold-specialized Antarctic notothenioid, *Notothenia coriiceps* (Fago *et al.* 1992). Moreover, the levels of ubiquitin-conjugated proteins, saturated lipid in brain cellular membranes, and heat tolerance capacity in *P. angustata* are intermediate between cold-specialized notothenioids and basal or tropical fish families (Logue *et al.* 2000; Todgham *et al.* 2007; Bilyk and Devries 2012). *P. angustata* is a diploid species with 26 pairs of chromosomes. Most of its chromosomes are meta- and submeta-centric (Pisano *et al.* 2003). This karyotype of *P. angustata* differs from that of primarily temperate notothenioids. For example, *E. maclovinus* has 48 predominantly telocentric diploid chromosomes (Mazzei *et al.* 2008). The cold-adapted *T. borchgrevinki* (bald notothen) is cryopelagic, inhabiting the spaces between ice platelets beneath the surface of fast ice in Antarctica. *T. borchgrevinki* exhibits a circum-Antarctic distribution (Eastman and DeVries 1985). It suffers from heat stress at approximately 6 degrees Celsius above its usual ambient temperature (Somero and DeVries 1967) and is susceptible to oxidative damage at higher temperatures (Carney Almroth *et al.* 2015). *T. borchgrevinki* exhibits a sex-specific diploid chromosome number, with males possessing 45 chromosomes and females having 46 chromosomes, most of which are acrocentric (Morescalchi *et al.* 1992).

To gain insights into the secondary temperate adaptations in *P. angustata,* I conducted genome sequencing of both *P. angustata* (as a focal species) and *T. borchgrevinki* (as an outgroup representing cold-specialized notothenioids). For both species, continuous long-reads (CLRs) were generated from the Pacific Biosciences (PacBio) Sequel II platform. I produced high quality *de novo* chromosome-level assemblies by scaffolding long-read contigs with chromosome conformational capture data (Hi-C reads) while also manually correcting errors within these assemblies. Additionally, I conducted assembly annotation and characterized unique repeat content patterns specific to *P. angustata.* Using conserved synteny and gene neighborhoods, we delineated the chromosomal fusions and structural variations, including inversions and translocations, that are particular to *P. angustata.* Subsequently, I performed a genome scan based on differences in nucleotide diversity (π), as well as differentiation (F_{ST}) and divergence (D_{XY}) between the two species. Moreover, we explored linkage disequilibrium based on cross-population extended haplotype homozygosity (XP-EHH) using *P. angustata* as a target and *T. borchgrevinki* as a reference in search of signatures of positive selection specific to *P. angustata.* To pinpoint protein-coding genes subjected to positive selection, I estimated the lineage-specific ratio of non-synonymous changes (dN) to synonymous changes (dS) within *P. angustata* using both branch-site and branch models.

I found distinct differences in the genome structure of *P. angustata* compared to *T. borchgrevinki*. These disparities are primarily attributed to the expansion of DNA transposons and a series of chromosomal rearrangements, including fusions, inversions, and translocations. Furthermore, I identified chromosomal rearrangements such as fusions, inversions, and translocations potentially specific to *P. angustata*. The chromosomes' orientation in these fusions appears to be predominantly unique to *P. angustata.* In the case of inversions, one to three genes within these regions exhibited a significant dN/dS ratio. Based on my findings from π, D_{XY}, XP-EHH, and dN/dS ratios, I propose that the genes under selection, particularly associated with protein chaperoning, circadian rhythm, vision, erythrocyte differentiation and development, and heme metabolism, as well as mitochondria and ribosomes, likely play a pivotal role in the adaptations of *P. angustata* to temperate environment.

MATERIALS AND METHODS

Specimen collection and generation of long-read-based genome sequences, as well as Hi-C library preparation and sequencing

For *Trematomus borchgrevinki*, two populations were sampled (one from McMurdo Sound (West Antarctica) and another from Prydz Bay (East Antarctica)), located on the opposite side of Antarctica. Seventy-one individuals were collected (specifically, 53 from McMurdo Station and 18 from Prydz Bay). I retrieved the raw CLRs for the female individual from a prior study (Rayamajhi *et al.* 2022).

For *Paranotothenia angustata*, 41 individuals were collected from one population in Otago Harbor, South Island, New Zealand. High molecular weight (HMW) genomic DNAs were extracted from only one individual using an in-house protocol. The HMW gDNAs were used to construct libraries for PacBio Sequel II-based long-read sequencing. Those libraries were sequenced using two single-molecule real-time sequencing (SMRT) cells of the PacBio Sequel II platform and generated consensus long-reads (CLRs). The library construction and sequencing were conducted at the Genomics and Cell Characterization Core Facility, University of Oregon (**Figure 3.1.A**; step 1).

Moreover, the Hi-C library was constructed for each species using a single individual for which PacBio-based contig-level *de novo* assembly was built in this study. Phase Genomics Inc. generated each library with the commercialized scaffolding kit Proximo Hi-C. The restriction nuclease DpnII was used for chromatin fragmentation. The Hi-C library was quantified by qPCR and then sequenced on the NovaSeq6000 machine, an Illumina platform, to generate 2x150bp paired-end reads so that I could utilize long-range information to scaffold contig-level genome assemblies (**Figure 3.1.A**; step 2).

RADseq library preparation and sequencing

I generated the RAD library for each species and sequenced it to genotype all the sampled individuals of *P. angustata* and *T. borchgrevinki* at randomly sampled genomic regions (**Figure 3.1.A**; step 3). To accomplish RAD library preparation and sequencing, I extracted HWM gDNAs from ethanol-preserved muscle tissues of sampled individuals. I used the standard

GuSCN for *T. borchgrevinki* and phenol/chloroform protocols for *P. angustata* to extract HWM gDNAs. The quality and concentration of the DNAs were measured using a Qubit fluorometer (Thermo Fisher Scientific, USA), and their bands were visualized in 1% agarose gel. The RAD libraries were constructed from the high quality HMW DNAs, following the published protocol (Baird *et al.* 2008; Etter *et al.* 2011) with some modifications. In each sample, 1µg of gDNA was digested with the single restriction enzyme SbfI (8-base cutter) in 50 µl of reaction volume. The digestion reaction was carried out at 37ºC for 90 mins, followed by incubation at 80ºC for 20 mins to kill SbfI-HF. The reaction volume included 30µl solution with 33.3 ng/µl gDNAs, 1µl of diluted enzyme solution containing one part of SbfI-HF enzyme and seven parts of dilutant B, 5µl of 10x NEB cut-smart buffer and 14 µl of RNAase-free sterile water.

Next, unique 7-base-pair-barcode-labeled P1 adapters were ligated onto the genomic fragments generated from each digestion reaction with the recommended protocol. However, ligation reaction time was extended to 1 hour at room temperature, followed by overnight incubation at 4ºC to inactivate the exonuclease activity of the ligase. I pooled the P1-adapter-ligated fragments, each containing a unique barcode for individual identification. This multiplexing was repeated four times to create four replicates, and each replicate was processed separately. The pooled fragments were sheared independently for each replicate using a Covaris M220-focused ultrasonicator (Woburn, MA). Subsequently, the sheared fragments were subjected to size selection to recover the pieces within the length range of 300-600bp. For *T. borchgrevinki*, the size selection step was performed using an agarose gel-based method. However, for *P. angustata*, the size selection process was conducted using the AMPure XP magnetic beads (Beckman Coulter) method.

Sequentially, the size-selected fragments were repaired at the ends and A-tailed by adding a 3'-dA overhang. Subsequently, the P2 adapters were ligated to form the genomic RAD library. For *T. borchgrevinki*, approximately 100 ng of P2-ligated genomic RAD fragments per replicate were combined to generate 400 ng of DNA templates. For *P. angustata*, 100 ng of the fragments from two replicates and 150 ng from another two replicates were pooled together to produce a total of 500 ng of DNA templates. Both 400 and 500 ng of the pooled DNA templates were separately enriched in a 100 µl PCR reaction volume with 12 PCR cycles. Post-PCR cleanup was performed using 0.85x AMPure XP magnetic beads to obtain the final library, and its DNA concentration was estimated using Qubit. The library was then sent to the Roy J. Carver Biotechnology Center, the University of Illinois Urbana-Champaign, USA, for sequencing on an Illumina NovaSeq600 SP sequencer to generate 2x150 paired-end reads.

Generation of de novo contig- and chromosome-level genome assemblies

For *T. borchgrevinki*, two different strategies were used to create two separate *de novo* contig-level assemblies (**Figure 3.1.A**; step 4). First, raw PacBio CLRs were aligned to each other using minimap2 (v2.1; Li 2018) with an all-versus-all approach (using PacBio preset `ava-pb` and mapping option `-g 5000` to set maximum distance between seeds to generate overlap). I used Filter Pairwise Alignment software (fpa; v0.5.1; Marijon *et al.* 2020) with subcommand `drop` to filter alignments if a) the overlaps had length less than 2000 (`--length-lower 2000`) and b) they were formed by an internal match between reads (i.e., all the nucleotides in one read is contained in another read) (`--internalmatch`). Next, I used Yet Another Chimeric Read Detector for long-reads (yacrd; v0.6.2; Marijon *et al.* 2020) on alignment data from fpa to detect chimeric reads. Using the subcommand `filter,` I removed reads detected as chimeric and those having regions with coverage equal to or less than 3 (`--coverage 3`), accounting for 40% or greater of total length (`--not-coverage 0.4`). I assembled these filtered reads separately using Flye (v2.6; Kolmogorov *et al.* 2019) and WTDBG2 (v2.5; Ruan and Li 2020) algorithms, resulting in two independent *de novo* contig-level genome assemblies.

In a second strategy, I retrieved subsampled PacBio CLRs with read lengths ranging from a minimum of 10 Kb and a maximum of 40 Kb, totaling 70 Gb of data, used in Rayamajhi *et al*. 2022 to assemble the contig-level assembly for female *T. borchgrevinki*. I also obtained the pre-existed contig-level WTDBG2 assembly built using the same subsampled data and corrected with arrow module in GCpp (v2.0.0; Pacific Biosciences) (Rayamajhi *et al*. 2022). Moreover, the same subsampled raw reads were assembled with Flye.

Next, I estimated contiguity statistics for the Flye and WTDBG2-based assemblies from each of the two strategies using QUAST (v4.6.2; Gurevich *et al.* 2013) and compared them. Based on contiguity metrics, I retained Flye- and WTDBG2-based assembly obtained from the first and second strategy, respectively. However, I considered Flye- and WTDBG2-based assemblies primary and secondary contig-level assemblies, respectively. That's because the genome statistics for Flye-based assemblies were very similar to those for WTDBG2-based assemblies.

For *P. angustata,* I only employed a subsampling strategy (**Figure 3.1.A**; step 4) as Yacrd required ample disk space and a long time to process the large volume of long-read sequence data. I subsampled for a minimum of 15 Kb, a maximum of 40 Kb long-read length, and a total of ~80G size data. The subsampled reads were assembled with Flye and WTDBG2 assemblers separately (**Figure 3.1.A**; step 5). All the assemblies from both species were subjected to QUAST (v4.6.2; Gurevich *et al.* 2013) to estimate contiguity metrics. Since the contiguity of Flye-based assembly was higher than that of WTDBG2-based, I considered the former primary and the latter a secondary contig-level assembly (**Figure 3.1.A**; step 6). The WTDBG2-based assembly was polished with one round of arrow (**Figure 3.1.A**; step 7).

To generate chromosome-scale genome models or assemblies for two species, I aligned Hi-C reads from each species to their corresponding primary as well as secondary contig-level assemblies and generated lists of Hi-C contacts using Juicer (v1.6.2; Durand *et al.* 2016) (**Figure 3.1.A**; step 8). Each list of Hi-C contacts and its corresponding contig-level assembly were fed to Juicer's 3d-DNA pipeline for ordering, orienting, and joining the contigs to produce chromosome-level super-scaffolds. Moreover, for each assembly from 3d-DNA, the information on structural constituents of chromosomes (i.e., description of contigs or scaffolds organized in chromosomes) was stored in AGP file format using a custom Python script.

The chromosome-scale genome model derived from the scaffolding contigs in the primary contig-level assembly was labeled as a primary assembly for each species. In contrast, the one built from the secondary contig-level assembly was referred to as secondary assembly. Subsequently, the primary assemblies underwent QUAST analysis. Additionally, they were assessed for the completeness of 3,640 Actinopterygii-specific single-copy orthologs using Benchmarking Universal Single-Copy Ortholog (BUSCO) (v5.1.3; Simão *et al.* 2015) software with default parameters. BUSCO classifies orthologs into a) single copy and complete, b) complete but duplicated, c) fragmented, or d) missing categories.

Manual curation and annotation of de novo chromosomal-level assemblies

Annotation repeats and genes were conducted in primary and secondary assemblies per species (**Figure 3.1.A**; steps 9-10). For repeat annotation, a *de novo* custom repeat library was generated from the assembly of interest using RepeatModeler (v2.02a; Flynn *et al.* 2020). The known repeat library for teleost was obtained from Repbase (Bao *et al.* 2015) and combined with the *de novo* repeat library. This pooled library was used to identify and soft mask repetitive elements in the assembly with RepeatMasker (v4.1.2-p1; Smit and Hubley 2013). Moreover, I retrieved the chromosome-level assembly of *Notothenia rossii* (Clawson *et al.* 2023) and re-annotated the repeats for comparison (**Figure 3.1.A**; step 10). This is because the haploid number of chromosomes for *N. rossii* (12) and *P. angustata* (13) (reviewed in Amores *et al.* 2017) are very similar. These two species are more closely related to each other than they are to *T. borchgrevinki* or any other species from different genera of notothenioids (reviewed in Amores *et al.* 2017).

RNAseq reads were retrieved from previously published studies for gene annotation of *P. angustata* and *T. borchgrevinki* (**Figure 3.1.A**; step 11). The RNAseq reads for *T. borchgrevinki* were obtained from the same species (Bilyk and Cheng 2014), while those for *P. angustata* were obtained from the closely related Antarctic notothenioid species, *Notothenia coriiceps* (Shin *et al.* 2014). These RNAseq reads were mapped to the masked assembly using STAR (Spliced Transcripts Alignment to a Reference) (v2.7.1.a; Dobin *et al.* 2013). Additionally, RNAseq alignments and zebrafish proteins (obtained from OrthoDB (v10.1; Kriventseva *et al.* 2019)) were independently employed with the masked assembly to run BRAKER2 (Brůna *et al.* 2021) pipeline. The gene predictions from two BRAKER2 runs were processed using TSEBRA (Gabriel *et al.* 2021) to retain only gene annotations supported by both proteins and transcripts. The curated genes were annotated for their functions using InterProscan (Quevillon *et al.* 2005). The names of genes obtained from the functional annotation analysis were retained.

Utilizing the data in the genome annotation and the AGP files, conserved synteny analysis (described below in another section) was conducted between the primary and the secondary assemblies of a species of interest. This analysis was integral to the manual curation process for the primary assemblies (**Figure 3.1.A**; step 12). The secondary assemblies were used for comparison purposes in the curation process. Specifically, I searched for discrepancies in the genomic structures between the primary and the secondary assemblies for a species of interest. For example, I looked for contigs or scaffolds inverted or translocated in the primary but not in the secondary assembly. In the primary assembly, the orientation or location of the contigs or scaffolds was appropriately changed when the structure in the secondary assembly was supported by evidence, for instance, the boundaries of contigs or scaffolds. The error corrections were performed at various stages of the assembly process using a custom Python script with FASTA format sequences, genome annotation (Gene Transfer Format (GFF)) files, and AGP files.

Following the final curation process, the curated primary chromosome-level assemblies were subjected to QUAST and BUSCO analyses and genome annotation with slight modification (**Figure 3.1.A**; step 13). The same pipeline was followed as described above for the repeat annotation on final assemblies. Since the repeat content of *P. angustata* (notably the proportion of DNA transposons) was higher than that of *T. borchgrevinki* and *N. rossii*, I considered comparing it to that of other notothenioids to assess if an increase in DNA transposons is specific to *P. angustata*. For this purpose, I obtained the previously reported data on repeats in *Eleginops maclovinus* (non-Antarctic notothenioid fish; Cheng *et al.* 2023) and *Champsocephalus gunnari* (Antarctic notothenioid fish; Rivera-Colón *et al.* 2023) as the data were based on the same annotation method as mine. *E. maclovinus* is more ancestral, whereas *C. gunnari* is more derived notothenioids than *P. angustata* and *T. borchgrevinki*.

Finally, the gene annotation pipeline was conducted on the curated final primary assemblies with slide modification. InterProscan was removed from the pipeline due to the extended processing time. Instead, Synolog was used to identify gene homology between *P.*

angustata and zebrafish as well as between *T. borchgrevinki* and zebrafish. Synolog was fed with annotations of zebrafish (from the Ensembl database) as well as of *P. angustata* and *T. borchgrevinki*. The names of genes in the assemblies were assigned based on identified gene homology between the sequenced species in this study and the zebrafish. This name assignment process was conducted using a custom Python script (**Figure 3.1.A**; step 13).

Conserved synteny analysis for identifying and characterizing structural variations

The following steps were undertaken to detect structural variations (such as fusions, inversion, and translocations) specific to *P. angustata*. First, I retrieved annotated coding sequences from the primary chromosome-level assemblies for *P. angustata* and *T. borchgrevinki*, along with those from the previously published assemblies for *E. maclovinus*, *C. gunnari*, and *Notothenia coriiceps* (Shin *et al.* 2014) (**Figure 3.1.B**; steps 14-15). Subsequently, the coding sequences from one assembly were blasted against those from the other assemblies, independently, using the blastp program within BLAST+ (v2.4; Camacho *et al.* 2009) (**Figure 3.1.B**; steps 14-15) for all possible combinations of disparate assemblies, to obtain Reciprocal Best Hits (RBHs). Next, these RBHs, along with genome annotations (in Gene Transfer Format (GTF) or General Feature Format (GFF)) and AGP files, were used as input in Synolog (unpublished version of the Synteny Database (Catchen *et al.* 2009)). Finally, I tracked down and visualized conserved synteny blocks (i.e., orthologous chromosome regions that show considerable similarity in sequence and order of genes) among the assemblies using Synolog (**Figure 3.1.B**; step 16).

Specifically, first, I tracked down conserved syntenic gene neighborhoods among *P. angustata*, *T. borchgrevinki*, *E. maclovinus*, and *C. gunnari* to identify and characterize chromosomal rearrangements specific to *P. angustata* (**Figure 3.1.B**; step 17). Next, given that

P. angustata had drastically reduced chromosome number compared to the rest of other species due to the presence of chromosomal fusions, I compared its chromosomes with those of *N. coriiceps* using *E. maclovinus* (which is a single species in a clade sister to Antarctic notothenioid clade). This is because the haploid chromosome numbers of *P. angustata* (13) and *N. coriiceps* (11) are highly similar (reviewed in Amores *et al.* 2017), and the comparison between these two species would be useful to gain insights into how the ancestral chromosomes oriented to form fusions in one versus another species. I added *N. coriiceps* even though its genome was of lower quality because the genome was built with the aid of a genetic map, which would provide reliable information on the orientations of the chromosomes.

Identification of putative signatures of positive selection based on diversity, differentiation, divergence, and linkage analyses

To find positively selected genomic regions in *P. angustata*, first, I performed a RADseq-based genome scan that provides the patterns of within-species diversity (i.e., π) and between-species absolute divergence (D_{XY}) and differentiation (i.e., F_{ST}). I obtained Illumina-based raw paired-end reads produced by sequencing of RADseq libraries generated in this study for the genome scan analyses. I processed and analyzed RADseq data using three modules of Stacks (v2.60; Rochette *et al.* 2019) (**Figure 3.1.B**; step 18): process_radtags, gstacks, and populations. The reads from each species were demultiplexed, cleaned (for retaining reads without uncalled base), and filtered (for keeping sequences with high quality phred scores), as well as their cut site and barcodes were rescued with process_radtags. Next, the retained demultiplexed reads were aligned to the genome of *P. angustata* using bwa-mem (v0.7.17; Li 2013), and the alignments were sorted with Samtools (v1.12; Li *et al.* 2009). Moreover, after removing PCR duplicates, I executed the gstacks module on aligned sequences to build RAD loci and genotype SNPs in each individual with at least 10x effective coverage. I employed the populations module to retain loci present at least in 50% of samples in each species (i.e., a minimum 18 of 36 individuals of *P. angustata* and 32 of 65 individuals of *T. borchgrevinki*) as well as variants with a minimum of 3 allele counts. I also estimated population genetic parameters such as π, FST, and D_{XY} using the populations module. I estimated D_{xy} because, unlike F_{st}, its pattern is unaffected by any process that alters within-species π. For π and F_{ST} metrics, I relied on single nucleotide variation data, whereas the D_{XY} calculations were based on RAD haplotypes.

In addition, I instructed the populations module to kernel-smooth the estimates of π, $F_{ST,}$ and D_{XY} using a sliding window of size 900 kilobase pairs (Kpb) and export the output as sorted VCF. Using the bash command line, I subtracted the kernel smoothed π estimate for *P. angustata* from that for *T. borchgrevinki* at common sites with no missing data from either species. Such estimated difference in smoothed π at each site between the two species was referred to as delta π. This approach can capture the signature of environment-specific positive selection (Liu *et al.* 2022; Montejo-Kovacevich *et al.* 2022). For downstream analysis, I considered the bottom 0.5th percentile of the empirical distribution of delta π as outliers for stringency. Each window of size 900 Kbp for a given site with outlier delta π was considered an outlier window. For smoothed D_{xy} and F_{ST} estimates, I obtained p-values by bootstrapping windows 1000 times. Since F_{ST} was too high across the genome, I only used D_{XY} for downstream analysis. I considered the kernel-smoothed D_{XY} windows of RAD-haplotypes with p-value < $1x10^{-2}$ instead of $5x10^{-2}$ as outliers for stringency.

To assess if the drastic reduction in local genomic diversity (represented by delta π outlier) or significant increase in divergence was due to positive selection, I conducted a cross-population extended haplotype-homozygosity (XP-EHH) analysis, a haplotype-based linkage analysis that is sensitive to selected alleles near fixation or that is already fixed. For the XP-EHH analysis, I split single nucleotide variants (stored in the sorted VCF generated from previous analysis) by species using BCFtools (v1.12; Li 2011). With VCFTools (v0.1.15; Danecek *et al.* 2011), I divided the species-specific variant data by chromosome and retained genomic positions in which genotypes were called for at least 80% of the samples (max-missing 0.8), i.e., a minimum of 14 from *P. angustata* and 26 individuals from *T. borchgrevinki*. Moreover, the variants were phased chromosome-wise with Beagle (v5.4; Browning *et al.* 2021) (**Figure 3.1.B**; step 19). For each species, the phased variants per chromosome were concatenated using BCFtools.

Next, each species' concatenated, phased variant data was analyzed with the R package, rehh (v3.2.1; Gautier and Vitalis 2012) to perform XP-EHH analysis (**Figure 3.1.B**; step 20). Specifically, for each species, the phased variants were uploaded independently using the function data2haplohh() with `polarize_vcf=FALSE` parameter as I did not have any information on ancestral or derived alleles. In addition, extended haplotype homozygosity (EHHS) was calculated between a focal site and its flanking markers. Then, the EHHS integral (iES) was estimated to measure the decay of haplotype homozygosity in the region surrounding the focal site. EHHS and iES were quantified utilizing function scan_hh() by setting `polarized=FALSE` due to the lack of information on ancestral or derived alleles. To account for the gaps present in the RADseq data because of the unsequenced region from the genome, I allowed the maximum allowed distance in base pairs between markers (`maxgap`) to be 172.618 kb (i.e., 95th percentile of inter-locus physical distance), scaled the gaps (`scalegap`) to 27.029 kb (i.e., median physical distance (in base pairs) between RAD loci), and stopped integration of

EHH of sites only when the distance between markers was greater than the `maxgap` (`discard_integration_at_border=FALSE`).

Furthermore, the standardized ratios of iES across common genomic nucleotide positions (i.e., XP-EHH scores) between two species were independently calculated using the ies2xpehh() function with *P. angustata* as target and *T. borchgrevinki* as reference. I set `p.adjust.method=BH` parameter to account for multiple testing and generate adjusted p-value with the 'Benjamini and Hochberg' approach. Positive and negative XP-EHH scores were obtained for *P. angustata* and *T. borchgrevinki*, respectively. The variants having a positive smoothed XP-EHH score with a statistically significant adjusted p-value (i.e., q-value less than 0.05 or $\log_{10}$(q-value) > 1.30310) were considered as XP-EHH outlier. Additionally, I kernel-smoothed XP-EHH scores using 900 kilobase pairs sized sliding window, and in each window, the corresponding variant site was in the middle position. The windows with statistically significant variants were considered as XP-EHH outlier windows.

Next, I identified the overlapping regions between XP-EHH and either D_{XY} or delta π outlier windows to obtain robust signals of selection (**Figure 3.1.B**; step 21). The overlapping regions were only retained when the variant sites in the middle of the XP-EHH outlier windows were also located within D_{XY} or delta π outlier windows. The merging of the adjacent overlapping regions from XP-EHH and D_{XY} or delta π outlier windows was only performed when five or more consecutive overlaps were found. The genes extracted from the overlapping regions between a) XP-EHH and D_{XY} outlier windows were named "*Dxy&linkage*", and b) XP-EHH and delta π outlier windows were referred to as "*deltapi&linkage*" candidates. However, all these candidates were referred to as "*scan&linkage*" candidates. All the processes related to identifying overlapping windows and gene extraction were performed using a custom Python script. Also, I determined if any of the "*scan&linkage*" candidates were located within or boundaries of putative structural variation specific to *P. angustata*.

Non-synonymous to synonymous substitution ratio (dN/dS) analyses on protein-coding genes

To identify the protein-coding genes of *P. angustata* under positive selection in the temperate environment but not in cold-specialized notothenioids, I performed non-synonymous to synonymous substitution ratio (dN/dS) analysis (**Figure 3.1.B**; steps 22-26) using the following steps. First, I obtained a phylogenetic tree for the group of species of interest. I trimmed a previously published tree for notothenioids by Near *et al.* (2018). However, the tree did not contain *P. angustata;* it had a closely related species, *N. coriiceps*. I replaced *N. coriiceps* with *P. angustata* as I did not intend to use branch length in the CODEML module in the PAML (v4; Yang 2007) package for dN/dS analysis.

Next, extracted coding sequences of 15,501 single-copy orthologs -- present in the species of interest, including *P. angustata* (secondarily temperate non-Antarctic notothenioid) and four cold-specialized Antarctic notothenioids (*T. borchgrevinki*, *Trematomus bernachii* (Bista *et al.* 2023)*, Gymnodraco actuiceps* (Bista *et al.* 2023)*,* and *Pseudochaenichthyus georgianus* (Bista *et al.* 2023). To obtain single-copy orthologs, I implemented a custom Python script on a) ortholog gene clusters identified by Synolog software and b) coding sequences generated by BRAKER2 for all five species. Note that the primary inputs for Synolog were RBHs, along with genome annotations (in Gene Transfer Format (GTF) or General Feature Format (GFF)) and AGP files (**Figure 3.1.B**; step 22). The obtained orthologs were aligned using PRANK (Löytynoja 2014) (**Figure 3.1.B**; step 23), and the alignments were filtered with Gblocks (Castresana 2000). In addition, the alignments with less than 450 base pairs (i.e., 150 amino acid sites) were removed (**Figure 3.1.B**; step 25).

Additionally, I independently estimated the dN/dS ratio per alignment based on branch and branch-site models, assigning *P. angustata* as foreground and the rest of the species as background (**Figure 3.1.B**; step 26). The dN/dS analysis per alignment was performed using a wrapper for the CODEML program GWideCodeml (Macías *et al.* 2020). From branch model-based analysis, I retained the orthologs with a significant likelihood ratio test (LRT) (hereafter referred to as *branch_set*). Furthermore, when CODEML determines the presence of codon under positive selection in the foreground lineage based on LRT using branch-site model, it also implements the Bayes Empirical Bayes (BEB) method to determine the BEB score of a site (i.e., probability of a site being under positive selection). Thus, I also retained orthologs showing significant LRT and having sites with BEB scores above 95% (hereafter referred to as *branch-site_set*) if the proportion for a site in class 2a and 2b was not zero.

Moreover, I analyzed two groups, *branch_set,* and *branch-site_set*, both independently and together. The genes within *branch_set* and *branch-site_set* were independently ranked based on significant p-values from LRT. From each group, the biological functions of the top ten highly-ranked genes were assessed using the Zebrafish Information Network (ZFIN) database (https://zfin.org/). When the gene from either group had no orthologs in zebrafish, we searched its coding sequence against reference sequences in the NCBI (National Center for Biotechnology Information) database (https://www.ncbi.nlm.nih.gov/) by using the blastn program within BLAST+ software to see if the queried sequence finds a match or remain uncharacterized. Further, I analyzed if the two groups had any genes in common. Finally, I combined all the unique genes from *branch_set* and *branch-site_set* and referred to this set as "*dN/dS*" *candidates*. I assessed if there were cases in which one or more members of "*dN/dS*" *candidates* were in the putative structural variation specific to *P. angustata*. For downstream functional enrichment analysis, I retained all "*dN/dS*" *candidates* regardless of their rank in the group they belonged to or their location in the genome.

Analysis for identifying significant enriched biological functions

I performed the following steps to categorize the biological functions of "*scan&linkage*" and "*dN/dS*" candidates together. I obtained the Ensemble IDs (https://www.ensembl.org/index.html) for the homologous genes between *P. angustata* and zebrafish. Then, the IDs of candidates were identified and uploaded to the web tool of the Ensemble (https://www.ensembl.org/index.html), known as BioMart, using zebrafish as a reference, such that I could retrieve gene descriptions as well as external IDs specific to Zebrafish Information Network (ZFIN) database (https://zfin.org/). Additionally, I uploaded the retrieved ZFIN IDs to the PANTHER database (v17.0; Mi *et al.* 2021) (**Figure 3.1.B**; steps 27) and selected zebrafish as a `reference`. Next, I chose the `PANTHER GO-slim Biological Process` option as an annotation dataset for a statistical overrepresentation test.

Further, I selected `Fisher's Exact` as a test type and `Calculate False Discovery Rate` as a correction option. I could not find any statistical significance in enriched biological functions, so I explored the observed categories of biological functions and genes within them. Additionally, since some gene ZFIN IDs were not mapped in the PANTHER database, I manually mapped those IDs in the ZFIN database (**Figure 3.1.B**; steps 27) to assess the biological functions.

RESULTS

Scaffolding of PacBio long-read-based contigs with Hi-C reads produced chromosomal-level genome assemblies

For *P. angustata*, PacBio sequencing generated 196.25 gigabase pairs (Gb) of data, comprising 12.23 million reads with a mean read length of 16.0 kilobase pairs (Kb) and an N50 30.1Kb (**Table 3.1**). Paired-end sequencing of the Hi-C library produced 319.23 million reads. In the final *de novo* assembly, I identified 13 putative chromosomes ranging in size from 15.6 to 96.9 megabase pairs (Mb), a total assembled length of 987.55 Mb, an N50 of 87.09 Mb, and total bases of 968.61 Mb (i.e., 98.08% of full assembly length) (**Table 3.2**). Furthermore, out of 3640 BUSCO genes (Actinopterygii-specific single-copy orthologs), 3507 (96.4%) were classified as complete. Among these complete BUSCO genes, 3468 were single-copy, and only 39 were duplicated. The total number of predicted protein-coding genes in the assembly was 27,096.

For *T. borchgrevinki*, PacBio sequencing yielded 181.42 Gb data, consisting of 7,651,558 reads with a mean length of 23.7 Kb and an N50 of 33.4 Kb (**Table 3.1**). Additionally, Hi-C library sequencing resulted in 209.01 million reads. The final *de novo* assembly comprised 24 putative chromosomes (ranging in size from 17.82 to 48.28 Mb), a total length of 935.09 Mb, an N50 41.31 Mb, and a total base length of 912.23 Mb (covering 97.56% of the total length). The top 23 chromosomes (sorted by size) varied from 27.11 to 48.27 Mb, encompassing 95.65% of the assembly. Moreover, 3523 (96.8%) out of 3640 BUSCO orthologs were complete. Of those complete BUSCO genes, 3481 were single-copy, and only 42 were duplicated. The total number of predicted protein-coding genes was 28,561 (**Table 3.2**).

The proportion of DNA transposon in P. angustata differed from cold-specialized and primarily temperate notothenioids

Repetitive elements made up a large proportion of the genome for both species. **Figure 3.2** and **Tables A.1**, **A.2**, and **A.3** show the breakdown of the interspersed repeats due to DNA transposons, retroelements, SINEs (short Interspersed Nuclear Elements), LINEs (Long Interspersed Nuclear Elements), LTR (Long Terminal Repeats), and unclassified elements both in terms of proportion of the genomes and absolute length in base pairs. The repeat contents of both *P. angustata* (57.67%) and *T. borchgrevinki* (54.61%) were higher than that of *E. maclovinus* (33.43%; Cheng *et al.* 2023) but lower than that of *N. rossii* (60.83%) and *C. gunnari* (59.45%; Rivera-Colón *et al.* 2023). However, the proportion of DNA transposons in *P. angustata* was notably the highest (**Figure 3.2; Table A.1**). The length occupied by DNA transposons in the genomes of both *P. angustata* and *T. borchgrevinki* was higher than that in the genomes of *E. maclovinus* (Cheng *et al.* 2023) (**Table A.2**). However, while such estimate for *T. borchgrevinki* was lower than those of *N. rossii* and *C. gunnari* (Rivera-Colón *et al.* 2023), it was highest for *P. angustata* (**Table A.2**). Additionally, the number of DNA transposons in *P. angustata* and *T. borchgrevinki* was higher than *E. maclovinus* but fewer than *C. gunnari*. Unlike in *T. borchgrevinki*, the number of DNA transposons in *P. angustata* was higher than *N. rossii* as well (**Table A.3**).

Compared to *T. borchgrevinki*, *P. angustata* had notably a 3.06% higher total repeat content and 5.45% more DNA transposons (**Table A.1**). Specifically, it harbored 77,477 more DNA transposons, accounting for approximately 66.15 Mb of genome length (**Table A.2**). Additionally, *P. angustata* showed only 0.76% higher proportion of retroelements (**Table A.1**) (contributing 16.05 Mbps of genome length (**Table A.2**)), although the retroelements was 35,793 fewer in number (**Table A.3**). In terms of unclassified elements, *P. angustata* had a 2.35% lower proportion (**Table A.1**) (corresponding to 17.83 Mb of genome length (**Table A.2**)) and 63,117 fewer in number (**Table A.3**).

Conserved synteny unveiled distinct patterns of chromosomal fusions as well as a few intra-chromosomal structural changes specific to the genome of P. angustata

Paranotothenia angustata showed differences with outgroups in large-scale genome organization. For example, genome-wide conserved synteny analyses between *P. angustata* and outgroups *C. gunnari*, *T. borchgrevinki,* and *E. maclovinus* showed evidence of chromosomal fusions in *P. angustata*. Out of 13 haploid chromosomes of *P. angustata*, two and eleven exhibited 1:1 and 1:2 homologous relationships, respectively, with chromosomes of both *C. gunnari* and *E. maclovinus* (**Figure 3.3**). Additionally, two and six chromosomes of *P. angustata* showed a 1:2 and 1:1 relationship, respectively, with the corresponding homologous chromosomes of *T. borchgrevinki* (**Figures 3.4 & 3.5**). The remaining five chromosomes exhibited five cases of complex relationships with homologous chromosomes in *T. borchgrevinki* (**Figure 3.4**)

Specifically, the five cases of complex relationships resulted from the disruption of conserved synteny between the species due to inter-chromosomal translocations in the genome of *T. borchgrevinki* (**Figure 3.4**). In these five cases, generally, most parts of one chromosome of *T. borchgrevinki* mapped to the homologous chromosome in *P. angustata*, whereas the remaining small part of the chromosome in *T. borchgrevinki* mapped to its homologous region in another chromosome in *P. angustata*. For example, small portions of chromosomes 1, 2, 3, 10, and 23 in *T. borchgrevinki* had their homologous regions in chromosomes 3, 14, 1, 14, and 2 in *P. angustata*, respectively (**Figure 3.4**). In two of these five cases, one whole chromosome and a major portion of another chromosome in *T. borchgrevinki* also showed evidence of fusion in *P. angustata* (**Figure 3.4**). Moreover, I found distinct chromosomal fusion patterns in *P. angustata* compared to *N. corriiceps*. Nine and four chromosomes of *P. angustata* had a 1:1 and 1:2 homologous relationship, respectively, with the chromosomes of *N. coriiceps*. However, seven chromosomes of *P. angustata* (represented by numbers one, two, three, five, six, twelve, and fourteen) and their homologous chromosomes in *N. coriiceps* exhibited differences in how their corresponding homologous chromosome pairs in *E. maclovinus* oriented or positioned themselves when they mapped to the genome of *N. coriiceps* versus *P. angustata* (**Figure 3.6**).

Regarding local changes in the genomic structure of *P. angustata*, many genomic rearrangements were identified between most of the chromosomes in *P. angustata* and *T. borchgrevinki*. Ten out of 13 fused chromosomes in *P. angustata* showed local chromosomal rearrangements with their homologs in *T. borchgrevinki*. These rearrangements included inversions, translocations (including non-inverted and inverted), and complex structural variations (i.e., changes that cannot be distinguished as one type of structural variation). Most of these rearrangements were specific to *T. borchgrevinki* or shared by *P. angustata* and *C. gunnari*. At the same time, only a few (12 in number with a size greater than 100 Kb) were potentially specific to *P. angustata*. Nine of those 12 structural changes were considerably large, exceeding 1 Mb (**Table 3.3**). Out of 12 structural changes, seven were inversions (**Figures A.1, A.2, A.3, A.4, A.5, A.6, & A.7**), four were translocations (as shown in **Figures A.6**, **A.7, & A.8**), and one was complex a change (i.e., structural change for which defining boundaries of inversion or translocation was difficult) (**Figure A.9**; **Table 3.3**).

A limited convergence between signatures of selection from genome scan and that from linkage (XP-EHH) analysis

From RAD sequences of two species, I identified 32,669 homologous RAD loci between species and 567,413 variants in those loci. The mean locus length was 763.51 bp. The total number of sites for both species was about 24.94 million. The average genome-wide estimate of nucleotide diversity (π) was 0.045 for *P. angustata* and 0.032 for *T. borchgrevinki*. The average F_{ST} between the two species was 0.736. The estimated mean absolute divergence (D_{XY}) between the two species was 0.0187. Additionally, 277 D_{XY} outlier windows were found between the species. Next, I identified 1,075 kernel-smoothed XP-EHH outlier windows specific to *P. angustata*. I detected 58 of those 1,075 XP-EHH outlier windows overlapping with D_{XY} outlier windows (**Figures 3.7 & 3.8**). Within these overlapped regions, I found 30 genes ("*Dxy&linkage*"). These "*Dxy&linkage*" candidates (**Tables A.4**) were located on chromosomes 5 and 15 (**Tables A.5**). Among cases in which the XP-EHH outlier windows did not overlap with D_{XY} outlier windows, I found two instances in which cluster of XP-EHH outliers coincided with inversion, specifically, on chromosomes 2 (**Figure 3.9**) and 14 (**Figure A.10**). Moreover, I found 72 variants exhibiting delta π outliers. Among these 72 delta π outliers, 11 had windows that overlapped with XP-EHH outlier windows (**Figures 3.9, A.11**). These overlapped regions encompassed 29 genes ("*deltapi&linkage*") (**Tables A.6 & A.7**). The "*deltapi&linkage*" candidates were distributed on chromosomes 2 and 6. From the combination of "*Dxy&linkage*" and "*deltapi&linkage*" candidates, I obtained 59 total unique genes, which were referred to as "*scan&linkage*" candidates.

Moreover, three of 59 "*scan&linkage*" candidates (i.e., *mrpl4*, *g_3470*, and *mdn1* genes from chromosomes 5 (**Figure 3.7**), 6, and 15 (**Figure 3.8**), respectively) contained XP-EHH outliers. In addition, another three candidates (i.e., *dnajc24*, *g_9481*, and *g_2999* genes from chromosomes 2 (**Figure 3.9**), 5, and 6, respectively) were adjacent to the XP-EHH outliers in the intergenic region. The *dnajc24* gene was located within 2 Mb distance from the inversion on chromosome 2 (**Figure 3.9**). Additionally, two of the "*scan&linkage*" candidates (*dla* and *abcc10* genes) were also members of the "*dN/dS*" candidates.

Accelerated molecular evolution in protein-coding genes

Based on the branch model, 138 genes exhibited significant LRT and were assigned to *branch_set* (**Table A.8** & **A.9**). The top ten highly-ranked candidates (in descending order) were *kcnc1b*, *ubtd1b*, *zgc:110626*, *mybbp1a*, *lyplal1*, *ccdc62*, *entpd2a.1*, *g_16386*, *enpp1*, and *ciartb*. The human ortholog of *kcnc1b* is predicted to participate in the transmembrane transport of potassium ions (Zhao *et al.* 2013). In humans, *ubtd1b* regulates cellular senescence (Zhang *et al.* 2015), and *zgc:110626* participates in innate immune response (Meng *et al.* 2017). In mice, *mybbp1a* acts as a co-repressor on the clock gene *Period2* (Hara *et al.* 2009). Moreover, *lyplal1* is associated with protein depalmitoylation (Tian *et al.* 2012), and *ccdc62* is linked to human spermatid development (Oud *et al.* 2020). Further, *entpd2a.1* respond to copper (Rosemberg *et al.* 2007) and ethanol (Rico *et al.* 2008) in zebrafish. The BLAST analysis revealed that the coding sequence of *g_16386* matches with a *mixed lineage kinase domain-like protein* (*mlkl)*. In humans, *mlkl* is involved in necroptosis (Cai *et al.* 2014). While the ortholog of the *g_16386* gene was not found, *enpp1* is linked to phosphate ion homeostasis and regulation of bone mineralization (Apschner *et al.* 2014) in zebrafish. In the ZFIN database, the *ciartb* is predicted to be involved in the circadian regulation of gene expression and negative regulation of transcription of DNA-template. Its homolog in mammals (*ciart* or CHORON) is known to regulate core proteins of the circadian feedback loop (*BMAL1* and *CLOCK*) (Goriki *et al.* 2014).

Based on the branch-site model, 210 genes showed significant LRT and their codon(s) displayed BEB scores greater than 95%. These 210 genes were assigned to *branch-site_set* (**Table A.10** & **A.11**). The top ten highly-ranked candidates (in descending order) were *clocka*, *g_24544*, *cnot4b*, *g_30555*, *scml2*, *lmnl3*, *sin3aa*, *g_25237*, *g_22086*, and *tcf7l2*. The gene *clocka* is linked to photoperiodism (Whitmore *et al.* 1998; reviewed in Vatine *et al.* 2011) in zebrafish, apart from several other processes. *cnot4b* is involved in protein ubiquitination in humans (Wang *et al.* 2018), and *scml2* plays a role in the regulation of transcription in mammals (Menon *et al.* 2019). Additionally, the human ortholog of *sin3aa* is involved in transcriptional response to hypoxia (Tiana *et al.* 2018); however, information on the biological function of *lmnl3* was found. The *tcf7l2* has multiple functions, including the regulation of lipolysis and lipogenesis in mice (Geoghegan *et al.* 2019). The BLAST analysis predicted that *g_24544*, *g_30555*, *g_25237*, and *g_22086* match to *caldesmmon 1a*, *leucine-rich repeat neural protein 3-like*, *synergin gamma*, and *DENN domain-containing protein 2A-like mRNAs*, respectively. *Caldesmmon 1a* is linked to peristalsis in zebrafish (Abrams *et al.* 2012); however, either ortholog or functions of *g_24544*, *g_30555*, *g_25237*, and *g_22086* were not found.

From *branch_set* and *branch-site_set*, I identified a total of 317 unique "*dN/dS*" candidates across the groups and 31 common candidates between the groups (**Table A.12**). Among these 317 unique candidates, 11 were located within structural variations specific to *P. angustata*. Specifically, two of 11 genes (mentioned within the parenthesis following chromosome number) were found within translocations, specifically on chromosomes 8 (*g_1982*) and 24 (*g_15022*) (**Tables 3.3** & **A.8**-**11**). Additionally, nine of 11 genes were distributed within inversions on chromosomes 2 (*cmip*, *g_31324*, and *ZNF276*), 4 (*il16*), 6 (*g_30547 and g_16712*), 14 (*si:dkey-106g10.7* and *spatal6*), and 15 (*nus1*). Of these nine candidates, three genes, including *ZNF276*, *il16*, and *spatal6* were part of the common members between *branch_set* and *branch-site_set* (**Table A.12**). The gene *cmip* is involved in the negative regulation of T cell signaling in mice, as demonstrated by Oniszczuk *et al.* (2020). The candidate gene *il16* is believed to have multiple functions, including serving as an immunomodulatory and proinflammatory cytokine (reviewed in Wilson *et al.* 2003; Mathy *et al.* 2000). Its ortholog in mice is associated with upregulation of immunoglobin E (Hessel *et al.* 1998). According to the ZFIN database, the candidates *ZNF276* and *spata6l* are thought to be involved in both the regulation of transcription by RNA polymerase II and spermatogenesis. However, I did not find information on the biological functions of these genes. The gene *nus1* is related to movement in zebrafish (Yu *et al.* 2021). Unfortunately, I did not find the actual or predicted function of the gene *si:dkey-106g10.7* (*zfta*).

Absence of significant enrichment in biological functions for identified candidates, yet, the presence of some genes with functional relevance to temperate environment

I obtained 374 unique candidates from the combination of 59 "*scan&linkage*" and 317 *"dN/dS"* candidates, with two groups having two genes in common. Of the 374 unique genes, 99 had no orthologs in zebrafish based on Synolog. From the remaining 275 candidates, 241 were mapped to the PANTHER database, and I did not observe statistically significant enriched categories of biological functions, indicating a wide variety of biological functions may have been involved in the adaptation of *P. angustata* to temperate conditions. Despite observing no statistically significant biological categories, I observed some genes having functions that could be important for *P. angustata* in its adaptation to a temperate environment. Specifically, I found genes related to protein chaperoning (*dnajc24*), erythrocyte development and differentiation (*rasa3, numb, etv7*), heme metabolism (*cpox*), circadian rhythm (*atxn2l*, *clocka*, and *cipcb*), visual system development (*ift172*, *dhdds*, *itag5*, *dnase1l1l*, and *get1*), mitochondria (*sdhaf2*, and *ndufv3*), and ribosomes (*mrpl4*, *mrpl30*, *mrps10*, *mrps34*, *mdn1,* and *nsun4*).

DISCUSSION

I generated high quality *de novo* chromosome-level genome assemblies for the focal species, *P. angustata,* and outgroup species, *T. borchgrevinki,* to identify structural changes and genes with an accelerated non-synonymous substitution that are specific to the genome of *P. angustata*. I produced population-level RADseq data for these two species to detect *P. angustata*-specific signals of positive selection based on differences in the nucleotide diversity (π), differentiation (F_{ST}), and divergence (D_{XY}) between the species, as well as haplotype homozygosity. I found a high proportion of DNA transposons, a unique pattern of chromosomal fusions, inversions, and translocations in the genome of *P. angustata*. A few of the genes with accelerated molecular evolution co-localized with inversions. I propose that genes related to protein chaperoning, circadian rhythm, vision, erythrocyte differentiation and development, heme metabolism, and vision, as well as mitochondria and ribosomes, may have contributed to adaptations of *P. angustata* in the temperate environment.

De novo chromosome-level assemblies of P. angustata and T. borchgrevinki are of highly quality

The genus *Paranotothenia* consists of a monophyletic clade of secondarily temperate notothenioids (*Paranotothenia microlepidota*, *P. magellenica,* and *P. angustata*) (Cheng 2003; Dettai *et al.* 2012) each with 13 haploid chromosomes (reviewed in Amores *et al.* 2017). I found about 98% of the total bases in the assembly of *P. angustata* were covered by 13 chromosomes, suggesting that the *de novo* chromosome-level assembly for the species is highly complete in length. The diploid chromosome number of *T. borchgrevinki* depends on sex (i.e., 2n=45 for males and 2n=46 for females) (Pisano *et al.* 2003; Auvinet *et al.* 2020). For *T. borchgrevinki*, we observed 24 haploid chromosomes (instead of 23, given the sampled individual was female), covering about 97.5% of the total bases of the assembly. However, 23 chromosomes covered approximately 95.65% of the total bases, indicating that the assembly for *T. borchgrevinki* is still highly complete in length.

The extra chromosome's presence (approximately 16.87 Mb or 1.85% of genome length) could be explained by one of the two phenomena. One possibility is that the highly complex repetitive region in one of the chromosomes could not be resolved during the *de novo* assembly process. Consequently, it may have resulted in the fragmentation of that chromosome. The assembly of highly complex repeats can cause different types of issues. For example, they can generate collapsed, fragmented, or chimeric assemblies (Kong *et al.* 2023). However, biological variation in chromosome number among individuals within species is another possibility. This possibility cannot be ruled out because the intraspecific polymorphism in chromosome number has been observed in another couple of species within the *Trematomus* genus, including *T. hasoni* (2n=45/46, 46, and 48 (Morescalchi *et al.* 1992; Ozouf-Costaz *et al.* 1991; Ozouf-Costaz *et al.* 1999b)) and *T. loennbergii* (2n=26, 27, 28, 29, 30, 31, 33, and 48 (Morescalchi *et al.* 1992; Ozouf-Costaz *et al.* 1999b; Ghigliotti *et al.* 2015)). Future genomic or cytogenetic analyses would be necessary to shed light on these possibilities. For example, conserved synteny analysis between the chromosome-level assembly for male and female *T. borchgrevinki* may tell if the additional chromosome in the observed data is due to the fragmentation or the actual intraspecific chromosome number.

In terms of the number of protein-coding genes, my assemblies consist of 27-28K genes, which are comparable to those in assemblies of other notothenioids (with about 20-29 thousand genes) (Bargelloni *et al.* 2019; Bista *et al.* 2020, 2023; Rivera-Colón *et al.* 2023; Cheng *et al.*

2023). This suggests that my annotations effectively captured at least most of the protein-coding genes in the genomes of each species. Genome assemblies may sometimes exhibit a high count of complete BUSCO genes due to the inadvertent increase in complete but duplicated BUSCO genes (Rayamajhi *et al.* 2022). However, in the case of both species' genome assemblies, the proportion of BUSCO genes with a complete status was notably high, at approximately 96%. In contrast, the proportion with a duplicated status was minimal, around 1%. This observation strongly suggests that the assemblies are of high quality. These well-constructed assemblies hold great potential for facilitating genome-based research in polar and non-polar notothenioids.

High repeat content in both P. angustata and T. borchgrevinki but lineage-specific expansion of DNA transposons only in P. angustata

The repeat contents of *P. angustata* (57.67%) and *T. borchgrevinki* (54.61%) were in between those of *E. maclovinus* (33.43%; Cheng *et al.* 2023) and *N. rossii* (60.83%). However, the total repeat content of 16 notothenioid species (including three from non-Antarctic and 13 from the Antarctic region) ranges from 13% to 54% (Bista *et al.* 2023). Recent studies have shown that repetitive elements can contribute to a small to large fraction of the fish genome. For example, a recent study on 39 fish species reported the degree of contribution of transposable elements (TEs) in the genome ranged from 5% (in pufferfish) to 56% (in zebrafish) (Shao *et al.* 2019). Based on this evidence, the repeat contents of *P. angustata* and *T. borchgrevinki* can be considered high.

Compared to *E. maclovinus, T. borchgrevinki*, *N. rossii*, and *C. gunnari*, I observed that *P. angustata* consisted of the highest proportion of DNA transposons and length of genome occupied by the transposons. These differences could be due to the lineage-specific expansion of DNA transposons in the lineage of *P. angustata* after it diverged from *N. rossii*, given species within *Paranotothenia* and *Notothenia* genus are more closely related to each other than any other species across notothenioid clade. The two species, *P. angustata* and *N. rossii* share a common ancestor about 8 million years ago (reviewed in Amores *et al.* 2017). Also, the total assembled length of *P. angustata* (987.55 Mb) was higher than that of *E. maclovinus* (606.28Mb; Cheng *et al.* 2023) and *T. borchgrevinki* (935.08Mb) but lower than those of *N. rossii* (1042.90 Mb; Clawson *et al.* 2023), *C. gunnari* (994.20 Mb; Rivera-Colón *et al.* 2023). The largest length and the highest repeat content for assembly of *N. rossii* among the five notothenioids indicate that the collapse of repeats in *N. rossii* may not have contributed to the observed difference in DNA transposons between *N. rossii* and *P. angustata*. Moreover, I observed that *P. angustata* had fewer DNA transposons in number than *C. gunnari*. Such a discrepancy could result from fewer but larger copy sizes of DNA transposons in *P. angustata* compared to *C. gunnari*.

While the insertion of TEs can be harmful to organisms, such an effect could be mitigated by different mechanisms in the genome, enabling fitness maintenance of both TEs and their host. For example, if TEs prefer to be inserted into other pre-existing TEs or introns rather than exons, then the fitness of both the host and TEs will not be affected (Kidwell and Lisch 1997). Moreover, TE expansion can occur in response to environmental stresses, such as temperature (Carotti *et al.* 2022), and they can even facilitate evolutionary adaptation (González *et al.* 2008, 2010; Casacuberta and González 2013). A recent comparative study on 52 fish species suggested an association between repetitive elements and fish habitats (Yuan *et al.* 2018). Another study using 39 species of teleost – living in the cold waters of the Arctic, Antarctic, temperate regions and warm waters of tropical, subtropical, and temperate regions – showed a correlation between *Rex3* retroelements and temperature. Despite taxonomic differences among these teleost species, the phylogenetic analysis showed that *Rex3* retroelements from species inhabiting cold environments formed separate clusters compared to those from species residing in temperate environments (Carducci *et al.* 2019). TEs can also produce novel coding genes (Long *et al.* 2003) through functional changes through gene regulation by inserting into regulatory elements or alterations in protein function by directly inserting into the coding sequence of genes (Chuong *et al.* 2017).

For example, in Midas cichlids, intronic insertion of piggyBac transposons generated a color polymorphism (Kratochwil *et al.* 2022). The TEs that cause deleterious effects on organisms may be removed by purifying selection. However, when the effective population size is small, the efficacy of purifying selection becomes too low to remove mildly deleterious TEs, and genetic drift can fix them in the population. However, passively accumulated, slightly deleterious TEs could be secondarily adaptive as a novel genetic basis of adaptation (reviewed in Lynch *et al.* 2011). Thus, it is possible that the lineage-specific expansion of DNA transposons observed in *P. angustata* played a role in the adaptation of *P. angustata* to a temperate environment.

Chromosomal rearrangements may have independently occurred in P. angustata

The genome-wide conserved synteny between *P. angustata* and other notothenioids, including *E. maclovinus*, *N. corriceps*, *T. borchgrevinki*, and *C. gunnari*, support the presence of extensive chromosomal fusions in *P. angustata*. I found a similarity between *P. angustata* and *N. coriiceps* in terms of haploid chromosome number. However, in some instances, the chromosomes from *E. maclovinus* that mapped to their orthologs in *P. angustata* had different orientations or positioning compared to when they mapped to their corresponding ortholog in *N. coriiceps*. This observation suggests that certain chromosomal fusions evolved independently in the lineages of *P. angustata* and *N. coriiceps*. The distinct pattern of chromosomal fusions in the lineage of *P. angustata* compared to *N. coriiceps* is intriguing because of three reasons. First, these two species have adapted to different thermal environments. Second, the genus *Paranotothenia* forms a monophyletic clade of secondarily temperate notothenioids (Cheng 2003; Dettai *et al.* 2012) with the same chromosome number (Amores *et al.* 2017). Third, the theory predicts that the chromosomal fusions can facilitate adaptation through clustering of coadapted alleles at multiple loci (that were previously unlinked) and reducing recombination among those loci such that they are in linkage disequilibrium (Guerrero and Kirkpatrick 2014). Evidence from empirical studies such as on Atlantic salmon (Wellband *et al.* 2019) and threespine stickleback (Liu *et al.* 2022) has supported the notion that chromosomal fusions can facilitate adaptation. It is also crucial to recognize that chromosomal fusions can also alter the three-dimensional organization of the genome, resulting in changes in the position of the genome within the nucleus, which, in turn, can alter gene expression dynamics (Di Stefano *et al.* 2020) and contribute to phenotype divergence (Diament and Tuller 2017).

I also detected instances of intra-chromosomal translocations and inversions that appear to be specific to *P. angustata*. Translocations can potentially confer genetic adaptations, often influencing gene expression (Zimmer *et al.* 2014). On the other hand, inversions may directly or indirectly experience positive selection. For instance, if an inversion's breakpoint modifies gene expression in a way that generates an adaptive trait in the organism, direct positive selection becomes plausible. Alternatively, if selection operates on the recombination effects of the inversion, indirect selection on the inversion can occur. For instance, a new inversion might link pre-existing adaptive loci through recombination suppression, thereby preventing the reshuffling of co-adapted loci, including those engaged in local adaptation or epistatic interactions (Faria *et al.* 2019). Notably, nearly all putative inversions contained one or more positively selected genes (*cmip*, *il16*, *ZNF276*, *spata6l*, *nus1*, and *si:dkey-106g10.7*) in *P. angustata* (**Table 3.3**). These lines of evidence suggest the potential role of structural variation in the adaptation of *P. angustata* cannot be ruled out without further investigation.

Genome-wide patterns of differentiation and divergence primarily reflect the phylogenetic relationship between species

Given that the Antarctic Polar Front (APF) acts as both a thermal and physical barrier for species to its south and north, it is unsurprising that I observed a high mean F_{ST} between *P. angustata* and *T. borchgrevinki* (0.73). This estimate indicates the absence of gene flow between these species (Wright 1984). Rivera-Colón *et al.* 2023 also reported a similar pattern of F_{ST} between cold-specialized and secondary temperate notothenioids. Specifically, researchers reported a high mean F_{ST} between *C. gunnari* (cold-specialized) and its sister *C. esox* (secondarily temperate) (0.40), which recently diverged (approximately 1.6 million years ago). This estimate is about 1.825 times lower than I observed between *P. angustata* and *T. borchgrevinki*, a species pair with comparatively deeper divergence time. The recent time-calibrated phylogeny (species tree) of notothenioids depicted by Bista *et al.* 2023 suggests the mean age of divergence between the clades containing species from genus *Notothenia* (closely related to *P. angustata*) and *Trematomus* is 10.06 million years. This suggests that speciation and the cessation of gene flow between *P. angustata* and *T. borchgrevinki* have long been complete.

Moreover, the mean D_{XY} between these species (0.0187) was also about 4.45 times higher than the mean D_{XY} between *C. gunnari* and *C. esox* (0.0042; Rivera-Colón *et al.* 2023). In the absence of gene flow, species pair with more profound divergence is expected to have a higher D_{XY} than recently diverged species due to the accumulation of more fixed mutations (reviewed in Cruickshank and Hanh 2014; Chase *et al.* 2021). Overall, my data's genome-wide patterns of F_{ST} and D_{XY} largely reflect the phylogenetic relationships between the two species.

Potential secondarily temperate adaptations of P. angustata

Protein chaperoning

Researchers have consistently shown that cold-specialized notothenioids cannot increase or induce heat shock proteins (HSPs) as a coping mechanism to heat stress (Hoffman *et al.* 2000; 2005; Place *et al.* 2004; Place and Hoffman 2005; Bilyk *et al.* 2018). However, Hoffman *et al.* 2005 showed that *P. angustata* can induce mRNA from the HSP70 gene in response to elevated temperature. Even though they did not observe induction of HSP70 at the protein level, a prior study by Carpenter and Hofmann 2002 reported that *P. angustata* possesses a higher endogenous level of HSP70 or 70 kDa Hsps compared to three Antarctic *Trematomus* congers (*T. bernachii*, *T. hansonii*, and *T. pennellii*). More recently, Bilyk and Devries 2012 demonstrated that *P. angustata* exhibits significantly greater heat tolerance capacity than a cold-specialized notothenioid (*N. coriiceps*) even though lower than basal New notothenioid (*Bovichtus variegatus*). These pieces of evidence from prior studies suggest that HSPs could contribute to the higher thermal-stress tolerance capacity of *P. angustata* in temperate environments compared to cold-specialized notothenioids. In this study, one of the candidates, *dnajc24*, is related to human heat shock proteins (Thakur *et al.* 2012). This gene is from the family of DNAJ/HSP40, in humans. It acts as a co-chaperone to heat shock proteins (from family HSP70), mediated by its iron-binding properties (Thakur *et al.* 2012).

Hsps play a crucial role in recruiting client proteins to the HSP70 machinery and enhance the stability of the interaction between HSP70 proteins and their clients by stimulating ATP hydrolysis (Qiu *et al.* 2006; Kampinga and Craig 2010; Wan *et al.* 2020; Hu *et al.* 2022; Cyr and

Ramos 2023). The signals of positive selection related to the HSPs in *P. angustata* were observed close to *dnajc24* but not within its coding sequence. Hence, it is possible that genetic changes may have occurred in the regulatory region of the *dnajc24* gene. These changes may have enhanced the thermal tolerance capacity of *P. angustata* by influencing the regulation of co-chaperoning HSP70 proteins. However, the SNPs indicating these genetic changes could also be neutral markers associated with the adaptive locus of *dnajc24*. A suite of DNAJ/HSP40 paralogs has been reported under positive selection in another secondary temperate notothenioid, *C. esox* (Rivera-Colón *et al.* 2023).

Circadian rhythm and visual system development

Antarctica experiences unique light/dark cycles with several months of continuous daylight and darkness each year, which differs from New Zealand's. In Otago Harbor of NZ, light or dark periods in a day would be between 8 and 16 hours (Stuart 1998; reviewed in Dean and Hurd 2007). Disparate lighting environments could exert different selection pressures on circadian rhythms (Hut *et al.* 2013). I found that *clocka* gene, with the highest rank among members of *branch-site_set*, is known for being one of the major rhythm-setting genes of the circadian core feedback loop in zebrafish (Whitmore *et al.* 1998; reviewed in Vatine *et al.* 2011). This gene is also involved in the circadian rhythms of opsin gene expression (Li *et al.* 2008). Another candidate gene is *cipcb.* In mammals, the ortholog of *cipcb* acts as a negative feedback regulator of the circadian clock (Zhao *et al.* 2007). Another candidate gene (*atxn2l)* is also involved in regulating the activity of circadian clocks in mammals (Zhuang *et al.* 2023).

Moreover, variation in photoperiod can also change light sensitivity for circadian response to light (Glickman *et al.* 2012). A candidate *ift172* plays a role in maintaining retinal photoreceptors (Gross *et al.* 2005) and transporting light-sensitive, opsin proteins located in rod and cone cells (Sukumaran and Perkins 2009). This gene has also been shown to be under positive selection in *C. esox* (Rivera-Colón *et al.* 2023). Additionally, the candidate *dhdds* is crucial for retina formation, which affects the expression of light-sensitive rhodopsin protein in drosophila (Brandwine *et al.* 2021). The candidate *get1* is essential for synaptic functions in retinal photoreceptors (Lin *et al.* 2016). The positive selection of light-sensitive, retinal photoreceptors- and circadian rhythm-related genes indicates that they could have a role in the adaptive entrainment of internal clocks in *P. angustata* for proper timing of physiology and behaviors.

A study on the thermal stability of eye lenses among 12 vertebrates (including Antarctic icefishes and occurring in temperatures ranging from -2°C to 47°C) has shown the direct correlation between the resistance of the lens to thermal stress (leading to, for example, loss of lens transparency or cataract formation) and environmental temperature in which those vertebrate naturally occur (McFall-Ngai and Horwitz 1990). Fluctuation in water temperature can also cause cataract development (Bjerkås *et al.* 2001). While there is no information on how specifically the eye lens of *P. angustata* differs from cold-specialized species, I identified *itga5* and *dnase1l1l* as candidate genes. Previous work shows that they affect lens fiber morphogenesis and that their mutation can cause cataracts in zebrafish (Hayes *et al.* 2012; Zhang *et al.* 2020). These genes may have been involved in adaptive structural changes of the lens in the cold-water ancestor of *P. angustata*, contributing to the thermal stability of the lens in the temperate environment.

Erythrocyte differentiation and development, as well as heme metabolism

P. angustata has higher hematocrit (the proportion red blood cells occupy in total blood volume) than six Antarctic, cold-specialized notothenioids (Macdonald and Wells 1991). *P.*

angustata also has higher hemoglobin content and mean cellular hemoglobin concentration, except in comparison to *T. centronotus* and *D. mawsoni,* for which data was unavailable (Macdonald and Wells 1991). In agreement with prior studies, I observed three candidate genes (*rasa3, numb,* and *etv7*) with non-synonymous mutations that are related to red blood cells and another candidate gene with non-synonymous mutations (*cpox*) that are known to be involved in heme metabolism. The gene *rasa3* is known to be involved in a critical function in vertebrate erythropoiesis (Blanc *et al.* 2012). In mice, *rasa3* regulates the cell cycle during erythropoiesis (Brindley *et al.* 2021). In zebrafish, the *numb* gene plays a role in primitive erythrocyte differentiation (Bresciani *et al.* 2010), whereas *etv7* modulates red blood cell development during erythropoiesis by changing expression of lanosterol synthase, which is essential in cholesterol synthesis pathway (Quintana *et al.* 2013). The candidate gene list includes *cpox*, which encodes proteins that catalyze the reaction, converting coproporphyrinogen III to protoporphyrinogen IX, which is required for heme biosynthesis (reviewed in Zhang and Hamza 2019). Collectively, this evidence suggests that the candidates (*rasa3, numb, etv7*, and *cpox*) in *P. angustata* may play a role in ensuring proper production and maintenance of functional red blood cells as well as hemoglobin suited to temperate environments, leading to a higher oxidative capacity in *P. angustata*, which is required in warmer habitats.

Mitochondria and ribosomes

In a prior study, *P. angustata* had higher metabolic demand than *N. coriiceps* (Campbell *et al.* 2007) suggesting that *P. angustata* needs relatively more production of adenosine triphosphate (ATP) than *N. coriiceps* to sustain its physiological functions. Previous work by Bilyk *et al.* (2023) showed that the genes involved in a wide range of functions in mitochondria – including those related to the biosynthesis of mitoribosome proteins (MRPs) and components of electron transport chain (ETC) – are under relaxed selection pressure in Antarctic fish relative to tropic and temperate fish (Bilyk *et al.* 2023). Conversely, the genes with functions related to mitochondrial morphology, cellular respiration, and organization of ETC were suggested to be under positive selection in *C. esox*, a notothenioid that recently underwent secondary adaptation to temperate conditions (Rivera-Colón *et al.* 2023). In this study, I found that the candidates *sdhaf2* and *ndufv3,* the sub-units of mitochondrial respiratory complexes I and II (Zhu *et al.* 2016; Sharma *et al.* 2020), are under positive selection in *P. angustata*. These complexes are major components of the ETC that play a central role in oxidative phosphorylation (OXPHOS), i.e., ATP synthesis (Hirst 2013; Sharma *et al.* 2020). Moreover, genes such as *mrpl4*, *mrpl30*, *mrps10*, and *mrps34* were also found to be under positive selection in *P. angustata*. These genes' human orthologs are related to mitoribosome proteins (MRPs), which are structural components of the mitochondrial ribosome (Brown *et al.* 2014; Amunts *et al.* 2015; Lake *et al.* 2017). Mitoribosomes facilitate protein synthesis in eukaryotes (Amunts *et al.* 2015). Chen *et al.* (2008) reported that in the Antarctic cold-specialized notothenioid *Dissostichus mawsoni*, genes related to ribosome biogenesis are upregulated compared to warm-water teleosts, including temperate/tropical fishes. Based on this evidence, they suggested that *D. mawsoni* has comparatively enhanced protein synthesis capacity compared to temperate/tropical fishes. Another study on Antarctic and temperate fish of Zoarcidae has shown that protein synthesis capacity could correlate with temperature (Storch *et al.* 2005). My candidate genes, *mdn1,* and *nsun4,* have a role in human ribosome biogenesis (Spåhr *et al.* 2012; Chen *et al.* 2018). My findings suggest the observed candidates related to mitochondria and ribosomes have undergone adaptive genetic changes that may have enabled *P. angustata* for proper energy production and protein synthesis needed to adapt in temperate environments.

CONCLUSION

In this study, I present chromosome-level genome assemblies with high quality annotations for two species: *P. angustata* and *T. borchgrevinki*. I found that *P. angustata* has a high proportion of DNA transposons and a set of unique structural variants. Several candidate regions showed signals of positive selection including genes related to protein chaperoning, erythrocyte development and differentiation, heme metabolism, circadian rhythm, vision, mitochondria, and ribosomes. These results provide a compelling line of evidence of how secondarily temperate adaptations in *P. angsustata* may have evolved. They also contribute valuable genomic resources for polar biologists to conduct functional, comparative, and population genomics studies in the future, especially considering the existence of other secondary temperate notothenioids that may or may not share the same adaptive genetic changes found in *P. angustata*

TABLES

Table 3.1 Summary of PacBio data from sequenced libraries for *Paranotothenia angustata* and *Trematomus borchgrevinki*

	P. angustata			***T. borchgrevinki***
Sequencing library	library 1	library 2	library 1 + 2	library 1
Read count	6,389,651	5,846,611	12,236,262	7,651,558
Total Length (Mb)	100,633.23	95,622.72	196,255.96	181,428.53
Mean Length (bp)	15,749	16,355.24	16,038.88	23,711.32
N50 Length (bp)	29,602	30,648	30,107	33,463
Cnt >20kb	1,792,907	1,726,621	3,519,528	4,132,522
Cnt >50kb	312,889	309,731	622,620	492,063

Table 3.2 Summary of genome assembly and BUSCO statistics for *Paranotothenia angustata and Trematomus borchgrevinki*

Genome assembly statistics	***P. angustata***	***T. borchgrevinki***
Number of chromosomes	13	24
Total scaffold length	987,554,504	935,086,594
Number of fragments	1,888	2,095
Scaffold N50	87,087,854	41,310,500
Scaffold L50	6	11
Largest Scaffold	96,963,040	48,277,306
Total bases in chromosomes	968,615,351	912,238,485
Percentage of Assembly in chromosomes	98.08%	97.56%
Protein coding genes	27,096	28,561
BUSCOs statistics		
Complete	3507 (96.4%)	3523 (96.8%)
Complete and single-copy	3468 (95.3%)	3481 (95.6%)
Complete and duplicated	39(1.1%)	42(1.2%)
Fragmented	7(0.2%)	8(0.2%)
Missing	126 (3.4%)	109(3.0%)
Total	3640	3640

Table 3.3 Summary of potential structural variation (SV) with length greater than 100 kilobase pairs and specific to chromosomes (Chr.) of *Paranotothenia angustata* (indicated as Pang)

Chr.	**SV specific to Pang**[+]	**Start position**	**End position**	**Size (Megabase pairs)**		***branch_set***	***branch-site_set***
1	Complex	5,791,117	7,254,034	1,462,917	(1.46 Mb)		
2	Inversion	87,669,862	93,028,066	5,358,204	(5.35 Mb)	*ZNF276***	*cmip*, *NA*(*g_31324*)
4	Inversion	50,315,097	51,807,681	1,492,584	(1.49 Mb)	*il16***	
6	Inversion	44,791,198	48,621,198	3,830,000	(3.83 Mb)		*NA* (*g_30547*), *NA* (*g_16712*)
6	Inversion	49,569,590	49,692,661	123,071	(0.12 Mb)		
14	Inversion	214,127	4,026,751	3,812,624	(3.81 Mb)	*spata6l***	*si:dkey-106g10.7*
15	Inversion	40,474,519	40,947,389	472,870	(0.47 Mb)	*nus1*	
24	Inversion	95,954	840,964	745,010	(0.74 Mb)		
3	Translocation	130,494	2,302,870	2,172,376	(2.17 Mb)		
8	Translocation	59,073,956	61,104,117	2,030,161	(2.03 Mb)	*NA* (*g_1982*)	
24	Translocation	2,198,895	4,994,940	2,796,045	(2.79 Mb)		*NA* (*g_15022*)
24	Translocation	5,017,256	7,490,159	2,472,903	(2.47 Mb)		

[+] indicates that *P. angustata* was compared to *Champsocephalus gunnari*, *Trematomus borchgrevinki*, and *Eleginops maclovinus*

******indicates that the gene was significant in both branch and branch-model based dN/dS analysis.

NA indicates the genes without recognizable orthologs in Zebrafish.

Table 3.4 Gene candidates and their biological functions as well as rank for *dN/dS* candidates based on P-value from likelihood ratio test

Biological Functions	**Name of Gene Candidates**	**Candidate Source**	**Rank/Total**	**P-value**
Protein chaperoning	*dnajc24 (DnaJ heat shock protein family (Hsp40))***	*deltapi&linkage candidates*		
Circadian rhythm	*clocka* (*clock circadian regulator a*)	*dN/dS candidates (branch-site_set)*	1/210	0
	cipcb (*CLOCK-interacting pacemaker b*)	*dN/dS candidates (branch_set)*	61/138	0.015
	atxn2l (ataxin 2-like)	*dN/dS candidates (branch-site_set)*	109/210	0
Vision	*ift172 (intraflagellar transport 172)*	*dN/dS candidates (branch-site_set)*	210/210	0.046
	itga5 (integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	*dN/dS candidates (branch-site_set)*	138/210	0.001
	dhdds (dehydrodolichyl diphosphate synthase)	*dN/dS candidates (branch-site_set)*	205/210	0.036
	dnase1l1l (deoxyribonuclease I-like 1-like)	*dN/dS candidates (branch_set)*	99/138	0.031
	get1(guided entry of tail-anchored proteins factor 1)	*dN/dS candidates (branch_set)*	50/138	0.012
Erythrocyte differentiation and development	*etv7 (ETS variant transcription factor 7)*	*dN/dS candidates (branch-site_set)*	156/210	0.003

Table 3.4 – Continued

	numb (NUMB endocytic adaptor protein)	*dN/dS candidates (branch-site_set)*	190/210	0.021
	rasa3 (RAS p21 protein activator 3)	*dN/dS candidates (branch-site_set)*	18/210	0
Heme metabolism	*cpox (coproporphyrinogen oxidase)*	*dN/dS candidates (branch-site_set)*	173/210	0.008
Mitochondria and Ribosomes	*mrpl4 (mitochondrial ribosomal protein L4)****	*Dxy&linkage candidates*		
	mrps10 (mitochondrial ribosomal protein S10)	*dN/dS candidates (branch-site_set)*	125/210	0
	mrps34 (mitochondrial ribosomal protein S34)	*dN/dS candidates (branch-site_set)*	97/210	0
	mrpl30 (mitochondrial ribosomal protein L30)	*dN/dS candidates (branch_set)*	95/138	0.029
	ndufv3 (NADH:ubiquinone oxidoreductase subunit V3)	*dN/dS candidates (branch_set)*	66/138	0.017
	sdhaf2 (succinate dehydrogenase complex assembly factor 2)	*dN/dS candidates (branch_set)*	41/138	0.01
	*mdn1(midasin AAA ATPase 1)****	*Dxy&linkage candidates*		
	nsun4 (NOP2/Sun RNA methyltransferase 4)	*dN/dS candidates (branch-site_set)*	105/210	0

** indicates that the gene was closest to XP-EHH outlier; ***indicates that the genes contained XP-EHH outliers

FIGURES

A.

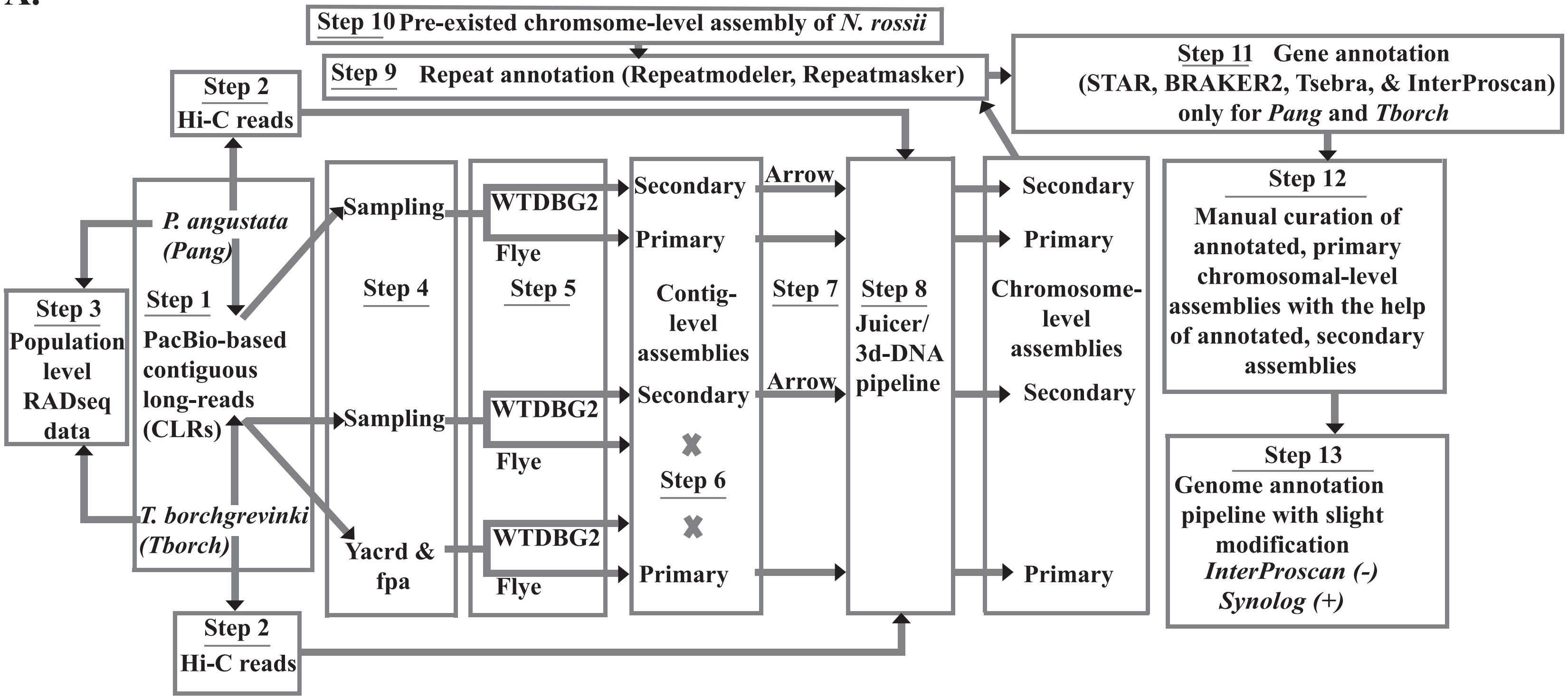

Step 10 Pre-existed chromsome-level assembly of *N. rossii*
Step 9 Repeat annotation (Repeatmodeler, Repeatmasker)
Step 11 Gene annotation (STAR, BRAKER2, Tsebra, & InterProscan) only for *Pang* and *Tborch*
Step 2 Hi-C reads
P. angustata (Pang)
Step 1 PacBio-based contiguous long-reads (CLRs)
T. borchgrevinki (Tborch)
Step 3 Population level RADseq data
Step 4
Sampling
Sampling
Yacrd & fpa
Step 5
WTDBG2
Flye
WTDBG2
Flye
WTDBG2
Flye
Secondary
Primary
Contig-level assemblies
Secondary
Step 6
Primary
Arrow
Step 7
Arrow
Step 8 Juicer/ 3d-DNA pipeline
Secondary
Primary
Chromosome-level assemblies
Secondary
Primary
Step 12 Manual curation of annotated, primary chromosomal-level assemblies with the help of annotated, secondary assemblies
Step 13 Genome annotation pipeline with slight modification *InterProscan (-)* *Synolog (+)*
Step 2 Hi-C reads

Figure 3.1 provides an overview of the methods implemented in this study. A) Steps 1, 2, and 3 involved PacBio-based long-read and Illumina-based Hi-C sequencing and population-level RADseq data collection for each species. Steps 4-13 encompassed the construction of contigs and scaffolding them into chromosome-level assemblies as well as the conduction of annotations, correction of structural errors, and re-annotation (for which *Interproscan (-)* was replaced with *Synolog (+)*). Also, it consisted of repeat annotation on the pre-existed chromosome-level assembly of *N. rossii*.

B.

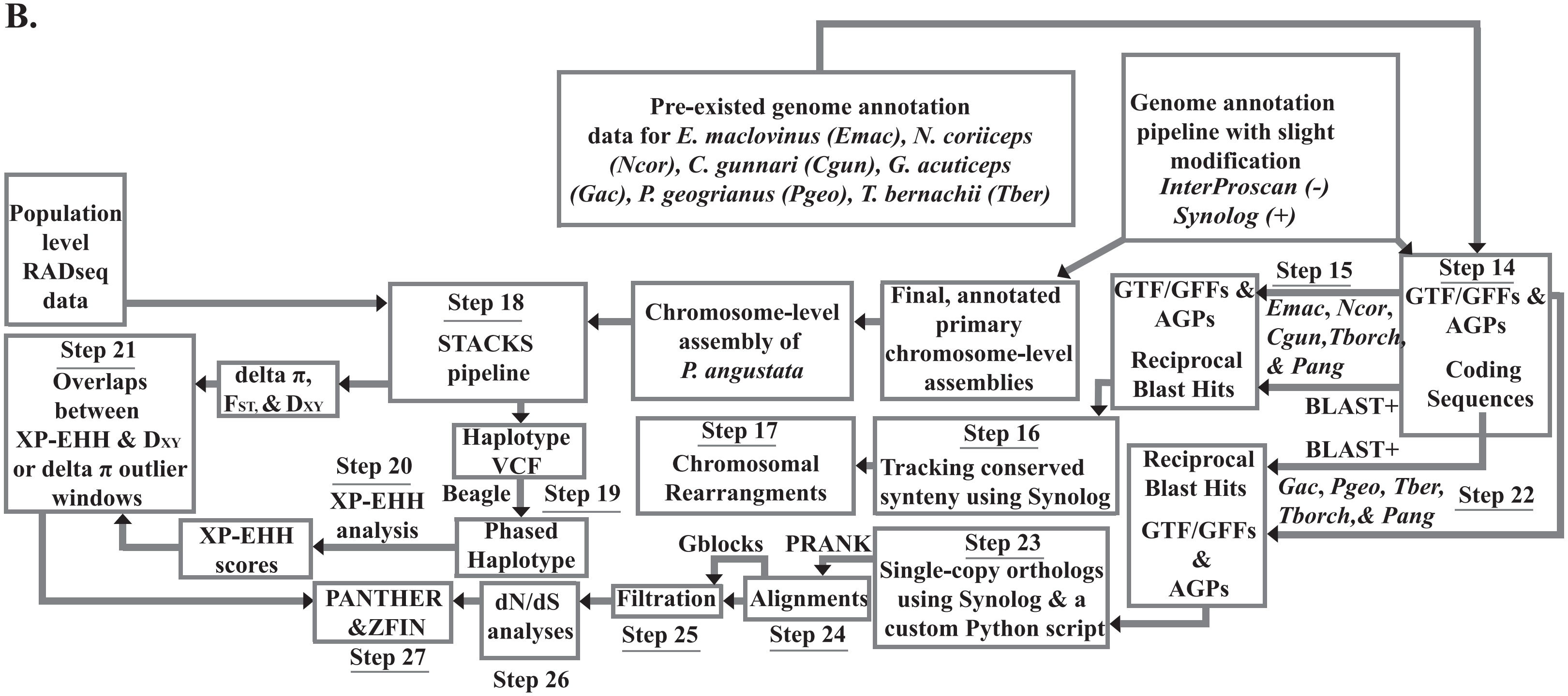

Pre-existed genome annotation data for E. maclovinus (Emac), N. coriiceps (Ncor), C. gunnari (Cgun), G. acuticeps (Gac), P. geogrianus (Pgeo), T. bernachii (Tber)
Genome annotation pipeline with slight modification InterProscan (-) Synolog (+)
Population level RADseq data
Step 18
STACKS pipeline
Chromosome-level assembly of P. angustata
Final, annotated primary chromosome-level assemblies
GTF/GFFs & AGPs
Reciprocal Blast Hits
Step 15
Emac, Ncor, Cgun,Tborch, & Pang
BLAST+
Step 14
GTF/GFFs & AGPs
Coding Sequences
Step 21
Overlaps between XP-EHH & DXY or delta π outlier windows
delta π, FST, & DXY
Haplotype VCF
Beagle
Step 19
Step 20
XP-EHH analysis
Phased Haplotype
XP-EHH scores
Step 17
Chromosomal Rearrangements
Step 16
Tracking conserved synteny using Synolog
Reciprocal Blast Hits
GTF/GFFs & AGPs
BLAST+
Gac, Pgeo, Tber, Tborch,& Pang
Step 22
Gblocks
PRANK
Step 23
Single-copy orthologs using Synolog & a custom Python script
PANTHER &ZFIN
Step 27
dN/dS analyses
Step 26
Filtration
Step 25
Alignments
Step 24

Figure 3.1 – Continued. B) Steps 14-17 included a procedure to obtain Reciprocal Blast Hits, GTF/GFF, and AGP files and identifying and characterizing chromosomal rearrangements specific to *P. angustata* using a conserved synteny approach. Except for *P. angustata* and *T. borchgrevinki*, data on genome annotations already existed for the remaining six fishes (*Emac*, *Ncor*, *Cgun*, *Gac*, *Pgeo*, and *Tber*). Steps 18-20 included F_{ST}, DXY, delta π-based genome scans, and XP-EHH-based linkage analyses with the population level RADseq data. Steps 22-26 entailed extraction of single-copy orthologs, their alignments, filtration, and both branch and branch-model based dN/dS analyses. Step 27 included combining the genes from dN/dS analyses and those from the overlapping windows between XP-EHH and a) D_{XY} and b) delta π outliers. Additionally, it involved analyzing genes for assessing biological functions using PANTHER and ZFIN databases.

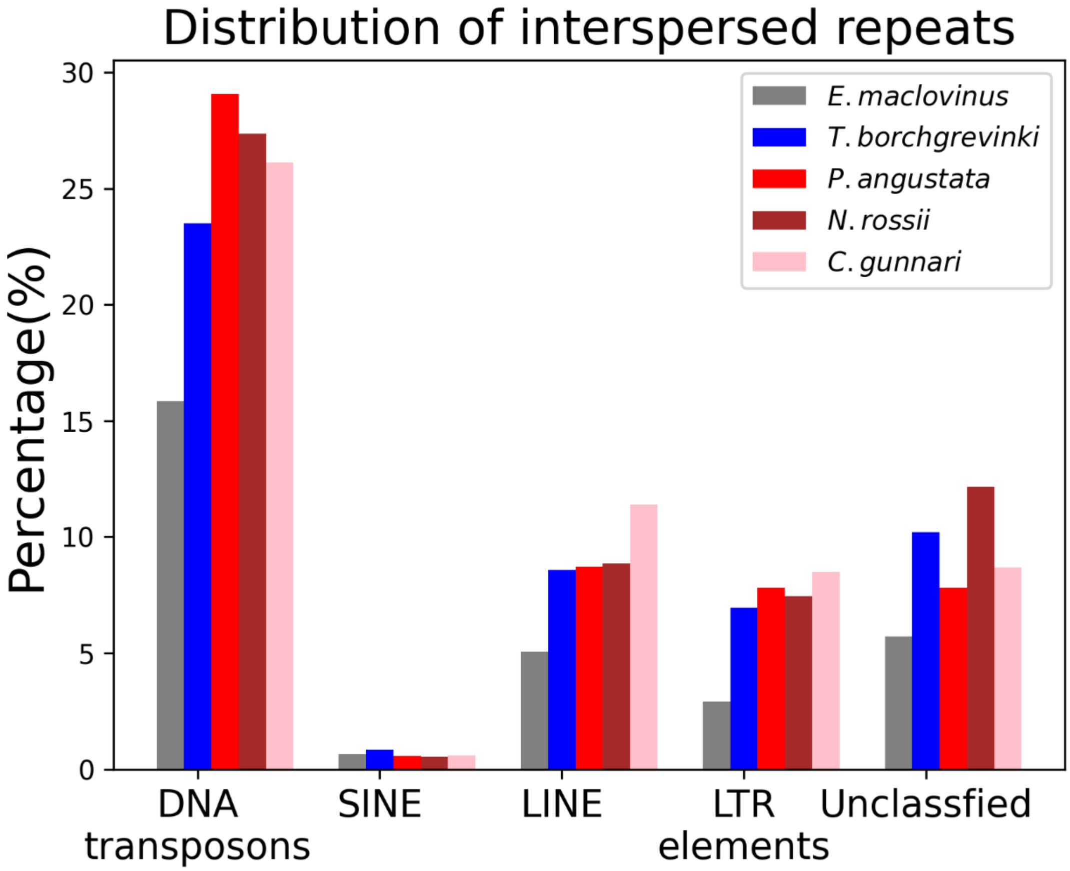


Figure 3.2 displays the percentage of interspersed repeats (DNA transposons, SINE, LINE, LTR, and Unclassified elements) in five notothenioids, including *Eleginops maclovinus, Trematomus borchgrevinki, Paranotothenia angustata*, *Notothenia rossii*, and *Champsocephalus gunnari. P. angustata* possesses a higher percentage of DNA transposons than the rest of the three species.

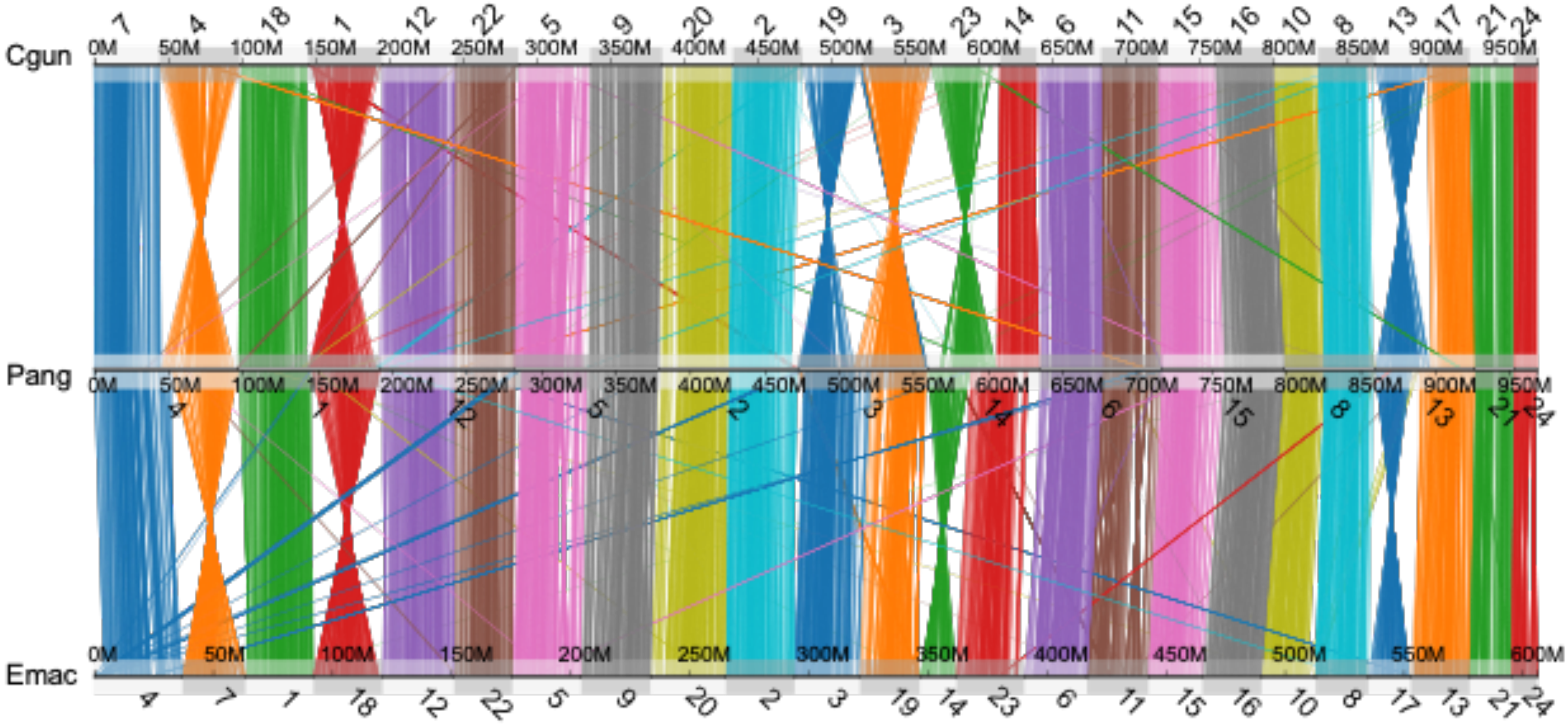


Figure 3.3 illustrates the pattern of conserved synteny among genomes of non-Antarctic and Antarctic notothenioid species. This figure exhibits genome-wide conserved synteny between genomes of *Paranotothenia angustata* (Pang; middle) and *Champsocephalus gunnari* (Cgun; top), as well as that between genomes of *P. angustata* and *Eleginops maclovinus* (Emac; bottom). Each line between any pair of genomes represents the orthologous gene between the corresponding species. The lines between any pair of genomes are color-coded according to the chromosome of their origin. This figure shows 1:2 and 1:1 relationships between chromosomes of *P. angustata* and their corresponding homologs in any other species. The chromosomes 21 & 24 of *P. angustata* show a 1:1 homologous relationship with chromosomes 21 & 24, respectively, of *C. gunnari* and *E. maclovinus*. Chromosomes 1, 2, 3, 4, 5, 6, 8, 12, 13, 14, and 15 of *P. angustata* exhibit 1:2 homologous relationship with chromosome pairs a) 1 & 18, b) 2 & 20, c) 3 & 19, d) 4 & 7, e) 5 & 9, f) 6 & 11, g) 8 & 10, h) 12 & 22, i) 13 & 17, j) 14 & 23, and k) 15 & 16, respectively, of *C. gunnari* as well as with those of *E. maclovinus*.

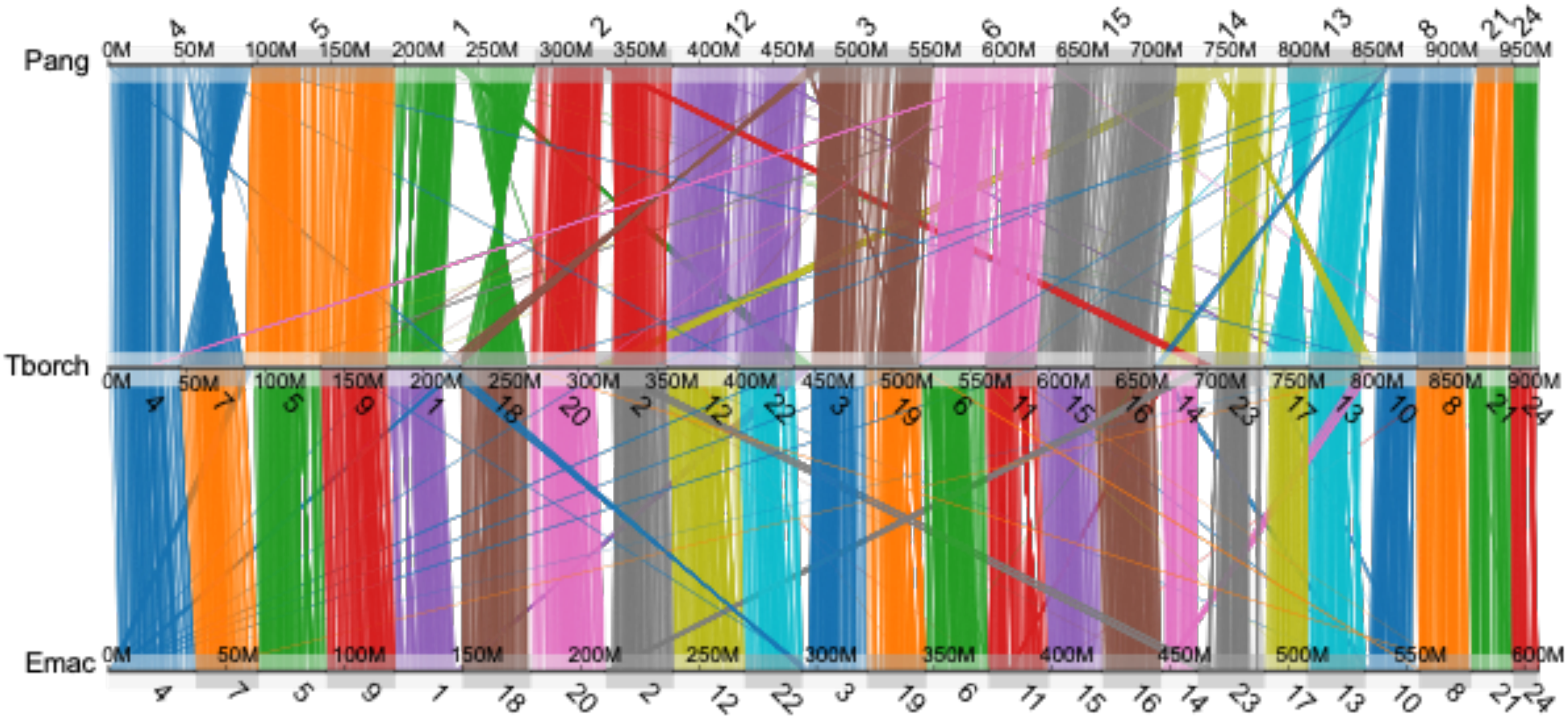


Figure 3.4 illustrates the pattern of conserved synteny among non-Antarctic and Antarctic notothenioid species. Each line between the chromosomes of any given species pair represents orthologous genes between the species. These lines are colored-coded according to the chromosome of origin. Specifically, this figure exhibits genome-wide conserved synteny between *Paranotothenia angustata* (Pang; top) and *Trematomus borchgrevinki* (Tborch; middle), as well as that between *T. borchgrevinki* and *Eleginops maclovinus* (Emac; bottom). Three different types of relationships, including 1:1, 1:2, and complex, are observed between the homologous chromosomes of Tborch and Pang and those of Tborch and Emac. For example, chromosomes 21 & 24 in Tborch displayed a 1:1 relationship with their corresponding homologous chromosomes 21 & 24 in Pang. Chromosome pairs a) 4 & 7, b) 5 & 9, c) 6 & 11, d) 15 & 16, e) 13 & 17, f) 8 & 10 of Tborch exhibit 2:1 relationship with homologous chromosomes 4, 5, 6, 15, 13, & 8 in Pang, respectively. While a single chromosome 14 and most of the portion of chromosome 23 in Tborch mapped to chromosome 14 in Pang, the remaining small portion of chromosome 23 in Tborch had a homologous region in chromosome 2 of Pang, exhibiting complex relationship between chromosomes of *T. borchgrevinki* and *P. angustata*.

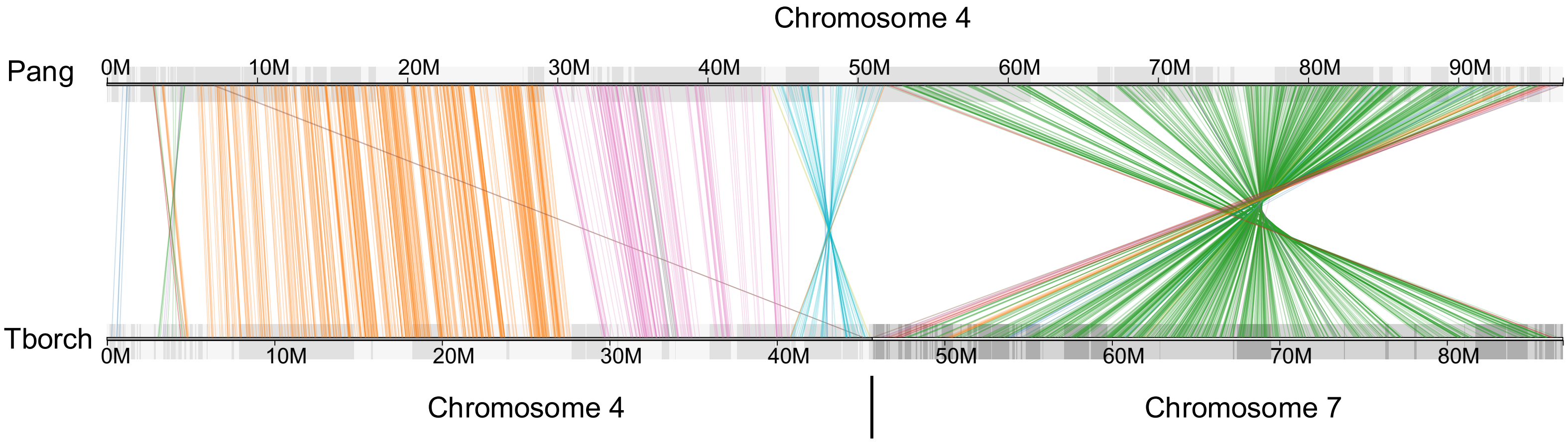
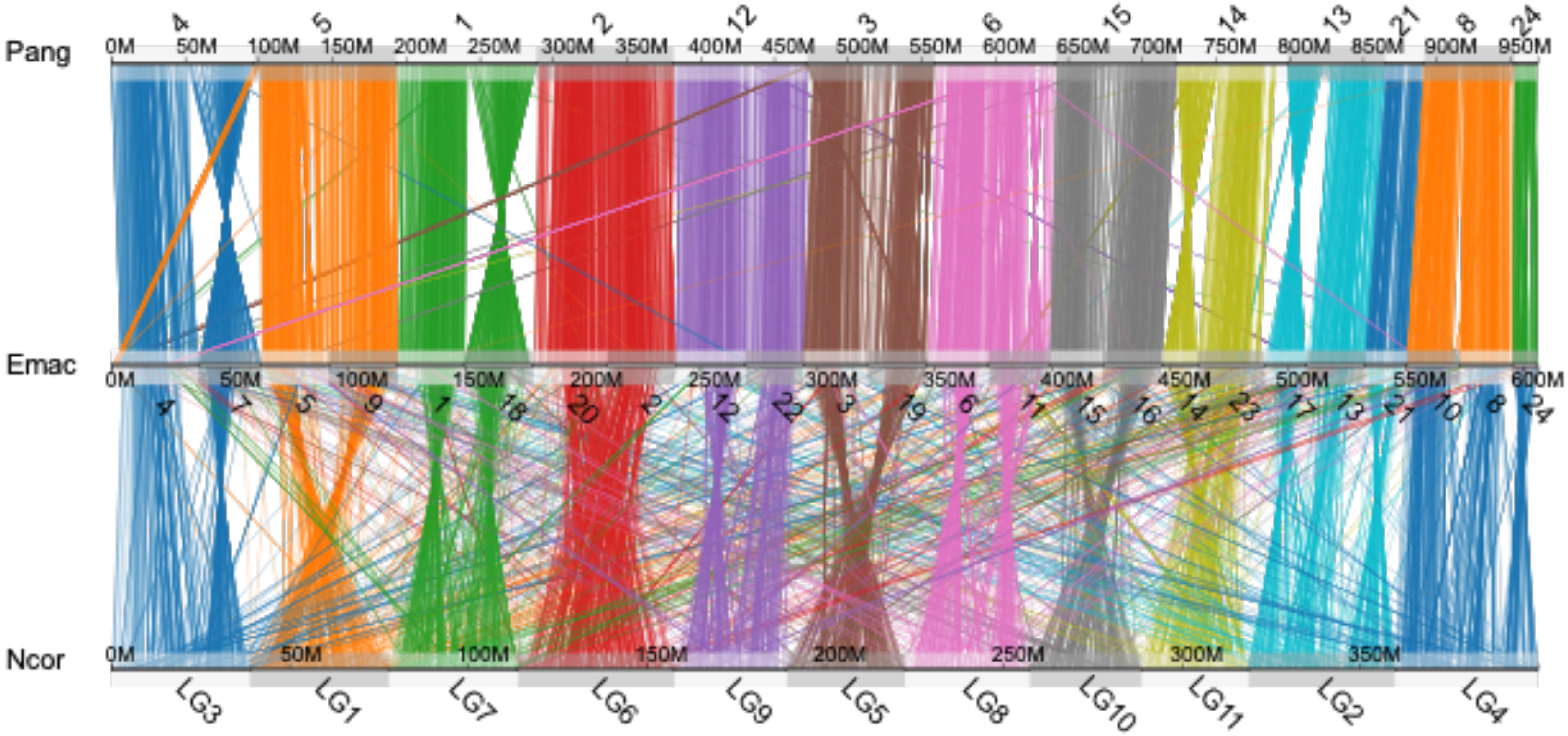


Figure 3.5 This figure shows the conserved synteny between chromosome 4 of *P. angustata* (Pang; top) and chromosomes 4 (bottom left) and 7 (bottom right) of *T. borchgrevinki* (Tborch; bottom), as well as shows an example of evidence of chromosomal fusion in *P. angustata*. The lines between chromosomes of two species are colored-coded based on conserved synteny between genomic regions. The vertical black line demarcates the boundary between chromosomes 4 and 7 of *T. borchgrevinki*.



.

Figure 3.6 illustrates the pattern of conserved synteny among genomes of non-Antarctic and Antarctic notothenioid species. This figure exhibits genome-wide conserved synteny between genomes of *Eleginops maclovinus* (Emac; middle) and *Paranotothenia angusta* (Pang; top), as well as that between genomes of *E. maclovinus* and *Notothenia coriiceps* (Ncor; bottom). Each line between any pair of genomes represents the orthologous gene between the corresponding species. The lines between any pair of genomes are color-coded according to the chromosome of their origin. This plot also exhibits a difference in the pattern of orientation of chromosomes of ancestral proxy (*E. maclovinus*) when mapped to chromosomes of *P. angustata* versus *N. coriiceps*. For example, chromosomes 5 and 9 of *E. maclovinus* mapped to chromosome 5 in tandem but without change in the orientation, unlike to when they mapped to chromosome LG1 of *N. coriiceps*.

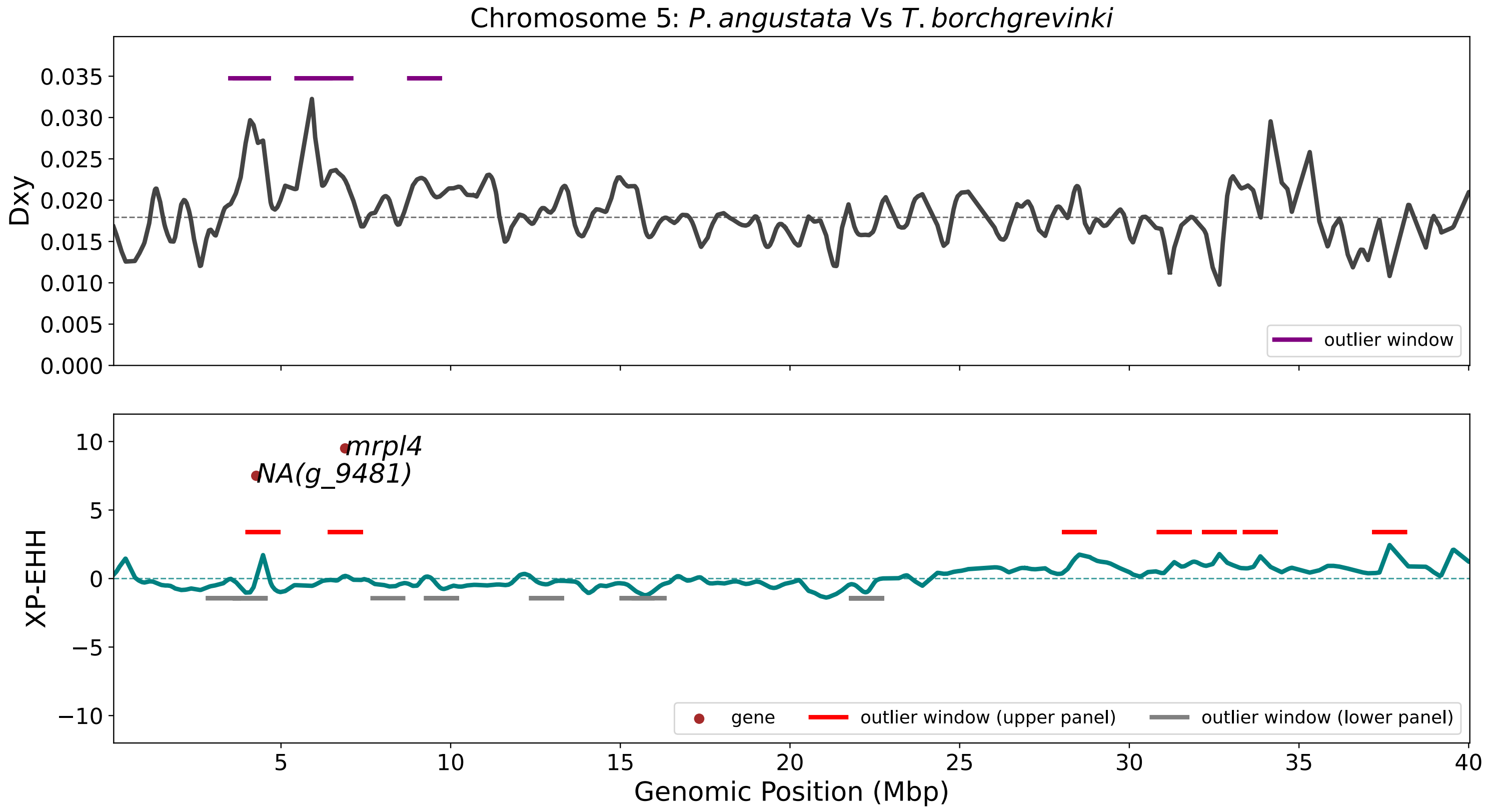


Figure 3.7 The figure shows an example of patterns of kernel-smoothed genetic divergence (D_{XY}) (top subplot) and the cross-population extended haplotype homozygosity (XP-EHH) scores (bottom subplot) for genomic positions between 0 and 40 megabase pairs (Mbp) in chromosome 5. For XP-EHH analysis, *P. angustata* and *T. borchgrevinki* are the target and reference, respectively. In the top subplot, purple horizontal solid lines denote the DXY outlier windows. The black-colored dashed line in the genetic divergence-based plot represents the mean $D_{XY.}$ In the bottom subplot, the significant signals of positive selection specific to *P. angustata* and *T. borchgrevinki* are represented by outlier windows (red and grey horizontal solid lines, respectively). In addition, this figure displays the genes (*NA(with gene identifier g_9481)* and *mrpl4,* denoted by brown dots) that a) are located within the overlapping region between D_{XY} and XP-EHH outlier windows and b) either contain or reside nearest to the XP-EHH outliers. The green-colored dash line in the XP-EHH-based plot separates the upper and lower panels, which consist of positive and negative scores, respectively.

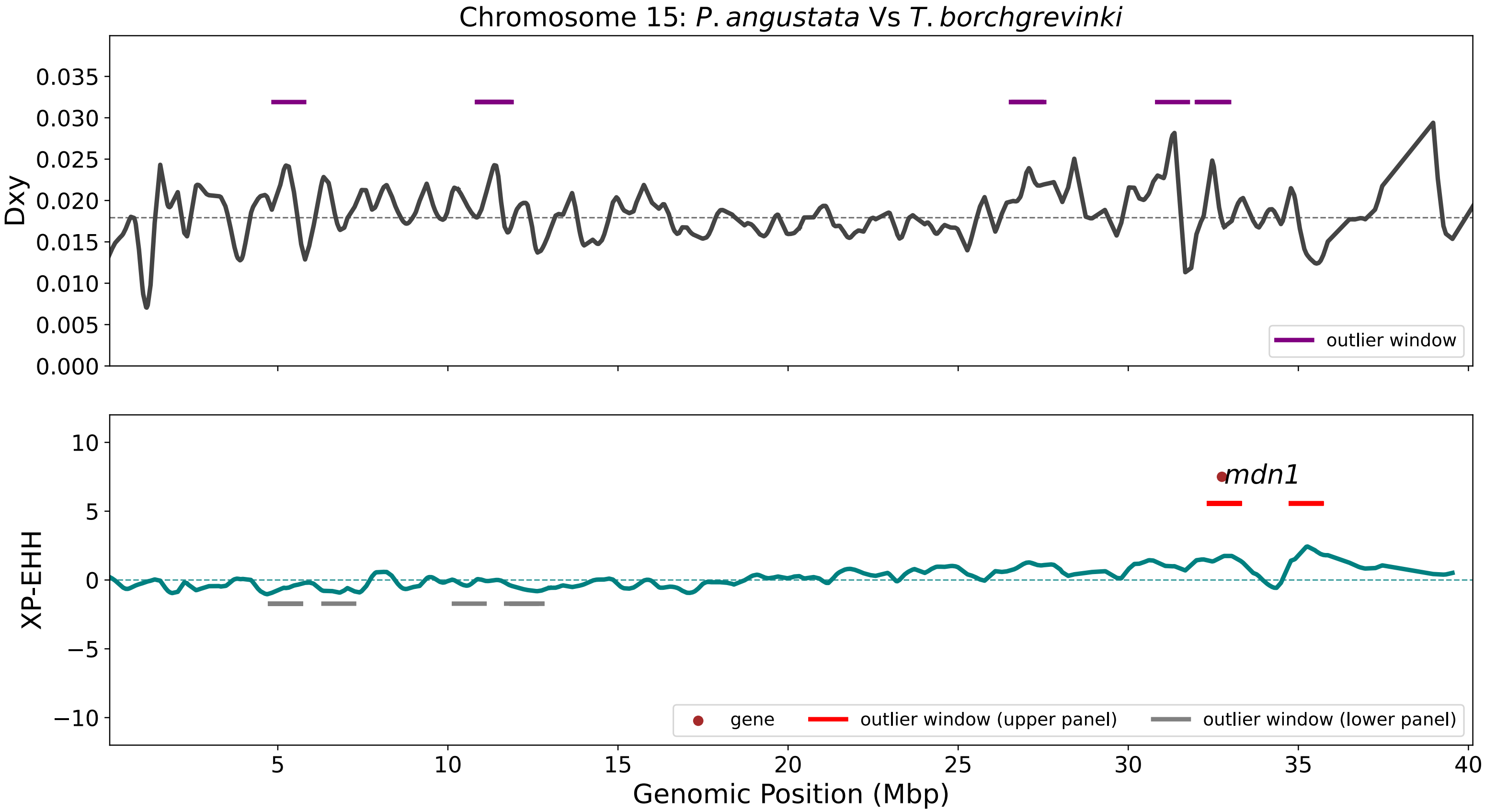


Figure 3.8 The figure shows an example of patterns of kernel-smoothed genetic divergence (D_{XY}) (top subplot) and the cross-population extended haplotype homozygosity (XP-EHH) scores (bottom subplot) for genomic positions between 0 and 40 megabase pairs (Mbp) in chromosome 15. For XP-EHH analysis, *P. angustata* and *T. borchgrevinki* are the target and reference, respectively. In the top subplot, purple horizontal solid lines denote the DXY outlier windows. The black-colored dashed line in the genetic divergence-based plot represents the mean D_{XY}. In the bottom subplot, the significant signals of positive selection specific to *P. angustata* and *T. borchgrevinki* are represented by outlier windows (red and grey horizontal solid lines, respectively). In addition, this figure displays the gene *mdn1* (denoted by brown dots) that a) resides within the overlapping region between D_{XY} and XP-EHH outlier windows and b) contains the XP-EHH outliers. The green-colored dash line in the XP-EHH-based plot separates the upper and lower panels, which consist of positive and negative scores, respectively.

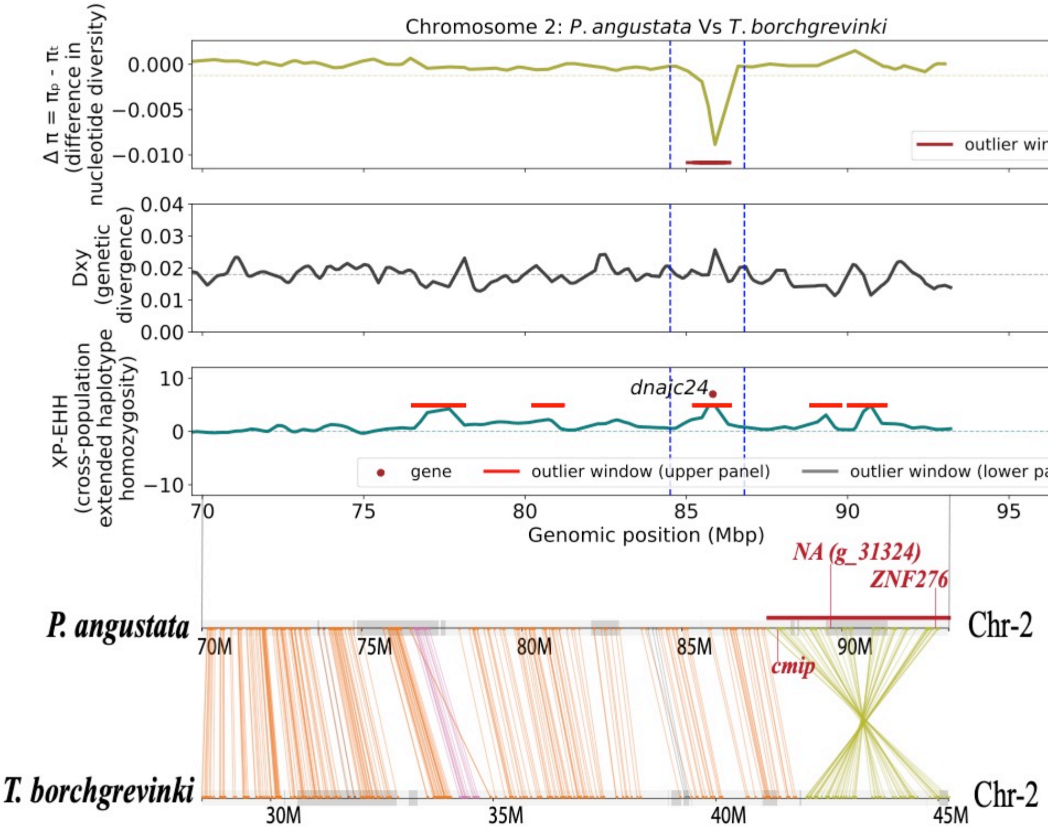


Figure 3.9 This figure illustrates an example of patterns of the difference (Δ) in nucleotide diversity (π), the genetic divergence (D_{XY}), XP-EHH scores, and a local conserved synteny between *P. angustata* and *T. borchgrevinki* on the region beyond 70 megabase pairs (Mbp) genomic position on chromosome 2. Specifically, the first subplot displays the distribution of Δ π estimated by subtracting the kernel-smoothed nucleotide diversity of *T. borchgrevinki* (π_t) from *P. angustata* (π_p) (y-axis). The olive-colored dashed horizontal line represents the bottom 0.5th percentile of Δ π. The window of the variant site at which Δ π is less than a threshold is shown as a brown, solid horizontal line. The second subplot exhibits the distribution of kernel-smoothed D_{XY} between the species (y-axis). The black-colored dashed line represents the genome-wide mean D_{XY}. The third subplot demonstrates the distribution of kernel-smoothed XP-EHH scores, and the red solid horizontal lines indicate outlier windows. The dashed, blue verticle lines indicate boundaries for a genomic region within which the overlap between Δ π and XP-EHH outlier windows and XP-EHH outliers are contained in the *dnajc24* gene. For clarity, a thin, solid black line connects the third to the fourth subplots. The fourth subplot shows the local conserved synteny between the two species from genomic region 70 to 93.33 megabase pairs (Mbp) on chromosome 2 (Chr-2), containing the putative *P. angustata*-specific inversion (marked by dark red solid, horizontal block spanning genomic positions 87.66-93.02 Mbs). The inversion contains three "*dN/dS*" candidates: *cmip*, *NA* (with gene id *g_31324*), and *ZNF276*. The solid verticle lines of dark red and blue above the conserved synteny plot represent the positions of the genes. The plots demonstrate the coincidence between XP-EHH outlier windows and inversion specific to *P. angsustata*.

CHAPTER 4: EXAMINATION OF GENEALOGICAL TREES WITHIN AND BETWEEN *PARANOTOTHENIA ANGUSTATA* AND *TREMATOMUS BORCHGREVINKI*

ABSTRACT

Understanding when traits evolved and whether these time periods coincide with known geological events or speciation time is critical in understanding past selection. Here, I focused on two fish species of the Antarctic notothenioid clade, *Trematomus borchgrevinki* (a cold-specialized species) and *Paranotothenia angustata* (a secondarily temperate species that evolved from an Antarctic ancestor) to present data on times of origin of potential adaptations of *P. angustata* using gene trees. In this study, most gene trees, including those near or within candidate loci or those contained within structural variations in *P. angustata*, exhibited reciprocally monophyletic patterns between species. The average time to the most recent common ancestor (T_{MRCA}) of alleles between species appears to be lower than the time required for a genome-wide reciprocally monophyletic pattern to form under neutrality. Species-specific selection may partly explain the observed pattern as it accelerates lineage sorting. I found no local distinct peaks of inter-species T_{MRCA}, suggesting that adaptations of *P. angustata* evolved after the divergence of *P. angusta* and *T. borchgrevinki*. An intra-species T_{MRCA} outlier was found within a candidate inversion, but none was found within my candidate loci. Also, intra-species T_{MRCA} distributions within and outside candidate loci (exhibiting accelerated molecular evolution) and structural variations showed no significant difference, supporting a substantial contribution of *de novo* mutations in the temperate adaptation of *P. angustata*. Intra-species T_{MRCA} outliers, however, were identified within translocations specific to *T. borchgrevinki* suggesting structural changes contributed to adaptations within *T. borchgrevinki*.

INTRODUCTION

Advancements in genomics have enabled scientists to use genetic variation across the genome to identify regions under selection for a given species (Martinez Barrio *et al.* 2016) and to estimate the age at which two species diverged (Tiley *et al.* 2023). This has allowed researchers to correlate the timing of genetic adaptations with known historical environmental changes, geological events, or speciation times. Such information is important to gain insights into the past effect of selection. Patterns of genealogies and characteristics of gene trees are useful for finding potential adaptive loci and their time of origin (Dopman *et al.* 2005; Nelson and Cresko 2018). For example, a non-recombining, homologous genomic block between two populations under divergent selection would exhibit a reciprocal monophyletic relationship (i.e., the haplotypes of the individuals from one population and those of another population would each form distinct monophyletic clades) (Dopman *et al.* 2005; Nelson and Cresko 2018). This is because a reciprocally monophyletic pattern takes a long time to accumulate (i.e., 9-12 N_e generations, where N_e is the historical effective population size after the initial divergence) under neutrality, however, directional selection can accelerate the process (Hudson and Coyne 2002). Estimating the time to the most recent common ancestor (T_{MRCA}) for a given gene tree with a reciprocally monophyletic pattern can therefore provide insights into the time of origin of adaptive loci (Nelson and Cresko 2018).

Genetic adaptation in organisms can arise through pre-existing, ancestral standing genetic variation, or via *de novo* mutation in response to environmental changes and corresponding selection pressures (Chan *et al.* 2010; Nelson and Cresko 2018; Lai *et al.* 2019). However, the extent to which standing genetic variation and *de novo* mutation have contributed to the adaptation of organisms is a subject of ongoing research (reviewed in Bomblies and Peichel

2022). T_{MRCA} is an informative measure in understanding the contributions of standing genetic variation and *de novo* mutations in the adaptation of organisms. For example, when an adaptive process results in the use of standing genetic variation, the interspecific T_{MRCA} for the genomic region would be higher than the coalescence time of the two taxa (Nelson and Cresko 2018). Moreover, when *de novo* mutations enable a taxon to adapt to a novel environment, then intuitively, the derived, adaptive loci are expected to be taxon-specific, show a monophyletic pattern, and have a T_{MRCA} shorter than the coalescence time of the taxon from its sister taxon.

In this study, I focus on the Antarctic notothenioids, a group of teleost fish primarily found in the Southern Ocean surrounding Antarctica which remain cold year-round (Eastman 1993; Beers and Jayasundara 2015). The Antarctic notothenioids evolved from a non-Antarctic, temperate ancestor approximately 10.7 million years ago (MYA) (Bista *et al.* 2023). Most notohenioids possess Anti-Freeze Glycoproteins (AFGPs) (e.g., *Trematomus borchgrevinki*) as a key adaptation, allowing them to avoid freezing (DeVries 1988). These cold-water inhabiting notothenioids are stenothermal and have become cold-specialized. For example, *T. borchgrevinki* suffers from thermal heat stress at ~6 ^{o}C and incurs oxidative damage at higher temperatures (Almroth *et al.* 2015). Remarkably, a few species within the Antarctic clade of notothenioids are secondarily temperate, meaning that they evolved from an ancestor that originated in the cold waters of Antarctica but later migrated and re-adapted to warmer waters in temperate regions, such as the coastal waters of New Zealand, Australia, and South America (Eastman 1993; Beers and Jayasundara 2015). *Paranotothenia angustata* is a secondarily temperate notothenioid which lives in the temperate waters surrounding New Zealand (ranging from 6-8 to 15-18 ^{o}C (reviewed in Lau *et al.* 2001)). While *P. angustata* exhibits a lower critical thermal maximum than the basal New Zealand notothenioid *Bovichtus variegatus*, its heat tolerance capacity is higher than the cold-specialized notothenioid, *Notothenia coriiceps* (Bilyk and Devries 2012).

Chapter 3 identified candidate loci and structural variants that might contribute to the adaptation of *P. angustata* to temperate environments. Here, I present data on the time of origin of these adaptations. To answer this question, I proposed to test two complementary hypotheses. First, I tested whether there was rapid evolution in *P. angustata* (due to a novel environment) compared to *T. borchgrevinki* (a close relative that remained in the cold waters of Antarctica). Second, I tested whether loci contributing to adaptation in temperate conditions arose from standing genetic variation or *de* novo mutations that occurred after the split between these two species. I found that, in general, gene trees built from the haplotypes at each orthologous RAD locus between the species were reciprocally monophyletic, reflecting the presence of complete lineage sorting and high population structure between the species throughout the genome. Hence, distinguishing genomic regions under divergent selection versus neutrality was not possible.

However, the accelerated reciprocal monophyly for genome-wide gene trees between the species, as well as the presence of such pattern within previously identified candidate loci and structural variation, suggests that strong divergent selection may have contributed to the observed pattern. Additionally, the fact that (a) none of the estimated T_{MRCA} between the two species was greater than the assumed divergence time between the two species and b) no distinct peak(s) of inter-species T_{MRCA} were found suggests the origination of the secondarily temperate adaptations in *P. angustata* may have occurred after the divergence of two species.

While I observed only one T_{MRCA} outlier within *P. angustata*-specific structural variation, none were associated with candidate genes or regions under selection. Additionally, I found a highly similar distribution of intra-species T_{MRCA} within and outside of candidate loci and structural variation, indicating a larger contribution of *de novo* mutations than standing variation in adaptations of *P. angustata*. Between populations of *T. borchgrevinki*, I found pervasive incomplete lineage sorting, indicating high gene flow. I identified a cluster of intra-species T_{MRCA} outliers within translocations specific to *T. borchgrevinki*.

MATERIALS AND METHODS

Historical effective population size (N_e) inference

For both *Paranotothenia angustata* and *Trematomus borchgrevinki*, I inferred the trajectory of historical effective population sizes using pairwise sequentially Markovian coalescent (PSMC) software (Li and Durbin 2011). First, I conducted self-error correction of raw PacBio continuous long reads (CLRs) using Canu (v2.2; Koren *et al.* 2017). Second, the corrected reads were aligned to their corresponding genomes using minimap (v2.24; Li 2018). I performed the alignment even though the reference genomes had a small number of known errors related to the orientation and location of contig/scaffolds. These errors were not expected to affect the overall demographic inference based on the coalescence approach. Third, the alignments were sorted with samtools (v1.2; Li *et al.* 2009). Fourth, each site on the sorted aligned reads was genotyped using samtools mpileup, and subsequently, consensus sequences were produced from genotyped reads using bcftools (v1.12; Danecek *et al.* 2021). The consensus sequences were converted to FASTQ format using vcfutils.pl (a Perl script in the samtools suite). The FASTQ file was converted to FASTA format with fq2psmcfa (a PSMC utility).

Next, I used the FASTA file as input on PSMC (v0.6.5) by setting a maximum number of iterations (`-N`) to 25, maximum coalescence time (`-t`) to 15, and an initial diversity recombination ratio (`-r`) to 5, and an atomic interval (`-p`) to “1*12+25*2+4+6”. I re-ran PSMC with 100 bootstrap parameter (`-b`) settings to obtain a profile of pseudo-replicates. Moreover, I assumed a mutation rate per site per generation of 5.32x10^{-9} for *P. angustata* and 4.27x10^{-9} generation for *T. borchgrevinki*, along with a generation time of seven years for both species.

I assumed the generation time and mutation rate for both species based on estimates from closely related species. Specifically, *Notothenia rossii* and *N. coriiceps* are cold-specialized notothenioids but are closely related to *P. angustata*. For these species, juveniles can take seven years to reach sexual maturity, depending upon sex (Calì *et al.* 2017). Furthermore, female *T. borchgrevinki* older than six years had been reported to have exhibited signs of ovulation, but the exact age was unknown. Hence, I assumed the generation times for *P. angustata* and *T. borchgrevinki* to be seven years.

Daane *et al.* 2019 also estimated a substitution rate per base per year of 0.76x10^{-9} for *N. coriiceps* and 0.61x10^{-9} for *Trematomus scotia* (closely related species to *T. borchgrevinki*). After adjusting for generation, the estimates were 5.32x10^{-9} for *N. coriiceps* and 4.27x10^{-9} for *Trematomus scotia*. A recent study showed that the mean germline mutation rate per generation based on eight different fish species and 19 trios was 5.97x10^{-9} (95% confidence interval from 4.39x10^{-9} to 7.55x10^{-9}) (Bergeron *et al.* 2023). Also, the reported mutation rate per generation for one of the Antarctic notothenioids, *Champsocephalus aceratus,* was 3.28x10^{-9} (Kim *et al.* 2019). Hence, the assumed mutation rate per generation for this study is reasonable.

RADseq data analyses

A single-digest, *sbfI* RADseq library protocol (Baird *et al.* 2008) generates two adjacent stretches of DNA per restriction site along a given chromosome. Each adjacent DNA pair is known as sister RAD-tags or -haplotype pair (hereafter, each tag is referred to as RAD-haplotype). The library of RAD-haplotypes sampled across the genome can be prepared and sequenced (**Figure 4.1.A.i**) on a short-read sequencing platform like Illumina. Variants can be called on these RAD-haplotypes; however, their presence may vary (**Figure 4.1.A.ii**). The variants in the same RAD-haplotype are from the same DNA fragment, because of which they can be considered phased. However, variants between RAD-haplotype pairs cannot be deemed phased (**Figure 4.1.B**) if the genome of a species is not haploid. Consider a cut-site from diploid species, which would have a homologous pair. In general, two RAD-haplotype pairs are expected from a given homologous pair of cut-site (**Figure 4.1.A.i & ii**). Sequencing alone provides no information on which RAD-haplotype pair originated from the same cut-site or homologous site. The original phased state between the variants within the RAD-haplotype pair is hidden in the sequence data, and the phasing process is required to infer that state.

Establishing the accurate pairing of RAD-haplotypes originating from the same cut-sites, the length of RAD-haplotypes can be increased by merging the pairs and producing a single, longer locus (hereafter referred to as merged RAD-haplotypes pair) from each pair (**Figure 4.C**). Genotype data based on variants on each RAD-locus (i.e. either a RAD-haplotype or a merged RAD-haplotype pair) across samples (**Figure 4.D**) can be utilized to infer gene topology and most recent common ancestor (T_{MRCA}) per locus (**Figure 4.E**) along the genome using tree sequence analysis. To perform such analyses within and between *P. angustata* and *T. borchgrevinki*, I retrieved the pre-existed RADseq and genome data. These data were utilized to obtain alignments, which were categorized and processed for downstream analyses. This includes estimating contemporary effective population sizes, reconstructing gene trees, and the inference of genome-wide T_{MRCA} within and between the species using unmerged and merged RAD-haplotypes pairs.

Generation of RADseq alignments followed by their categorization and processing

I used the RADseq and genome data described in chapter 3 of this dissertation for both species. The data consisted of paired-end reads generated through sequencing the single-digest *SbfI* RAD-seq libraries (Baird *et al.* 2008), prepared separately for 71 *T. borchgrevinki* and 41 *P. angustata*. Of the 71 individuals of *T. borchgrevinki*, 53 were from McMurdo Station, and 18 were from Prydz Bay, located on the opposite side of Antarctica. All the individuals of *P. angustata* were sampled from a single location, Otago Harbor, South Island of New Zealand. I used the process_radtags module of Stacks (v2.60; Rochette *et al.* 2019) on RAD-seq data of both species for demultiplexing (to separate reads per sample), cleaning (to keep reads without ambiguous base) and filtering (to discard low quality reads), and rescuing (to save mutated cut sites and barcodes of reads whenever possible). Next, I performed the alignment of the retained reads in two ways. First, the reads from both species were aligned to the same reference genome of *P. angustata* using bwa-mem (v0.7.17; Li 2013). For clarity, these alignments were referred to as *Pang-Tborch-combo-align*. In the second strategy, the reads from the samples of each species were aligned to their corresponding reference genomes using bwa-mem (v0.7.17; Li 2013). I referred to the set of alignments for *P. angustata* as *Pang-align* and that for *T. borchgrevinki* as *Tborch-align.*

I established three distinct groups of alignments to perform specific downstream analyses by tailoring the total alignment data. The first group was a subset of *Pang-align* and *Tborch-align* sets. Specifically, the first group contained alignments from 36 *P. angustata* and 49 *T. borchgrevinki* (from McMurdo Station). The second group was a subset of the *Pang-Tborch-combo-align* set. It consisted of alignments from 36 *P. angustata* and 36 *T. borchgrevinki* (from McMurdo Station). The third group was again a subset of *Pang-align* and *Tborch-align* sets but contained a modified set of alignments for *T. borchgrevinki* compared to the first group. This group had the exact alignments of 36 *P. angustata*; however, it contained alignments of 32 *T. borchgrevinki* across two populations (i.e., 16 each from McMurdo Station and Prydz Bay). I sorted all alignments across the groups using Samtools (v1.12; (Li *et al.* 2009a). The sets of the sorted alignments obtained from the data in the first, second, and third groups were referred to as *GrpI*, *GrpII*, and *GrpIII,* respectively. Next, I removed PCR duplicates and built separate catalogs of genotyped RAD loci per species from *GrpI* and *III* datasets using the gstacks module of Stacks. However, I created a single catalog from *GrpII* data, which included loci from both species.

All of the catalogs were filtered using the populations module of Stacks. From each species-specific catalog derived from the *GrpI* dataset, I retained a) RAD loci present across 80% of individuals of species, b) one SNP (single nucleotide polymorphism) per locus, and c) variants with three minimum allele counts. I also exported i) the Gene transfer format (GTF) file (*populations.gtf*) containing genomic coordinates of each RAD loci and ii) the SNP-based Variant Call Format file (*populations.snps.vcf*) with genotype information for each site per sample. After the filtration process was applied to either the catalog developed from *GrpII* or the species-specific catalogs derived from *GrpIII*, I retained the loci present in 100% of samples per species and the variants with a minimum allele count of one. Moreover, I pruned unshared SNPs to reduce haplotype-wise missing data. Additionally, I exported the haplotype-based VCF file (*populations.haps.vcf*) and the FASTA file with consensus sequences (*populations.loci.fa*).

Estimating contemporary effective population size (N_e)

I utilized SMC++ (Terhorst *et al.* 2017) software with RADseq-based genome-wide genotypic data for both species to quantify the contemporary effective population size. To run

SMC++, I first retrieved species-specific SNP-based VCF (*populations.snps.vcf*) and the GTF (*populations.gtf*) files previously exported from the catalogs built from the *GrpI* dataset. Next, I performed two conversions: I) the GTF file was converted to a BED format file (with genomic coordinates in which RAD loci are absent) using a custom Python script, and II) the VCF with genotypes was changed into SMC++ input format by using vcf2smc module. Moreover, using a BED file, I masked sites without RAD loci and estimated the contemporary effective population size with SMC++ using genotypic data. Further, I generated replicates by applying standard 25 bootstraps. I assumed the same mutation rate per site per generation and generation times utilized in the prior PSMC-based analyses.

Constructing tree sequence with unmerged RAD-haplotypes among species

I retrieved the previously exported haplotype-based VCF (*populations.haps.vcf*) file corresponding to the catalog derived from the *GrpII* dataset. On this VCF, I implemented a custom Python script, stacks_haps_to_tsinfer.py, written by Rivera-Colón (2022) to infer the tree sequence. This script was previously developed and implemented on the RADseq data of icefishes to infer tree sequence (Rivera-Colón 2022). It uses the tsinfer software (Kelleher *et al.* 2019), which, in turn, employs functionalities from the tskit library for loading, evaluating, and manipulating tree sequences and applying methods for estimating genetic statistics.

Briefly, tsinfer determines the chronological order of when a mutation (derived allele) in each site of a haplotype evolved and uses the allele frequency as a proxy of its age. It iterates over all sites with derived alleles (youngest to oldest). It infers the ancestor's state around a given focal site in each iteration using a pattern of genetic variation among samples per site. This repetitive process generates putative ancestral haplotypes corresponding to genetic variation in the sampled sequences. Next, tsinfer compares ancestral haplotypes to relatively older ancestors and matches contemporary samples to inferred ancestors. Such comparisons allow tsinfer to determine the immediate ancestor for each segment of a given focal haplotype and to identify break points in the haplotype (if any) due to recombination. Finally, after inferring the path of inheritance of the segments along the length of all ancestral and sampled haplotypes, tsinfer generates a genealogical tree sequence spanning the genomic region. The sequence of trees with different topologies for the genomic region accounts for the recombination events between variants (Kelleher *et al.* 2019).

The stacks_haps_to_tsinfer.py performed the following tasks using *populations.haps.vcf* file. First, it removed loci (RAD-haplotypes) with the number of variant sites equal to or less than two. Next, it converted haplotype data per locus into the tsinfer format (**Figure 4.D**). The ancestral state for each variant site per locus was determined using parsimony based on allele frequency. Furthermore, for a given locus of samples, it created an empty object with tsinfer.SampleData() to hold the metadata. It filled the object with information that linked individuals to the populations using the SampleDate.add_population() method and defined individuals with their population code, name, and ploidy level using the SampleDate.add_individual() method. In addition, it iteratively added the variant site position and its corresponding data (the array of genotypes and that of ancestral and derived alleles) to the object using SampleDate.add_site() method. Finally, it implemented the metadata in tsinfer.infer() to infer the genealogy of each locus (**Figure 4.E**) and saved the individual tree sequences in a file using the dump() method. It also produced a summary table with detailed information on each locus, including the genomic coordinates, number and length of haplotypes, and span of each marginal tree (i.e., an individual tree in a given tree sequence reflecting recombination events).

Establishing tree sequence per species with phased and merged RAD-haplotype pairs

To obtain merged RAD-haplotype pairs for *P. angustata* and *T. borchgrevinki* independently, I retrieved the previously exported species-specific, haplotype-based VCF (*populations.haps.vcf*) and FASTA (*populations.loci.fa*) files corresponding to the catalogs derived from the *GrpIII* dataset. For each species, I parsed the VCF and FASTA files with a custom Python script, phase_rad_loci.py (Rivera-Colón 2022), to phase variants within RAD-haplotype pairs using PHASE (v.2.1.1; Stephens and Scheet 2005) and join each sister pair. Internally, the phase_rad_loci.py set PHASE parameters `-l 2` (to divide data into two consecutive loci), `-MR` (to utilize the recombination model), `-d 1` (to specify not to assume stepwise mutation for multiallelic loci), `-x 5` (to run the algorithm for five times in total) with 1,000 iterations and thinning intervals as well as 100 rounds of burn-in. The script only considered RAD haplotypes with adjacent pairs. Moreover, while default values were used for most parameters, `--min-phase_prob` was set to 0.9 to retain a haplotype exhibiting a 90% probability of being phased correctly. The program joined together RAD-haplotypes to their sister pairs per chromosome and produced one consensus sequence per merged RAD-haplotype pairs. Finally, the program recoded the alleles and coordinates of variants for each merged RAD-haplotypes pair per chromosome, along with the other locus-specific information (for example, the start positions and IDs of loci). The recoded information was stored in a new VCF file (hereafter *merged_haps.vcf*). Next, I implemented the stacks_haps_to_tsinfer.py script on *merged_haps.vcf* file. The script removed merged RAD-haplotype pairs with variant sites equal to or less than 2. Subsequently, it generated tree sequence files from merged RAD-haplotype pairs and stored them in a directory. It also produced a summary table with detailed information on each locus, including genomic coordinates, the number and length of haplotypes, and the span of each marginal tree.

Inferring genome-wide patterns of genealogical nearest neighbour (GNN) and time to the most recent common ancestor (T_{MRCA})

The Genealogical Nearest Neighbour (GNN, Kelleher *et al.* 2019) is a topology-based statistic that can be implemented on a tree sequence using tsinfer software to assess how haplotypes of the same population are related to each other compared to those from another population(s) in a given gene tree. In the GNN analysis, the child nodes of a gene tree represent the haplotypes of sampled individuals. Additionally, a reference set consists of an array of sets of all haplotypes from different populations. The nearest neighbours of the focal child node or haplotype are determined in two steps. First, the focal haplotype's immediate ancestor (i.e., parental haplotype) is identified. Second, other haplotypes descended from the same ancestor are detected and considered the nearest neighbors of the focal haplotype. The GNN estimates of a focal haplotype represent the proportion of the nearest neighbors from each population in the reference set. These estimates are calculated for each child node one at a time. This process forms a matrix of all-populations-versus-all-haplotypes in a gene tree. Each entry in the matrix, corresponding to a specific population (e.g., X) and haplotype (e.g., A), indicates the GNN estimate of the particular haplotype (e.g., A) based on a proportion of the nearest neighbours from the specific population (e.g., X) to the haplotype (e.g., A).

Moreover, this matrix could be condensed into an all-populations-versus-all-populations matrix by summarizing the GNN estimates of all haplotypes from the same population with respect to each population in the reference set. These GNN estimates could be summarized as mean, median, and standard deviation. For each population, the average GNN per locus across the genome can provide insights into the pattern of reciprocal monophyly between populations, population structure, and lineage sorting. For example, at a specific locus, if the mean GNN for haplotypes of (a) Population A relative to those of Population A is equal to 1, and (b) Population B relative to those of Population B is also equal to 1, then it suggests that all nearest neighbors of haplotypes in Population A are from Population A, and similarly, all nearest neighbors of haplotypes in Population B are from Population B. This implies that the gene tree generated from such a locus should exhibit a reciprocally monophyletic pattern, where individuals from Population A and Population B form distinct monophyletic clades.

The T_{MRCA} within and between species could be computed from tree sequence data using software such as tsdate (Wohns *et al.* 2022), which implements the approximate Bayesian method to infer the ages of nodes. For a given tree sequence, tsdate generates a prior distribution of age (with mean and variance from conditional coalescent approach) per node based on the number of tips that have descended from the focal node. By default, the mean and variance of the prior distribution per node are fitted to a lognormal distribution for approximation. To infer the ages of nodes (hidden states), the priors are updated by tsdate, using an inside-outside algorithm (a belief propagation approach) based on a Hidden Markov Model. The algorithm traverses tree sequences from contemporary sampled nodes to its MRCA ("inside-pass") and updates the prior estimate of the age of each node. This update is based on the summation of likelihood (with Poisson distribution) for the number of mutations observed on the edge from the focal node to its child node (at a given time interval, span of edge, and population-scale mutation rate). Then, the algorithm proceeds from the root towards sampled nodes ("outside-pass"). Next, it estimates the final posterior of the child's age based on the parent's updated prior, which is not accounted for during the "inside pass" step.

For unmerged RAD-haplotype data, I implemented the run_ts_statistics.py script, written by Rivera-Colón (2022), to estimate genome-wide GNN per species and T_{MRCA} within and between species. Specifically, the script was applied to the tree sequence files and the summary table previously produced by stacks_haps_to_tsinfer.py. I ran the run_ts_statistics.py script, allowing a maximum number of two subtrees. I set the effective population size parameter to 41,411 (average of contemporary effective population size of the two species) and the mutation rate parameter to $4.7X10^{-9}$ (average mutation rate of the two species).

For merged RAD-haplotype pairs, I implemented the same run_ts_statistics.py script on species-specific data independently to estimate genome-wide a) T_{MRCA} within *P. angustata* and *T. borchgrevinki*, b) GNN per population of *T. borchgrevinki*, and c) T_{MRCA} within and between populations of *T. borchgrevinki*. Specifically, I utilized the script on the tree sequence files and the summary table previously generated by stacks_haps_to_tsinfer.py. I set the parameters to allow a maximum of two subtrees for a given RAD-locus. I specified the contemporary effective population sizes as 36,778 for *P. angustata* and 46,046 for *T. borchgrevinki*. I assigned the mutation rates as $5.32x10^{-9}$ for *P. angustata* and $4.27x10^{-9}$ for *T. borchgrevinki*.

Next, I kernel-smoothed GNN and T_{MRCA} estimates across analyses. Next, I plotted and assessed the distribution of intra-species smoothed T_{MRCA}. I implemented the interquartile range method for each species to find potential outliers and their locations in the genome because the distribution was noticeably skewed at the upper tail. I considered the standard upper bound or threshold, i.e., the sum of the third quartile and 1.5 times the difference between the distribution's third and first quartiles. The T_{MRCA} was considered an outlier when its value exceeded the threshold. I also compared the distribution of intra-specific T_{MRCA} in two ways. I compared the distributions of intra-species T_{MRCA} for RAD-loci found a) within and outside dN/dS genes, and

b) within and outside structural variation specific to *P. angustata* (previously identified in the chapter 3 of this dissertation).

RESULTS

Ancient and recent demographic history (change in N_e over time)

PSMC infers past changes in N_e by analyzing heterozygous site distribution (within the unphased genome of a single diploid individual) using coalescent hidden-Markov model (HMM) (reviewed in Webster *et al.* 2023). Based on PSMC analysis, I found that the highest historical N_e for *P. angustata* (between 400 and 500K) and *T. borchgrevinki* (between 700 and 800K) was between 20 and 10 MYA (**Figure 4.2**) through population expansion. However, the N_e for both species continuously dropped and reached tens of thousands between 0.3 and 0.1 MYA. Notably, the population contraction scale from 10 to 0.1 MYA was lower for *P. angustata* compared to *T. borchgrevinki*. During the period from 0.1 to 0.01 MYA, the N_e for *P. angustata* stabilized at around 60K, whereas that for *T. borchgrevinki* sharply expanded and stabilized to a size between 400 and 500K.

SMC++ is an extension of PSMC but it additionally incorportates analysis of the site-frequency spectrum and can take multiple unphased genomes. Compared to PSMC, SMC++ is more informative for recent demographic changes (reviewed in Moorjani and Hellenthal 2023; Webster *et al.* 2023). Based on SMC++ analysis, the contemporary N_e for *P. angustata* and *T. borchgrevinki* overall showed a pattern of population expansion from 1 MYA to the present. The mean contemporary N_e for *P. angustata* and *T. borchgrevinki* was 36,778 and 46,046, respectively (**Table 4.1; Figure 4.3**). Despite the difference in the mean contemporary N_e between the species, the overall population size remained constant from the recent past (i.e., 0.01 MYA) until the present (**Figure 4.3**).

However, I observed a difference in the number of individuals constituting N_e based on SMC++ and PSMC. At one MYA, SMC++ indicated that N_e for both species consisted of only a few individuals, whereas PSMC estimated an N_e of these species to be in the thousands. In the recent past, for *P. angustata*, the N_e based on SMC++ (36K) was about half of that based on PSMC (around 60K). For *T. borchgrevinki*, the SMC++-based N_e estimate (46K) was about an order of magnitude lower than the PSMC-based N_e (400K).

High genome-wide genealogical nearest neighbour (GNN) statistics for each species suggested the presence of gene trees with reciprocally monophyletic pattern across the genome

I obtained 24,079 unmerged, orthologous RAD-loci between *P. angustata* and *T. borchgrevinki*, and the haplotype data from 23,051 of those loci were converted to tsinfer format. I retained 20,564 of 23,051 loci after filtering 2,487 loci (each of which generated more than two gene trees). Out of 20,564 retained loci, 17,920 produced one gene tree per locus, whereas 2,644 generated two gene trees per locus due to recombination events (i.e., total of 23,208 gene trees). These 20,564 loci and 23,208 gene trees were used for GNN and T_{MRCA} analyses. The mean smoothed-kernel GNN was 0.97 for *P. angustata* and 0.96 for *T. borchgrevinki*, indicating that 97% and 96% of the genealogical nearest neighbours of a given haplotype of *P. angustata* and *T. borchgrevinki* belonged to the same species. While variation in kernel-smoothed mean GNN existed throughout the genome, in some instances, it drastically lowered the mean, as seen in the regions between 45 and 46 Mb on chromosome 4 (**Figure 4.4**). I also noticed that such drastic changes in GNN can occur when the root has a polytomy (i.e., more than two descendants). However, the kernel-smoothed GNN was generally considered high because the average unsmoothed GNN for both species was one (the highest possible value) in gene trees of 17,282 (i.e., 84.04%) of 20,564 loci. These results indicate that complete or near complete lineage sorting has generated reciprocally monophyletic patterns in most gene trees across the genome.

Of 17,282 loci (that generated gene trees with reciprocally monophyletic patterns), I found 211 were within 92 dN/dS genes (previously reported to be potentially under selection in chapter 3 of this dissertation). Some of these dN/dS genes were related to previously reported potential secondarily temperate adaptations of *P. angustata*, including circadian rhythm (*clocka*), red blood cell differentiation and development (*rasa3, and numb*), vision (*ift172* and *itga5*), mitochondria (*ndufv3*). A couple of those genes (*il16* and *si:dkey-106g10.7*) located within *P. angustata*-specific inversions also contained RAD-loci used in constructing gene trees. I found that 451 of those 17,282 loci were also within all the structural variations reported as potentially specific to *P. angustata*.

No distinct peaks of inter-species coalescence time to the most recent common ancestor (T_{MRCA})

Based on unmerged RAD-loci, the mean kernel-smoothed T_{MRCA} (in generations) was about 43K for *P. angustata* and 46K for *T. borchgrevinki*. The mean inter-species smoothed T_{MRCA} was 274K, indicating a substantial time gap between intra- and inter-species T_{MRCA}. The T_{MRCA} between the two species varied along the genome, but I found some cases where the value was drastically lower than the mean. These sharp drops in T_{MRCA} coincided with a drastic reduction in GNN for either one or both species due to polytomy gene tree structure, as seen for the regions between 45 and 46Mb on chromosome 4 (**Figure 4.4**). While I observed kernel-smoothed T_{MRCA} between species above the mean, I did not or was unable to identify any clear peaks. The maximum values of unsmoothed inter-species T_{MRCA} within and outside dN/dS genes were highly similar (336,423.86 generations for the former and 337,213.87 for the latter).

Similarly, the maximum values of unsmoothed inter-species T_{MRCA} within and outside structural variation specific to *P. angustata* were 337,213.87 and 337,029.3084, respectively.

Patterns of intra-species T_{MRCA} distribution for P. angustata

For *P. angustata*, I obtained 20,879 merged RAD-loci, and the haplotype data from 16,281 of those loci were converted to tsinfer format. I retained 12,902 of 16,281 loci after filtering 3,379 loci, each generating more than two gene trees. Of 12,902 retained loci, 10,418 were found to have generated a single gene tree per locus, whereas 2,484 were observed to have produced two gene trees per locus due to recombination. In total, 15,386 gene trees generated from 12,902 retained loci were used for estimating T_{MRCA} (units in generations). The distribution of kernel-smoothed T_{MRCA} had a mean of 43.72K and a median of 43.43K, which was noticeably skewed at the upper tail. I found 292 potential intra-species T_{MRCA} outliers (across multiple chromosomes) with values greater than the threshold of 63.05K (**Figure 4.5**).

The most notable T_{MRCA} outliers were found in an intergenic region within 45-55 Mbp on chromosome 4 (**Figure 4.6**). Specifically, these T_{MRCA} outliers were associated with two RAD-loci and their windows partly spanned the inside and outside portions of a previously identified *P. angustata*-specific inversion on chromosome 4. The T_{MRCA} values of these outliers (67-68K) were drastically higher than those estimated for other nearby loci of the region, encompassing not only inversion but also the chromosomal fusion point adjacent to it on chromosome 4 (chapter 3). The windows of these outliers contained the *il16* candidate gene, which is also located inside the inversion (**Figure 4.6**). The values of these outliers differed the most from the threshold of the T_{MRCA} distribution compared to the rest of the outliers with windows containing dN/dS genes. None of the outliers belonged to RAD-loci located within dN/dS genes. Moreover, I observed no significant difference between distributions of T_{MRCA} for RAD-loci found within and outside dN/dS genes (**Figure 4.7**), as well as those for RAD-loci found within and outside *angustata*-specific structural variation (**Figure 4.8**).

Low intra-population GNN but co-localization between intra-species T_{MRCA} outliers and translocations specific to T. borchgrevinki

I next looked at gene genealogies within two populations of *T. borchgrevinki*. I obtained 22,659 merged RAD-haplotype pairs, and the haplotype data from 19,477 of those loci were transformed into tsinfer format. These 19,477 loci produced a total of 33,493 gene trees. After filtration, I retained 16,355 loci and 19,262 gene trees. For both populations of *T. borchgrevinki*, the average kernel-smoothed GNN was about 0.49 (**Figure 4.9**). I did not observe high GNN (i.e.,> 0.90) across the genome for one or both populations. For each population, the mean smoothed T_{MRCA} was about 50K generations. The average smoothed T_{MRCA} between populations was about 51K, and the estimated T_{MRCA} varied along the genome (**Figure 4.9**). The distribution of smoothed inter-population T_{MRCA} had a median of 50K and was noticeably skewed at the upper tail. I found 510 potential T_{MRCA} outliers with values greater than the threshold of 65.41K generations (**Figure 4.10**).

Moreover, I found two cases in which intra-species smoothed T_{MRCA} outliers co-localized within translocations specific to *T. borchgrevinki* on chromosomes 1 (**Figure 4.11** & **4.12**) and 23 (**Figure 4.9** & **4.13**). The translocation in each case disrupted the conserved synteny between *T. borchgrevinki* and *P. angustata,* and represented a structural change specific to *T. borchgrevinki*. The T_{MRCA} values within these translocations ranged from 75 to 83K for three loci on chromosome 1, whereas they ranged from 66 to 97K for 11 loci on chromosome 23, indicating the signal of elevated T_{MRCA} was most robust on chromosome 23.

DISCUSSION

Higher historical N_e in the distant past and moderate contemporary N_e for P. angustata and T. borchgrevinki

From the PSMC analysis, I observed a prolonged decline of *Ne* for both species but at different scales, followed by stabilization of N_e for *P. angustata* but aberrant N_e expansion for *T. borchgrevinki* before its stabilization. While the overall N_e decreased over time, N_e reached 60K for *P. angustata* and 400K for *T. borchgrevinki* in the recent past. These estimates were different from those inferred by SMC++. However, *P. angustata* had more similar N_e across the two different methods than *T. borchgrevinki*. This indicates some agreement between methods for *P. angustata* but more variability for *T. borchgrevinki*. PSMC is less accurate in inferring recent demographic history (Liu *et al.* 2022) because its accuracy depends on the density of coalescence events, which tend to be lower in the recent past but increase in a deeper time scale. However, SMC++ jointly analyzes a larger sample size in each genetic region than PSMC, which has comparatively increased its accuracy in inferring demographic changes, especially in the recent past (Terhorst *et al.* 2017; Liu and Fu 2020). An increase in the sample size decreases the expected time to the first coalescence, which is specifically informative for recent coalescence and demographic events (reviewed in Moorjani and Hellenthal 2023). In summary, the pattern of PSMC- and SMC++-based N_e shows that both species had higher historical N_e in the distant past compared to the recent past, and they have experienced significant population contractions.

Our estimates of contemporary N_e for both *P. angustata* (~36K) and *T. borchgrevinki* (~46K) are lower than those of secondarily temperate species *Champsocephalus esox* (~50K) and cold-specialized notothenioid *C. gunnari* (~54K) (Rivera-Colón 2022). Additionally, the estimates of the populations of both species are higher than those of multiple populations of another cold-specialized notothenioid, *C. hamatus* (ranging from about 2K to 10K; inferred from Figure 2 of Lu *et al.* 2022). However, they were smaller than contemporary N_e of one population of *C. hamatus* (between 100-200K; Lu *et al.* 2022). This suggests that *P. angustata* and *T. borchgrevinki* have a moderate contemporary N_e in the context of other notothenioids.

Strong divergent selection may partly explain the reciprocally monophyletic patterns between the species

The high mean GNN (> 95%) for haplotypes of *P. angustata* and *T. borchgrevinki* suggests a strong population structure and no gene flow between species. Such an interpretation is supported by the observed high genome-wide F_{ST} (0.736) between these two species (Chapter 3). The presence of genome-wide reciprocal monophyly in most gene trees suggests little to no maintenance of ancestral diversity within the focal species due to complete lineage sorting. It is well established that genetic drift and directional selection can reduce genetic variability within a population and produce reciprocally monophyletic patterns between populations if drift or divergent selection fixes different alleles in different populations.

According to Hudson and Coyne (2002), when an ancestral population splits into two populations of equal size with no gene flow between them, mathematically, the genetic drift alone (in the absence of selection) can cause complete lineage sorting between descendants at 95% of loci in 9-12 N_e generations (where N_e is the historical effective population size of each descendent). However, directional selection shortens the time to achieve a reciprocally monophyletic pattern (Hudson and Coyne 2002). This framework has also been applied in other divergent species (Vijay *et al.* 2017; Zhou *et al.* 2017).

While there is no information on the exact divergence time between *P. angustata* and *T. borchgrevinki*, the recent time-calibrated phylogeny by Bista *et al.* (2023) suggested that the mean age of divergence between the clade containing *Notothenia rossii* (which is closely related to *P. angustata*) and that consisting of *T. bernachii* is 10.06 million years. This means that the divergence time of lineages for *P. angustata* and *T. borchgrevinki* could be similar to that for *N. rossii* and *T. bernachii.* The historical N_e for both species in the time interval between 10 and 20 MYA is greater than 300,000 (**Figure 4.2**). With an assumption of divergence time between *P. angustata* and *T. borchgrevinki* being 10.06 MY as well as the historical N_e of 300,000, the number of generations that may have passed since divergence between the lineages of two species would be 4.7 N_e generations, which is less than 9-12 N_e generations. Given the magnitude of the difference between the expected and observed measures, substantial variation would be expected to be shared between the species. My result indicates that most of the shared variation between species was lost earlier than expected.

Moreover, the observed mean T_{MRCA} between haplotypes of the two species was 274K generations, based on the average contemporary N_e (41,411) and the mutation rate 4.7X10^{-9} of the two species. The average T_{MRCA} is also lower than 9-12 N_e (i.e., 372-496K generations based on the same average contemporary N_e, which is lower than or similar to historical N_e). I suggest that the lineage sorting between *P. angustata* and *T. borchgrevinki* has occurred on average at a much faster pace than what would be expected from genetic drift alone. Along with the evidence presented here, it is also essential to underscore that these two species live in drastically different thermal environments, and strong environment-specific selection between species is expected. Gene trees with reciprocally monophyletic patterns between species were also found within a) 92 genes exhibiting a significant dN/dS ratio (indicating selection in *P. angustata*) and b) the *P. angustata*-specific structural variations (previously presented in Chapter 3). Thus, the contribution of the divergent selection between species on the observed complete lineage sorting cannot be completely ruled out.

Standing genetic variation that existed before the split of two species may not have contributed to the genetic adaptation of P. angustata

Based on unmerged RAD-loci, I observed that intra- and inter-species coalescence time varied along the genome. This result reflects the amount of sequence divergence among haplotypes within and between species (Cruickshank and Hanh 2014). I observed that the mean time of coalescence within species was lower than between species. This indicates that the sequence divergence between species is higher than with species, which is expected because individuals within the same species are more closely related to each other genetically compared to those from different species. When the old variation contributes to the adaptation of the taxon, inter-taxa T_{MRCA} of adaptive loci would be at least greater than the time of the taxa split. For example, three adaptive inversions in threespine sticklebacks have been shown to exhibit a drastic increase in T_{MRCA} between populations (Nelson and Cresko 2018). In this study, I did not observe distinct peaks of inter-species coalescence time that were greater than probable speciation time between *P. angustata* and *T. borchgrevinki* (10.06 MYA, i.e., 1,514K generations). I also did not find any distinct peaks of inter-species coalescence time compared to the background (i.e., mean T_{MRCA}). This suggests that mutations that contributed to temperate adaptations in *P. angustata* may have evolved after the split of a lineage of *P. angustata* from that of *T. borchgrevinki*.

Patterns of intra-species T_{MRCA} within P. angustata reinforce that de novo mutations may have primarily contributed to the genetic adaptation of P. angustata

Based on merged RAD-loci, I found that the T_{MRCA} outlier windows contained the *il16* gene, which is under positive selection in *P. angustata*. These windows spanned a portion of an inversion specific to *P. angustata* and had higher values than their flanking regions, including those of other nearby loci spanning a chromosomal fusion, suggesting that these haplotypes arose before the evolution of the fusion on chromosome 4. Such variation in an intergenic region bordering an inversion could be due to their linkage to older adaptive loci. However, whether the inversion captured the old adaptive variation or itself has adaptive significance in *P. angustata* cannot be concluded from this study. This is because it would require information on when the inversion arose compared to the variants within it. To understand the timing of the origin of inversion, it is crucial to know where the breakpoints of the inversions are. Conserved synteny analysis cannot provide the exact location of the breakpoints. The absence of T_{MRCA} outliers within previously identified dN/dS genes (Chapter 3) and the highly similar distribution of the T_{MRCA} within and outside dN/dS genes indicate that *de novo* mutations in protein-coding sequences may have made major contributions to the adaptation of *P. angustata* to temperate environment.

GNN suggests incomplete lineage sorting between populations of T. borchgrevinki, but its translocations require further investigation

Based on the merged RAD-loci for *T. borchgrevinki*, I found a low mean GNN (~50%) for each population across the genome. This means that there is a high incomplete lineage sorting between two populations. My result suggests that although these two populations are from opposite sides of the Antarctic continent, there is high gene flow and low divergence between them. Additionally, we found that a few intra-species T_{MRCA} outliers co-localized with two large translocations specific to *T. borchgrevinki*. Translocations can generate adaptation through changes in gene expressions (reviewed in Gorkovskiy and Verstrepen 2021).

CONCLUSION

Here, I show the pattern of gene genealogies within and between *P. angustata* and *T. borchgrevinki*. I found a higher effective population size in the distant past compared to recent times for both species. I observed a genome-wide reciprocally monophyletic pattern between species. Also, the average time to the most recent common ancestor (T_{MRCA}) of alleles between species appears to be lower than the time required for a genome-wide reciprocally monophyletic pattern to form under neutrality. This piece of evidence, in addition to the presence of completely sorted gene trees within candidate loci and structural variation of *P. angustata*, suggests that divergent selection can explain the observed pattern to some extent. A lack of distinct, prominent peaks of inter-species T_{MRCA} for *P. angustata* indicates that the adaptive mutations generated after the split of the two species may have enabled temperate re-adaptation in *P. angustata*. I found no intra-species T_{MRCA} outlier within candidate loci. I observed insignificant differences in the distribution of intra-species T_{MRCA} within and outside of these loci, reinforcing that *de novo* mutations may have played a major role in adaptations of *P. angustata*. While there is pervasive incomplete lineage sorting between populations of *T. borchgrevinki*, the co-localization of species-specific translocations with potential intra-species T_{MRCA} outliers calls for further investigation to understand the role of these structural changes in the continuing cold adaptation of *T. borchgrevinki*.

TABLES

Table 4.1 Contemporary effective population sizes of *Paranotothenia angustata* and *Trematomus borchgrevinki*

Species	Number of populations	Effective Population size (N_e)
P. angustata	1	36,778
T. borchgrevinki	1	46,046

Table 4.2 The genome-wide average of mean Genealogical Nearest Neighbours (GNN) estimates from RAD-haplotypes of *Paranotothenia angustata* and *Trematomus borchgrevinki* as well as those from merged RAD-haplotype pairs of two populations of *T. borchgrevinki* (i.e. McMurdo Sound and Prydz Bay)

Species or population	**Average of mean GNNs**
P. angustata	0.97
T. borchgrevinki	0.96
McMurdo Sound	0.49
Prydz Bay	0.49

Table 4.3 The genome-wide mean time to the most recent common ancestor (T_{MRCA}) (in generations) for unmerged RAD-haplotypes (indicated by *) as well as merged RAD-haplotype pairs (denoted by *) from _Paranotothenia angustata_ and _Trematomus borchgrevinki_

	Mean TMRCA
Within _P. angustata_*	43,551.8
Within _T. borchgrevinki_*	46,811.5
Between _P. angustata_ and _T. borchgrevinki_*	274,190
Within _P. angustata_**	43,723.13
Within _T. borchgrevinki_ or between its McMurdo Sound and Prydz Bay populations**	51,381.21
Within McMurdo Sound population of _T. borchgrevinki_**	50,949.57
Within Prydz Bay population of _T. borchgrevinki_**	50,914.80

FIGURES

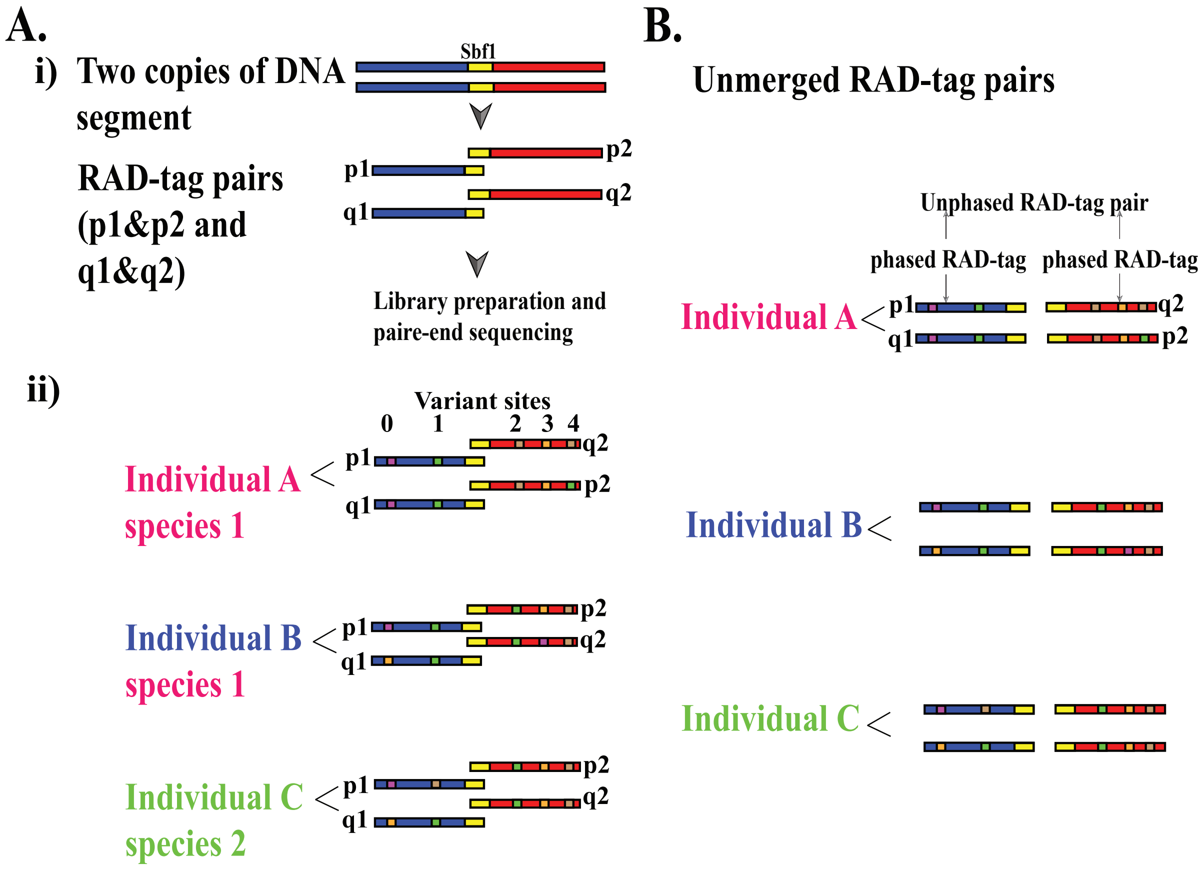


Figure 4.1 Simplified outline of tree sequence analysis using Sbf1-based RADseq data. **A.i**.) shows that the process involves DNA cutting at the restriction enzyme recognition site (Sbf1). Each Sbf1 cut-site in a homologous DNA segment is expected to produce a RAD-tag pair in a diploid individual. For example, a) p1 and p2, as well as b) q1 and q2, are shown as two pairs of RAD-tags from two homologous regions. These tags are sampled across the genomes of multiple individuals affiliated with the same or different population or species. The sampled tags are used for RAD library preparation and sequencing. **A.ii)** illustrates variant calling and genotyping of variant sites (shown as 0, 1, 2, 3, and 4) on each tag from each individual per population or species. Here, variants are depicted as small squares with pink, green, brown, and orange colors. **B).** displays an example of unmerged RAD-tag pairs. Variants in each tag are phased, but variants across RAD-tag pairs are not.

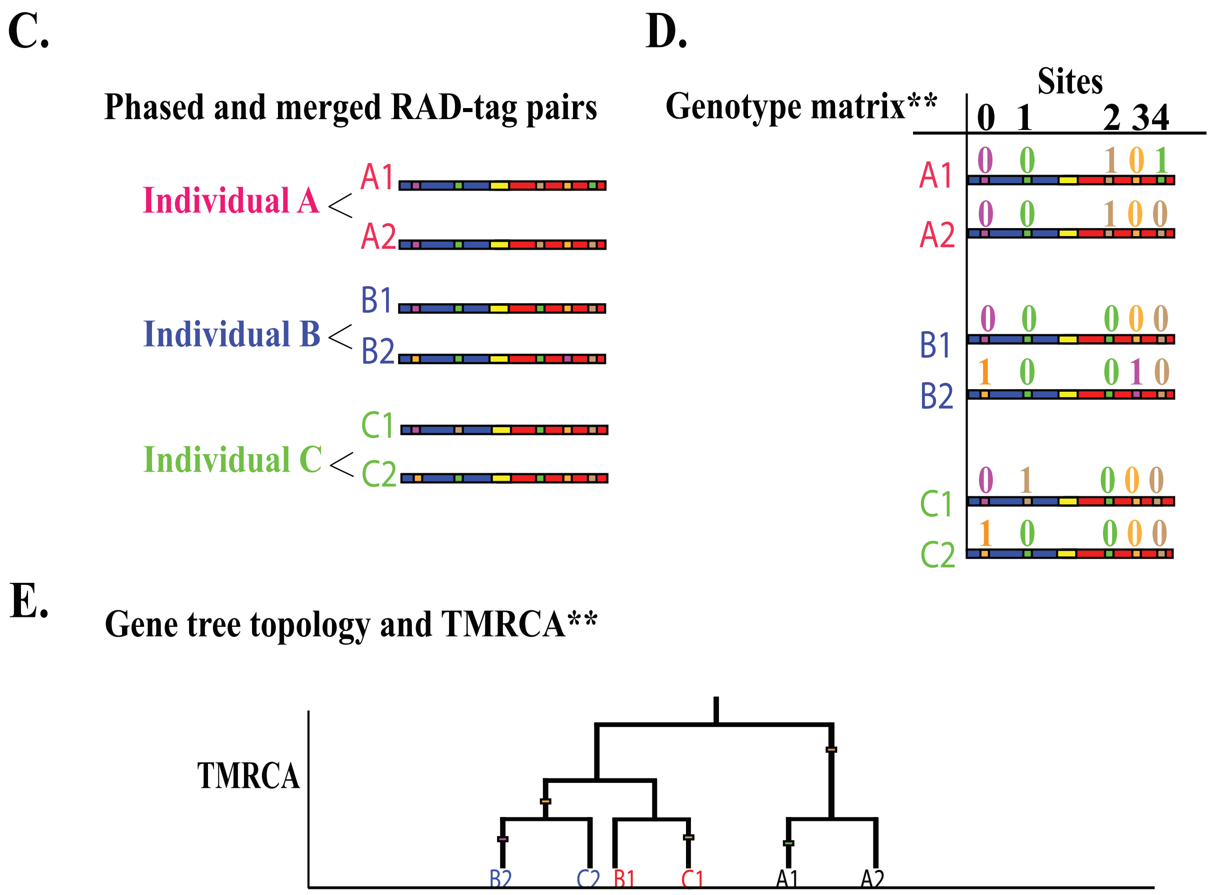


Figure 4.1 – Continued. C). demonstrates phasing of variants within each RAD-tag pair per individual per population or species and merging of the sister tags to generate a longer merged RAD-haplotype pair (e.g., A1, A2, B1, B2, C1, and C2). **D).** shows that a genotypic matrix can be built from the genotype data at each variant site across the merged RAD-haplotype pairs of each individual. Genotypic data are encoded as 0 and 1 (colored according to variants in tags) in the matrix. **E.** illustrates that the encoded genotypes can be utilized to produce gene tree topology for each merged RAD-haplotype pair occupying a specific genomic region in a chromosome (X-axis). Time to the most recent common ancestor (T_{MRCA}) (Y-axis) in a gene tree can also be estimated. The RAD-haplotypes (A1, A2, B1, B2, C1, and C2) of the gene tree having the same color denotes genealogical nearest neighbours of each other. The average GNN for haplotypes from each population/species can also be estimated. ** denotes that encoding of genotypes, gene tree construction, and estimation of T_{MRCA} can be performed without merging RAD-tag pairs.

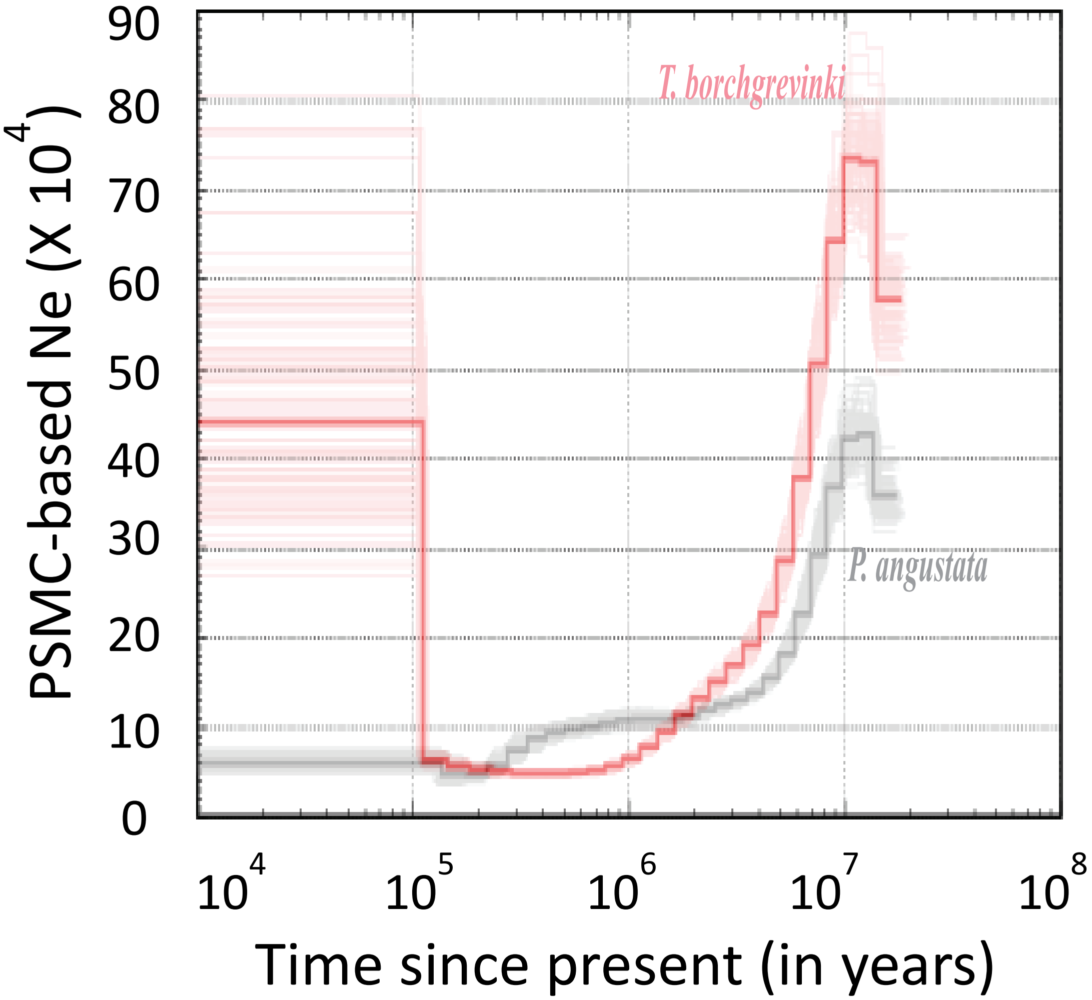


Figure 4.2 depicts the PSMC-based temporal trajectory of effective population sizes (N_e) for *P. angusta* and *T. borchgrevinki*. The inferred N_e trajectories are indicated by a dark red colored line for *T. borchgrevinki* and by a dark grey colored line for *P. angustata*. The light grey and red lines represent N_e estimates from the 100 bootstrap replicates. The Y-axis represents $N_{e,}$ and the X-axis indicates the time before present in years. Dashed black, vertical, and horizontal lines represent grids of the plots. The N_e was highest for both species between 10 and 20 million years ago (MYA). Subsequently, it steadily decreased before 0.1 MYA. However, the N_e for *P. angustata* stabilized, but that of *T. borchgrevinki* expanded before reaching stabilization. The time is scaled based on generation time of 7 years for both species but 4.27 x10^{-9} and 5.32 x10^{-9} substitutions per base per generation for *P. angustata* and *T. borchgrevinki*, respectively.

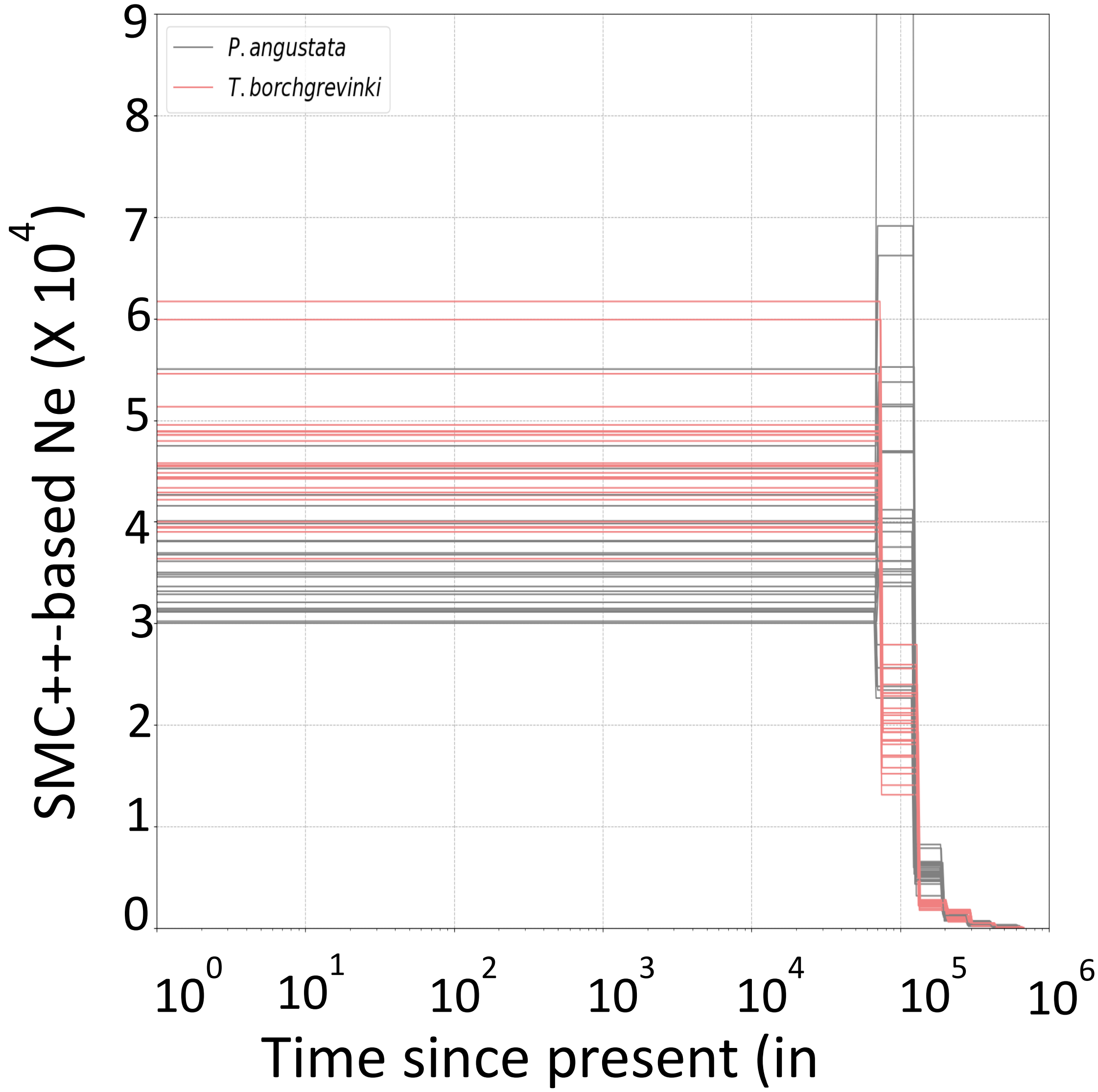


Figure 4.3 shows the SMC++-based temporal trajectory of effective population sizes (N_e) for *P. angustata* (indicated by a grey-colored solid line) and *T. borchgrevinki* (indicated by a colored solid line) with 25 bootstrap replicates. The time scale with generation time of 7 years for both species but 4.27 x10^{-9} and 5.32 x10^{-9} substitutions per base per generation for *P. angustata* and *T. borchgrevinki*, respectively. The Y-axis represents $N_{e,}$ and the X-axis indicates the time before present in years. Dashed grey vertical and horizontal lines represent the grids of the plots. It shows that the two species had a constant effective population size from the recent past (0.01 MYA) to the present.

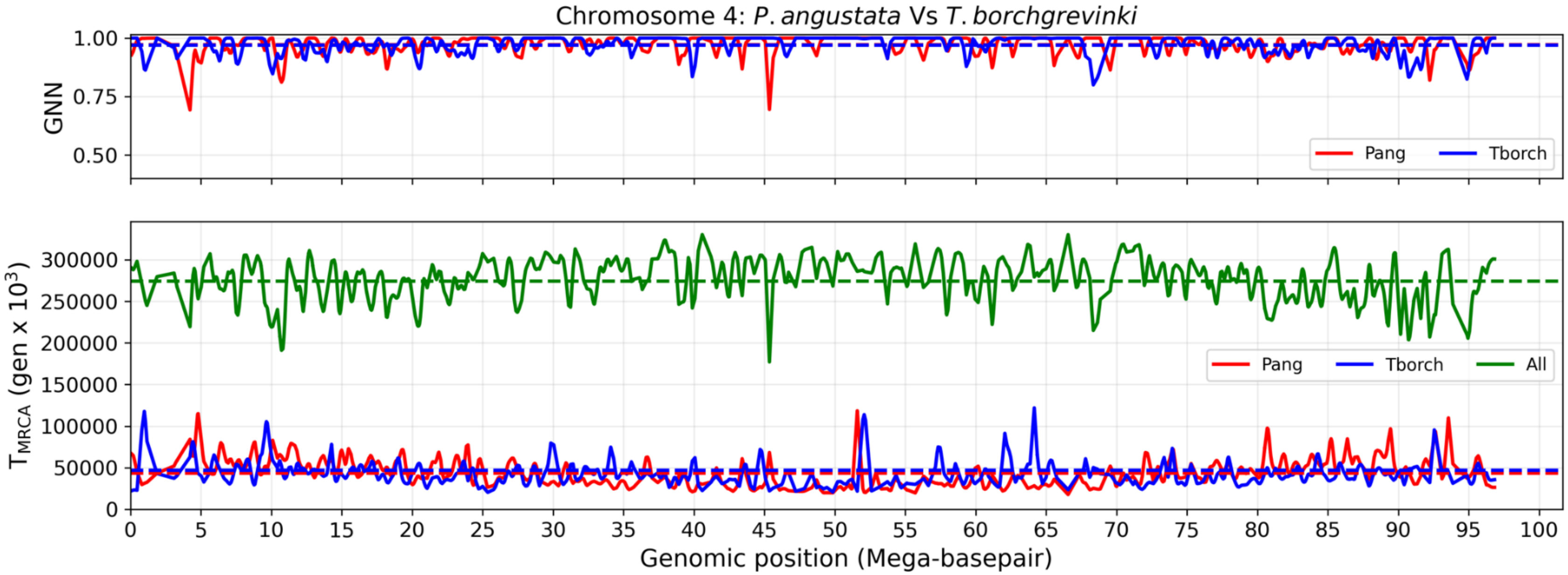


Figure 4.4 shows an example of the pattern of genealogical nearest neighbours (GNN) and time to the most recent common ancestor (TMRCA) along chromosome 4 for *P. angustata* (Pang) and *T. borchgrevinki* (Tborch). The figure consists of two panels, an upper and a lower. In both panels, the x-axis represents the genomic position (mega-basepair) along chromosome 4. In the upper panel, the y-axis represents the GNN. The solid red and blue lines represent the kernel-smoothed estimates of GNN within *P. angustata* and *T. borchgrevinki*, respectively. The dashed red and blue lines denote the genome-wide average of mean GNN for *P. angustata* (0.97) and *T. borchgrevinki* (0.96). In the lower panel, the y-axis represents T_{MRCA} in thousands (K) of generations (gen). The solid red and blue lines represent the kernel-smoothed estimates of T_{MRCA} within *P. angustata* and *T. borchgrevinki*, respectively. The solid green line denotes smoothed T_{MRCA} between species. The dashed red, blue, and green lines represent a genome-wide average of T_{MRCA} within *P. angustata* (43K), within *T. borchgrevinki* (46K), and between the two species (274K), respectively.

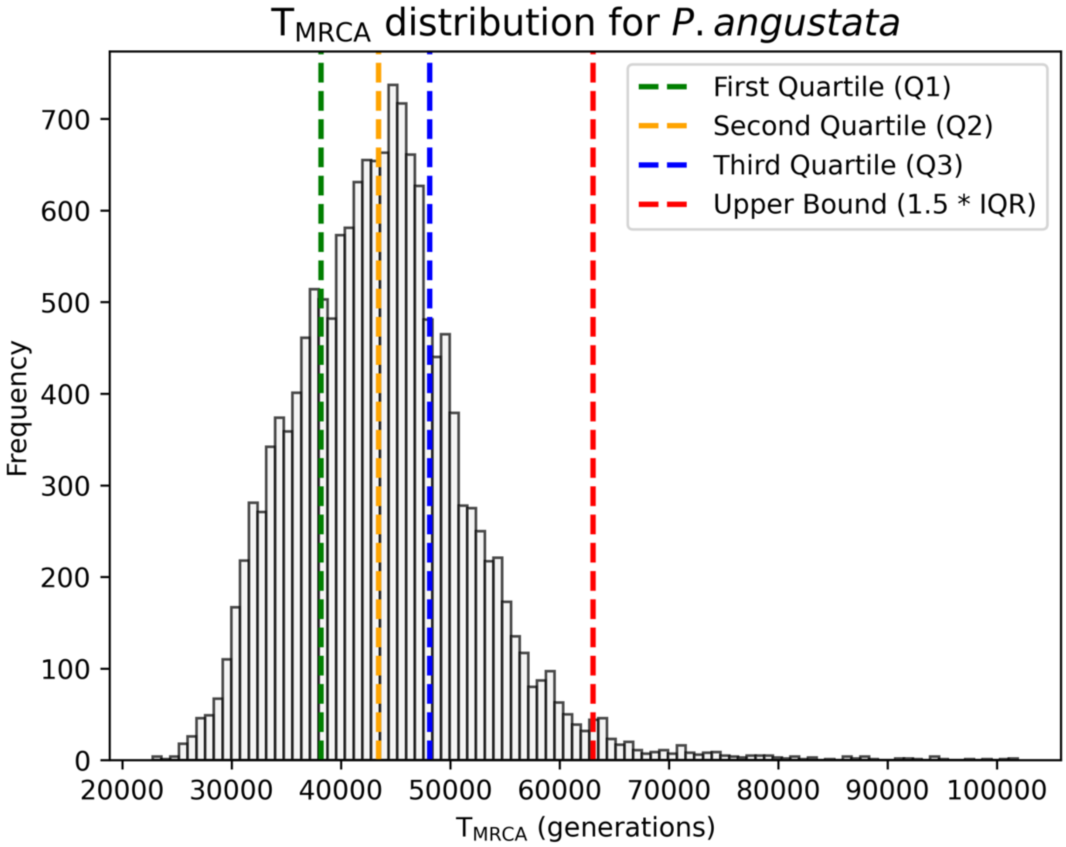


Figure 4.5 The histogram depicts the distribution of smoothed T_{MRCA} obtained from merged RAD-haplotype pairs of *P. angustata*. The bottom X-axis represents the smoothed T_{MRCA} (in generations), while the Y-axis indicates the frequency of the observed T_{MRCA}. The vertical, green, yellow, and blue dashed lines represent the first (Q1), second (Q2), and third (Q3) quartiles, whereas the red dashed line indicates the upper bound of the distribution. The upper bound is the sum of Q3 and 1.5 times the interquartile range (IQR, i.e., the difference between Q3 and Q1). The value of this distribution's upper bound or threshold is 63,052.434 generations.

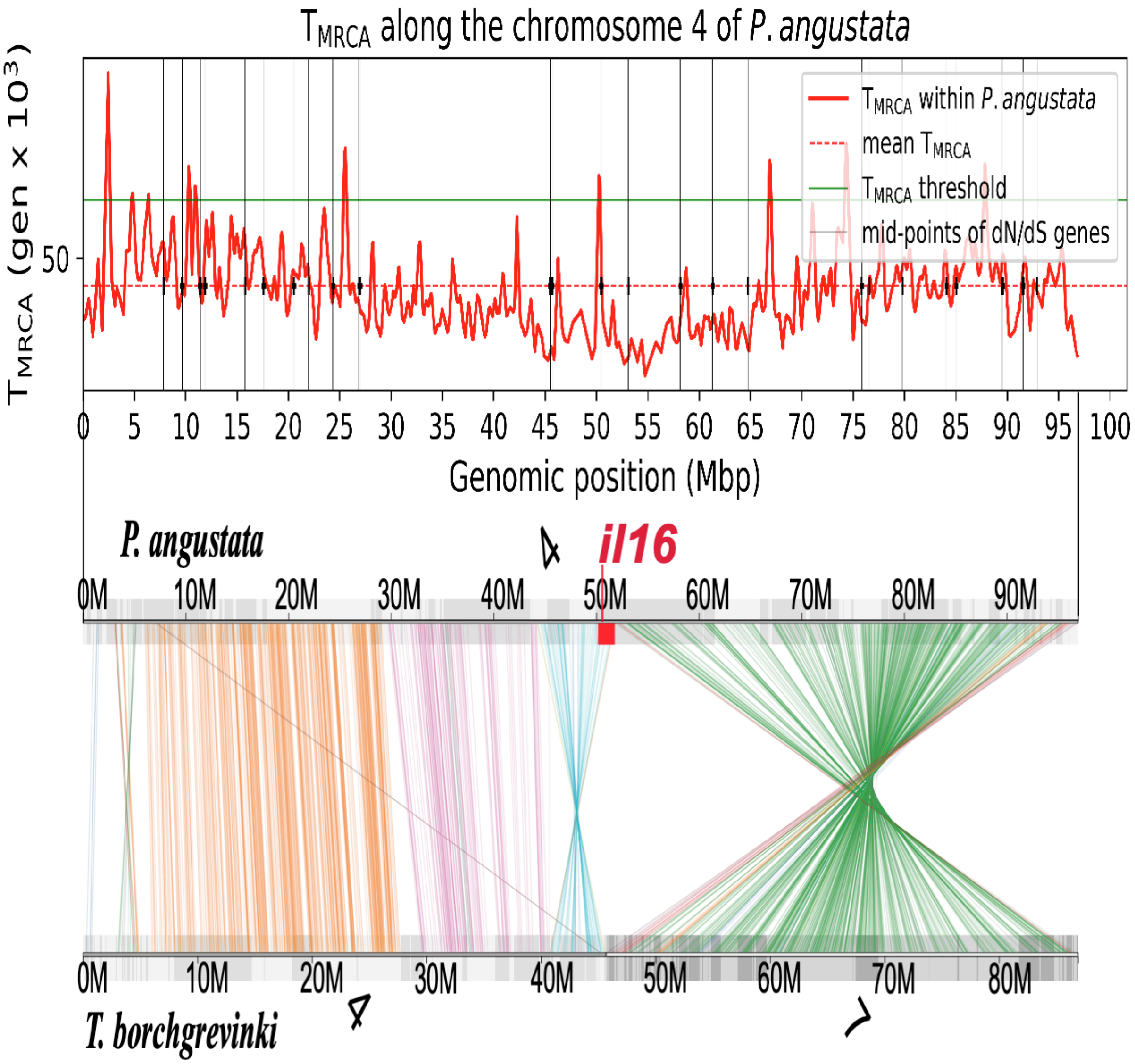


Figure 4.6 shows a) the pattern of TMRCA within *P. angustata* along chromosome 4, where the X-axis represents genomic positions in mega base-pair (Mbp), as well as b) the conserved synteny between chromosome 4 of *P. angustata* (top) and chromosomes 4 (bottom left) and 7 (bottom right) of *T. borchgrevinki*. The figure exhibits that *P. angustata*-specific inversions (red block) on chromosome 4 consist of the positively selected *il16* gene in the species. The inversion spanned partially by sharp TMRCA peaks of windows centered at 50.23 and 50.33Mb genomic positions. The second window contained the *il16* gene as well.

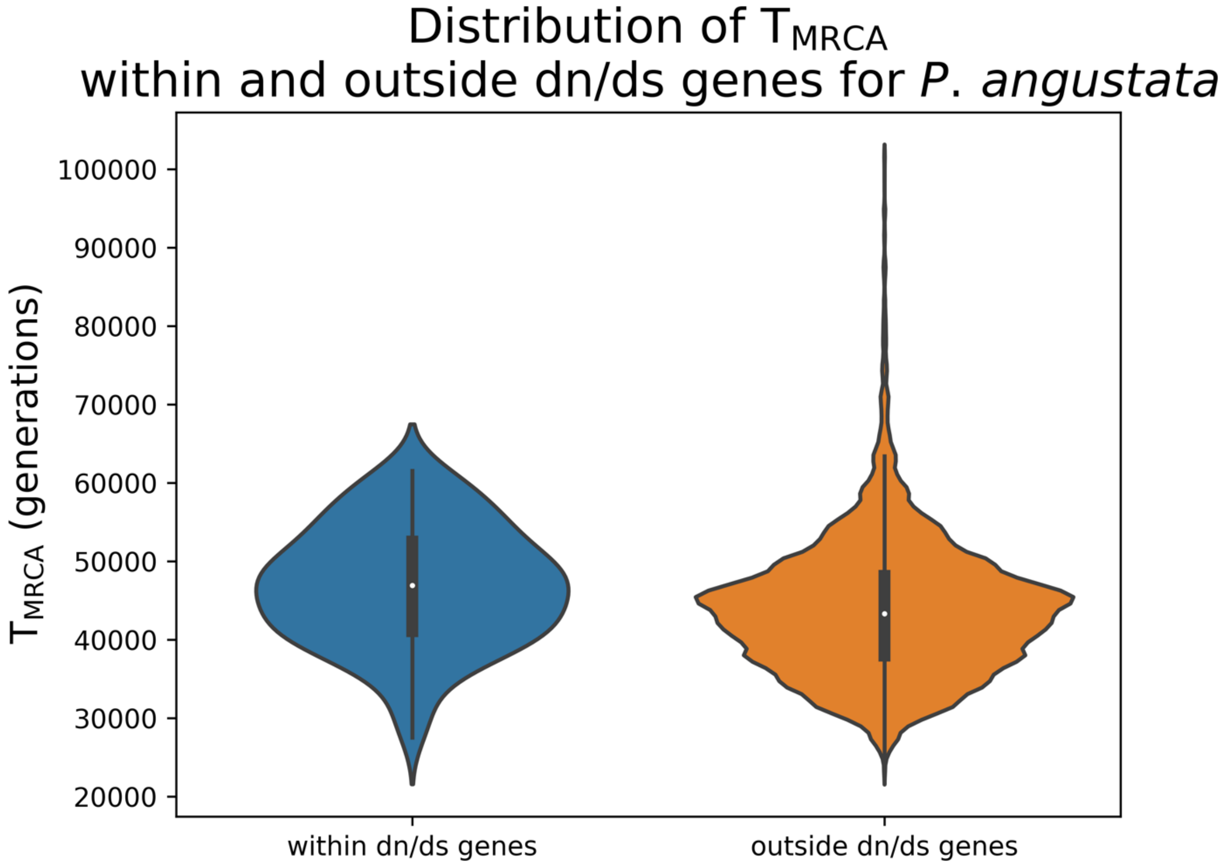


Figure 4.7 The distribution of smoothed T_{MRCA} within and outside of dN/dS genes under positive selection in *P. angustata*.

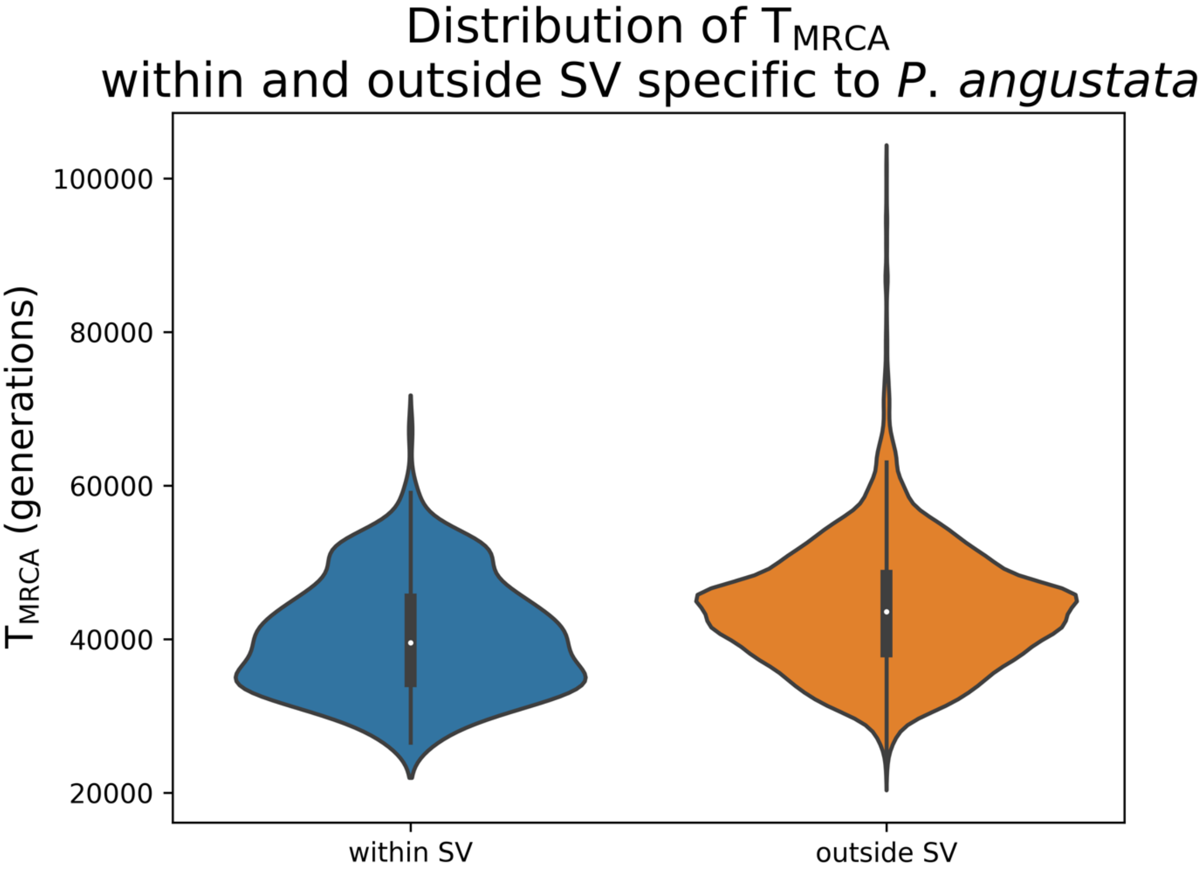


Figure 4.8 The distribution of smoothed T_{MRCA} within and outside of structural variation (SV) specific to *P. angustata*.

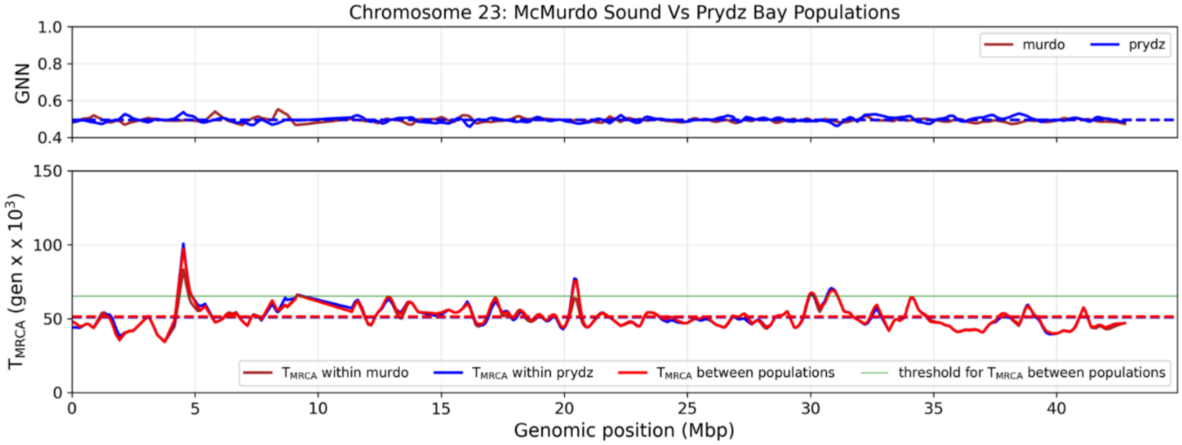


Figure 4.9 shows an example of the pattern of genealogical nearest neighbours (GNN) within McMurdo Sound and Prydz Bay populations of *Trematomus borchgrevinki* and T_{MRCA} within and between the populations along chromosome 23. The figure consists of two panels, an upper and a lower. In both panels, the x-axis represents the genomic position (megabase pairs (Mbp)). In the upper panel, the y-axis represents the GNN. In the upper panel, the solid brown and blue lines represent the kernel-smoothed estimates of GNN within McMurdo Sound and Prydz Bay populations, respectively. The dashed brown and blue lines denote the genome-wide average of mean GNN for population McMurdo Sound (0.49) and Prydz Bay (0.49), respectively. In the lower panel, the y-axis represents T_{MRCA} in generations (gen). The solid brown and blue lines represent the kernel-smoothed estimates of T_{MRCA} within McMurdo Sound and Prydz Bay populations, respectively. The solid red line denotes smoothed T_{MRCA} between populations. The dashed brown, blue, and red lines represent a genome-wide average of T_{MRCA} within McMurdo Sound (approximately 50.94K), within Prydz Bay (about 50.91K), and between the two populations (about 51.38K), respectively. The green solid line represents the interquartile range-based threshold for the distribution of the smoothed T_{MRCA} within *T. borchgrevinki*. The genomic region between 0-5 Mbp consists of T_{MRCA} outliers with values greater than the threshold.

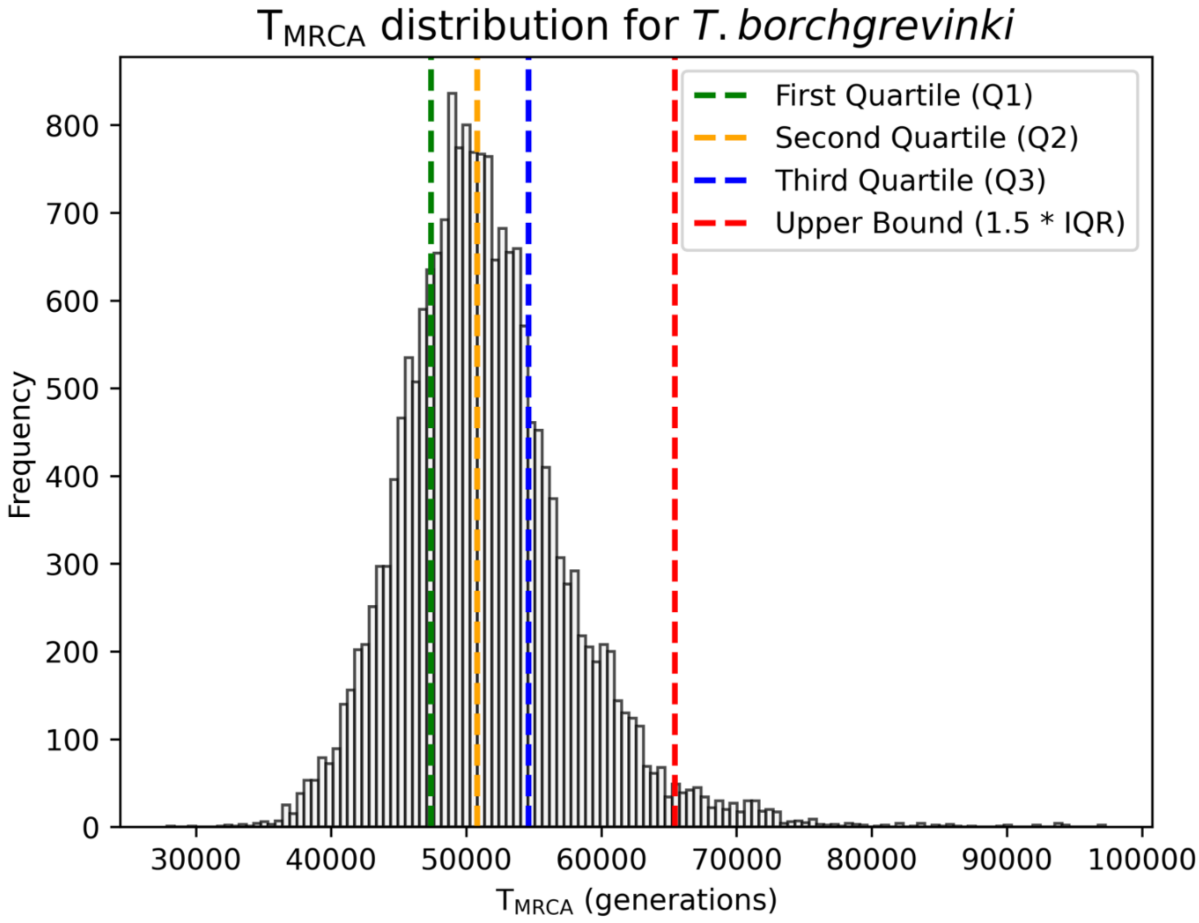


Figure 4.10 The histogram shows the distribution of smoothed T_{MRCA} obtained from merged RAD-haplotype pairs of *T. borchgrevinki*. The bottom X-axis represents the smoothed T_{MRCA} (in generations), and the Y-axis indicates the frequency of the observed T_{MRCA}. The vertical, green, yellow, and blue dashed lines represent the first (Q1), second (Q2), and third (Q3) quartiles, whereas the red dashed line indicates the upper bound of the distribution. The upper bound is the sum of Q3 and 1.5 times the interquartile range (IQR, i.e., the difference between Q3 and Q1). The value of the upper bound or threshold of this distribution is 65,413.712 generations.

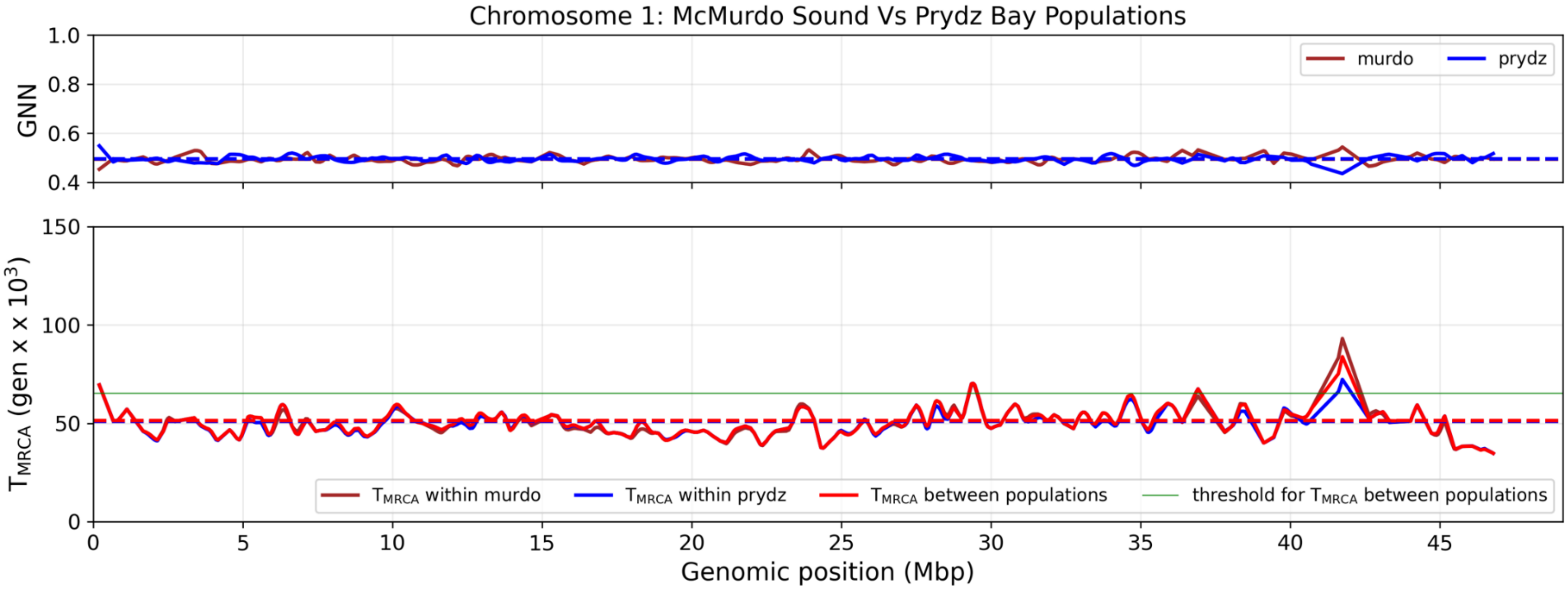


Figure 4.11 shows an example of the genealogical nearest neighbours (GNN) pattern within McMurdo Sound and Prydz Bay populations of *Trematomus borchgrevinki* and T_{MRCA} within and between the populations along chromosome 1. The figure consists of two panels, an upper and a lower. In both panels, the x-axis represents the genomic position (mega-basepair (Mbp)). In the upper panel, the y-axis represents the GNN. In the upper panel, the solid brown and blue lines represent the kernel-smoothed estimates of GNN within McMurdo Sound and Prydz Bay populations, respectively. The dashed brown and blue lines denote the genome-wide average of mean GNN for population McMurdo Sound (0.49) and Prydz Bay (0.49), respectively. In the lower panel, the y-axis represents T_{MRCA} in generations (gen). The solid brown and blue lines represent the kernel-smoothed estimates of T_{MRCA} within McMurdo Sound and Prydz Bay populations, respectively. The solid red line denotes smoothed T_{MRCA} between populations. The dashed brown, blue, and red lines represent a genome-wide average of T_{MRCA} within McMurdo Sound (approximately 50.94K), within Prydz Bay (about 50.91K), and between the two populations (about 51.38K), respectively. The green solid line represents the interquartile range-based threshold for the distribution of the smoothed T_{MRCA} within *T. borchgrevinki*. The genomic region between 40-45 Mbp consists of T_{MRCA} outliers with values greater than the threshold.

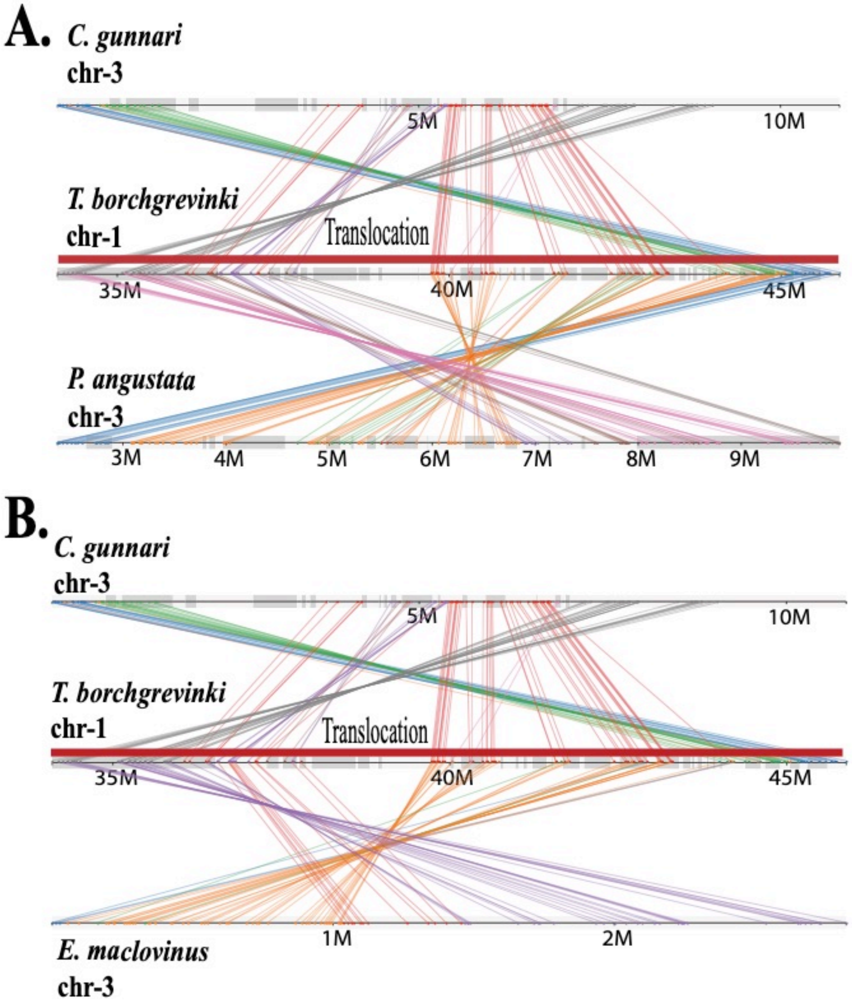


Figure 4.12 shows the translocation specific to chromosome (chr) 1 of *T. borchgrevinki*. Panels A and B display the local conserved synteny among chr-3 of *Champsocephalus gunnari*, chr-1 of *T. borchgrevinki borchgrevinki*, chr-3 of *P. angustata*, and chr-3 of *Eleginops maclovinus.* The red blocks on chromosome 1 on *T. borchgrevinki* in both panels represent the same translocation.

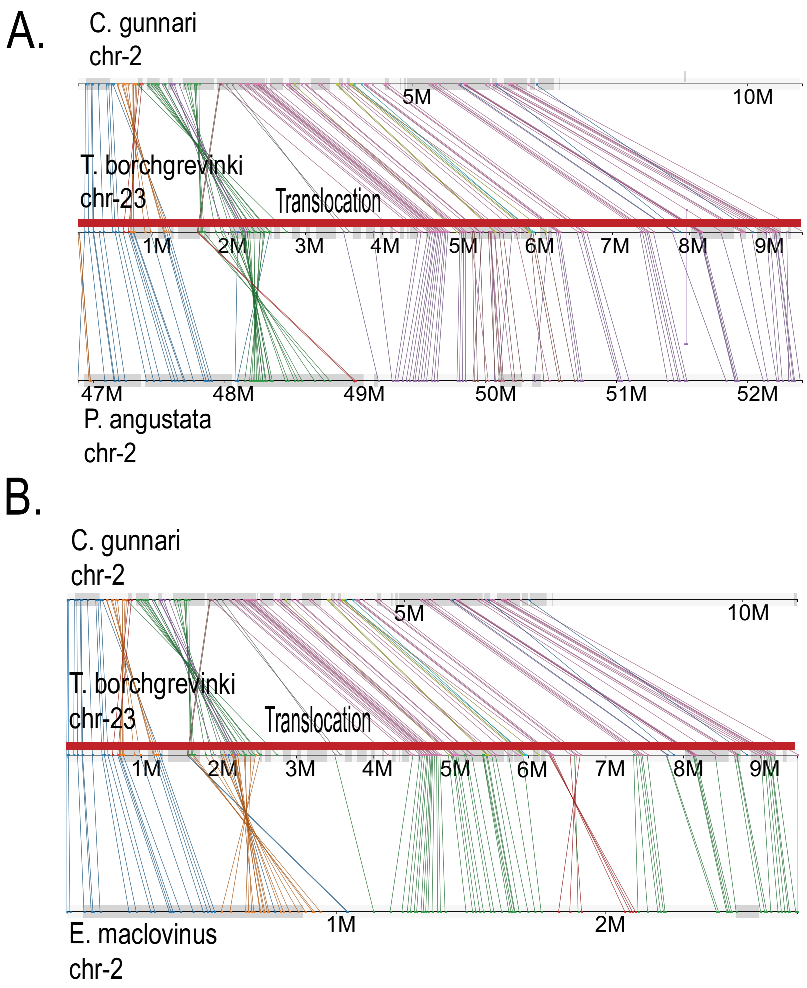


Figure 4.13 shows the translocation specific to chromosome (chr) 1 of *T. borchgrevinki*. Panels A and B display the local conserved synteny among chr-2 of *Champsocephalus gunnari*, chr-23 of *T. borchgrevinki borchgrevinki*, chr-1 of *P. angustata*, and chr-1 of *Eleginops maclovinus*. The red blocks on chromosome 1 on *T. borchgrevinki* in both panels represent the same translocati

CHAPTER 5: CONCLUSIONS

In chapter 2 of this dissertation, I compared Illumina-, Nanopore-, and PacBio-based *de novo* genome assembly strategies to identify the optimal strategy for notothenioids. The strategies I compared in this chapter mimic at least three phases of genome assembly approaches that adapted to changes in DNA sequencing technologies. Phase I strategy utilizes a high-volume of short-reads only, whereas the phase II approach implements a hybrid of a high-volume of short-reads and a low-volume of long-reads. Phase III utilizes a high-volume of long-reads only. From my findings in the first research chapter (chapter 2), I conclude that the phase III strategy is the current-state-of-art and can be optimized through a subsampling approach. In contrast, assemblies from phase I and II approaches are of low quality. Specifically, in the phase I strategy, the inclusion of mate-pair reads may enhance the assembly contiguity (e.g., N50); however, it can introduce hidden scaffolding errors, which, in turn, could lead to inaccurate measures of BUSCO gene completeness or fragmentation. Moreover, there is no optimal combination of mate-pair libraries of different insert sizes, as they can interfere with each other and affect the assembly quality. While a k-mer-based contig replacement strategy can enhance the completeness of the BUSCO genes in the assembly, its overall effectiveness could be constrained by inconsistencies present in the alternative assemblies.

Moreover, in the phase II strategy, the merging between contigs generated from the low-volume long-reads and those from phase I could fail due to sequence errors or small repeat alignments. Consequently, the quality of hybrid assembly degrades. It is essential to optimize the alignment parameters used for the merging process. Moreover, the hybrid assembly further suffers in terms of quality if it is produced using phase I assembly having hidden scaffolding errors. Hybrid assembly with high contiguity may not be of high quality, and a thorough examination of its BUSCO scores could reveal its quality. Finally, in the phase III strategy, long-reads generate highly contiguous assemblies. However, the presence of chimeric long-reads or excessive coverage can lower the contiguity of the assembly. A random sampling approach could improve the contiguity of assembly. I recommend critically evaluating the quality of phase I and II assemblies before their usage, even if those assemblies seem to have high contiguity. Only reporting BUSCO scores for a publication's genome assembly is insufficient; these metrics must be interpreted. Since the change in orientation of contigs or scaffolds can generate spurious BUSCO gene completeness or fragmentation, it is also possible that a) annotations of other genes could also have been impacted in phase I and II assemblies, and b) Hi-C scaffolding could also have the same effect as mate-pairs. These ideas need to be assessed in future studies.

In chapter 3 of this dissertation, I performed a genome-based investigation to find potential secondary temperate adaptations in *P. angustata* by using *T. borchgrevinki* as a part of the outgroup. I presented high quality chromosome-level genome assemblies with well-represented gene space for both *P. angustata* and *T. borchgrevinki*. I delineated the presence of lineage-specific DNA transposons in *P. angustata*. I identified, characterized, and described the *P. angustata*-specific structural changes, including chromosomal fusions, inversions, and translocations. I showed that the orientations of chromosomes that formed the fusions are predominantly unique to *P. angustata,* and inversions had one to three genes with an accelerated rate of change of non-synonymous to synonymous substitutions. For *P. angustata*, I proposed that potential secondarily temperate adaptations are related to protein chaperoning, circadian rhythm, vision, erythrocyte development and differentiation, heme metabolism, mitochondria, and ribosomes. My results provide compelling evidence of how secondarily temperate adaptations in *P. angsustata* may have evolved. Future functional studies should validate the role of candidates in a temperate adaptation of *P. angustata*. Data from this dissertation contribute valuable genomic resources for polar biologists to conduct future functional, comparative, and population genomics studies, especially considering the existence of other secondary temperate notothenioids that may or may not share the same adaptive genetic changes located in *P. angustata*.

From the exploration of gene genealogical patterns within and between *P. angustata* and *T. borchgrevinki* (chapter 4), I inferred that these species had a higher effective population size in the distant past compared to recent times. I observed a genome-wide reciprocally monophyletic pattern between species. Also, the average time to the most recent common ancestor (T_{MRCA}) of alleles between species appears to be lower than the time required for a genome-wide reciprocally monophyletic pattern to form under neutrality. This piece of evidence, in addition to the presence of completely sorted gene trees within candidate loci and structural variation of *P. angustata*, suggests that divergent selection can explain the observed pattern to some extent. A lack of distinct, prominent peaks of inter-species T_{MRCA} for *P. angustata* indicates that the adaptive mutations generated after the split of the two species may have enabled temperate re-adaptation in *P. angustata*. I found no intra-species T_{MRCA} outlier within and outside the candidate loci and structural variations. These results suggest that *de novo* mutations may have played a major role in the adaptations of *P. angustata*. While there is pervasive incomplete lineage sorting between populations of *T. borchgrevinki*, the co-localization of translocations specific to *T. borchgrevinki* with potential intra-species T_{MRCA} outliers calls for further investigation to understand the role of this structural change in minor adaptations. In future studies, it would be more appropriate to use Antarctic notothenioid species such as *N. rossii* or *N. coriiceps* with *P. angustata* (instead of *T. borchgrevinki*) for proper interpretation of the contribution of *de novo* and standing variation in temperate adaptation. This is because more extended haplotypes could be generated in sufficient amounts because the loss of RAD-tag pairs will be less due to lower divergence between species. Also, it is crucial to estimate the mutation rate and generation time for these species to interpret the demographic history of these species more accurately.

REFERENCES

Abrams, J., G. Davuluri, C. Seiler, and M. Pack, 2012 Smooth muscle caldesmon modulates peristalsis in the wild type and non-innervated zebrafish intestine. J Neurogastroenterol Motil. 24: 288–299.

Alkan, C., S. Sajjadian, and E. E. Eichler, 2011 Limitations of next-generation genome sequence assembly. Nat. Methods 8: 61–65.

Almroth, B. C., N. Asker, B. Wassmur, M. Rosengren, F. Jutfelt *et al.*, 2015 Warmer water temperature results in oxidative damage in an Antarctic fish, the bald notothen. J. Exp. Mar. Biol. Ecol. 468: 130–137.

Amarasinghe, S. L., S. Su, X. Dong, L. Zappia, M. E. Ritchie, Q Gouil, 2020 Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21: 1–16.

Amores, A., C. A. Wilson, C. A. H. Allard, H. W. Detrich, and J. H. Postlethwait, 2017 Cold Fusion: Massive Karyotype Evolution in the Antarctic Bullhead Notothen *Notothenia coriiceps*. G3 (Bethesda) 7: 2195–2207.

Amunts, A., A. Brown, J. Toots, S. H. Scheres, and V. Ramakrishnan, 2015 The structure of the human mitochondrial ribosome. Science 348: 95-98.

Apschner, A., L. F. Huitema, B. Ponsioen, J. Peterson-Maduro, and S. Schulte-Merker, 2014 Zebrafish enpp1 mutants exhibit pathological mineralization, mimicking features of generalized arterial calcification of infancy (GACI) and pseudoxanthoma elasticum (PXE). DMM 7: 811–822.

Auvinet, J., P. Graça, A. Dettai, A. Amores, J. H. Postlethwait *et al.*, 2020 Multiple independent chromosomal fusions accompanied the radiation of the Antarctic teleost genus Trematomus (Notothenioidei:Nototheniidae). BMC Evol. Biol. 20: 39.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver *et al.*, 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3: e3376.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin M *et al*., 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Bio. 19: 455–477.

Bao, E., and L. Lan, 2017 HALC: High throughput algorithm for long read error correction. BMC Bioinformatics 18, 1–12.

Bao, W., K. K. Kojima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6: 11.

Bargelloni, L., M. Babbucci, S. Ferraresso, C. Papetti, N. Vitulo *et al.*, 2019 Draft genome assembly and transcriptome data of the icefish Chionodraco myersi reveal the key role of mitochondria for a life without hemoglobin at subzero temperatures. Commun. Biol. 2: 443.

Brown, A., A. Amunts, X. C. Bai, Y. Sugimoto, P. C. Edwards, G. Murshudov, S. H. Scheres, and V. Ramakrishnan, 2014 Structure of the large ribosomal subunit from human mitochondria. Science, 346: 718-722.

Beers, J. M., and N. Jayasundara, 2015 Antarctic notothenioid fish: what are the future consequences of 'losses' and 'gains' acquired during long-term evolution at cold and stable temperatures? (J. E. Podrabsky, J. H. Stillman, & L. Tomanek, Eds.). J. Exp. Biol. 218: 1834–1845.

Bergeron, L. A., S. Besenbacher, J. Zheng, P. Li, M. F. Bertelsen *et al.*, 2023 Evolution of the germline mutation rate across vertebrates. Nature 615: 285–291.

Berglund, E. C., A. Kiialainen, and A. C. Syvänen, 2011 Next-generation sequencing technologies and applications for human genetic history and forensics. Investig. Genet. 2: 1–15.

Bilyk, K. T., and A. L. Devries, 2012 Heat tolerance of the secondarily temperate Antarctic notothenioid, *Notothenia angustata*. Antarct. Sci. 24: 165–172.

Bilyk, K. T., and C.-H. C. Cheng, 2014 RNA-seq analyses of cellular responses to elevated body temperature in the high Antarctic cryopelagic nototheniid fish Pagothenia borchgrevinki. Mar. Genomics 18: 163–171.

Bilyk, K. T., X. Zhuang, and C. Papetti, 2023 Positive and Relaxed Selective Pressures Have Both Strongly Influenced the Evolution of Cryonotothenioid Fishes during Their Radiation in the Freezing Southern Ocean. Genome Biol. Evol. 15: evad049.

Bilyk, K.T., L. Vargas-Chacoff, and C.-H.C, Cheng, 2018 Evolution in chronic cold: varied loss of cellular response to heat in Antarctic notothenioid fish. BMC Evol. Biol. 18:1–6.

Bista, I., J. M. D. Wood, T. Desvignes, S. A. McCarthy, M. Matschiner *et al.*, 2023 Genomics of cold adaptations in the Antarctic notothenioid fish radiation. Nat. Commun. 14: 3412.

Bista, I., S. A. McCarthy, J. Wood, Z. Ning, H. W. Detrich Iii *et al.*, 2020 The genome sequence of the channel bull blenny, Cottoperca gobio (Günther, 1861). Wellcome Open Res. 5: 148.

Bjerkås, E., E. Bjørnestad, O. Breck, and R. Waagbø, 2001 Water temperature regimes affect cataract development in smolting Atlantic salmon, Salmo salar L. J. Fish Dis. 24: 281–291.

Blanc, L., S. L. Ciciotte, B. Gwynn, G. J. Hildick-Smith, E. L. Pierce *et al.*, 2012 Critical function for the Ras-GTPase activating protein RASA3 in vertebrate erythropoiesis and megakaryopoiesis. Proc. Natl. Acad. Sci. 109: 12099–12104.

Bomblies, K., and C. L. Peichel, 2022 Genetics of adaptation. Proc. Natl. Acad. Sci. 119: e2122152119.

Brandwine, T., R. Ifrah, T. Bialistoky, R. Zaguri, E. Rhodes-Mordov *et al.*, 2021 Knockdown of Dehydrodolichyl Diphosphate Synthase in the Drosophila Retina Leads to a Unique Pattern of Retinal Degeneration. Front. Mol. Neurosci. 14: 693967

Bresciani, E., S. Confalonieri, S. Cermenati, S. Cimbro, E. Foglia *et al.*, 2010 Zebrafish Numb and Numblike Are Involved in Primitive Erythrocyte Differentiation (M. C. Capogrossi, Ed.). PLoS ONE 5: e14296.

Brindley, E. C., J. Papoin, L. Kennedy, R. F. Robledo, S. L. Ciciotte *et al.*, 2021 Rasa3 regulates stage-specific cell cycle progression in murine erythropoiesis. Blood Cells. Mol. Dis. 87: 102524.

Browning, B. L., X. Tian, Y. Zhou, and S. R. Browning, 2021 Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. 108: 1880–1890.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 3: lqaa108.

Burton, J. N., A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, *et al*. 2013 Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31: 1119-1125.

Cai, Z., S. Jitkaew, J. Zhao, H. C. Chiang, S. Choksi, J Liu, Y. Ward, L. G. Wu, and Z. G. Liu, 2014 Plasma membrane translocation of trimerized MLKL protein is required for TNF-induced necroptosis. Nat. Cell Biol. 16: 55-65.

Calì, F., E. Riginella, M. La Mesa, and C. Mazzoldi, 2017 Life history traits of Notothenia rossii and N. coriiceps along the southern Scotia Arc. Polar Biol. 40: 1409–1423.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Campbell, H. A., K. P. P. Fraser, L. S. Peck, C. M. Bishop, and S. Egginton, 2007 Life in the fast lane: The free-ranging activity, heart rate and metabolism of an Antarctic fish tracked in temperate waters. J. Exp. Mar. Biol. Ecol. 349: 142–151.

Carducci, F., M. A. Biscotti, M. Forconi, M. Barucca, and A. Canapa, 2019 An intriguing relationship between teleost Rex3 retroelement and environmental temperature. Biol. Lett. 15: 20190279.

Carotti, E., F. Carducci, A. Canapa, M. Barucca, and M. A. Biscotti, 2022 Transposable Element Tissue-Specific Response to Temperature Stress in the Stenothermal Fish Puntius tetrazona. Animals 13: 1.

Casacuberta, E., and J. González, 2013 The impact of transposable elements in environmental adaptation. Mol. Ecol. 22: 1503–1517.

Castresana, J., 2000 Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. 17: 540–552.

Catchen, J. M., J. S. Conery, and J. H. Postlethwait, 2009 Automated identification of conserved synteny after whole-genome duplication. Genome Res. 19: 1497–1505.

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson, 2016 Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44: e147.

Chan, Y. F., M. E. Marks, F. C. Jones, G. Villarreal, M. D. Shapiro *et al.*, 2010 Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. Science 327: 302–305.

Chapman, J. A., I. Ho, S. Sunkara, S. Luo, G. P. Schroth, D. S. Rokhsar, 2011. Meraculous: de novo genome assembly with short paired-end reads. PLoS ONE: 6: e23501.

Chen, Y., F. Nie, S. Q. Xie, Y. F. Zheng, Q. Dai, T. Bray, Y. X. Wang, J. F. Xing, Z. J. Huang ZJ, D. P. Wang *et al.*, 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. 12: 1–10.

Chen, Z., C. -H. C. Cheng, J. Zhang, L. Cao, L. Chen, L. Zhou, Y. Jin, H. Ye, C. Deng, Z. Dai *et al.*, 2008 Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. USA 105: 12944–12949.

Chen, L., A. L. DeVries, and C. H. C. Cheng, 1997 Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. USA 94: 3811–3816

Chen, Z., C.-H. C. Cheng, J. Zhang, L. Cao, L. Chen *et al.*, 2008 Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. 105: 12944–12949.

Chen, Z., H. Suzuki, Y. Kobayashi, A. C. Wang, F. DiMaio *et al.*, 2018 Structural Insights into Mdn1, an Essential AAA Protein Required for Ribosome Biogenesis. Cell 175: 822–834.

Cheng, C.-H. C., 2003 Functional Antifreeze Glycoprotein Genes in Temperate-Water New Zealand Nototheniid Fish Infer an Antarctic Evolutionary Origin. Mol. Biol. Evol. 20: 1897–1908.

Cheng, C.-H. C., A. G. Rivera-Colón, B. F. Minhas, L. Wilson, N. Rayamajhi *et al.*, 2023 Chromosome-Level Genome Assembly and Circadian Gene Repertoire of the Patagonia Blennie Eleginops maclovinus—The Closest Ancestral Proxy of Antarctic Cryonotothenioids. Genes 14: 1196.

Cheng, C.-H. C., L. Chen, T. J. Near, and Y. Jin, 2003 Functional antifreeze glycoprotein genes in temperate-water New Zealand nototheniid fish infer an Antarctic evolutionary origin. Mol. Biol. Evol. 20: 1897–1908.

Chikhi, R., and P. Medvedev, 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30 31–37.

Chin, C. S., P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz *et al.* 2016 Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13: 1050–1054.

Christmas, M. J., A. Wallberg, I. Bunikis, A. Olsson, O. Wallerman *et al.*, 2019 Chromosomal inversions associated with environmental adaptation in honeybees. Mol. Ecol. 28: 1358–1374.

Chuong, E. B., N. C. Elde, and C. Feschotte, 2017 Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18: 71–86.

Cirulli, E. T., and D. B. Goldstein, 2010 Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11: 415–425.

Clarke, A., R. B. Aronson, J. A. Crame, J.-M. Gili, and D. B. Blake, 2004 Evolution and diversity of the benthic fauna of the Southern Ocean continental shelf. Antarct. Sci. 16: 559–568.

Claros, M. G., R. Bautista, D. Guerrero-Fernández, H. Benzerki, P. Seoane, N. Fernández-Pozo, 2012 Why assembling plant genome sequences is so challenging. Biology 1: 439–459.

Clawson, H., B. T. Lee, B. J. Raney, G. P. Barber, J. Casper, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, and L. R. Nassar, 2023 GenArk: towards a million UCSC genome browsers. Genome Biol. 24: 217.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson *et al.*, 2005 Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science 307: 1928–1933.

Compeau, P. E., P. A. Pevzner, and G. Tesler, 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29: 987–991.

Coppes Petricorena, Z. L., and G. N. Somero, 2007 Biochemical adaptations of notothenioid fishes: Comparisons between cold temperate South American and New Zealand species and Antarctic species. Comp. Biochem. Physiol. A. Mol. Integr. Physiol. 147: 799–807.

Cruickshank, T. E., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23: 3133–3157.

Cyr, D. M., and C. H. Ramos, 2023 Specification of Hsp70 Function by Hsp40 Co-chaperones, pp. 127–139 in The Networking of Chaperones by Co-Chaperones, edited by A. L. Edkins and G. L. Blatch. Subcellular Biochemistry, Springer International Publishing, Cham.

Daane, J. M., and H. W. Detrich, 2022 Adaptations and Diversity of Antarctic Fishes: A Genomic Perspective. Annu. Rev. Anim. Biosci. 10: 39–62.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan *et al.*, 2021 Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

Das, D., S. K. Singh, J. Bierstedt, A. Erickson, G. L. J. Galli, D. A. Crossley, T. Rhen, 2020 Draft genome of the common snapping turtle, Chelydra serpentina, a model for phenotypic plasticity in reptiles. G3 (Bethesda)10:4299–4314.

Dean, P. R., and C. L. Hurd, 2007 Seasonal growth, erosion rates, and nitrogen and photosynthetic ecophysiology of Undaria pinnatifida (Heterokontophyta) in southern New Zealand1. J. Phycol. 43: 1138–1148.

Deng, C., C.-H. C. Cheng, H. Ye, X. He, and L. Chen, 2010 Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc. Natl. Acad. Sci. 107: 21593–21598.

Dettai, A., M. Berkani, A.-C. Lautredou, A. Couloux, G. Lecointre *et al.*, 2012 Tracking the elusive monophyly of nototheniid fishes (Teleostei) with multiple mitochondrial and nuclear markers. Mar. Genomics 8: 49–58.

DeVries, A. L., 1988 The role of antifreeze glycopeptides and peptides in the freezing avoidance of antarctic fishes. Comp. Biochem. Physiol. Part B Comp. Biochem. 90: 611–621.

Devries, A.L., 1971. Glycoproteins as biological antifreeze agents in Antarctic fishes. Science. 172:1152–1155.

Di Stefano, M., F. Di Giovanni, V. Pozharskaia, M. Gomar-Alba, D. Baù *et al.*, 2020 Impact of Chromosome Fusions on 3D Genome Organization and Gene Expression in Budding Yeast. Genetics 214: 651–667.

Diament, A., and T. Tuller, 2017 Tracking the evolution of 3D gene organization demonstrates its connection to phenotypic divergence. Nucleic Acids Res. 45: 4330–4343.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15.

Dopman, E. B., L. Pérez, S. M. Bogdanowicz, and R. G. Harrison, 2005 Consequences of reproductive barriers for genealogical discordance in the European corn borer. Proc. Natl. Acad. Sci. U. S. A. 102: 14706–14711.

Dorant, Y., H. Cayuela, K. Wellband, M. Laporte, Q. Rougemont *et al.*, 2020 Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. Mol. Ecol. 29: 4765–4782.

Durand, N. C., M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley *et al.*, 2016 Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 3: 95–98.

Eastman, J. T., 1993 Antarctic Fish Biology: Evolution in a Unique Environment. Academic Press.

Eastman, J. T., and A. L. DeVries, 1985 Adaptations for cryopelagic life in the antarctic notothenioid fish *Pagothenia borchgrevinki*. Polar Biol. 4: 45–52.

Eastman, J. T., and A. L. DeVries, 1986. Antarctic fishes. Sci. Am. 254:106–114

Eastman, J. T., and A. R. McCune, 2000 Fishes on the Antarctic continental shelf: evolution of amarine species flock?*. J. Fish Biol. 57: 84–102.

Eastman, J. T., and R. R. Eakin, 2021 Checklist of the species of notothenioid fishes. Antarct. Sci. 33: 273–280.

Eastman, J. T., 2005 The nature of the diversity of Antarctic fishes. Polar Biol. 28:93–107

Eddy, S. R., 2011. Accelerated profile HMM searches. PLoS comput. Biol. 7: e1002195.

Ekblom, R., and J. B. Wolf, 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. *7*: 1026–1042.

English, A. C, S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley *et al.*, 2012 Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE: 7: e47768.

Etter, P. D., S. Bassham, P. A. Hohenlohe, E. A. Johnson, and W. A. Cresko, 2011 SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol. Biol. Clifton NJ 772: 157–178.

Fago, A., R. D'avino, and G. Di Prisco, 1992 The hemoglobins of Notothenia angustata, a temperate fish belonging to a family largely endemic to the Antarctic Ocean. Eur. J. Biochem. 210: 963–970.

Faria, R., K. Johannesson, R. K. Butlin, and A. M. Westram, 2019 Evolving Inversions. Trends Ecol. Evol. 34: 239–248.

Faure, G., and T. M. Mensing, 2010 Break-up of Gondwana and assembly of Antarctica. In: The transantarctic mountains: Rocks, ice, meteorites and water, pp.491–515.

Fichot, E. B, and R. S. Norman, 2013 Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome 1: 1–5.

Fierst, J. L. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front. Genet. 6: 220.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark *et al.*, 2020 RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. 117: 9451–9457.

Fulgione, A., C. Neto, A. F. Elfarargi, E. Tergemina, S. Ansari *et al.*, 2022 Parallel reduction in flowering time from de novo mutations enable evolutionary rescue in colonizing lineages. Nat. Commun. 13: 1461.

Futuyma, D. J., and M. Kirkpatrick, 2017 *Evolution*. Sinauer Associates is an imprint of Oxford University Press, Sunderland, Massachusetts.

Gabriel, L., K. J. Hoff, T. Brůna, M. Borodovsky, and M. Stanke, 2021 TSEBRA: transcript selector for BRAKER. BMC Bioinformatics 22: 566.

Gautier, M., and R. Vitalis, 2012 rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28: 1176–1177.

Geoghegan, G., J. Simcox, M. M. Seldin, T. J. Parnell, C. Stubben, S. Just, L. Begaye, A. J. Lusis, and C. J. Villanueva, 2019. Targeted deletion of Tcf7l2 in adipocytes promotes adipocyte hypertrophy and impaired glucose metabolism. Mol. Metab. 24: 44–63

Giani, A. M., G. R. Gallo, L. Gianfranceschi, G. Formenti, 2020 Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput. Struct. Biotechnol. J 18: 9-19.

Glickman, G., I. C. Webb, J. A. Elliott, R. M. Baltazar, M. E. Reale *et al.*, 2012 Photic sensitivity for circadian response to light varies with photoperiod. J. Biol. Rhythms 27: 308–318.

González, J., K. Lenkov, M. Lipatov, J. M. Macpherson, and D. A. Petrov, 2008 High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster (M. A. F. Noor, Ed.). PLoS Biol. 6: e251.

Goriki, A., F. Hatanaka, J. Myung, J. K. Kim, T. Yoritaka, S. Tanoue, T. Abe, H. Kiyonari, K. Fujimoto, Y. Kato, and T. Todo, 2014 A novel protein, CHRONO, functions as a core component of the mammalian circadian clock. PLoS Biol. 12: e1001839.

Gorkovskiy, A. and K. J. Verstrepen, 2021 The role of structural variation in adaptation and evolution of yeast and other fungi. Genes 12:699.

Gross, J. M., B. D. Perkins, A. Amsterdam, A. Egaña, T. Darland *et al.*, 2005 Identification of Zebrafish Insertional Mutants With Defects in Visual System Development and Function. Genetics 170: 245–261.

Guerrero, R. F., and M. Kirkpatrick, 2014 Local adaptation and the evolution of chromosome fusions. Evolution 68: 2747–2756

Gurevich, A., S. Vladislav, V. Nikolay, T. Glenn, 2013 QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075.

Hara, Y., Y. Onishi, K. Oishi, K. Miyazaki, A. Fukamizu, and N. Ishida, 2009 Molecular characterization of Mybbp1a as a co-repressor on the Period2 promoter. Nucleic Acid Res. 37:1115-1126.

Hayes, J. M., A. Hartsock, B. S. Clark, H. R. L. Napier, B. A. Link *et al.*, 2012 Integrin α5/fibronectin1 and focal adhesion kinase are required for lens fiber morphogenesis in zebrafish. Mol. Biol. Cell 23: 4725–4738.

Heather, J. M., and B. Chain, 2016 The sequence of sequencers: The history of sequencing DNA. Genomics 107: 1-8.

Hessel, E. M., W. W. Cruikshank, I. Van Ark, J. J. De Bie, B. Van Esch *et al.*, 1998 Involvement of IL-16 in the Induction of Airway Hyper-Responsiveness and Up-Regulation of IgE in a Murine Model of Allergic Asthma. J. Immunol. 160: 2998–3005.

Hill, J., E. D. Enbody, M. E. Pettersson, C. G. Sprehn, D. Bekkevold *et al.*, 2019 Recurrent convergent evolution at amino acid residue 261 in fish rhodopsin. Proc. Natl. Acad. Sci. 116: 18473–18478.

Hirst, J., 2013 Mitochondrial Complex I. Annu. Rev. Biochem. 82: 551–575.

Hofmann, G. E., B. A. Buckley, S. Airaksinen, J. E. Keen, and G. N. Somero, 2000 Heat-shock protein expression is absent in the antarctic fish Trematomus bernacchii (family Nototheniidae). J. Exp. Biol. 203: 2331–2339.

Hu, C., J. Yang, Z. Qi, H. Wu, B. Wang *et al.*, 2022 Heat shock proteins: Biological functions, pathological roles, and therapeutic opportunities. MedComm 3: e161.

Hudson, R. R., and J. A. Coyne, 2002 Mathematical consequences of the genealogical species concept. Evolution 56: 1557–1565.

Hut, R. A., S. Paolucci, R. Dor, C. P. Kyriacou, and S. Daan, 2013 Latitudinal clines: an evolutionary view on biological rhythms†,‡. Proc. R. Soc. B Biol. Sci. 280: 20130433.

Kampinga, H. H., and E. A. Craig, 2010 The Hsp70 chaperone machinery: J-proteins as drivers of functional specificity. Nat. Rev. Mol. Cell Biol. 11: 579–592.

Kelleher, J., Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers *et al.*, 2019 Inferring whole-genome histories in large population datasets. Nat. Genet. 51: 1330–1338.

Keller, O., M. Kollmar, M. Stanke, S. Waack, 2011 A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27: 757–763.

Kennett, J. P., 1977 Cenozoic evolution of Antarctic glaciation, the circum-Antarctic Ocean, and their impact on global paleoceanography. J. Geophys. Res. 82: 3843–3860.

Kidwell, M.G., and D. Lisch, 1997. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci. 94: 7704–7711.

Kim, B.-M., A. Amores, S. Kang, D.-H. Ahn, J.-H. Kim *et al.*, 2019 Antarctic blackfin icefish genome reveals adaptations to extreme environments. Nat. Ecol. Evol. 3: 469–478.

Klages, J. P., U. Salzmann, T. Bickert, C.-D. Hillenbrand, K. Gohl *et al.*, 2020 Temperate rainforests near the South Pole during peak Cretaceous warmth. Nature 580: 81–86.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37: 540–546.

Kong, W., Wang, Y., Zhang, S., Yu, J. and Zhang, X., 2023. Recent Advances in Assembly of Complex Plant Genomes. Genom. Proteom. Bioinform. 21:427–439.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Res. 27: 722–736.

Kratochwil, C. F., A. F. Kautt, A. Nater, A. Härer, Y. Liang *et al.*, 2022 An intronic transposon insertion associates with a trans-species color polymorphism in Midas cichlid fishes. Nat. Commun. 13: 296.

Kriventseva, E. V., D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias *et al.*, 2019 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47: D807–D811.

Küpper, C., M. Stocks, J. E. Risse, N. Dos Remedios, L. L. Farrell *et al.*, 2016 A supergene determines highly divergent male reproductive morphs in the ruff. Nat. Genet. 48: 79–83.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: 1–9.

La Mesa, M., J. T. Eastman, and M. Vacchi, 2004 The role of notothenioid fish in the food web of the Ross Sea shelf waters: a review. Polar Biol. 27: 321–338.

Lai, Y.-T., C. K. L. Yeung, K. E. Omland, E.-L. Pang, Y. Hao *et al.*, 2019 Standing genetic variation as the predominant source for adaptation of a songbird. Proc. Natl. Acad. Sci. 116: 2152–2157.

Lake, N.J., B. D. Webb, D. A. Stroud, T. R. Richman, B. Ruzzenente, A. G. Compton, H. S. Mountford, J. Pulman, C. Zangarelli, M. Rio, and N. Boddaert, 2017. Biallelic mutations in MRPS34 lead to instability of the small mitoribosomal subunit and Leigh syndrome. Am. J. Hum. Genet. 101: 239-254.

Lamichhaney, S., G. Fan, F. Widemo, U. Gunnarsson, D. S. Thalmann *et al.*, 2016 Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nat. Genet. 48: 84–88.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature, 409: 860–921.

Lau, D. T., A. Saeed-Kothe, S. K. Parker, and H. William Detrich, 2001 Adaptive Evolution of Gene Expression in Antarctic Fishes: Divergent Transcription of the 5′-to-5′ Linked Adult α1- and β-Globin Genes of the Antarctic Teleost *Notothenia coriiceps* is Controlled by Dual Promoters and Intergenic Enhancers. Integr. Comp. Biol. 41: 113–132.

Leinonen, M., and L. Salmela, 2020 Optical map guided genome assembly. BMC Bioinformatics, 21: 1–19.

Levy, S. E., and R. M. Myers, 2016 Advancements in next-generation sequencing. Annu. Rev. Genomics Hum. Genet. 17: 95–115.

Li, G., L. Wang, J. Yang, H. He, H. Jin, X. Li, T. Ren, Z. Ren, F. Li, X. Han *et al.*, 2021 A high-quality genome assembly highlights rye genomic characteristic and agronomically important genes. Nat. Genet. 53: 574–584.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin R, 2009 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics, 25: 2078–2079.

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34: 3094–3100.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences (I. Birol, Ed.). Bioinformatics 34: 3094–3100.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

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

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

Li, P., S. S. Chaurasia, Y. Gao, A. L. Carr, P. M. Iuvone *et al.*, 2008. CLOCK Is Required for Maintaining the Circadian Rhythms of Opsin mRNA Expression in Photoreceptor Cells. J. Biol. Chem. 283: 31673–31678.

Liao, X., M. Li, Y. Zou, F. X. Wu, and J. Wang, 2019 Current challenges and solutions of de novo assembly. Quant. Biol. 7: 90–109.

Lin, S.-Y., M. A. Vollrath, S. Mangosing, J. Shen, E. Cardenas *et al.*, 2016 The zebrafish pinball wizard gene encodes WRB, a tail-anchored-protein receptor essential for inner-ear hair cells and retinal photoreceptors. J. Physiol. 594: 895–914.

Liu, J., X. Ji, and H. Chen, 2022 Beta-PSMC: uncovering more detailed population history using beta distribution. BMC Genomics 23: 785.

Liu, X., and Y.-X. Fu, 2020 Stairway Plot 2: demographic history inference with folded SNP frequency spectra. Genome Biol. 21: 280.

Liu, Z., M. Roesti, D. Marques, M. Hiltbrunner, V. Saladin *et al.*, 2022 Chromosomal Fusions Facilitate Adaptation to Divergent Environments in Threespine Stickleback (M. O'Connell, Ed.). Mol. Biol. Evol. 39: msab358.

Logsdon, G. A., M. R. Vollger, E. E. Eichler, 2020 Long-read human genome sequencing and its applications. Nat. Rev. Genet. 21: 597-614.

Logue, J. A., A. L. de Vries, E. Fodor, and A. R. Cossins, 2000 Lipid compositional correlates of temperature-adaptive interspecific differences in membrane physical structure. J. Exp. Biol. 203: 2105–2115.

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

Löytynoja, A., 2014 Phylogeny-aware alignment with PRANK, pp. 155–170 in Multiple Sequence Alignment Methods, edited by D. J. Russell. Methods in Molecular Biology, Humana Press, Totowa, NJ.

Lu, Y., W. Li, Y. Li, W. Zhai, X. Zhou *et al.*, 2022 Population genomics of an icefish reveals mechanisms of glacier-driven adaptive radiation in Antarctic notothenioids. BMC Biol. 20: 231.

Lynch, M., L.-M. Bobay, F. Catania, J.-F. Gout, and M. Rho, 2011 The Repatterning of Eukaryotic Genomes by Random Genetic Drift. Annu. Rev. Genomics Hum. Genet. 12: 347–366.

Macdonald, J. A., and R. M. G. Wells, 1991 Viscosity of Body Fluids From Antarctic Notothenioid Fish, pp. 163–178 in *Biology of Antarctic Fish*, edited by G. di Prisco, B. Maresca, and B. Tota. Springer Berlin Heidelberg, Berlin, Heidelberg.

Macías, L. G., E. Barrio, and C. Toft, 2020 GWideCodeML: A Python Package for Testing Evolutionary Hypotheses at the Genome-Wide Level. G3 (Bethesda)10: 4369–4372.

Marijon, P., R. Chikhi, and J.-S. Varré, 2020 yacrd and fpa: upstream tools for long-read genome assembly. Bioinformatics 36: 3894–3896.

Martinez Barrio, A., S. Lamichhaney, G. Fan, N. Rafati, M. Pettersson, H. E. Zhang, J. Dainat, D. Ekman, M. Höppner, P. Jern, and M. Martin *et al.*, 2016 The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. elife. 5: e12081.

Mathy, N. L., W. Scheuer, M. Lanzendörfer, K. Honold, D. Ambrosius *et al.*, 2000 Interleukin-16 stimulates the expression and production of pro-inflammatory cytokines by human monocytes: *IL-16 and pro-inflammatory cytokines*. Immunology 100: 63–69.

Mazzei, F., L. Ghigliotti, J.-P. Coutanceau, H. W. Detrich, V. Prirodina *et al.*, 2008 Chromosomal characteristics of the temperate notothenioid fish *Eleginops maclovinus* (Cuvier). Polar Biol. 31: 629.

McFall-Ngai, M. J., and J. Horwitz, 1990 A comparative study of the thermal stability of the vertebrate eye lens: Antarctic ice fish to the desert iguana. Exp. Eye Res. 50: 703–709.

Meng, J., Z. Yao, Y. He, R. Zhang, Y. Zhang, X. Yao, H. Yang, L. Chen, Z. Zhang, H. Zhang, and X. Bao, 2017. ARRDC4 regulates enterovirus 71-induced innate immune response by promoting K63 polyubiquitination of MDA5 through TRIM65. Cell Death Dis. 8: e2866.

Menon, D.U., Y. Shibata, W. Mu, and T. Magnuson, 2019. Mammalian SWI/SNF collaborates with a polycomb-associated protein to regulate male germline transcription in the mouse. Development. 146: dev174094.

Metzker, M. L., 2005 Emerging technologies in DNA sequencing. Genome Res. 15: 1767–1776.

Mi, H., D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou *et al.*, 2021 PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 49: D394–D403.

Montejo-Kovacevich, G., J. I. Meier, C. N. Bacquet, I. A. Warren, Y. F. Chan *et al.*, 2022 Repeated genetic adaptation to altitude in two tropical butterflies. Nat. Commun. 13: 4676.

Moorjani, P., and G. Hellenthal, 2023 Methods for Assessing Population Relationships and History Using Genomic Data. Annu. Rev. Genomics Hum. Genet. 24: 305–332.

Moran, R. L., J. M. Catchen, and R. C. Fuller, 2019 Genomic resources for darters (Percidae: Etheostominae) provide insight into postzygotic barriers implicated in speciation. Mol. Biol. and Evol., 37: 711–729.

Morescalchi, A., E. Pisano, R. Stanyon, and M. A. Morescalchi, 1992 Cytotaxonomy of antarctic teleosts of the Pagothenia/Trematomus complex (Nototheniidae, Perciformes). Polar Biol. 12: 553–558

Murigneux, V., S. K. Rai, A. Furtado, T. J. C. Bruxner, W. Tian, I. Harliwong, H. Wei, B. Yang, Q. Ye, E. Anderson *et al*., 2020 Comparison of long-read methods for sequencing and assembly of a plant genome. GigaScience, 9: giaa146.

Myers, E.W., 2005 The fragment assembly string graph. Bioinformatics, 21: ii79-ii85.

Near, T. J., D. J. MacGuigan, E. Parker, C. D. Struthers, C. D. Jones, A. Dornburg, 2018 Phylogenetic analysis of Antarctic notothenioids illuminates the utility of RADseq for resolving Cenozoic adaptive radiations. Mol. Phylogenet. Evol. 129: 268-279.

Near, T. J., J. J. Pesavento, and C.-H. C. Cheng, 2004 Phylogenetic investigations of Antarctic notothenioid fishes (Perciformes: Notothenioidei) using complete gene sequences of the mitochondrial encoded 16S rRNA. Mol. Phylogenet. Evol. 32: 881–891.

Nelson, T. C., and W. A. Cresko, 2018 Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evol. Lett. 2: 9–21.

O'Connell, J., O. Schulz-Trieglaff, E. Carlson, M. M. Hims, N. A. Gormley, 2015 NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics, 31: 2035-2037.

Oniszczuk, J., K. Sendeyo, C. Chhuon, B. Savas, E. Cogné *et al.*, 2020 CMIP is a negative regulator of T cell signaling. Cell. Mol. Immunol. 17: 1026–1041.

Ou, C., F. Wang, J. Wang, S. Li, Y. Zhang, M. Fang, L. Ma, Y. Zhao, S. Jiang, 2019 A de novo genome assembly of the dwarfing pear rootstock Zhongai 1. Scientific Data 6: 1–8.

Oud, M.S., Ö. Okutman, L. A. J. Hendricks, P.F. de Vries, B. J. Houston, L. E. M. Vissers, M. K. O'Bryan, L. Ramos, H. E. Chemes, S. Viville, and J. A. Veltman, 2020. Exome sequencing reveals novel causes as well as new candidate genes for human globozoospermia. Hum. Reprod. 35:240–252.

Patarnello, T., C. Verde, G. Di Prisco, L. Bargelloni, and L. Zane, 2011 How will fish that evolved at constant sub-zero temperatures cope with global warming? Notothenioids as a case study. BioEssays 33: 260–268.

Pisano, E., L. Ghigliotti, F. Mazzei, and C. Ozouf-Costaz, 2003 Cytogenetic features of Notothenia angustata Hutton, 1875, an Antarctic fish living in non-Antarctic water. Antarct. Biol. Glob. Context Ed. Huiskes AHL WWC Gieskes J Rozema 117–120.

Place, S. P., M. L. Zippay, and G. E. Hofmann, 2004 Constitutive roles for inducible genes: evidence for the alteration in expression of the inducible *hsp70* gene in Antarctic notothenioid fishes. Am. J. Physiol.-Regul. Integr. Comp. Physiol. 287: R429–R436.

Qiu, X.-B., Y.-M. Shao, S. Miao, and L. Wang, 2006 The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. Cell. Mol. Life Sci. 63: 2560–2570.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder *et al.*, 2005 InterProScan: protein domains identifier. Nucleic Acids Res. 33: W116–W120.

Quintana, A. M., F. Picchione, R. I. Klein Geltink, M. R. Taylor, and G. C. Grosveld, 2013 Zebrafish *etv7* regulates red blood cell development through the cholesterol synthesis pathway. Dis. Model. Mech. dmm.012526.

Rayamajhi, N., C.-H. C. Cheng, and J. M. Catchen, 2022 Evaluating Illumina-, Nanopore-, and PacBio-based genome assembly strategies with the bald notothen, Trematomus borchgrevinki. G3 (Bethesda) 12: jkac192.

Rhie, A., S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim *et al*., 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592: 737-746.

Rice, E. S., and R. E. Green, 2019. New approaches for genome assembly and scaffolding. Annu. Rev. Anim. Biosci. 7: 17–40.

Rico, E.P., D. B. Rosemberg, M. R. Senger, M. de Bem Arizi, R. D. Dias, A. A. Souto, M. R. Bogo, and C. D. Bonan, 2008. Ethanol and acetaldehyde alter NTPDase and 5′-nucleotidase from zebrafish brain membranes. Neurochemistry International 52:290–296.

Rivera-Colón, A. G., 2022 Comparative and evolutionary genomics of secondarily temperate adaptation in a non-Antarctic icefish [Thesis]: University of Illinois, Urbana-Champaign.

Rivera-Colón, A. G., N. Rayamajhi, B. F. Minhas, G. Madrigal, K. T. Bilyk *et al.*, 2023 Genomics of Secondarily Temperate Adaptation in the Only Non-Antarctic Icefish. Mol. Biol. Evol. 40: msad029.

Rochette, N. C., A. G. Rivera-Colón, and J. M. Catchen, 2019 Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. 28: 4737–4754.

Rothberg, J. M., and L. H. Leamon, 2008 The development and impact of 454 sequencing. *Nat. Biotechnol.* 26:1117–1124.

Ruan, J., and H. Li, 2020 Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17: 155–158.

Rubenstein, D. R., J. A. Ågren, L. Carbone, N. C. Elde, H. E. Hoekstra *et al.*, 2019 Coevolution of Genome Architecture and Social Behavior. Trends Ecol. Evol. 34: 844–855.

Sahlin, K., R. Chikhi, and L. Arvestad, 2016 Assembly scaffolding with PE-contaminated mate-pair libraries. Bioinformatics 32: 1925-1932.

Salmela, L., R. Walve, E. Rivals, and E. Ukkonen, 2017 Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics, 33: 799–806.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg, 2010 Assembly of large genomes using second-generation sequencing. Genome Res. 20: 1165–1173.

Scher, H. D., and E. E. Martin, 2006 Timing and climatic consequences of the opening of Drake Passage. Science 312:428–430.

Sedlazeck, F. J., H. Lee, C. A. Darby, and M. C. Schatz, 2018 Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19: 329-346.

Shao, F., M. Han, and Z. Peng, 2019 Evolution and diversity of transposable elements in fish genomes. Sci. Rep. 9: 15399.

Sharma, P., E. Maklashina, G. Cecchini, and T. M. Iverson, 2020 The roles of SDHAF2 and dicarboxylate in covalent flavinylation of SDHA, the human complex II flavoprotein. Proc. Natl. Acad. Sci. 117: 23548–23556.

Shin, S. C., D. H. Ahn, S. J. Kim, C. W. Pyo, H. Lee *et al.*, 2014 The genome sequence of the Antarctic bullhead notothen reveals evolutionary adaptations to a cold environment. Genome Biol. 15: 468.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Simpson, J., and R. Durbin R., 2012 Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22: 549–556.

Simpson, J. T., and M. Pop, 2015 The theory and practice of genome sequence assembly. Annu. Rev. Genomics Hum. Genet., 16, 153–172.

Sohn, J. I., and J. W. Nam, 2018 The present and future of de novo whole-genome assembly. Brief Bioinformatics 19: 23–40.

Somero, G. N., and A. L. DeVries, 1967 Temperature Tolerance of Some Antarctic Fishes. Science 156: 257–258.

Spåhr, H., B. Habermann, C. M. Gustafsson, N.-G. Larsson, and B. M. Hallberg, 2012 Structure of the human MTERF4–NSUN4 protein complex that regulates mitochondrial ribosome biogenesis. Proc. Natl. Acad. Sci. 109: 15253–15258.

Stankovic, A., K. Spalik, E. Kamler, P. Borsuk, and P. Weglenski, 2002 Recent origin of sub-Antarctic notothenioids. Polar Biol. 25: 203–205.

Stickley, C.E., H. Brinkhuis, S.A. Schellenberg, A. Sluijs, U. Röhl *et al.*, 2004 Timing and nature of the deepening of the Tasmanian Gateway. Paleoceanography. 19(4)

Storch, D., G. Lannig, and H. O. Pörtner, 2005 Temperature-dependent protein synthesis capacities in Antarctic and temperate (North Sea) fish (Zoarcidae). J. Exp. Biol. 208: 2409–2420.

Storey, B.C., and R. Granot, 2021 Tectonic history of Antarctica over the past 200 million years. Geol. Soc. Lond. Mem. 55:9-17.

Stuart, M. D., 1998 The seasonal ecophysiology of Undaria pinnatifida (Harvey) Suringar in Otago Harbour, New Zealand [Thesis]: University of Otago.

Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley, 2011 Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. 43: 1160–1163.

Sukumaran, S., and B. D. Perkins, 2009 Early defects in photoreceptor outer segment morphogenesis in zebrafish ift57, ift88 and ift172 Intraflagellar Transport mutants. Vision Res. 49: 479–489.

Sullivan, M. J., N. L. B. Zakour, B. M. Forde, M. Stanton-Cook, S. A. Beatson, 2015. Contiguity: contig adjacency graph construction and visualisation. PeerJ PrePrints 3: e1037v1.

Sun, W., Y.-H. Shen, M.-J. Han, Y.-F. Cao, and Z. Zhang, 2014 An Adaptive Transposable Element Insertion in the Regulatory Region of the EO Gene in the Domesticated Silkworm, Bombyx mori. Mol. Biol. Evol. 31: 3302–3313.

Tao, Y., X. Zhao, E. Mace, R. Henry, D. Jordan, 2019. Exploring and exploiting pan-genomics for crop improvement. Molecular Plant, 12: 156-169.

Terhorst, J., J. A. Kamm, and Y. S. Song, 2017 Robust and scalable inference of population history from hundreds of unphased whole-genomes. Nat. Genet. 49: 303–309.

Thakur, A., B. Chitoor, A. V. Goswami, G. Pareek, H. S. Atreya *et al.*, 2012 Structure and Mechanistic Insights into Novel Iron-mediated Moonlighting Functions of Human J-protein Cochaperone, Dph4. J. Biol. Chem. 287: 13194–13205.

Tian, L., H. McClafferty, H. G. Knaus, P. Ruth, and M. J. Shipston, 2012. Distinct acyl protein transferases and thioesterases control surface expression of calcium-activated potassium channels. J. Biol. Chem. 287:14718–14725.

Tiana, M., B. Acosta-Iborra, L. Puente-Santamaría, P. Hernansanz-Agustin, R. Worsley-Hunt, N. Masson, F. García-Rio, D. Mole, P. Ratcliffe, W. W. Wasserman, and B. Jimenez, 2018. The SIN3A histone deacetylase complex is required for a complete transcriptional response to hypoxia. Nucleic Acids Res. 120–133.

Tigano, A., T. K. Reiertsen, J. R. Walters, and V. L. Friesen, 2018 A complex copy number variant underlies differences in both colour plumage and cold adaptation in a dimorphic seabird: Evol. Biol. preprint.

Tiley, G.P., T. Flouri, X. Jiao, J. W. Poelstra, B. Xu, T. Zhu, B. Rannala, A. D. Yoder, and Z. Yang, 2023. Estimation of species divergence times in presence of cross-species gene flow. Syst. Biol. 72:820–836.

Todgham, A. E., E. A. Hoaglund, and G. E. Hofmann, 2007 Is cold the new hot? Elevated ubiquitin-conjugated protein levels in tissues of Antarctic fish as evidence for cold-denaturation of proteins in vivo. J. Comp. Physiol. [B] 177: 857–866.

Treangen, T. J., and S. L. Salzberg, 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet, 13: 36-46.

Tvedte, E. S., M. Gasser, B. C. Sparklin, J. Michalski, C. E. Hjelmen, J. S. Johnston, X. Zhao, R. Bromley, L. J. Tallon, L. Sadzewicz *et al.*, 2021 Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. G3 (Bethesda) 11: jkab083.

Vakirlis, N., V. Sarilar, G. Drillon, A. Fleiss, N. Agier *et al.*, 2016 Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. Genome Res. 26: 918–932.

Van Dijk, E. L., Y. Jaszczyszyn, D. Naquin, and C. Thermes, 2018 The third revolution in sequencing technology. Trends in Genetics, 34: 666–681.

Vatine, G., D. Vallone, Y. Gothilf, and N. S. Foulkes, 2011 It's time to swim! Zebrafish and the circadian clock. FEBS letters 585:1485–1494.

Vijay, N., M. Weissensteiner, R. Burri, T. Kawakami, H. Ellegren, and J. B. Wolf, 2017. Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. Mol. Ecol. 26:4284–4295.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young *et al.*, 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS ONE, 9: e112963.

Wan, Q., D. Song, H. Li, and M. He, 2020 Stress proteins: the biological functions in virus infection, present and challenges for target-based antiviral drug development. Signal Transduct. Target. Ther. 5: 1–40.

Watson, M., and A. Warr, 2019 Errors in long-read assemblies can critically affect protein prediction. Nat. Biotechnol., 37: 124–126.

Webster, M.T., A. Beaurepaire, P. Neumann, and E. Stolle, 2023 Population genomics for insect conservation. Annu. Rev. Anim. Biosci. 11: 115–140.

Wellband, K., C. Mérot, T. Linnansaari, J. A. K. Elliott, R. A. Curry *et al.*, 2019 Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. Mol. Ecol. 28: 1439–1459.

Wellenreuther, M., C. Mérot, E. Berdan, and L. Bernatchez, 2019 Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Mol. Ecol. 28: 1203–1209.

White, R., C. Pellefigues, F. Ronchese, O. Lamiable, D. Eccles, 2017 Investigation of chimeric reads using the MinION. F1000Research, 6.

Whitmore, D., N. S. Foulkes, U. Strähle, and P. Sassone-Corsi, 1998 Zebrafish Clock rhythmic expression reveals independent peripheral circadian oscillators. Nat. Neurosci. 1:701–707.

Wohns, A.W., Y. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, D. Reich, J. Kelleher, and G. McVean, 2022. A unified genealogy of modern and ancient genomes. Science 375:eabi8264

Woronik, A., K. Tunström, M. W. Perry, R. Neethiraj, C. Stefanescu *et al.*, 2019 A transposable element insertion is associated with an alternative life history strategy. Nat. Commun. 10: 5757.

Wright, S., 1984 Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations. University of Chicago Press.

Xu, W., J. R. Tucker, W. A. Bekele, F. M. You, Y. B. Fu, R. Khanal, Z. Yao, J. Singh, B. Boyle, A. D. Beattie *et al*., 2021. Genome assembly of the Canadian two-row malting barley cultivar AAC Synergy. G3 (Bethesda) 11: jkab031.

Yang, Z., 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24: 1586–1591.

Yuan, Z., S. Liu, T. Zhou, C. Tian, L. Bao *et al.*, 2018 Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics 19: 141.

Zachos, J., M. Pagani, L. Sloan, E. Thomas, and K. Billups, 2001 Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. Science 292: 686–693.

Zerbino, D., and E. Birney, 2008 Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821-829.

Zhang, H., C. Jain, and S. Aluru, 2020 A comprehensive evaluation of long read error correction methods. BMC Genomics, 21:1–15.

Zhang, J., and I. Hamza, 2019 Zebrafish as a model system to delineate the role of heme and iron metabolism during erythropoiesis. Mol. Genet. Metab. 128: 204–212.

Zhang, J., W. Cui, C. Du, Y. Huang, X. Pi *et al.*, 2020 Knockout of DNase1l1l abrogates lens denucleation process and causes cataract in zebrafish. Biochim. Biophys. Acta BBA - Mol. Basis Dis. 1866: 165724.

Zhang, J., Y. Zhang, and H. F. Rosenberg, 2002 Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. 30: 411–415.

Zhang, X.W., X. F. Wang, S. J. Ni, W. Qin, L. Q. Zhao, R. X. Hua, Y. W. Lu, J. Li, G. P. Dimri, and W. J. Guo, 2015. UBTD1 induces cellular senescence through an UBTD1–Mdm2/p53 positive feedback loop. The Journal of Pathology, 235:656–667.

Zhao, J., Zhu, J. and Thornhill, W.B., 2013. Spinocerebellar ataxia-13 Kv3. 3 potassium channels: arginine-to-histidine mutations affect both functional and protein expression on the cell surface. Biochemical Journal, 454:259–265.

Zhao, W.-N., N. Malinin, F.-C. Yang, D. Staknis, N. Gekakis *et al.*, 2007 CIPC is a mammalian circadian clock protein without invertebrate homologues. Nat. Cell Biol. 9: 268–275.

Zhou, Y., L. Duvaux, G. Ren *et al.*, 2017 Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. Heredity 118:211–220.

Zhu, J., K. R. Vinothkumar, and J. Hirst, 2016 Structure of mammalian respiratory complex I. Nature 536: 354–358.

Zhuang, X., C. Yang, K. R. Murphy, and C.-H. C. Cheng, 2019 Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proc. Natl. Acad. Sci. 116: 4400–4405.

Zhuang, Y., Z. Li, S. Xiong, C. Sun, B. Li *et al.*, 2023 Circadian clocks are modulated by compartmentalized oscillating translation. Cell 186: 3245-3260.e23.

Zimmer, A., C. Durand, N. Loira, P. Durrens, D. J. Sherman *et al.*, 2014 QTL Dissection of Lag Phase in Wine Fermentation Reveals a New Translocation Responsible for Saccharomyces cerevisiae Adaptation to Sulfite. PLOS ONE 9: e86298.

APPENDIX A: SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER 3

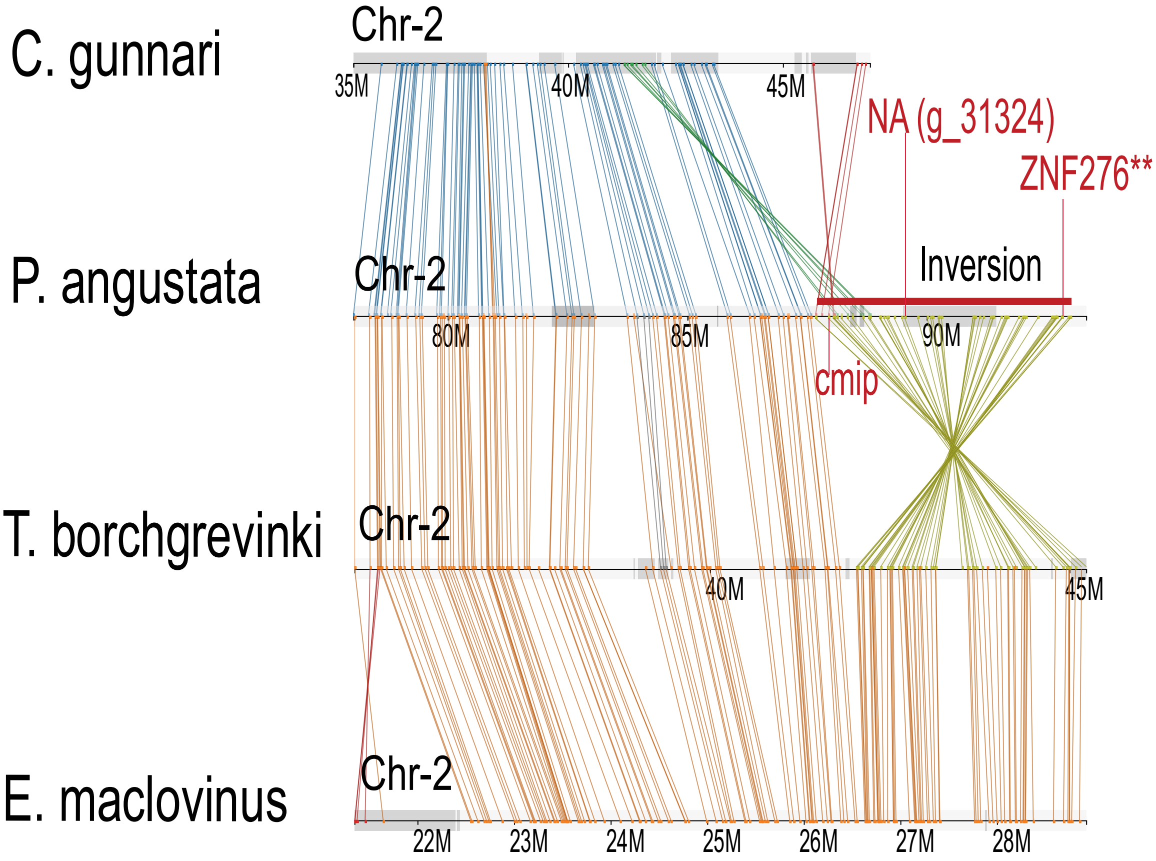


Figure A.1 This figure illustrates examples of putative *P. angustata*-specific inversion (indicated by red horizontal block) in chromosome 2 (Chr-2).

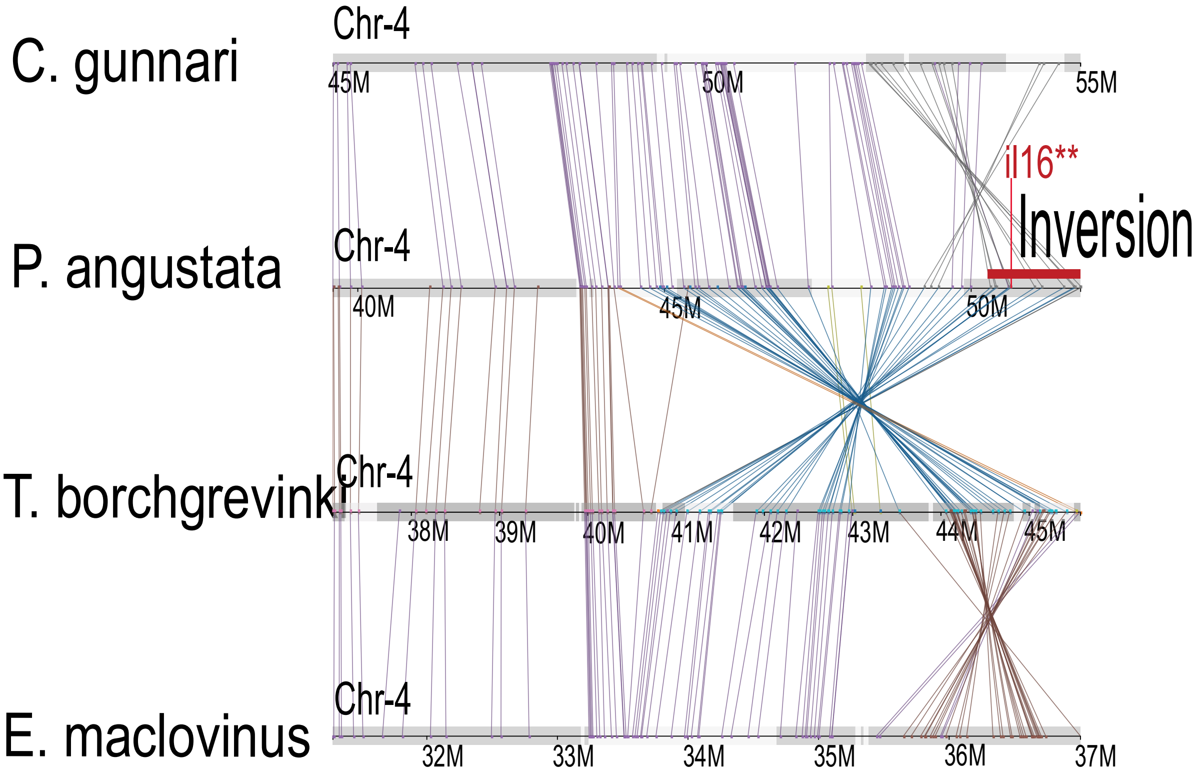


Figure A.2 illustrates examples of inversions (indicated by red blocks in plots **A, B**, **C**, and **D)** specific to *Paranotothenia angsustata*, each consisting of at least one of the "*dN/dS*" candidates. **A**) This figure shows that inversion between 87.66 and 93.02Mbs genomic positions on chromosome 2 also contains three candidates: *cmip*, *NA* (having annotation gene id *g_31324*), and *ZNF276*. **B**) This figure exhibits inversion between 50.31 and 51.80, 0.21 and 4.02, as well as 40.47 and 40.94 Mbs on chromosomes 4, 14, and 15, respectively. Plots **B**), **C**), and **D**) show that inversions on chromosomes 4, 14, and 15 consist of i) *il16*, ii) *si:dkey-106g10.7* and *spata6l*, as well as iii) *nus1* candidates, respectively. The double asterisk (**) indicates that branch and branch-site models identified the same gene under positive selection. *NA* denotes the gene for which ortholog in zebrafish is unavailable.

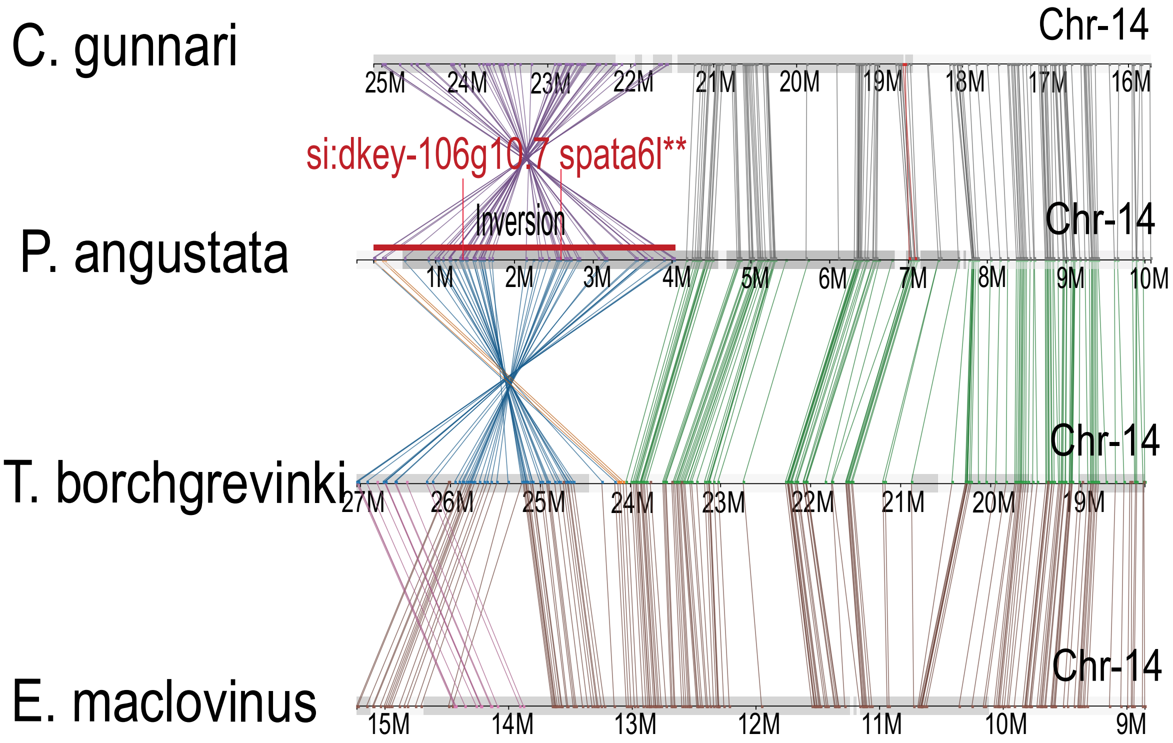


Figure A.3 illustrates examples of putative *P. angustata*-specific inversion (indicated by red horizontal block) in chromosome 14 (Chr-14).

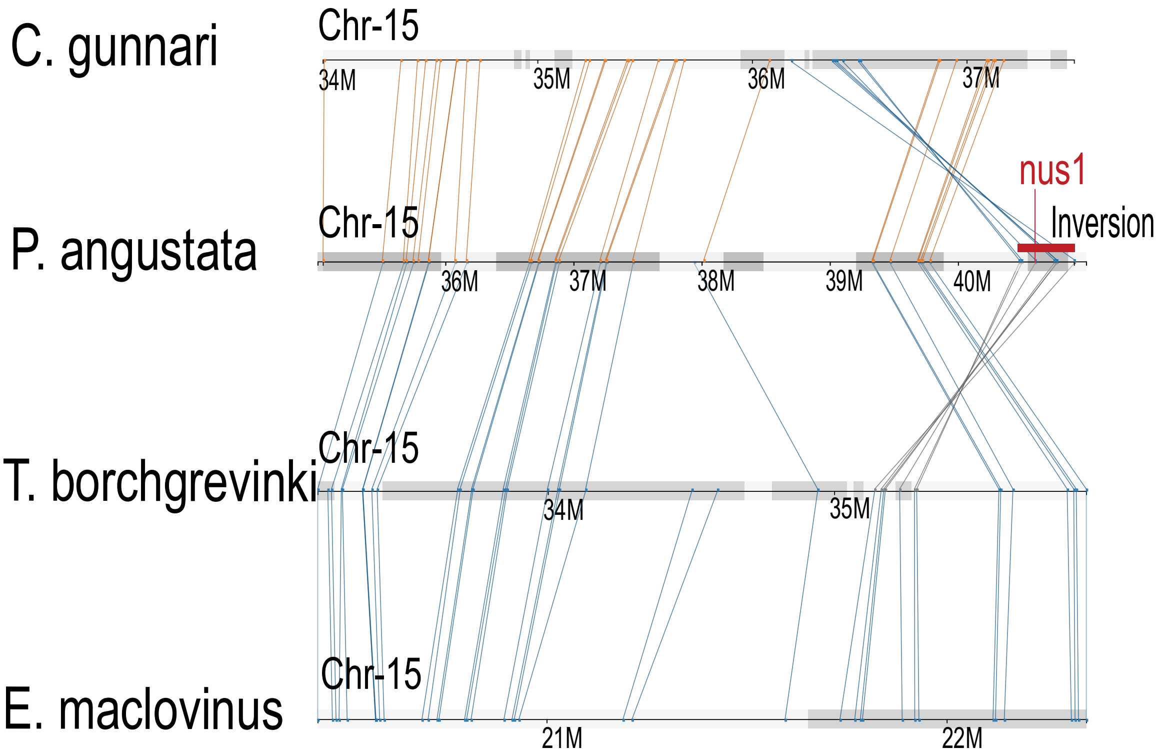


Figure A.4 illustrates examples of putative *P. angustata*-specific inversion (indicated by red horizontal block) in chromosome 15 (Chr-15).

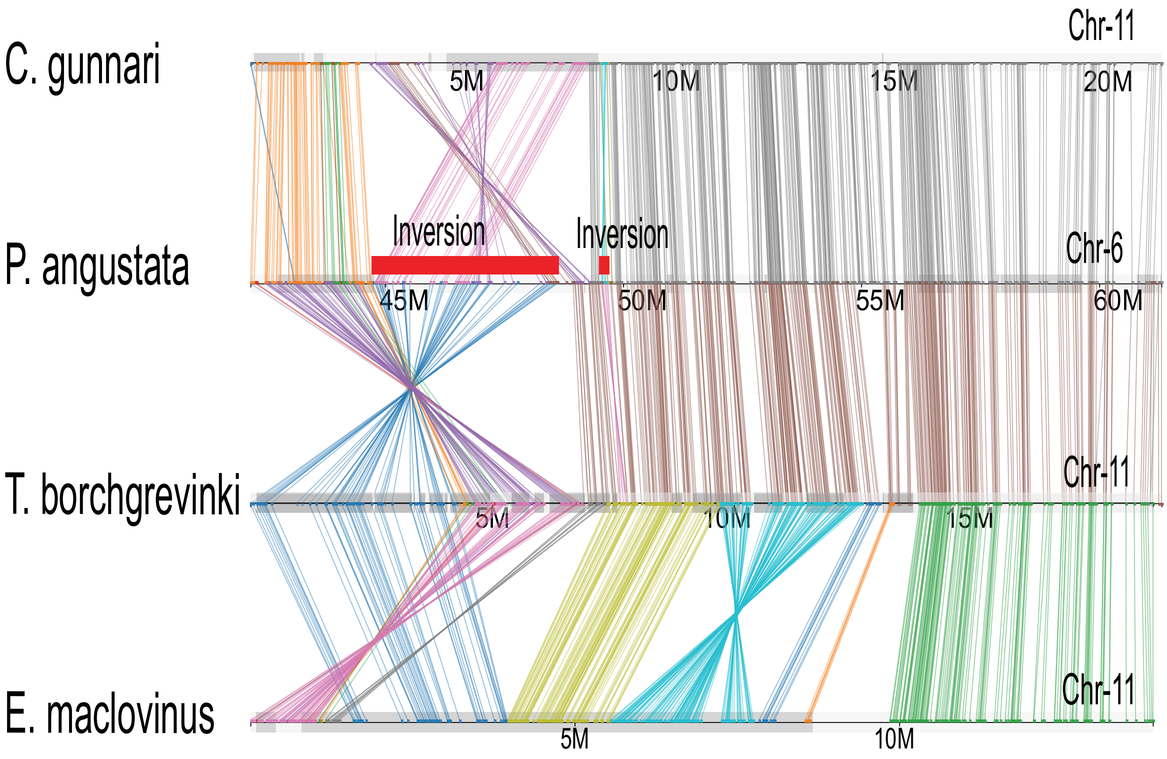


Figure A.5 illustrates examples of putative *P. angustata*-specific inversion (indicated by red horizontal block) in chromosome 11 (Chr-11).

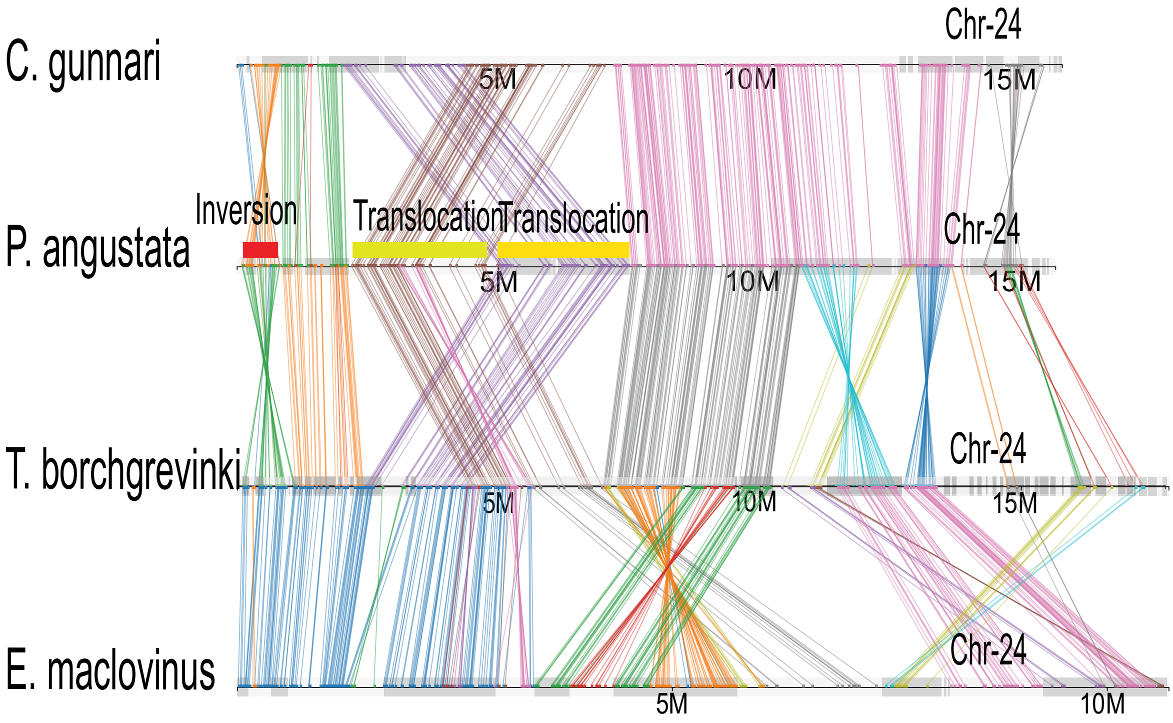


Figure A.6 This figure illustrates examples of one inversion (indicated by a red horizontal block) and two translocations (denoted by two separate green and orange blocks) putatively specific to *P. angustata* and located in chromosome 24 (Chr-

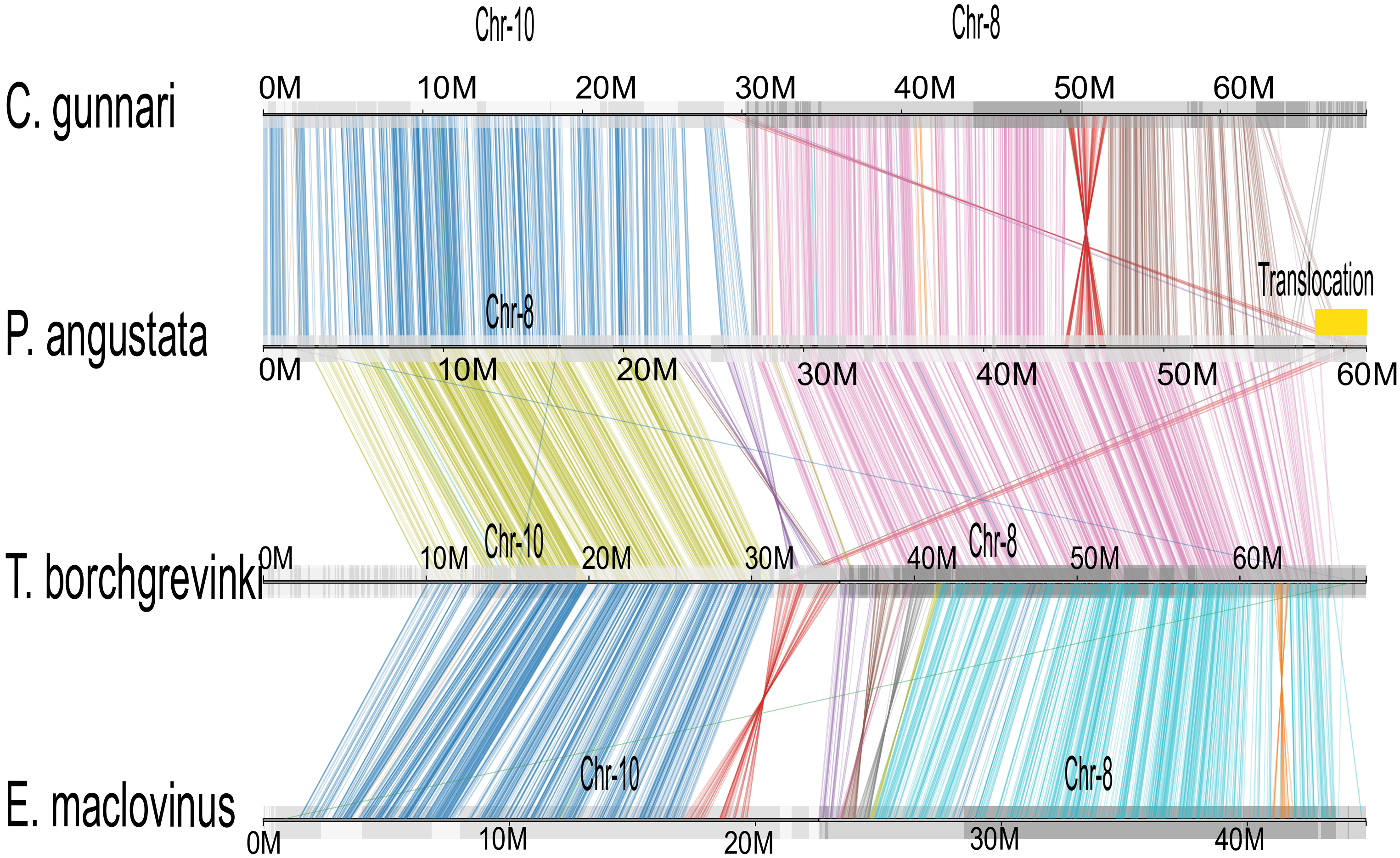


Figure A.7 illustrates examples of putative *P. angustata*-specific translocation (indicated by orange, horizontal block) in chromosome 8(Chr-8).

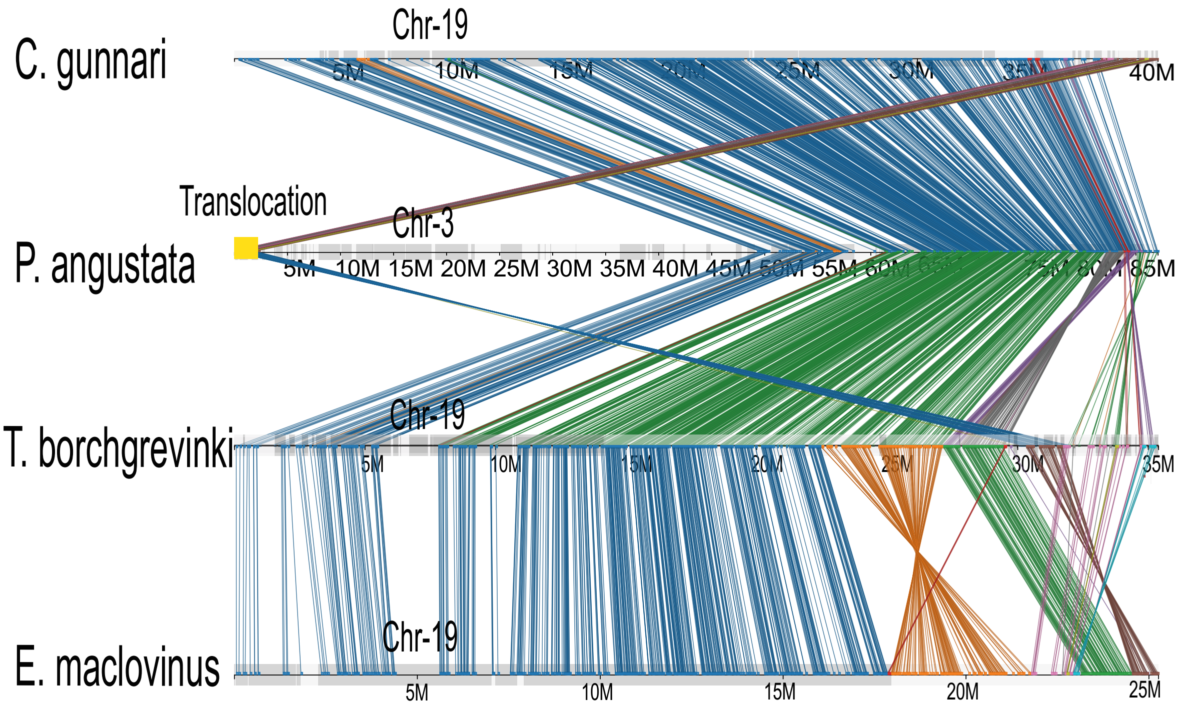


Figure A.8 illustrates examples of putative *P. angustata*-specific translocation (indicated by orange, horizontal block) in chromosome 3(Chr-3).

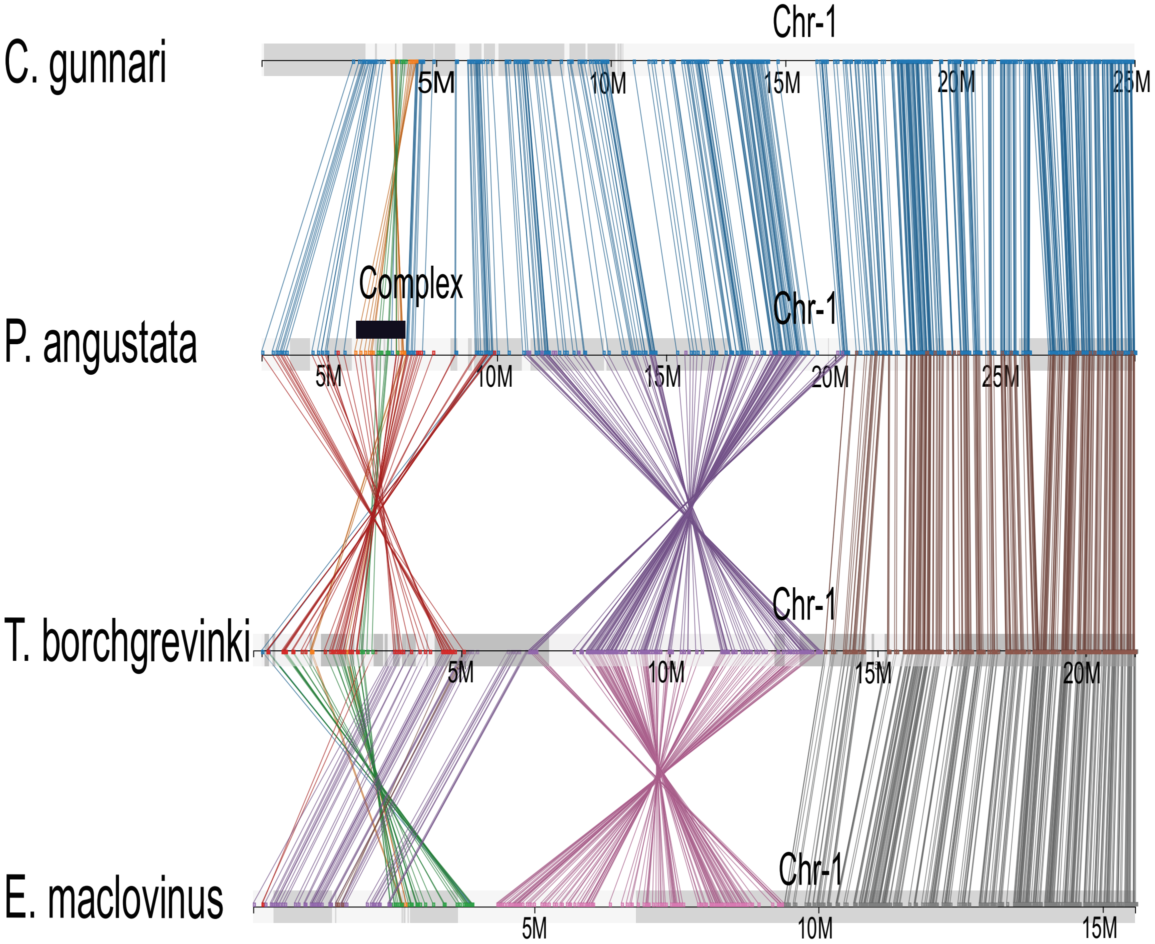


Figure A.9 This figure shows complex structural change (indicated by black block) specific to *P. angustata* within 5.79 and 7.25 M (megabase pairs) genomic positions on chromosome 1.

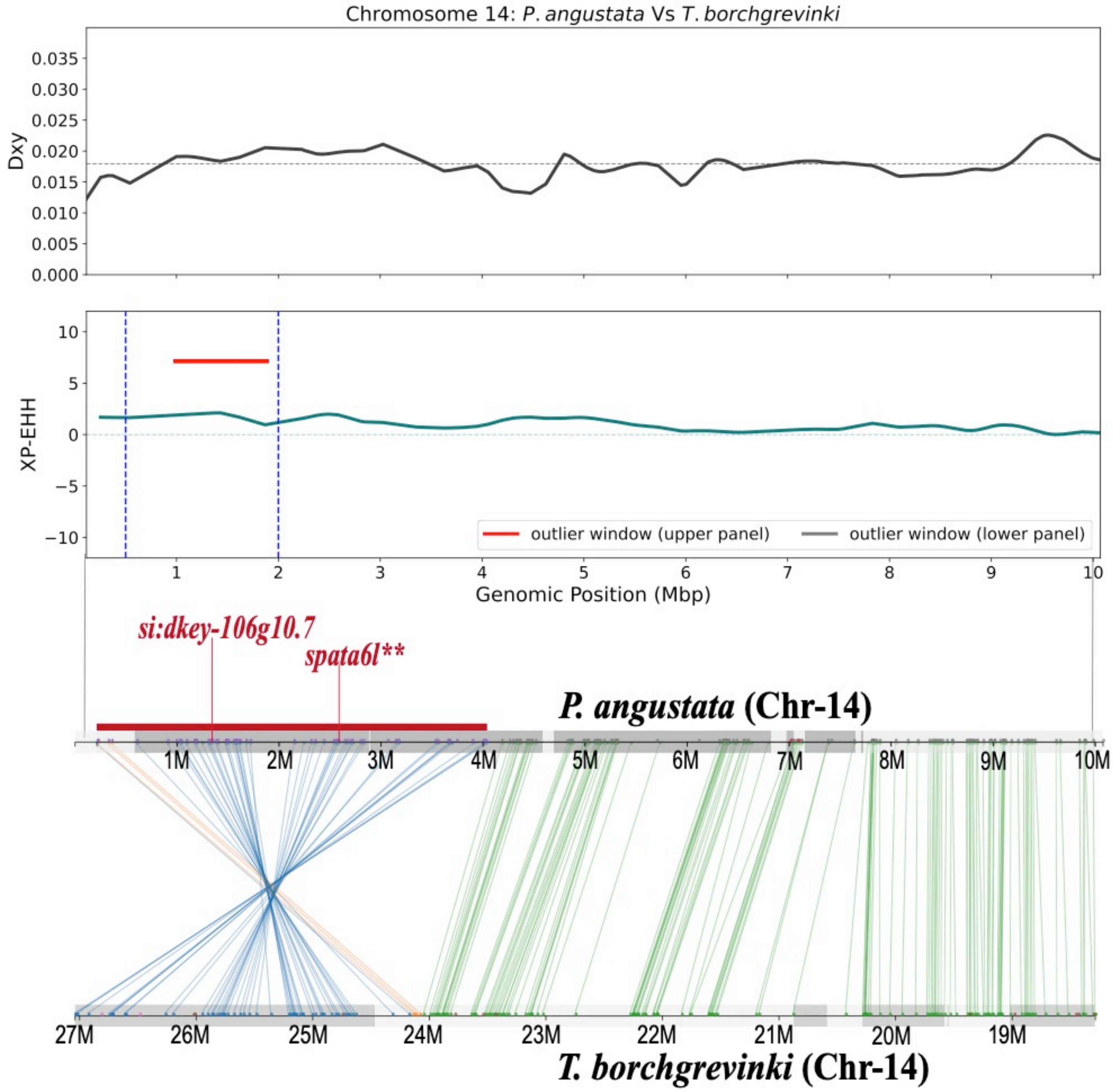


Figure A.10 This figure provides an illustration of patterns of the kernel-smoothed genetic divergence (D_{XY}) and the cross-population extended haplotype homozygosity (XP-EHH) scores between *P. angustata* and *T. borchgrevinki* within the 0-10 megabase pairs (Mbp) region of chromosome 14. The genome-wide mean D_{XY} is represented by a dashed black line. The plot displays inversion (marked by solid red block within genomic region 214,127-4,026,751 (3.81 Mbp size) on chromosome 14) specific to *P. angsustata* with two "*dN/dS*" candidates *si:dkey-106g10.7*, and *spata6l*. Additionally, it reveals the co-localization of *P. angustata*-specific XP-EHH outlier window and inversion. It depicts the presence of *P. angustata*-specific XP-EHH outlier windows (indicated by red solid horizontal lines) within genomic region 0.5-2Mbs on chromosome 14.

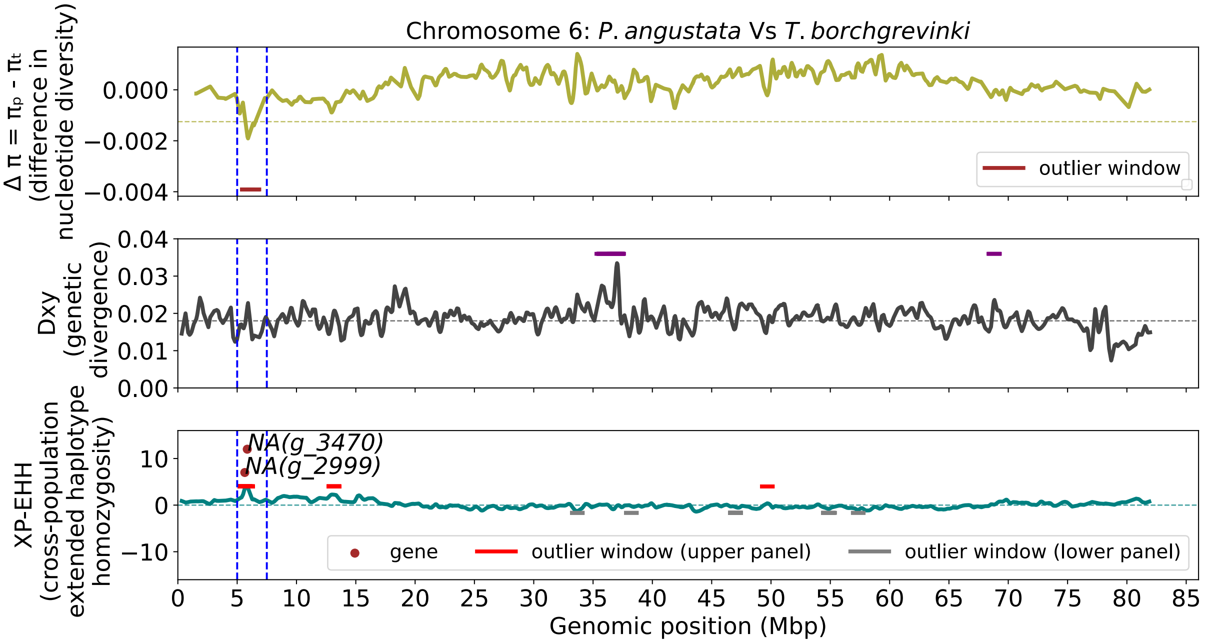


Figure A.11 This figure illustrates patterns of the difference (Δ) in nucleotide diversity (π), the genetic divergence (D_{XY}), and the XP-EHH scores between *P. angustata* and *T. borchgrevinki* on chromosome 6. Specifically, the first subplot displays the distribution of Δ π estimated by subtracting the kernel-smoothed nucleotide diversity of *T. borchgrevinki* (π_t) from *P. angustata* (π_p) at the same positions (y-axis). The olive-colored dashed horizontal line represents the bottom 0.5th percentile threshold of Δ π. The window of the variant site at which Δ π is less than a threshold is shown in a brown solid horizontal line. The second subplot exhibits the distribution of kernel-smoothed D_{XY} between species *P. angustata* and *T. borchgrevinki* (y-axis) with outliers (indicated by a solid purple line). The black-colored dashed line represents the genome-wide mean D_{XY}. The third subplot demonstrates the distribution of kernel-smoothed measure of XP-EHH scores, and the red solid horizontal line indicates the XP-EHH outlier window under *P. angustata*-specific positive selection. These plots reveal the overlap between Δ π and XP-EHH outlier windows, even without D_{XY} outliers, within genomic region 5-7.5 Mbs (i.e., represented by dashed, blue verticle lines) on chromosome 6. In addition, this figure displays the genes *NA(g_2999)* and *NA(g_3470)* (denoted by brown dots) that a) are located within the overlapping region between Δ π and XP-EHH outlier windows and b) either contain or reside nearest to the XP-EHH outliers.

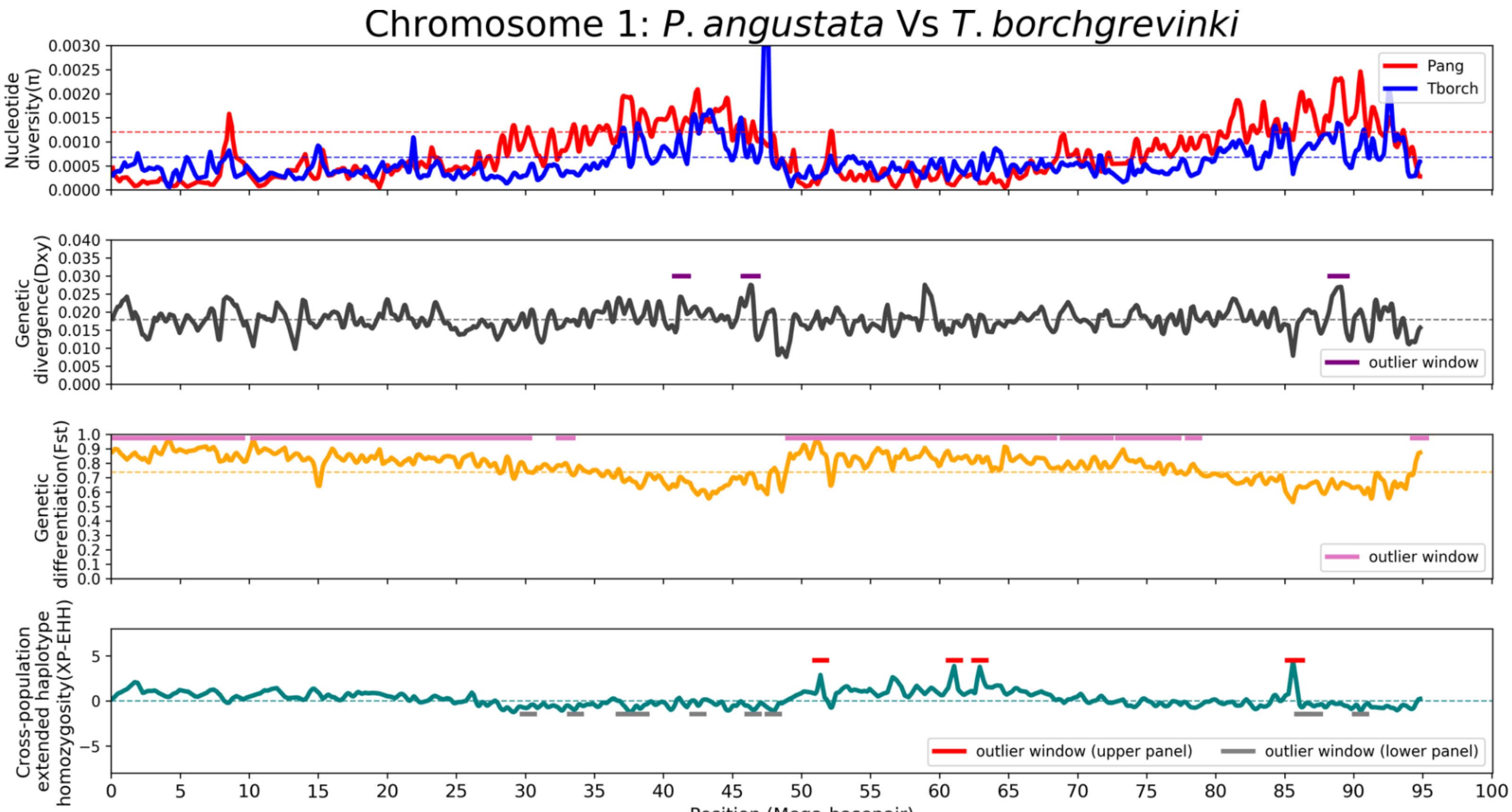


Figure A.12 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 1. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

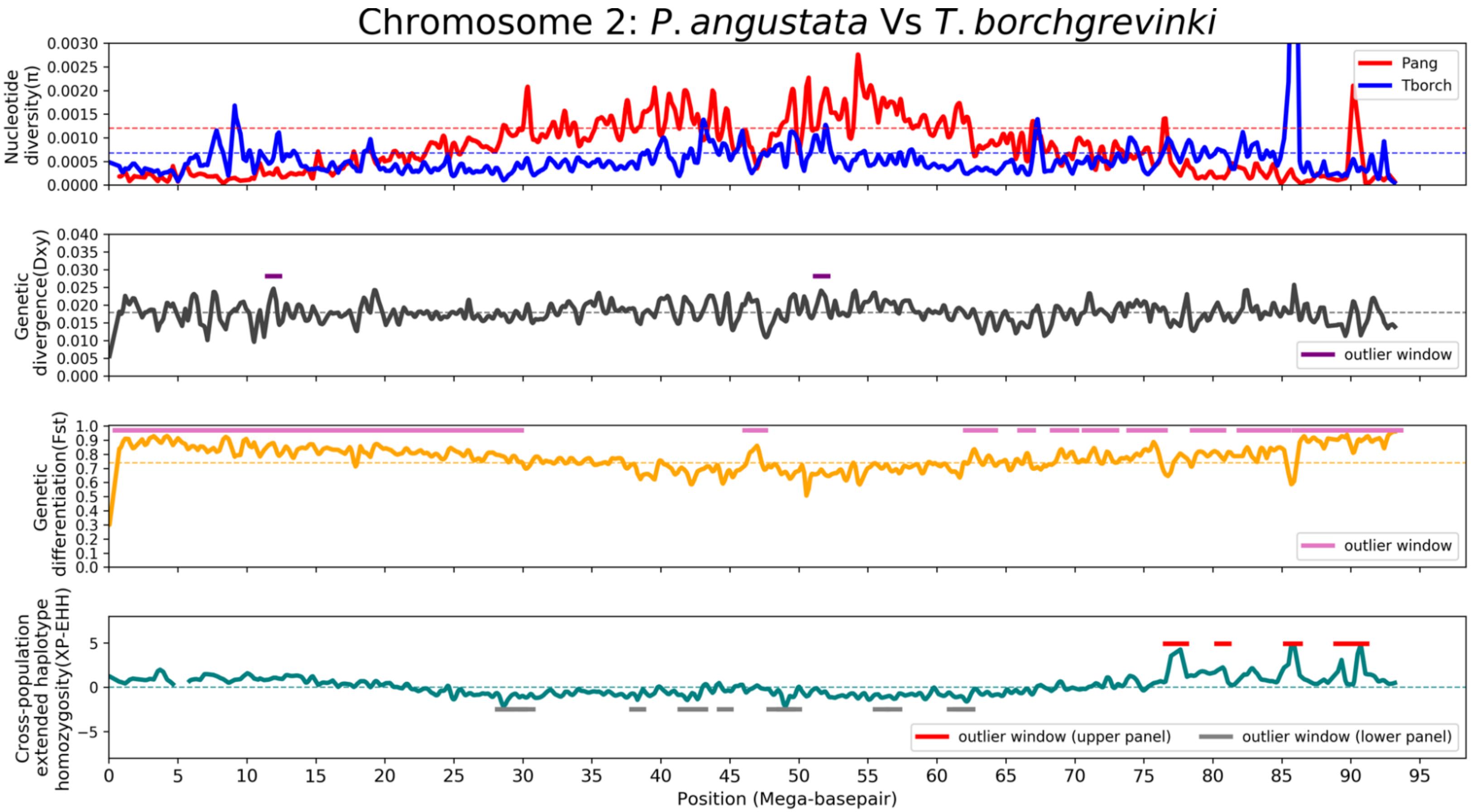


Figure A.13 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 2. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

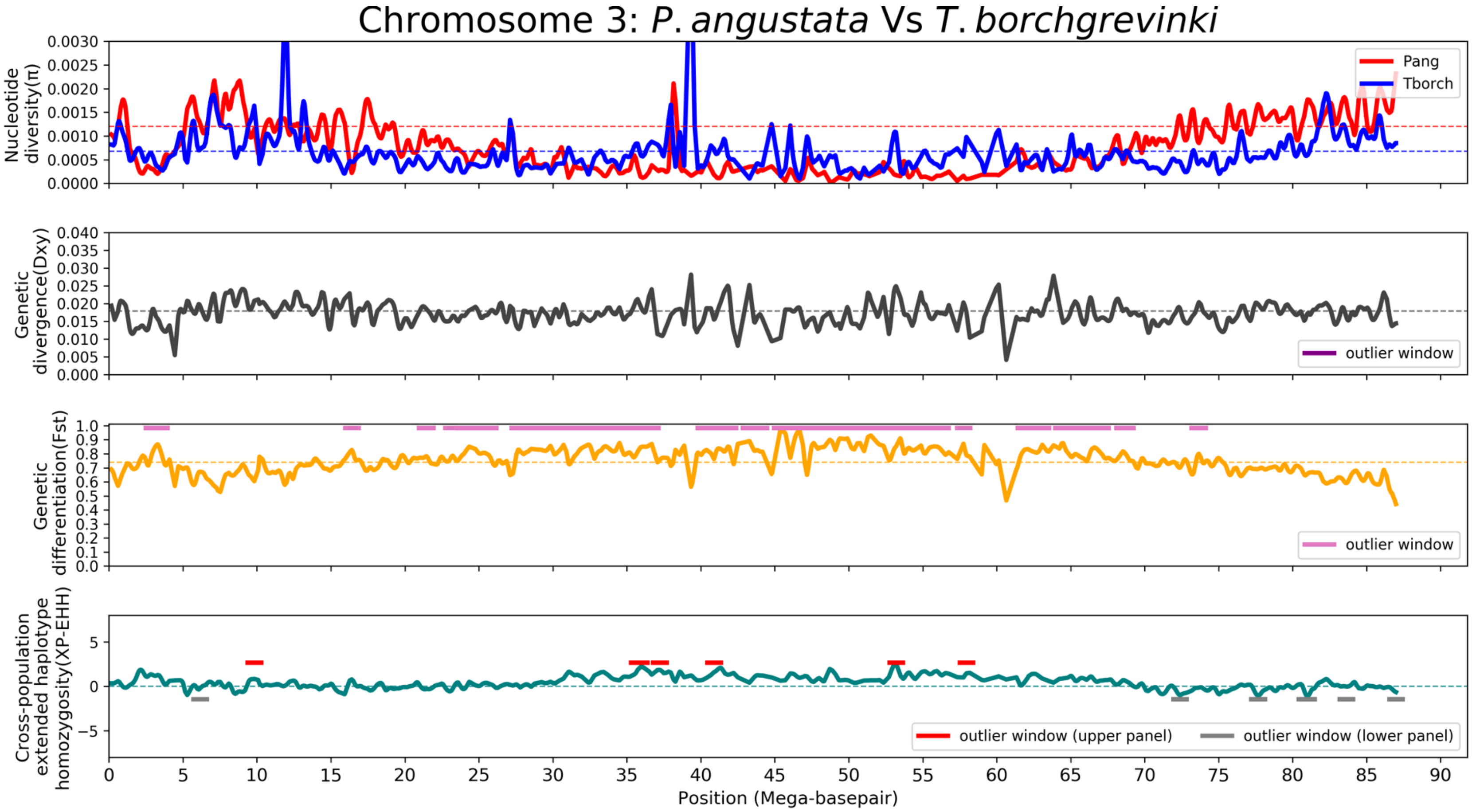


Figure A.14 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 3. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

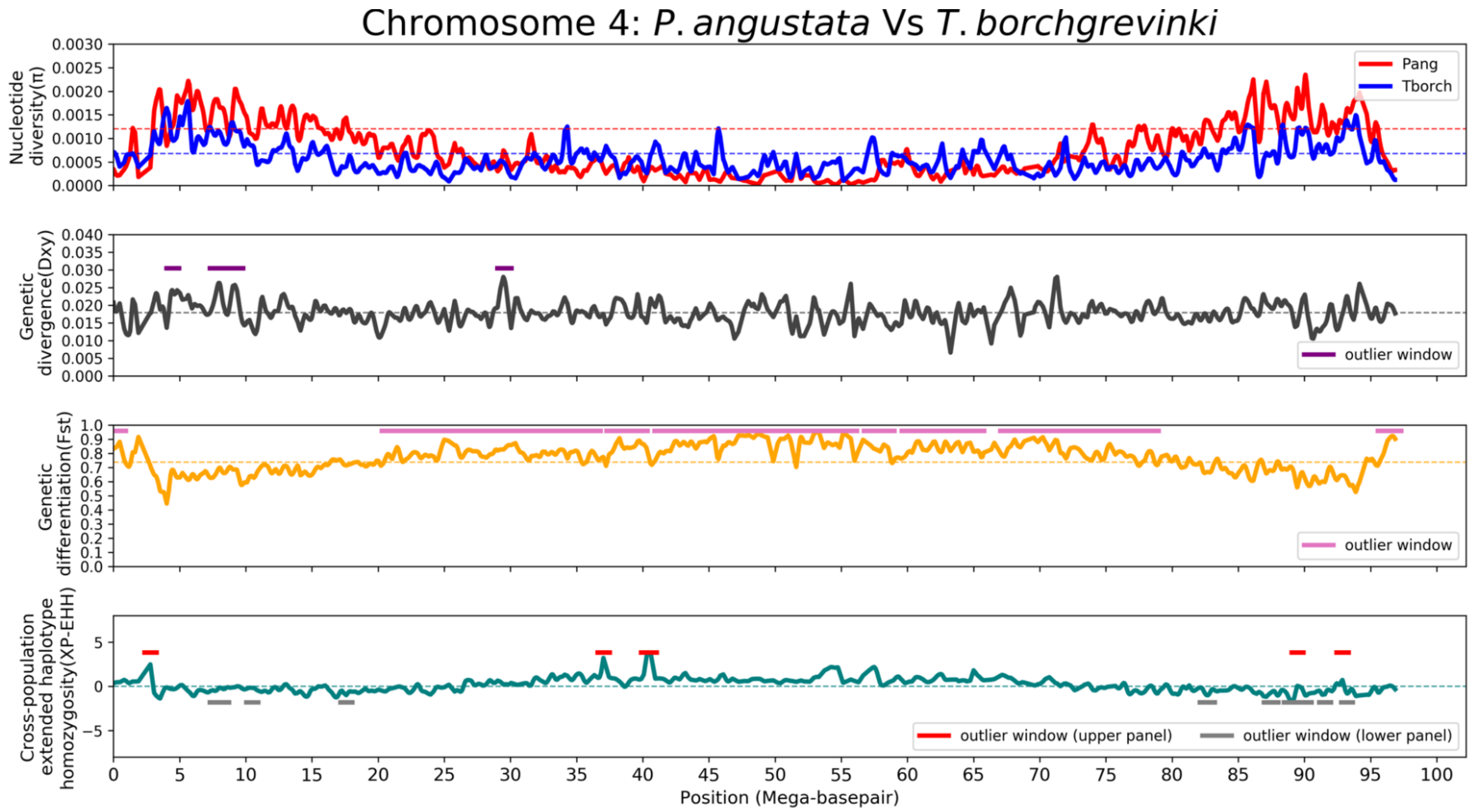


Figure A.15 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 4. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

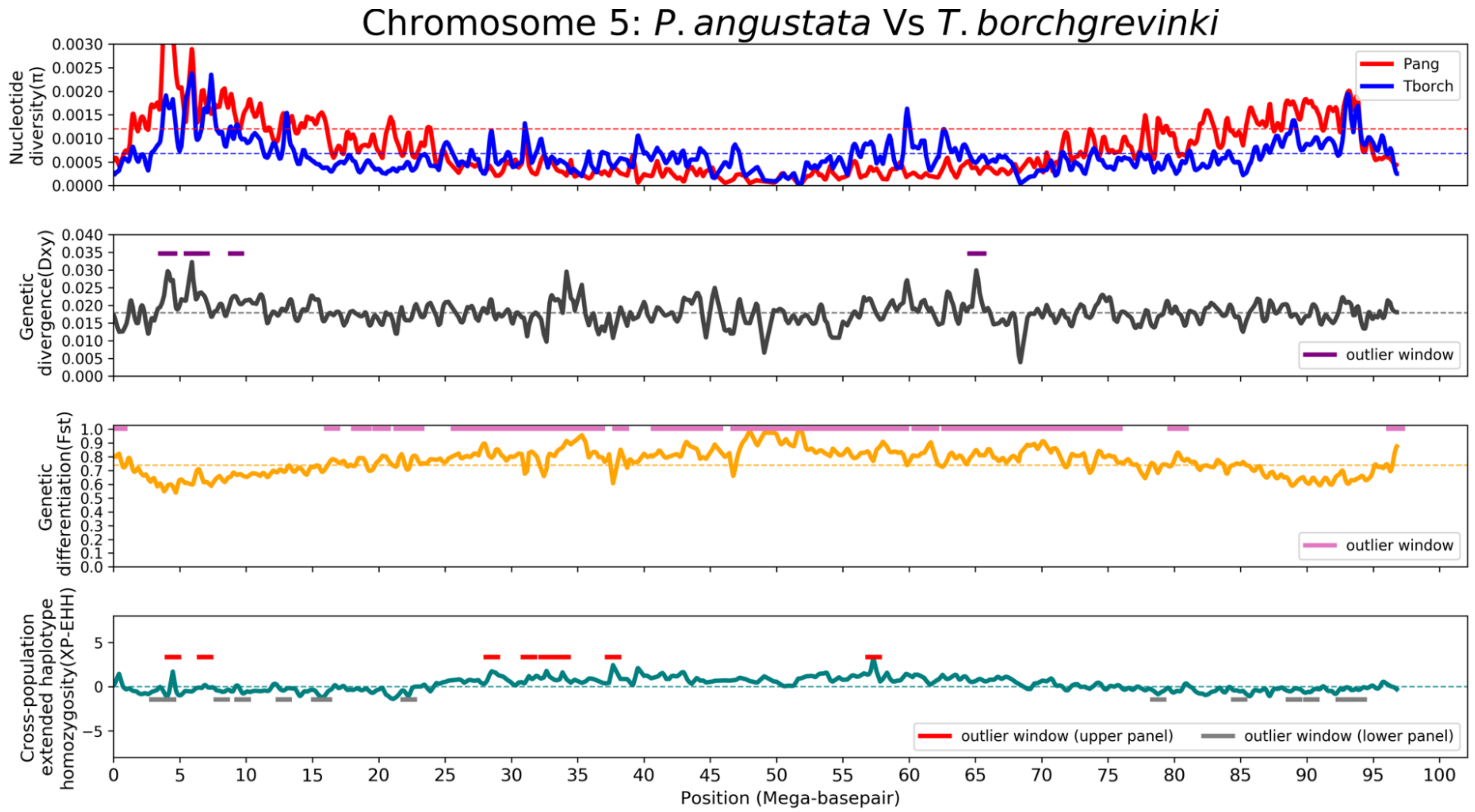


Figure A.16 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 5. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

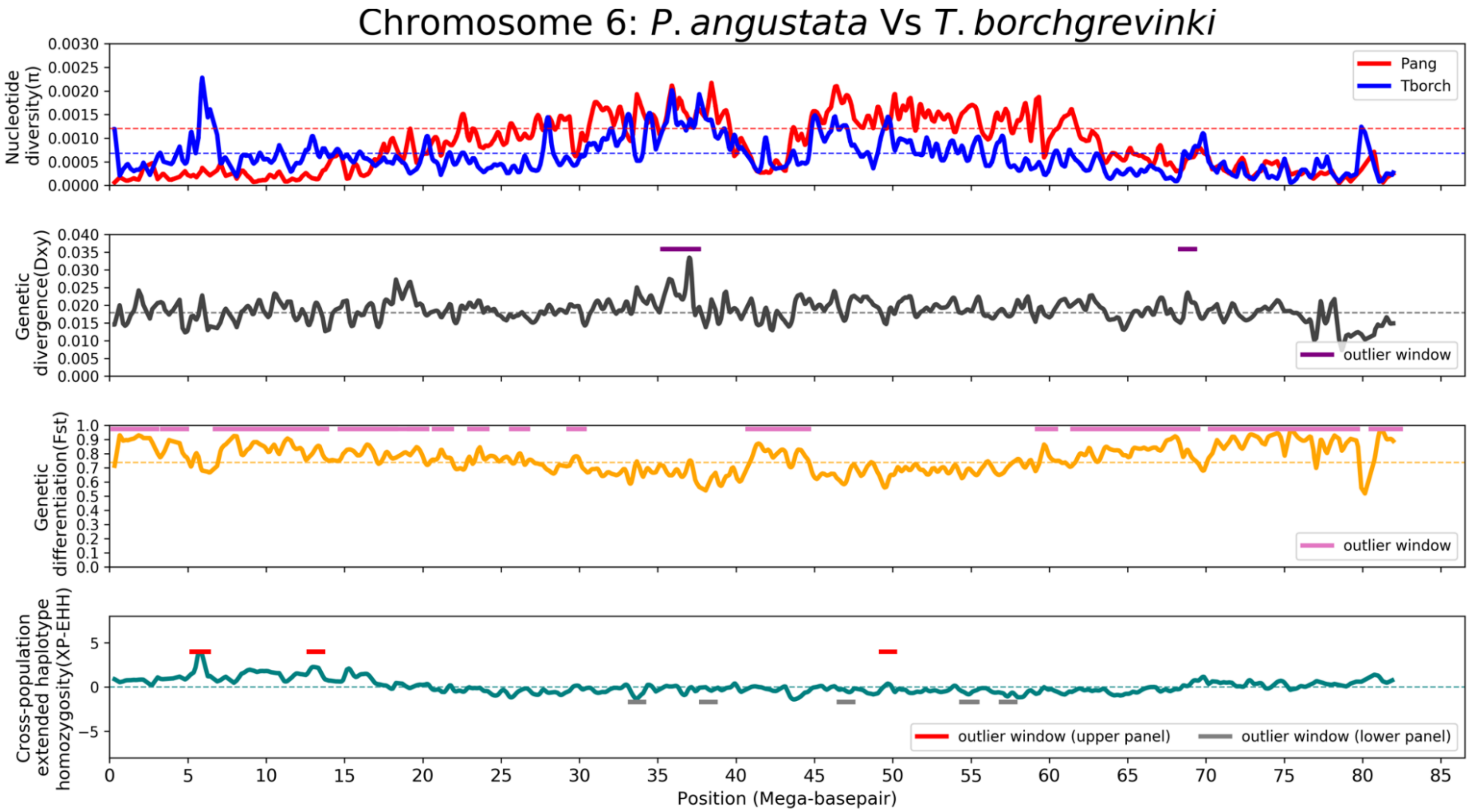


Figure A.17 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 6. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

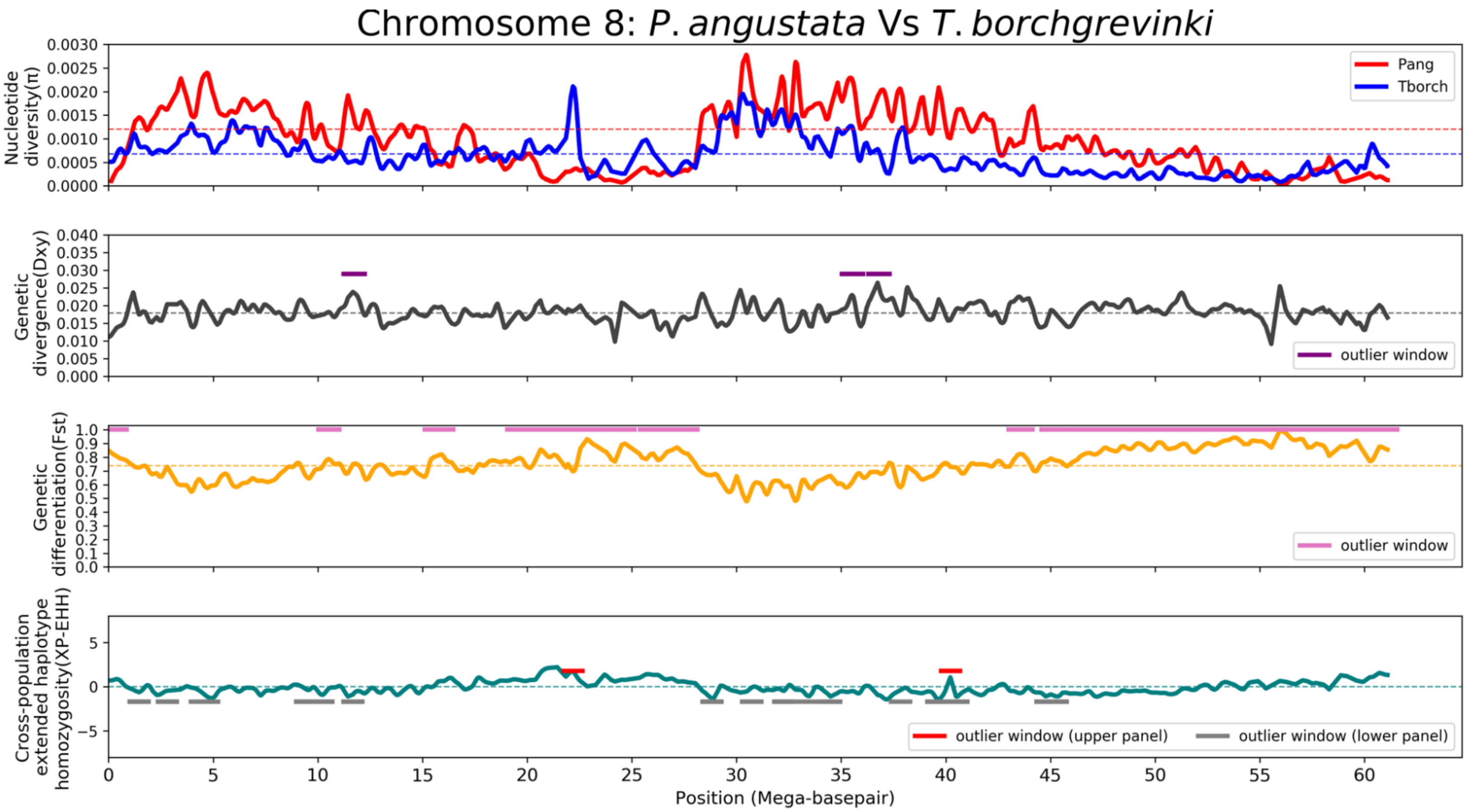


Figure A.18 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 8. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

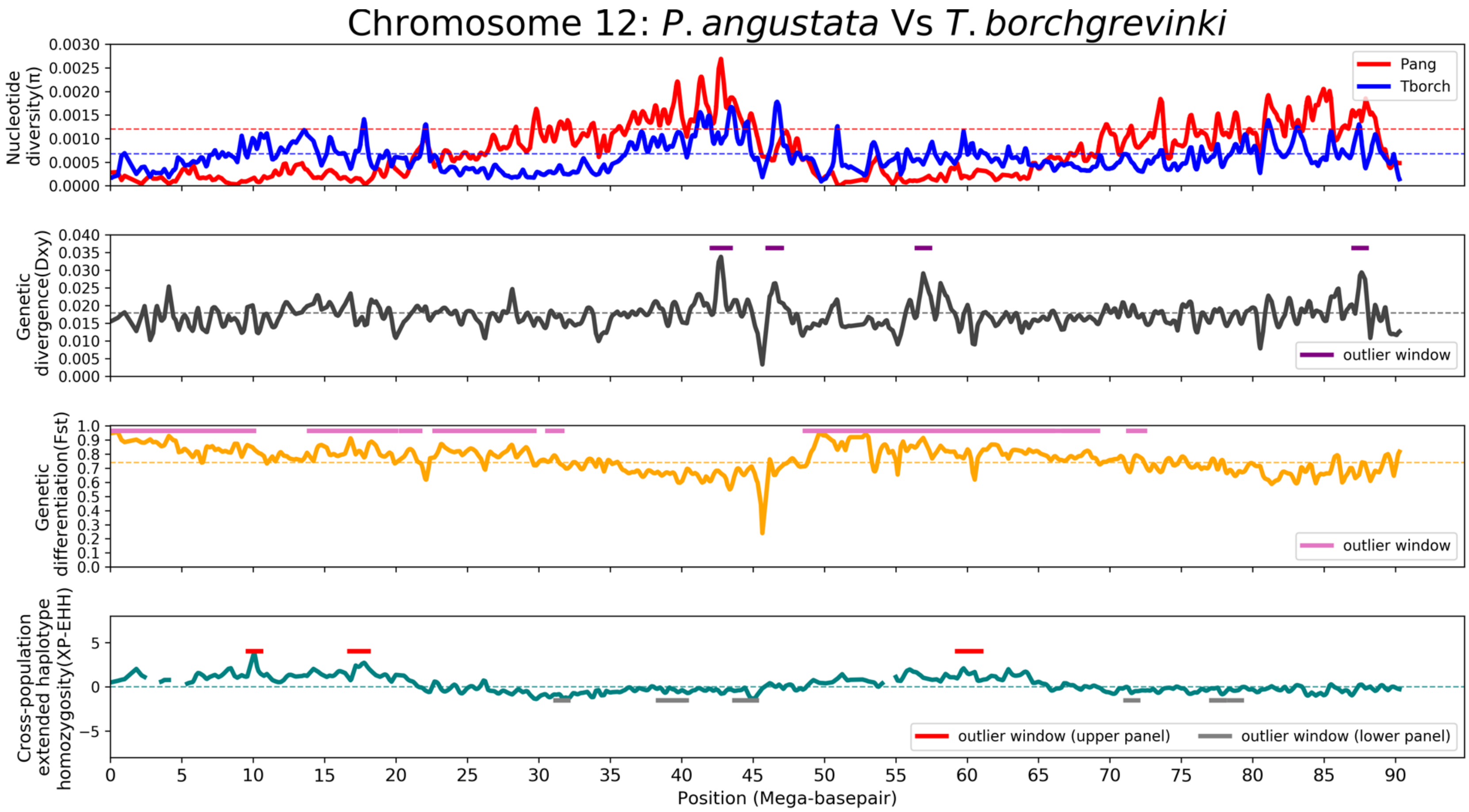


Figure A.19 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 12. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

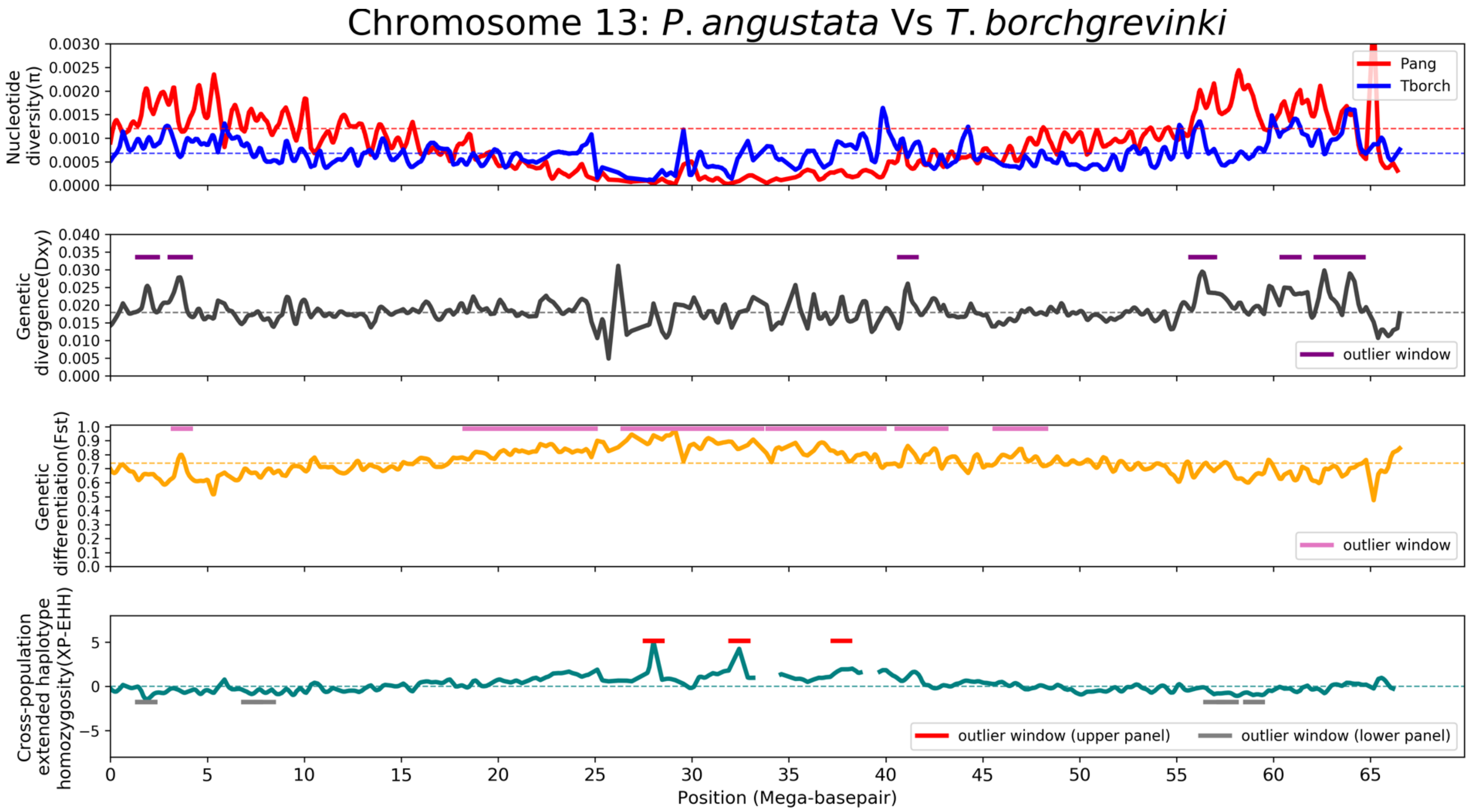


Figure A.20 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 13. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

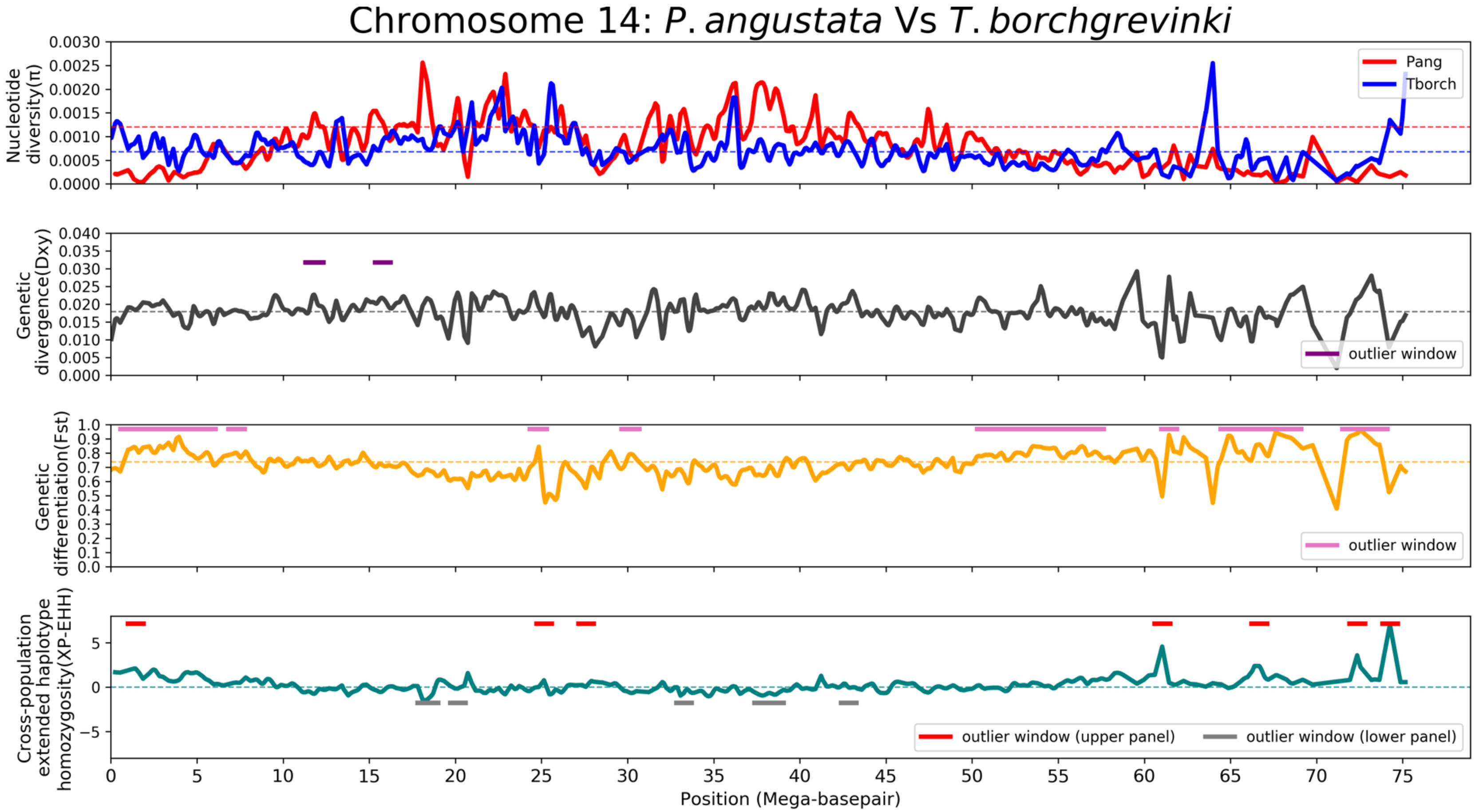


Figure A.21 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 14. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

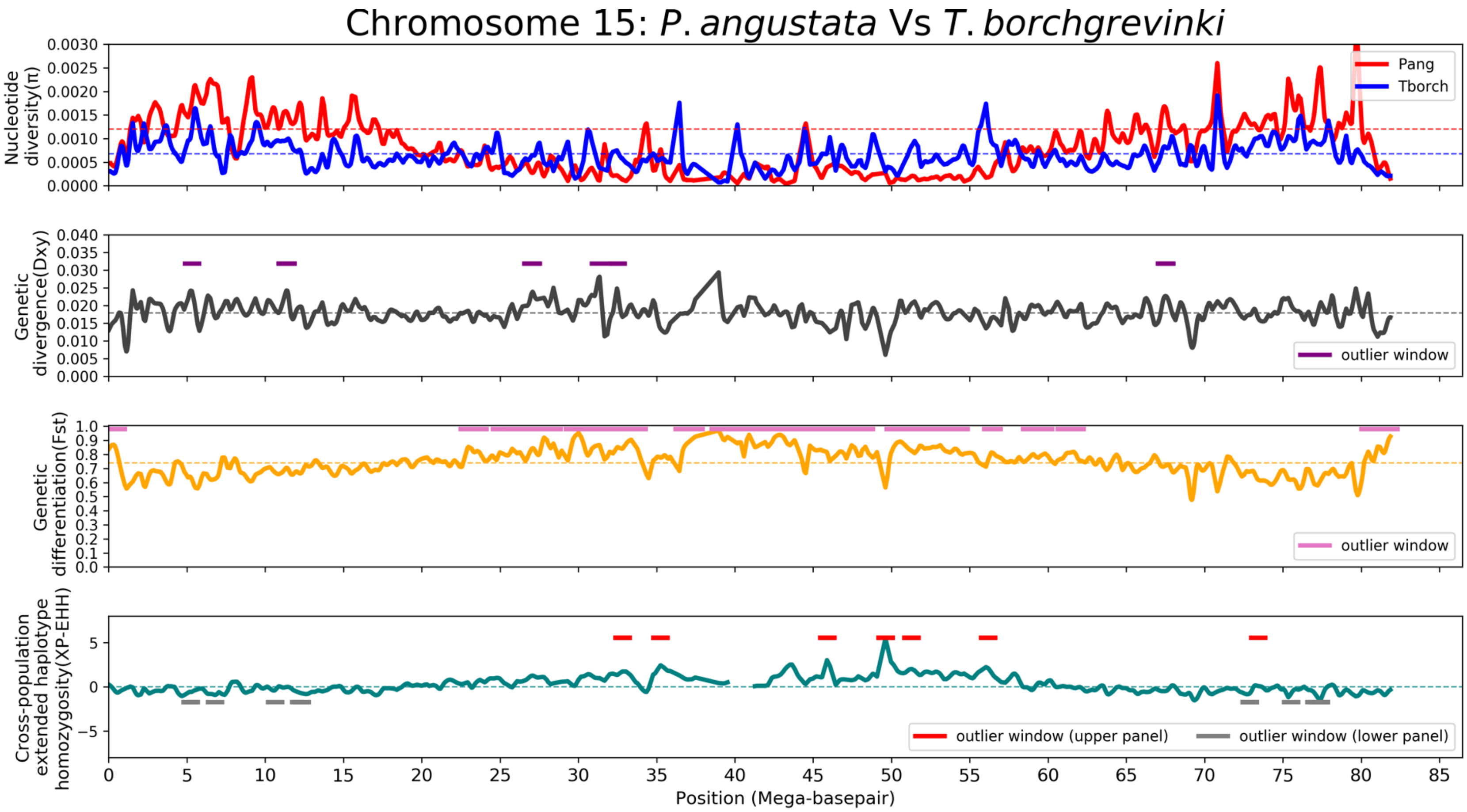


Figure A.22 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 15. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

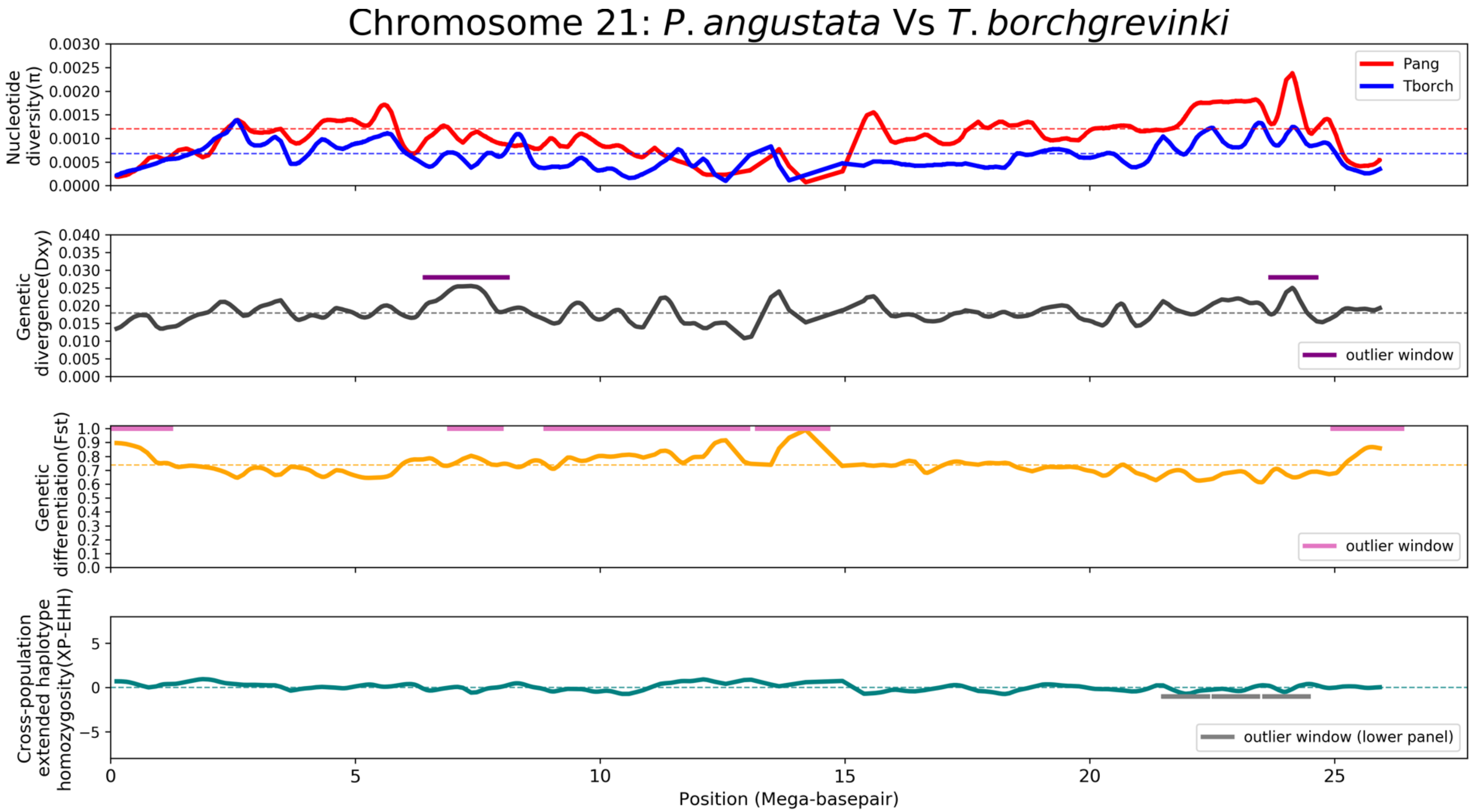


Figure A.23 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 21. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

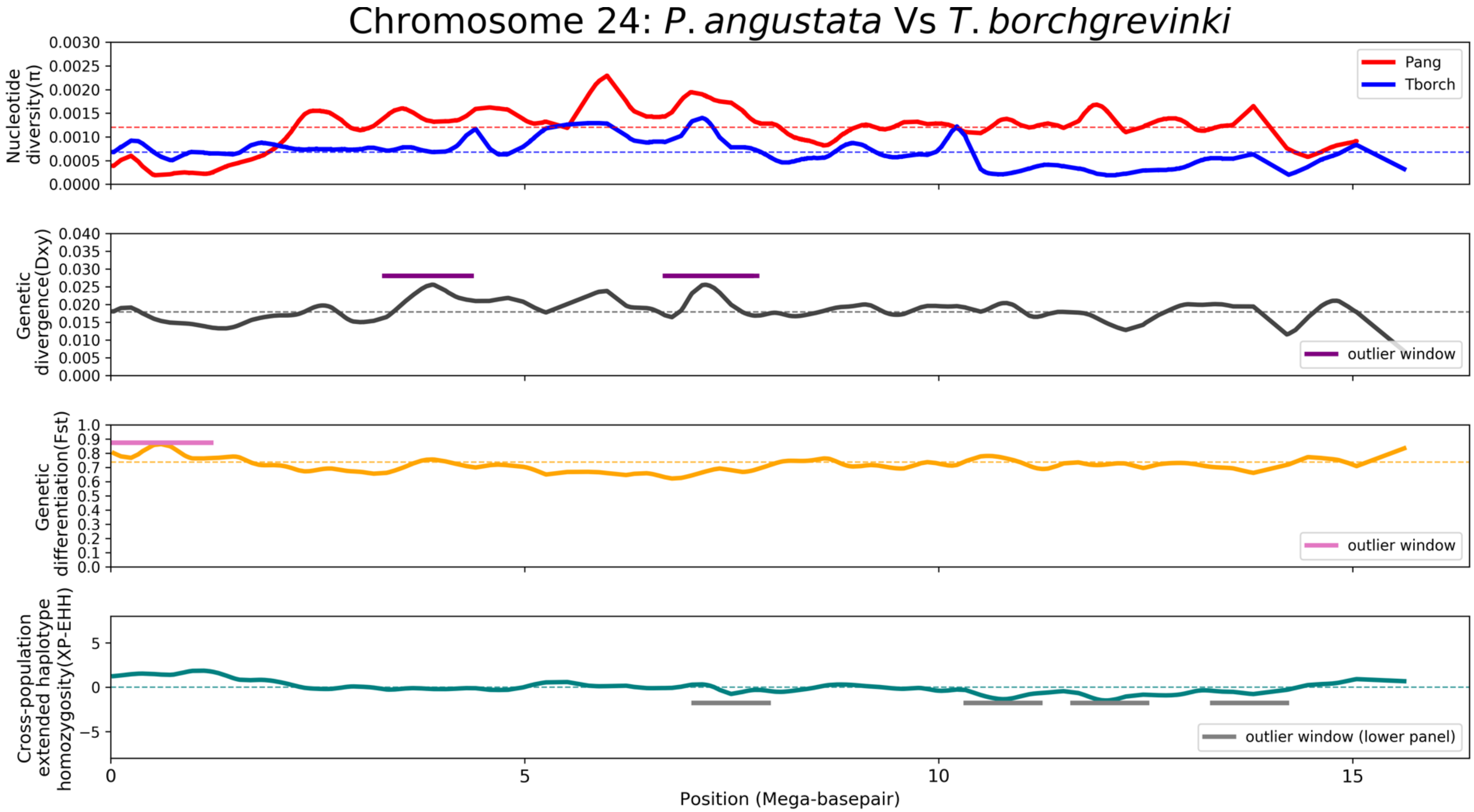


Figure A.24 This figure illustrates patterns of kernel-smoothed nucleotide diversity (π) within the species *P. angustata* and *T. borchgrevinki*, genetic divergence (D_{XY}), and differentiation (F_{ST}) between species, a signal of positive selection specific to *P. angustata* based on XP-EHH scores along the genomic positions of chromosome 24. Notably, D_{XY}, F_{ST}, and XP-EHH outlier windows are denoted by solid horizontal lines in purple, pink, and red, respectively.

Table A.1 Proportion of interspersed repeats of five notothenioids, including *Eleginops maclovinus*, *Trematomus borchgrevinki*, *Paranotothenia angustata*, *Notothenia rossii*, and *Champsocephalus gunnari*

Interspersed Repeats	***E. maclovinus***	***T. borchgrevinki***	***P. angustata***	***N. rossii***	***C. gunnari***
DNA transposons	15.84%	23.63%	29.08%	27.35%	26.13%
Retroelements	8.62%	16.34%	17.10%	16.84%	20.45%
SINE	0.64%	0.76%	0.56%	0.53%	0.59%
LINEs	5.06%	8.79%	8.71%	8.85%	11.38%
LTR elements	2.92%	6.79%	7.82%	7.45%	8.49%
Unclassified	5.72%	10.20%	7.81%	12.15%	8.69%

Table A.2 The length occupied by interspersed repeats (in base pairs) in the genomes of five notothenioids, including *Eleginops maclovinus*, *Trematomus borchgrevinki*, *Paranotothenia angustata*, *Notothenia rossii*, and *Champsocephalus gunnari*

Interspersed Repeats	***E. maclovinus***	***T. borchgrevinki***	***P. angustata***	***N. rossii***	***C. gunnari***
DNA transposons	96,041,897	221,004,185	287,158,168	285,275,742	259,784,750
Retroelements	52,256,784	152,796,101	168,840,490	175,583,116	203,333,903
SINE	389,009	7,070,725	5,514,139	5,556,071	5,865,787
LINEs	30,658,361	82,218,610	86,056,140	92,340,971	113,140,086
LTR elements	17,708,326	63,506,766	77,270,211	77,686,074	84,407,674
Unclassified	34,694,488	94,971,158	77,137,364	126,720,168	86,396,076

Table A.3 The number of interspersed repeats in the genomes of five notothenioids, including *Eleginops maclovinus*, *Trematomus borchgrevinki*, *Paranotothenia angustata*, *Notothenia rossii*, and *Champsocephalus gunnari*

Interspersed Repeats	***E. maclovinus***	***T. borchgrevinki***	***P. angustata***	***N. rossii***	***C. gunnari***
DNA transposons	528,922	1,023,627	1,094,143	1,071,904	1,147,758
Retroelements	259,919	564,839	529,046	528,213	646,995
SINE	33,668	52,331	46,936	46,986	52,807
LINEs	168,478	364,098	331,972	336,488	402,401
LTR elements	57,773	148,410	150,138	144,739	191,787
Unclassified	251,297	476,355	413,238	422,501	454,950

Table A.4 Thirty "*Dxy&linkage*" candidates and their ID, name, zebrafish orthologs' Ensembl ID, and description

Gene ID	Gene name	Ensembl ID	Gene description based on ZFIN
g_25000	si:dkey-85k7.12	ENSDARG00000078731	
g_19543	NA	Not available	Not available
g_9481	NA	Not available	Not available
g_18964	atpv0e2	ENSDARG00000059057	ATPase H+ transporting V0 subunit e2 [Source:ZFIN;Acc:ZDB-GENE-050522-135]
g_4847	NA	Not available	Not available
g_8549	NA	Not available	Not available
g_17157	NA	Not available	Not available
g_1751	NA	Not available	Not available
g_6060	dla	ENSDARG00000010791	deltaA [Source:ZFIN;Acc:ZDB-GENE-980526-29]
g_21751	NA	Not available	Not available
g_30411	hk1	ENSDARG00000039452	hexokinase 1 [Source:ZFIN;Acc:ZDB-GENE-040426-2848]
g_5471	myom2a	ENSDARG00000075433	myomesin 2a [Source:ZFIN;Acc:ZDB-GENE-030131-6201]
g_24011	NA	Not available	Not available
g_3454	lgsn	ENSDARG00000007715	lengsin, lens protein with glutamine synthetase domain [Source:ZFIN;Acc:ZDB-GENE-060312-26]
g_23710	ptp4a1	ENSDARG00000006242	protein tyrosine phosphatase 4A1 [Source:ZFIN;Acc:ZDB-GENE-041121-11]
g_6112	col9a1a	ENSDARG00000073699	collagen, type IX, alpha 1a [Source:ZFIN;Acc:ZDB-GENE-080721-25]
g_8891	si:dkey-23f9.4	ENSDARG00000098623	si:dkey-23f9.4 [Source:ZFIN;Acc:ZDB-GENE-141222-88]
g_25245	NA	Not available	Not available
g_10324	NA	Not available	Not available
g_22556	mrpl4	ENSDARG00000058824	mitochondrial ribosomal protein L4 [Source:ZFIN;Acc:ZDB-GENE-050522-388]
g_2976	si:ch211-195h23.3	ENSDARG00000068431	si:ch211-195h23.3 [Source:ZFIN;Acc:ZDB-GENE-070912-174]
g_4206	abcc10	ENSDARG00000077988	ATP-binding cassette, sub-family C (CFTR/MRP), member 10 [Source:ZFIN;Acc:ZDB-GENE-050517-24]
g_3936	ube2j1	ENSDARG00000033489	ubiquitin-conjugating enzyme E2, J1 [Source:ZFIN;Acc:ZDB-GENE-040426-2853]
g_24828	NA	Not available	Not available

Table A.4 – Continued

g_22624	ankrd6b	ENSDARG00000029370	ankyrin repeat domain 6b [Source:ZFIN;Acc:ZDB-GENE-030916-4]
g_33741	lyrm2	ENSDARG00000033138	LYR motif containing 2 [Source:ZFIN;Acc:ZDB-GENE-040914-27]
g_3275	mdn1	ENSDARG00000008976	midasin AAA ATPase 1 [Source:ZFIN;Acc:ZDB-GENE-04100-1381]
g_186	NA	Not available	Not available
g_15924	NA	Not available	Not available
g_1344	casp8ap2	ENSDARG00000022718	caspase 8 associated protein 2 [Source:ZFIN;Acc:ZDB-GENE-030826-8]

Table A.5 Thirty "*Dxy&linkage*" candidates and their ID, name, zebrafish orthologs' Ensembl ID, chromosome location, as well as start and end position in the genome

Gene ID	Gene name	Ensembl ID	Chromosome	Start	End
g_25000	*si:dkey-85k7.12*	*ENSDARG00000078731*	5	4032901	4037403
g_19543	*NA*	*Not available*	5	4056767	4060372
g_9481	*NA*	*Not available*	5	4270481	4278123
g_18964	*atpv0e2*	*ENSDARG00000059057*	5	6477080	6484388
g_4847	*NA*	*Not available*	5	6493173	6519002
g_8549	*NA*	*Not available*	5	6524867	6525292
g_17157	*NA*	*Not available*	5	6570228	6573092
g_1751	*NA*	*Not available*	5	6618808	6619218
g_6060	*dla*	*ENSDARG00000010791*	5	6645436	6653479
g_21751	*NA*	*Not available*	5	6668597	6684754
g_30411	*hk1*	*ENSDARG00000039452*	5	6691651	6735010
g_5471	*myom2a*	*ENSDARG00000075433*	5	6750532	6802340
g_24011	*NA*	*Not available*	5	6809579	6819619
g_3454	*lgsn*	*ENSDARG00000007715*	5	6822629	6826038
g_23710	*ptp4a1*	*ENSDARG00000006242*	5	6836056	6842942
g_6112	*col9a1a*	*ENSDARG00000073699*	5	6851369	6879677
g_8891	*si:dkey-23f9.4*	*ENSDARG00000098623*	5	6880910	6889658
g_25245	*NA*	*Not available*	5	6884408	6885139
g_10324	*NA*	*Not available*	5	6885357	6886103
g_22556	*mrpl4*	*ENSDARG00000058824*	5	6889868	6898382
g_2976	*si:ch211-195h23.3*	*ENSDARG00000068431*	5	6974855	6982572
g_4206	*abcc10*	*ENSDARG00000077988*	5	7024098	7062567
g_3936	*ube2j1*	*ENSDARG00000033489*	15	32448567	32484852
g_24828	*NA*	*Not available*	15	32658285	32658494
g_22624	*ankrd6b*	*ENSDARG00000029370*	15	32686507	32735956
g_33741	*lyrm2*	*ENSDARG00000033138*	15	32745384	32750270
g_3275	*mdn1*	*ENSDARG00000008976*	15	32753723	32884098
g_186	*NA*	*Not available*	15	32906124	32906411
g_15924	*NA*	*Not available*	15	32906783	32907445
g_1344	*casp8ap2*	*ENSDARG00000022718*	15	*32926319*	*32941893*

Table A.6 Twenty-nine "*deltapi&linkage*" candidates and their ID, name, zebrafish orthologs' Ensembl ID, and description

Gene ID	**Gene name**	**Ensembl ID**	**Gene description based on ZFIN**
g_15915	*Not available*	*Not available*	*Not available*
g_33402	*Not available*	*Not available*	*Not available*
g_27594	*Not available*	*Not available*	*Not available*
g_21219	*Not available*	*Not available*	*Not available*
g_22508	*Not available*	*Not available*	*Not available*
g_22706	*dnajc24*	*ENSDARG00000023927*	*DnaJ (Hsp40) homolog, subfamily C, member 24 [Source:ZFIN;Acc: ZDB-GENE-040426-1153]*
g_7154	*Not available*	*Not available*	*Not available*
g_15335	*Not available*	*Not available*	*Not available*
g_23192	*Not available*	*Not available*	*Not available*
g_11968	*Not available*	*Not available*	*Not available*
g_26188	*Not available*	*Not available*	*Not available*
g_11538	*Not available*	*Not available*	*Not available*
g_16892	*fam151a*	*ENSDARG00000058218*	*family with sequence similarity 151 member A [Source:ZFIN;Acc: ZDB-GENE-070705-105]*
g_25862	*atg10*	*ENSDARG00000104846*	*ATG10 autophagy related 10 homolog (S. cerevisae) [Source:ZFIN;Acc: ZDB-GENE-051030-72]*
g_2300	*Not available*	*Not available*	*Not available*
g_21063	*Not available*	*Not available*	*Not available*
g_9875	*fam110b*	*ENSDARG00000088073*	*family with sequence similarity 110 member B [Source:ZFIN;Acc: ZDB-GENE-050626-70]*
g_9312	*Not available*	*Not available*	*Not available*
g_2999	*Not available*	*Not available*	*Not available*
g_33806	*Not available*	*Not available*	*Not available*
g_22141	*Not available*	*Not available*	*Not available*
g_3470	*Not available*	*Not available*	*Not available*
g_7133	*Not available*	*Not available*	*Not available*
g_12378	*ptf1a*	*ENSDARG00000014479*	*pancreas associated transcription factor 1a [Source:ZFIN;Acc: ZDB-GENE-030616-579]*
g_14879	*Not available*	*Not available*	*Not available*

Table A.6 – Continued

g_2576	*Not available*	*Not available*	*Not available*
g_13655	*Not available*	*Not available*	*Not available*
g_21992	*Not available*	*Not available*	*Not available*
g_4925	*Not available*	*Not available*	*Not available*

Table A.7 Twenty-nine "*deltapi&linkage*" candidates and their ID, name, zebrafish orthologs' Ensembl ID, chromosome location, as well as start and end position in the genome

Gene ID	Gene name	Ensembl ID	Chromosome	Start	End
g_15915	*Not available*	*Not available*	2	85359095	85360647
g_33402	*Not available*	*Not available*	2	85403655	85405790
g_27594	*Not available*	*Not available*	2	85610650	85610916
g_21219	*Not available*	*Not available*	2	85710828	85711889
g_22508	*Not available*	*Not available*	2	85727047	85728562
g_22706	*dnajc24*	*ENSDARG00000023927*	2	85821788	85825948
g_7154	*Not available*	*Not available*	2	85890824	85906026
g_15335	*Not available*	*Not available*	2	86130563	86130880
g_23192	*Not available*	*Not available*	2	86288874	86295951
g_11968	*Not available*	*Not available*	6	5431644	5433245
g_26188	*Not available*	*Not available*	6	5440521	5441483
g_11538	*Not available*	*Not available*	6	5446195	5447070
g_16892	*fam151a*	*ENSDARG00000058218*	6	5469242	5488702
g_25862	*atg10*	*ENSDARG00000104846*	6	5498704	5506522
g_2300	*Not available*	*Not available*	6	5540214	5540785
g_21063	*Not available*	*Not available*	6	5546173	5546772
g_9875	*fam110b*	*ENSDARG00000088073*	6	5596379	5611107
g_9312	*Not available*	*Not available*	6	5611019	5611345
g_2999	*Not available*	*Not available*	6	5642530	5643804
g_33806	*Not available*	*Not available*	6	5795139	5795693
g_22141	*Not available*	*Not available*	6	5812629	5839928
g_3470	*Not available*	*Not available*	6	5851045	5889576
g_7133	*Not available*	*Not available*	6	6020963	6021298
g_12378	*ptf1a*	*ENSDARG00000014479*	6	6049300	6050396
g_14879	*Not available*	*Not available*	6	6095601	6096744
g_2576	*Not available*	*Not available*	6	6158687	6159772
g_13655	*Not available*	*Not available*	6	6165013	6180818
g_21992	*Not available*	*Not available*	6	6199304	6202468
g_4925	*Not available*	*Not available*	6	6313479	6314186

Table A.8 Hundred and thirty-eight “*dN/dS*” candidates based on branch model and their ID, name, zebrafish orthologs’ Ensembl ID, and description

Gene ID	Gene name	Ensembl ID	Gene Description
g_2285	*NA*	*Not available*	*Not available*
g_17258	*dnase1l1l*	*ENSDARG00000023861*	*deoxyribonuclease I-like 1-like [Source:ZFIN;Acc:ZDB-GENE-040718-100]*
g_16307	*apobec2b*	*ENSDARG00000113992*	*apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 2b [Source:ZFIN;Acc:ZDB-GENE-090618-1]*
g_11406	*rnd1a*	*ENSDARG00000030547*	*Rho family GTPase 1a [Source:ZFIN;Acc:ZDB-GENE-040630-6]*
g_3197	*CABZ01040556.1*	*ENSDARG00000058869*	*NIMA related kinase 3 [Source:NCBI gene;Acc:100536894]*
g_8223	*NA*	*Not available*	*Not available*
g_18930	*tlcd5a*	*ENSDARG00000024920*	*TLC domain containing 5a [Source:ZFIN;Acc:ZDB-GENE-000607-58]*
g_19278	*yif1b*	*ENSDARG00000040505*	*Yip1 interacting factor homolog B (S. cerevisiae) [Source:NCBI gene;Acc:492462]*
g_29761	*blvra*	*ENSDARG00000059857*	*biliverdin reductase A [Source:ZFIN;Acc:ZDB-GENE-060929-312]*
g_14273	*get1*	*ENSDARG00000074271*	*guided entry of tail-anchored proteins factor 1 [Source:ZFIN;Acc:ZDB-GENE-030131-7696]*
g_2	*NA*	*Not available*	*Not available*
g_16937	*NA*	*Not available*	*Not available*
g_26570	*zgc:112163*	*ENSDARG00000017657*	*Not available*
g_31175	*NA*	*Not available*	*Not available*
g_19587	*NA*	*Not available*	*Not available*
g_27671	*NA*	*Not available*	*Not available*
g_5377	*lsp1a*	*ENSDARG00000027310*	*lymphocyte specific protein 1 a [Source:ZFIN;Acc:ZDB-GENE-131127-171]*
g_16386	*NA*	*Not available*	*Not available*

Table A.8 – Continued

g_15913	*C25H12orf29*	*ENSDARG00000045785*	*RNA 5'-phosphate and 3'-OH ligase 1 [Source:ZFIN;Acc:ZDB-GENE-041212-30]*
g_30710	*rassf9*	*ENSDARG00000074721*	*Ras association domain family member 9 [Source:ZFIN;Acc:ZDB-GENE-091204-470]*
g_23492	*NA*	*Not available*	*Not available*
g_9319	*NA*	*Not available*	*Not available*
g_18423	*kcnc1b*	*ENSDARG00000032959*	*potassium voltage-gated channel, Shaw-related subfamily, member 1b [Source:ZFIN;Acc:ZDB-GENE-080414-3]*
g_4725	*ciartb*	*ENSDARG00000088171*	*circadian associated repressor of transcription b [Source:ZFIN;Acc:ZDB-GENE-131127-285]*
g_21318	*ZNF276*	*ENSDARG00000110991*	*zgc:158366 [Source:ZFIN;Acc:ZDB-GENE-070209-176]*
g_5105	*bmper*	*ENSDARG00000101980*	*BMP binding endothelial regulator [Source:ZFIN;Acc:ZDB-GENE-030219-146]*
g_16924	*pou3f1*	*ENSDARG00000009823*	*POU class 3 homeobox 1 [Source:ZFIN;Acc:ZDB-GENE-980526-372]*
g_17745	*NA*	*Not available*	*Not available*
g_467	*rusc1*	*ENSDARG00000078125*	*RUN and SH3 domain containing 1 [Source:ZFIN;Acc:ZDB-GENE-100922-274]*
g_4231	*zgc:101716*	*ENSDARG00000010738*	*zgc:101716 [Source:ZFIN;Acc:ZDB-GENE-041114-135]*
g_19202	*tcaim*	*ENSDARG00000079881*	*T cell activation inhibitor, mitochondrial [Source:ZFIN;Acc:ZDB-GENE-160113-67]*
g_10648	*lyplal1*	*ENSDARG00000088764*	*lysophospholipase like 1 [Source:ZFIN;Acc:ZDB-GENE-050306-32]*
g_27166	*sfxn5b*	*ENSDARG00000026137*	*sideroflexin 5b [Source:ZFIN;Acc:ZDB-GENE-050706-107]*
g_7392	*TTC9*	*ENSDARG00000074363*	*si:ch211-259k16.3 [Source:ZFIN;Acc:ZDB-GENE-090312-172]*

Table A.8 – Continued

g_8911	arg2	ENSDARG00000039269	arginase 2 [Source:ZFIN;Acc:ZDB-GENE-030131-1334]
g_7284	NA	Not available	Not available
g_27153	wdr32	ENSDARG00000029600	WD repeat domain 32 [Source:ZFIN;Acc:ZDB-GENE-040426-2314]
g_23676	strn3	ENSDARG00000001729	striatin, calmodulin binding protein 3 [Source:ZFIN;Acc:ZDB-GENE-030616-405]
g_10797	CELF6	ENSDARG00000101933	si:dkey-205h23.2 [Source:ZFIN;Acc:ZDB-GENE-120215-101]
g_3976	cgref1	ENSDARG00000075444	cell growth regulator with EF-hand domain 1 [Source:ZFIN;Acc:ZDB-GENE-131121-137]
g_18744	sdhaf2	ENSDARG00000062971	succinate dehydrogenase complex assembly factor 2 [Source:ZFIN;Acc:ZDB-GENE-030131-7564]
g_20127	zgc:113276	ENSDARG00000056650	zgc:113276 [Source:ZFIN;Acc:ZDB-GENE-050522-7]
g_8116	il16	ENSDARG00000102908	interleukin 16 [Source:ZFIN;Acc:ZDB-GENE-130103-3]
g_8024	ndufv3	ENSDARG00000090389	NADH:ubiquinone oxidoreductase subunit V3 [Source:ZFIN;Acc:ZDB-GENE-030131-6500]
g_24962	rdh1	ENSDARG00000017882	retinol dehydrogenase 1 [Source:ZFIN;Acc:ZDB-GENE-030912-15]
g_8089	si:dkey-100n23.3	ENSDARG00000062148	si:dkey-100n23.3 [Source:ZFIN;Acc:ZDB-GENE-070912-345]
g_9496	mrpl30	ENSDARG00000069850	mitochondrial ribosomal protein L30 [Source:ZFIN;Acc:ZDB-GENE-050522-240]
g_22198	vasna	ENSDARG00000099266	vasorin a [Source:ZFIN;Acc:ZDB-GENE-050522-43]
g_6060	dla	ENSDARG00000010791	deltaA [Source:ZFIN;Acc:ZDB-GENE-980526-29]
g_4206	abcc10	ENSDARG00000077988	ATP-binding cassette, sub-family C (CFTR/MRP), member 10 [Source:ZFIN;Acc:ZDB-GENE-050517-24]
g_31294	NA	Not available	Not available

Table A.8 – Continued

g_17536	*snrpb2*	*ENSDARG00000039424*	*small nuclear ribonucleoprotein polypeptide B2 [Source:ZFIN;Acc:ZDB-GENE-060616-2]*
g_17232	*socs1b*	*ENSDARG00000089873*	*suppressor of cytokine signaling 1b [Source:ZFIN;Acc:ZDB-GENE-090313-141]*
g_19518	*sod3b*	*ENSDARG00000079183*	*superoxide dismutase 3, extracellular b [Source:ZFIN;Acc:ZDB-GENE-030131-8743]*
g_8712	*NA*	*Not available*	*Not available*
g_27012	*rhbdd2*	*ENSDARG00000092463*	*rhomboid domain containing 2 [Source:ZFIN;Acc:ZDB-GENE-091204-359]*
g_13030	*slx4*	*ENSDARG00000061414*	*SLX4 structure-specific endonuclease subunit homolog (S. cerevisiae) [Source:ZFIN;Acc:ZDB-GENE-050208-359]*
g_2363	*cenpk*	*ENSDARG00000039616*	*centromere protein K [Source:ZFIN;Acc:ZDB-GENE-090313-204]*
g_14353	*fgf8b*	*ENSDARG00000039615*	*fibroblast growth factor 8b [Source:ZFIN;Acc:ZDB-GENE-010122-1]*
g_12806	*uck2a*	*ENSDARG00000006074*	*uridine-cytidine kinase 2a [Source: ZFIN;Acc:ZDB-GENE-030131-7158]*
g_23054	*kifap3a*	*ENSDARG00000008639*	*kinesin-associated protein 3a [Source:ZFIN;Acc:ZDB-GENE-040912-74]*
g_15227	*prrx1b*	*ENSDARG00000042027*	*paired related homeobox 1b [Source:ZFIN;Acc:ZDB-GENE-030131-9033]*
g_15614	*NA*	*Not available*	*Not available*
g_15380	*uox*	*ENSDARG00000007024*	*urate oxidase [Source:ZFIN;Acc:ZDB-GENE-030826-24]*
g_27365	*hsd11b1la*	*ENSDARG00000071377*	*hydroxysteroid (11-beta) dehydrogenase 1-like a [Source:ZFIN;Acc:ZDB-GENE-040426-1002]*
g_14296	*creb3l3a*	*ENSDARG00000056226*	*cAMP responsive element binding protein 3-like 3a [Source:ZFIN;Acc:ZDB-GENE-030131-4298]*
g_23906	*mier2*	*ENSDARG00000071413*	*mesoderm induction early response 1, family member 2 [Source:ZFIN;Acc:ZDB-GENE-050208-795]*

Table A.8 – Continued

g_12699	*haus5*	*ENSDARG00000019156*	*HAUS augmin-like complex, subunit 5 [Source:ZFIN;Acc:ZDB-GENE-041114-150]*
g_589	*si:ch73-71c20.5*	*ENSDARG00000097696*	*si:ch73-71c20.5 [Source:ZFIN;Acc:ZDB-GENE-060810-58]*
g_18291	*aknad1*	*ENSDARG00000094414*	*AKNA domain containing 1 [Source:ZFIN;Acc:ZDB-GENE-070912-649]*
g_7793	*elovl1a*	*ENSDARG00000099960*	*ELOVL fatty acid elongase 1a [Source:ZFIN;Acc:ZDB-GENE-041010-66]*
g_23900	*NA*	*Not available*	*Not available*
g_5085	*si:dkeyp-7a3.1*	*ENSDARG00000090429*	*si:dkeyp-7a3.1 [Source:ZFIN;Acc:ZDB-GENE-091204-119]*
g_20870	*cx47.1*	*ENSDARG00000073896*	*connexin 47.1 [Source:ZFIN;Acc:ZDB-GENE-040912-134]*
g_30991	*si:dkey-32m20.1*	*ENSDARG00000075715*	*si:dkey-32m20.1 [Source:ZFIN;Acc:ZDB-GENE-070705-455]*
g_28295	*or115-2*	*ENSDARG00000053817*	*odorant receptor, family F, subfamily 115, member 2 [Source:ZFIN;Acc:ZDB-GENE-070806-6]*
g_3141	*mybbp1a*	*ENSDARG00000078214*	*MYB binding protein (P160) 1a [Source:ZFIN;Acc:ZDB-GENE-030131-9864]*
g_19432	*rab34b*	*ENSDARG00000010977*	*RAB34, member RAS oncogene family b [Source:ZFIN;Acc:ZDB-GENE-091118-61]*
g_31071	*sgcd*	*ENSDARG00000098573*	*sarcoglycan, delta (dystrophin-associated glycoprotein) [Source:ZFIN;Acc:ZDB-GENE-030131-3684]*
g_13619	*lyn*	*ENSDARG00000107511*	*LYN proto-oncogene, Src family tyrosine kinase [Source:ZFIN;Acc:ZDB-GENE-040912-7]*
g_26086	*NA*	*Not available*	*Not available*
g_7876	*gdf2*	*ENSDARG00000059173*	*growth differentiation factor 2 [Source:ZFIN;Acc:ZDB-GENE -100107-1]*
g_10758	*ubtd1b*	*ENSDARG00000079623*	*ubiquitin domain containing 1b [Source:ZFIN;Acc:ZDB-GENE-050913-62]*

Table A.8 – Continued

g_29167	*tmem130*	*ENSDARG00000103789*	*transmembrane protein 130 [Source:ZFIN;Acc:ZDB-GENE-080204-23]*
g_14608	*vwa2*	*ENSDARG00000075441*	*von Willebrand factor A domain containing 2 [Source:ZFIN;Acc:ZDB-GENE-100302-1]*
g_25655	*NA*	*Not available*	*Not available*
g_1660	*fbxl15*	*ENSDARG00000005284*	*F-box and leucine-rich repeat protein 15 [Source:ZFIN;Acc:ZDB-GENE-040426-2440]*
g_21651	*entpd2a.1*	*ENSDARG00000035506*	*ectonucleoside triphosphate diphosphohydrolase 2a, tandem duplicate 1 [Source:ZFIN;Acc:ZDB-GENE-040724-187]*
g_14269	*ptgdsa*	*ENSDARG00000069439*	*prostaglandin D2 synthase a [Source:ZFIN;Acc:ZDB-GENE-081022-118]*
g_26258	*surf2*	*ENSDARG00000112476*	*surfeit 2 [Source:ZFIN;Acc:ZDB-GENE-040801-86]*
g_22150	*NA*	*Not available*	*Not available*
g_14166	*ccdc62*	*ENSDARG00000111759*	*coiled-coil domain containing 62 [Source:ZFIN;Acc:ZDB-GENE-040718-71]*
g_14680	*kmt5aa*	*ENSDARG00000105231*	*lysine methyltransferase 5Aa [Source:NCBI gene;Acc:751629]*
g_25714	*tmem174*	*ENSDARG00000035388*	*transmembrane protein 174 [Source:ZFIN;Acc:ZDB-GENE-080819-2]*
g_375	*kyat1*	*ENSDARG00000023645*	*kynurenine aminotransferase 1 [Source:ZFIN;Acc:ZDB-GENE-040426-2676]*
g_1982	*NA*	*Not available*	*Not available*
g_4798	*nipsnap1*	*ENSDARG00000005320*	*nipsnap homolog 1 (C. elegans) [Source:ZFIN;Acc:ZDB-GENE-991008-17]*
g_9871	*NA*	*Not available*	*Not available*
g_5422	*adamts12*	*ENSDARG00000067549*	*ADAM metallopeptidase with thrombospondin type 1 motif, 12 [Source:ZFIN;Acc:ZDB-GENE-070705-471]*
g_11725	*prnprs3*	*ENSDARG00000003705*	*prion protein, related sequence 3 [Source:ZFIN;Acc:ZDB-GENE-041221-3]*

Table A.8 – Continued

g_10250	*hnrnpk*	*ENSDARG00000018914*	*heterogeneous nuclear ribonucleoprotein K [Source:ZFIN;Acc:ZDB-GENE-040426-1926]*
g_16135	*si:ch211-170d8.2*	*ENSDARG00000094887*	*si:ch211-170d8.2 [Source:ZFIN;Acc:ZDB-GENE-030328-34]*
g_3793	*zgc:110626*	*ENSDARG00000053159*	*zgc:110626 [Source:ZFIN;Acc:ZDB-GENE-050417-447]*
g_3510	*riok2*	*ENSDARG00000035264*	*RIO kinase 2 (yeast) [Source:ZFIN;Acc:ZDB-GENE-040426-2913]*
g_15909	*wbp1la*	*ENSDARG00000013245*	*WW domain binding protein 1-like a [Source:ZFIN;Acc:ZDB-GENE-030131-1961]*
g_11576	*prop1*	*ENSDARG00000039756*	*PROP paired-like homeobox 1 [Source:ZFIN;Acc:ZDB-GENE-081107-40]*
g_23466	*fam149b1*	*ENSDARG00000061215*	*family with sequence similarity 149 member B1 [Source:ZFIN;Acc:ZDB-GENE-070112-2102]*
g_28017	*atl2*	*ENSDARG00000057719*	*atlastin GTPase 2 [Source:ZFIN;Acc:ZDB-GENE-030131-6505]*
g_3251	*kif20ba*	*ENSDARG00000071009*	*kinesin family member 20Ba [Source:ZFIN;Acc:ZDB-GENE-041111-213]*
g_28922	*NA*	*Not available*	*Not available*
g_19655	*pkp2*	*ENSDARG00000023026*	*plakophilin 2 [Source:ZFIN;Acc:ZDB-GENE-041210-167]*
g_20494	*ccnd2b*	*ENSDARG00000070408*	*cyclin D2, b [Source:ZFIN;Acc:ZDB-GENE-050420-354]*
g_5673	*slc9a3r1a*	*ENSDARG00000000068*	*SLC9A3 regulator 1a [Source:ZFIN;Acc:ZDB-GENE-031006-7]*
g_19735	*edn1*	*ENSDARG00000036912*	*endothelin 1 [Source:ZFIN;Acc:ZDB-GENE-000920-1]*
g_30256	*NA*	*Not available*	*Not available*
g_4306	*spata6l*	*ENSDARG00000004874*	*spermatogenesis associated 6-like [Source:ZFIN;Acc:ZDB-GENE-040426-1369]*

Table A.8 – Continued

g_4547	mblac1	ENSDARG00000077314	metallo-beta-lactamase domain containing 1 [Source:ZFIN;Acc:ZDB-GENE-111102-2]
g_10983	dthd1	ENSDARG00000086452	death domain containing 1 [Source:ZFIN;Acc:ZDB-GENE-140106-180]
g_19245	NA	Not available	Not available
g_63	gfra4b	ENSDARG00000074582	GDNF family receptor alpha 4b [Source:ZFIN;Acc:ZDB-GENE-130530-757]
g_16878	srpx2	ENSDARG00000034559	sushi-repeat containing protein X-linked 2 [Source:ZFIN;Acc:ZDB-GENE-110411-231]
g_15197	NA	Not available	Not available
g_27260	gdf9	ENSDARG00000003229	growth differentiation factor 9 [Source:ZFIN;Acc:ZDB-GENE-050221-7]
g_10773	NA	Not available	Not available
g_17785	NA	Not available	Not available
g_26658	cx32.3	ENSDARG00000041787	connexin 32.3 [Source:ZFIN;Acc:ZDB-GENE-030131-1337]
g_21475	NA	Not available	Not available
g_19065	emc7	ENSDARG00000012144	ER membrane protein complex subunit 7b [Source:ZFIN;Acc:ZDB-GENE-041001-170]
g_10988	rars2	ENSDARG00000032277	arginyl-tRNA synthetase 2, mitochondrial [Source:ZFIN;Acc:ZDB-GENE-040426-1244]
g_16204	enpp1	ENSDARG00000005789	ectonucleotide pyrophosphatase/phosphodiesterase 1 [Source:ZFIN;Acc:ZDB-GENE-040724-172]
g_24670	cipcb	ENSDARG00000078095	CLOCK-interacting pacemaker b [Source:ZFIN;Acc:ZDB-GENE-091204-292]
g_24568	nus1	ENSDARG00000027813	NUS1 dehydrodolichyl diphosphate synthase subunit [Source:ZFIN;Acc:ZDB-GENE-040718-48]
g_26356	si:ch73-208g10.1	ENSDARG00000079808	si:ch73-208g10.1[Source:ZFIN;Acc:ZDB-GENE-040108-6]
g_9738	grapa	ENSDARG00000005414	GRB2 related adaptor protein a [Source:ZFIN;Acc:ZDB-GENE-050522-347]

Table A.8 – Continued

g_9913	*mettl4*	*ENSDARG00000088999*	*methyltransferase like 4 [Source:ZFIN;Acc:ZDB-GENE-130129-2]*
g_27514	*pex2*	*ENSDARG00000062421*	*peroxisomal biogenesis factor 2 [Source:ZFIN;Acc:ZDB-GENE-070530-2]*
g_22006	*terf1*	*ENSDARG00000058710*	*telomeric repeat binding factor (NIMA-interacting) 1 [Source:ZFIN;Acc:ZDB-GENE-090612-2]*

Table A.9 Hundred and thirty-eight "*dN/dS*" candidates based on branch model and their ID, zebrafish orthologs Ensembl ID, chromosome location, as well as start and end position in the genome

Gene ID	Gene name	Ensembl ID	Chromosome	Start	End
g_2285	*NA*	*Not available*	1	18924231	18946748
g_17258	*dnase1l1l*	*ENSDARG00000023861*	1	25706130	25707812
g_16307	*apobec2b*	*ENSDARG00000113992*	1	29204950	29207283
g_11406	*rnd1a*	*ENSDARG00000030547*	1	34765862	34772230
g_3197	*CABZ01040556.1*	*ENSDARG00000058869*	1	67480381	67487780
g_8223	*NA*	*Not available*	1	74466747	74469178
g_18930	*tlcd5a*	*ENSDARG00000024920*	1	76677379	76678495
g_19278	*yif1b*	*ENSDARG00000040505*	1	78854505	78858653
g_29761	*blvra*	*ENSDARG00000059857*	1	80797991	80798860
g_14273	*get1*	*ENSDARG00000074271*	1	83593229	83598565
g_2	*NA*	*Not available*	2	13654546	13683613
g_16937	*NA*	*Not available*	2	23131403	23132257
g_26570	*zgc:112163*	*ENSDARG00000017657*	2	27685987	27686778
g_31175	*NA*	*Not available*	2	28420128	28423955
g_19587	*NA*	*Not available*	2	37173210	37174715
g_27671	*NA*	*Not available*	2	37190250	37192374
g_5377	*lsp1a*	*ENSDARG00000027310*	2	49550177	49585415
g_16386	*NA*	*Not available*	2	49903270	49910813
g_15913	*C25H12orf29*	*ENSDARG00000045785*	2	53914302	53922325
g_30710	*rassf9*	*ENSDARG00000074721*	2	54281012	54292182
g_23492	*NA*	*Not available*	2	55246920	55267779
g_9319	*NA*	*Not available*	2	58337220	58352746
g_18423	*kcnc1b*	*ENSDARG00000032959*	2	62677850	62686447
g_4725	*ciartb*	*ENSDARG00000088171*	2	63965019	63970778
g_21318	*ZNF276*	*ENSDARG00000110991*	2	92808730	92813531

Table A.9 – Continued

g_5105	*bmper*	*ENSDARG00000101980*	3	7844495	7866533
g_16924	*pou3f1*	*ENSDARG00000009823*	3	14802296	14803417
g_17745	*NA*	*Not available*	3	15038304	15039767
g_467	*rusc1*	*ENSDARG00000078125*	3	20647418	20653270
g_4231	*zgc:101716*	*ENSDARG00000010738*	3	30926772	30931875
g_19202	*tcaim*	*ENSDARG00000079881*	3	36261731	36266995
g_10648	*lyplal1*	*ENSDARG00000088764*	3	37372350	37395181
g_27166	*sfxn5b*	*ENSDARG00000026137*	3	79711232	79729089
g_7392	*TTC9*	*ENSDARG00000074363*	3	83271545	83276842
g_8911	*arg2*	*ENSDARG00000039269*	3	83666658	83680053
g_7284	*NA*	*Not available*	3	84462711	84470890
g_27153	*wdr32*	*ENSDARG00000029600*	3	84543869	84554210
g_23676	*strn3*	*ENSDARG00000001729*	3	85476175	85503912
g_10797	*CELF6*	*ENSDARG00000101933*	4	11402538	11434738
g_3976	*cgrefl*	*ENSDARG00000075444*	4	15779380	15784151
g_18744	*sdhaf2*	*ENSDARG00000062971*	4	21979348	21983416
g_20127	*zgc:113276*	*ENSDARG00000056650*	4	26879862	26884267
g_8116	*il16*	*ENSDARG00000102908*	4	50413232	50479596
g_8024	*ndufv3*	*ENSDARG00000090389*	4	53107197	53111826
g_24962	*rdh1*	*ENSDARG00000017882*	4	76582775	76585451
g_8089	*si:dkey-100n23.3*	*ENSDARG00000062148*	4	84071605	84095814
g_9496	*mrpl30*	*ENSDARG00000069850*	4	85055479	85057715
g_22198	*vasna*	*ENSDARG00000099266*	5	1624625	1626742
g_6060	*dla*	*ENSDARG00000010791*	5	6645436	6653479
g_4206	*abcc10*	*ENSDARG00000077988*	5	7024098	7062567
g_31294	*NA*	*Not available*	5	7307841	7313475
g_17536	*snrpb2*	*ENSDARG00000039424*	5	9047893	9051543

Table A.9 – Continued

g_17232	*socs1b*	*ENSDARG00000089873*	5	16228489	16230906
g_19518	*sod3b*	*ENSDARG00000079183*	5	17823647	17824756
g_8712	*NA*	*Not available*	5	18054685	18073558
g_27012	*rhbdd2*	*ENSDARG00000092463*	5	20724975	20725958
g_13030	*slx4*	*ENSDARG00000061414*	5	27514960	27531111
g_2363	*cenpk*	*ENSDARG00000039616*	5	29987825	29992659
g_14353	*fgf8b*	*ENSDARG00000039615*	5	30119343	30122751
g_12806	*uck2a*	*ENSDARG00000006074*	5	56750761	56759093
g_23054	*kifap3a*	*ENSDARG00000008639*	5	64109811	64146444
g_15227	*prrx1b*	*ENSDARG00000042027*	5	64232024	64245806
g_15614	*NA*	*Not available*	5	71610136	71615985
g_15380	*uox*	*ENSDARG00000007024*	5	80985542	80993627
g_27365	*hsd11b1la*	*ENSDARG00000071377*	5	84984897	84989654
g_14296	*creb3l3a*	*ENSDARG00000056226*	5	85258983	85265610
g_23906	*mier2*	*ENSDARG00000071413*	5	87873689	87886919
g_12699	*haus5*	*ENSDARG00000019156*	5	92951946	92973929
g_589	*si:ch73-71c20.5*	*ENSDARG00000097696*	6	16080312	16081605
g_18291	*aknad1*	*ENSDARG00000094414*	6	22242646	22247314
g_7793	*elovl1a*	*ENSDARG00000099960*	6	24441163	24443234
g_23900	*NA*	*Not available*	6	25667377	25675026
g_5085	*si:dkeyp-7a3.1*	*ENSDARG00000090429*	6	26553353	26571655
g_20870	*cx47.1*	*ENSDARG00000073896*	6	27373833	27375074
g_30991	*si:dkey-32m20.1*	*ENSDARG00000075715*	6	29147415	29151006
g_28295	*or115-2*	*ENSDARG00000053817*	6	44133714	44134670
g_3141	*mybbp1a*	*ENSDARG00000078214*	6	49430050	49432985
g_19432	*rab34b*	*ENSDARG00000010977*	6	56087817	56095154
g_31071	*sgcd*	*ENSDARG00000098573*	6	62685518	62810147

Table A.9 – Continued

g_13619	*lyn*	*ENSDARG00000107511*	6	81931920	81967404
g_26086	*NA*	*Not available*	8	2466424	2478151
g_7876	*gdf2*	*ENSDARG00000059173*	8	5138702	5140758
g_10758	*ubtd1b*	*ENSDARG00000079623*	8	5240011	5243102
g_29167	*tmem130*	*ENSDARG00000103789*	8	7561514	7566176
g_14608	*vwa2*	*ENSDARG00000075441*	8	12251985	12279420
g_25655	*NA*	*Not available*	8	13317587	13319835
g_1660	*fbxl15*	*ENSDARG00000005284*	8	14300536	14302387
g_21651	*entpd2a.1*	*ENSDARG00000035506*	8	35797651	35801123
g_14269	*ptgdsa*	*ENSDARG00000069439*	8	41429739	41433268
g_26258	*surf2*	*ENSDARG00000112476*	8	44428548	44430347
g_22150	*NA*	*Not available*	8	47028632	47035286
g_14166	*ccdc62*	*ENSDARG00000111759*	8	48529329	48539535
g_14680	*kmt5aa*	*ENSDARG00000105231*	8	48571329	48576089
g_25714	*tmem174*	*ENSDARG00000035388*	8	52966007	52968209
g_375	*kyat1*	*ENSDARG00000023645*	8	53909056	53914043
g_1982	*NA*	*Not available*	8	59963331	59978876
g_4798	*nipsnap1*	*ENSDARG00000005320*	12	19196711	19220372
g_9871	*NA*	*Not available*	12	23228868	23234304
g_5422	*adamts12*	*ENSDARG00000067549*	12	23679470	23681078
g_11725	*prnprs3*	*ENSDARG00000003705*	12	24925120	24926679
g_10250	*hnrnpk*	*ENSDARG00000018914*	12	28730314	28735425
g_16135	*si:ch211-170d8.2*	*ENSDARG00000094887*	12	33480727	33484245
g_3793	*zgc:110626*	*ENSDARG00000053159*	12	35565969	35570274
g_3510	*riok2*	*ENSDARG00000035264*	12	36321334	36326705
g_15909	*wbp1la*	*ENSDARG00000013245*	12	50640964	50646931
g_11576	*prop1*	*ENSDARG00000039756*	12	63178898	63182101

Table A.9 – Continued

g_23466	*fam149b1*	*ENSDARG00000061215*	12	77666799	77675876
g_28017	*atl2*	*ENSDARG00000057719*	12	78510171	78526349
g_3251	*kif20ba*	*ENSDARG00000071009*	12	87482007	87568371
g_28922	*NA*	*Not available*	13	1598685	1618679
g_19655	*pkp2*	*ENSDARG00000023026*	13	4925200	4950169
g_20494	*ccnd2b*	*ENSDARG00000070408*	13	16838775	16847759
g_5673	*slc9a3r1a*	*ENSDARG00000000068*	13	28726774	28768294
g_19735	*edn1*	*ENSDARG00000036912*	13	37472074	37474455
g_30256	*NA*	*Not available*	13	54459102	54460373
g_4306	*spata6l*	*ENSDARG00000004874*	14	2581558	2591619
g_4547	*mblac1*	*ENSDARG00000077314*	14	5032484	5044987
g_10983	*dthd1*	*ENSDARG00000086452*	14	15251657	15255743
g_19245	*NA*	*Not available*	14	26725835	26726849
g_63	*gfra4b*	*ENSDARG00000074582*	14	42211303	42217174
g_16878	*srpx2*	*ENSDARG00000034559*	14	44229318	44237648
g_15197	*NA*	*Not available*	14	47635366	47649187
g_27260	*gdf9*	*ENSDARG00000003229*	14	51506960	51512549
g_10773	*NA*	*Not available*	14	52877841	52881790
g_17785	*NA*	*Not available*	15	4954679	4969209
g_26658	*cx32.3*	*ENSDARG00000041787*	15	7570366	7571226
g_21475	*NA*	*Not available*	15	9486605	9490345
g_19065	*emc7*	*ENSDARG00000012144*	15	9699083	9701941
g_10988	*rars2*	*ENSDARG00000032277*	15	10848323	10860765
g_16204	*enpp1*	*ENSDARG00000005789*	15	11952092	11982611
g_24670	*cipcb*	*ENSDARG00000078095*	15	21463323	21465996
g_24568	*nus1*	*ENSDARG00000027813*	15	40599803	40608129
g_26356	*si:ch73-208g10.1*	*ENSDARG00000079808*	15	62706160	62710782

Table A.9 – Continued

g_9738	*grapa*	*ENSDARG00000005414*	15	70530688	70559077
g_15165	*trir*	*ENSDARG00000104178*	15	76180516	76184685
g_9913	*mettl4*	*ENSDARG00000088999*	21	8948135	8954987
g_27514	*pex2*	*ENSDARG00000062421*	21	21062173	21068580
g_22006	*terf1*	*ENSDARG00000058710*	21	23944705	23951147

Table A.10 Two hundred and ten "*dN/dS*" candidates based on branch-site model *their* ID, name, zebrafish orthologs' Ensemble ID, and description

Gene ID	Gene name	Ensembl ID	Gene Description
g_17780	*sdcbp2*	*ENSDARG00000012513*	*syndecan binding protein (syntenin) 2 [Source:ZFIN;Acc:ZDB-GENE-030131-3727]*
g_32331	*pard6b*	*ENSDARG00000003865*	*par-6 partitioning defective 6 homolog beta (C. elegans) [Source:ZFIN;Acc:ZDB-GENE-090312-133]*
g_2285	*NA*	*Not available*	*Not available*
g_30769	*ccdc114*	*ENSDARG00000015010*	*coiled-coil domain containing 114 [Source:ZFIN;Acc:ZDB-GENE-041114-110]*
g_25897	*sypl2b*	*ENSDARG00000000690*	*synaptophysin-like 2b [Source:ZFIN;Acc:ZDB-GENE-050417-309]*
g_16307	*apobec2b*	*ENSDARG00000113992*	*apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 2b [Source:ZFIN;Acc:ZDB-GENE-090618-1]*
g_9141	*etv7*	*ENSDARG00000089434*	*ETS variant transcription factor 7 [Source:ZFIN;Acc:ZDB-GENE-070209-53]*
g_13402	*NA*	*Not available*	*Not available*
g_12217	*ptpdc1b*	*ENSDARG00000058873*	*protein tyrosine phosphatase domain containing 1b [Source:ZFIN;Acc:ZDB-GENE-090312-138]*
g_4229	*uhrf1bp1*	*ENSDARG00000077011*	*UHRF1 binding protein 1 [Source:ZFIN;Acc:ZDB-GENE-090312-82]*
g_2079	*plekhg5b*	*ENSDARG00000101752*	*pleckstrin homology domain containing, family G (with RhoGef domain) member 5b [Source:ZFIN;Acc:ZDB-GENE-090908-6]*
g_27323	*zbtb48*	*ENSDARG00000039263*	*zinc finger and BTB domain containing 48 [Source:ZFIN;Acc:ZDB-GENE-030131-4450]*
g_13829	*itga5*	*ENSDARG00000006353*	*integrin, alpha 5 (fibronectin receptor, alpha polypeptide) [Source:ZFIN;Acc:ZDB-GENE-031116-52]*
g_4376	*NA*	*Not available*	*Not available*
g_25302	*zgc:101731*	*ENSDARG00000040965*	*zgc:101731 [Source:ZFIN;Acc:ZDB-GENE-040912-57]*
g_9073	*si:ch211-137a8.4*	*ENSDARG00000078748*	*si:ch211-137a8.4 [Source:ZFIN;Acc:ZDB-GENE-030131-3742]*
g_25237	*NA*	*Not available*	*Not available*

Table A.10 – Continued

g_8223	*NA*	*Not available*	*Not available*
g_19278	*yif1b*	*ENSDARG00000040505*	*Yip1 interacting factor homolog B (S. cerevisiae) [Source:ZFIN;Acc:ZDB-GENE-041114-16]*
g_27423	*inppl1a*	*ENSDARG00000104222*	*inositol polyphosphate phosphatase-like 1a [Source:NCBI gene;Acc:325179]*
g_23555	*meis3*	*ENSDARG00000002795*	*myeloid ecotropic viral integration site 3 [Source:ZFIN;Acc:ZDB-GENE-010406-2]*
g_10008	*lpar5b*	*ENSDARG00000068638*	*lysophosphatidic acid receptor 5b [Source:ZFIN;Acc:ZDB-GENE-081022-116]*
g_10350	*cntn2*	*ENSDARG00000000472*	*contactin 2 [Source:ZFIN;Acc:ZDB-GENE-990630-12]*
g_2381	*atp2b4*	*ENSDARG00000044902*	*ATPase plasma membrane Ca2+ transporting 4 [Source:ZFIN;Acc:ZDB-GENE-061027-60]*
g_10766	*fer1l4*	*ENSDARG00000076952*	*fer-1 like family member 4 [Source:ZFIN;Acc:ZDB-GENE-130530-815]*
g_6097	*skib*	*ENSDARG00000008034*	*v-ski avian sarcoma viral oncogene homolog b [Source:ZFIN;Acc:ZDB-GENE-990715-10]*
g_5149	*suclg2*	*ENSDARG00000044914*	*succinate-CoA ligase, GDP-forming, beta subunit [Source:ZFIN;Acc:ZDB-GENE-030114-3]*
g_27671	*NA*	*Not available*	*Not available*
g_7401	*camkvb*	*ENSDARG00000005141*	*CaM kinase-like vesicle-associated b [Source:ZFIN;Acc:ZDB-GENE-040426-1140]*
g_30422	*cntn4*	*ENSDARG00000098161*	*contactin 4 [Source:ZFIN;Acc:ZDB-GENE-060929-776]*
g_15303	*si:dkey-156n14.3*	*ENSDARG00000052351*	*si:dkey-156n14.3 [Source:ZFIN;Acc:ZDB-GENE-030131-4816]*
g_3513	*cand2*	*ENSDARG00000005749*	*cullin-associated and neddylation-dissociated 2 (putative) [Source:ZFIN;Acc:ZDB-GENE-060503-645]*
g_30555	*NA*	*Not available*	*Not available*
g_11692	*tnnt3b*	*ENSDARG00000068457*	*troponin T type 3b (skeletal, fast) [Source:ZFIN;Acc:ZDB-GENE-030520-2]*
g_16386	*NA*	*Not available*	*Not available*

Table A.10 – Continued

g_1586	sox6	ENSDARG00000015536	SRY-box transcription factor 6 [Source:ZFIN;Acc:ZDB-GENE-081120-6]
g_4370	scamp2	ENSDARG00000010279	secretory carrier membrane protein 2 [Source:ZFIN;Acc:ZDB-GENE-040426-2702]
g_8976	sin3aa	ENSDARG00000079716	SIN3 transcription regulator family member Aa [Source:ZFIN;Acc:ZDB-GENE-070620-3]
g_7519	snupn	ENSDARG00000008395	snurportin 1 [Source:ZFIN;Acc:ZDB-GENE-030131-3464]
g_23708	sigirr	ENSDARG00000062204	single immunoglobulin and toll-interleukin 1 receptor (TIR) domain [Source:ZFIN;Acc:ZDB-GENE-080303-3]
g_22086	NA	Not available	Not available
g_8854	taf3	ENSDARG00000045513	TAF3 RNA polymerase II, TATA box binding protein (TBP)-associated facto [Source:ZFIN;Acc:ZDB-GENE-030131-6406]
g_18423	kcnc1b	ENSDARG00000032959	potassium voltage-gated channel, Shaw-related subfamily, member 1b [Source:ZFIN;Acc:ZDB-GENE-080414-3]
g_22865	lactb	ENSDARG00000040803	lactamase, beta [Source:ZFIN;Acc:ZDB-GENE-020111-1]
g_30963	kti12	ENSDARG00000054301	KTI12 chromatin associated homolog [Source:ZFIN;Acc:ZDB-GENE-060825-174]
g_19186	cd9a	ENSDARG00000005842	CD9 molecule a [Source:ZFIN;Acc:ZDB-GENE-030131-1175]
g_31300	NA	Not available	Not available
g_8751	ush1c	ENSDARG00000051876	Usher syndrome 1C [Source:ZFIN;Acc:ZDB-GENE-060312-41]
g_5125	bicd1a	ENSDARG00000079496	bicaudal D homolog 1a [Source:ZFIN;Acc:ZDB-GENE-081031-9]
g_3070	aars1	ENSDARG00000069142	alanyl-tRNA synthetase 1 [Source:ZFIN;Acc:ZDB-GENE-030131-3663]
g_5203	cmip	ENSDARG00000062933	c-Maf inducing protein [Source:ZFIN;Acc:ZDB-GENE-050419-50]
g_31324	NA	Not available	Not available
g_21318	ZNF276	ENSDARG00000110991	zgc:158366 [Source:ZFIN;Acc:ZDB-GENE-070209-176]
g_5105	bmper	ENSDARG00000101980	BMP binding endothelial regulator [Source:ZFIN;Acc:ZDB-GENE-030219-146]
g_10035	dhdds	ENSDARG00000039851	dehydrodolichyl diphosphate synthase [Source:ZFIN;Acc:ZDB-GENE-040426-2236]

Table A.10 – Continued

g_3152	*pbx2*	*ENSDARG00000019717*	*pre-B-cell leukemia homeobox 2 [Source:ZFIN;Acc:ZDB-GENE-000405-5]*
g_22274	*si:dkey-17m8.1*	*ENSDARG00000079530*	*si:dkey-17m8.1 [Source:ZFIN;Acc:ZDB-GENE-110411-225]*
g_13715	*NA*	*Not available*	*Not available*
g_14904	*tnxba*	*ENSDARG00000001760*	*tenascin XBa [Source:ZFIN;Acc:ZDB-GENE-070103-5]*
g_10648	*lyplal1*	*ENSDARG00000088764*	*lysophospholipase like 1 [Source:ZFIN;Acc:ZDB-GENE-050306-32]*
g_8654	*NA*	*Not available*	*Not available*
g_10131	*ppie*	*ENSDARG00000103234*	*peptidylprolyl isomerase E (cyclophilin E) [Source:ZFIN;Acc:ZDB-GENE-050417-167]*
g_17915	*NA*	*Not available*	*Not available*
g_5455	*atg2b*	*ENSDARG00000097650*	*autophagy related 2B [Source:ZFIN;Acc:ZDB-GENE-131121-626]*
g_5016	*fgfrl1a*	*ENSDARG00000032617*	*fibroblast growth factor receptor like 1a [Source:ZFIN;Acc:ZDB-GENE-040128-2]*
g_21129	*prlh2r*	*ENSDARG00000054700*	*prolactin releasing hormone 2 receptor [Source:ZFIN;Acc:ZDB-GENE-120411-41]*
g_7392	*TTC9*	*ENSDARG00000074363*	*si:ch211-259k16.3 [Source:ZFIN;Acc:ZDB-GENE-090312-172]*
g_27153	*wdr32*	*ENSDARG00000029600*	*WD repeat domain 32 [Source:ZFIN;Acc:ZDB-GENE-040426-2314]*
g_4655	*numb*	*ENSDARG00000027279*	*NUMB endocytic adaptor protein [Source:ZFIN;Acc:ZDB-GENE-060422-1]*
g_22608	*guca1g*	*ENSDARG00000045737*	*guanylate cyclase activator 1g [Source:ZFIN;Acc:ZDB-GENE-050120-1]*
g_31014	*brd7*	*ENSDARG00000008380*	*bromodomain containing 7 [Source:ZFIN;Acc:ZDB-GENE-040426-2687]*
g_28942	*chrna3*	*ENSDARG00000100991*	*cholinergic receptor, nicotinic, alpha 3 [Source:ZFIN;Acc:ZDB-GENE-070822-1]*
g_20923	*ppfibp2b*	*ENSDARG00000029168*	*PPFIA binding protein 2b [Source:ZFIN;Acc:ZDB-GENE-040718-54]*
g_5635	*tead1b*	*ENSDARG00000059483*	*TEA domain family member 1b [Source:ZFIN;Acc:ZDB-GENE-091013-5]*
g_11371	*rasa3*	*ENSDARG00000063371*	*RAS p21 protein activator 3 [Source:ZFIN;Acc:ZDB-GENE-090313-21]*

Table A.10 – Continued

g_1316	scml2	ENSDARG00000012949	Scm polycomb group protein like 2 [Source:ZFIN;Acc:ZDB-GENE-130530-546]
g_91	lmnl3	ENSDARG00000007751	lamin L3 [Source:ZFIN;Acc:ZDB-GENE-020424-4]
g_7582	dok4	ENSDARG00000073731	docking protein 4 [Source:ZFIN;Acc:ZDB-GENE-041008-91]
g_8116	il16	ENSDARG00000102908	interleukin 16 [Source:ZFIN;Acc:ZDB-GENE-130103-3]
g_22445	GTPBP8	ENSDARG00000075033	GTP binding protein 8 (putative) [Source:ZFIN;Acc:ZDB-GENE-070912-719]
g_25957	znf142	ENSDARG00000061373	zinc finger protein 142 [Source:ZFIN;Acc:ZDB-GENE-080512-2]
g_21492	lrrc3	ENSDARG00000078415	leucine rich repeat containing 3 Source:ZFIN;Acc:ZDB-GENE-080327-13]
g_9290	pofut2	ENSDARG00000045175	protein O-fucosyltransferase 2 [Source:ZFIN;Acc:ZDB-GENE-030131-3595]
g_7993	si:dkey-11f4.16	ENSDARG00000099799	si:dkey-11f4.16 [Source:ZFIN;Acc:ZDB-GENE-070912-357]
g_23837	NA	Not available	Not available
g_2129	NA	Not available	Not available
g_9432	rftn2	ENSDARG00000056078	raftlin family member 2 [Source:ZFIN;Acc:ZDB-GENE-040426-2760]
g_2217	efhc2	ENSDARG00000004204	EF-hand domain (C-terminal) containing 2 [Source:ZFIN;Acc:ZDB-GENE-031001-10]
g_3352	ifngr1	ENSDARG00000074771	interferon gamma receptor 1 [Source:ZFIN;Acc:ZDB-GENE-081022-158]
g_17536	snrpb2	ENSDARG00000039424	small nuclear ribonucleoprotein polypeptide B2 [Source:ZFIN;Acc:ZDB-GENE-060616-2]
g_14774	pex6	ENSDARG00000070958	peroxisomal biogenesis factor 6 [Source:ZFIN;Acc:ZDB-GENE-081104-252]
g_10167	cyp2u1	ENSDARG00000026548	cytochrome P450, family 2, subfamily U, polypeptide 1 [Source:ZFIN;Acc:ZDB-GENE-070730-1]
g_34736	casp6a	ENSDARG00000093405	caspase 6, apoptosis-related cysteine peptidase a [Source:ZFIN;Acc:ZDB-GENE-030825-4]

Table A.10 – Continued

g_10757	*psip1a*	*ENSDARG00000104710*	*PC4 and SFRS1 interacting protein 1a [Source:ZFIN;Acc:ZDB-GENE-050522-104]*
g_17232	*socs1b*	*ENSDARG00000089873*	*suppressor of cytokine signaling 1b [Source:ZFIN;Acc:ZDB-GENE-090313-141]*
g_19811	*primpol*	*ENSDARG00000033273*	*primase and polymerase (DNA-directed) [Source:ZFIN;Acc:ZDB-GENE-051113-100]*
g_10706	*NA*	*Not available*	*Not available*
g_17667	*pdcd4b*	*ENSDARG00000041022*	*programmed cell death 4b [Source:ZFIN;Acc:ZDB-GENE-030131-9847]*
g_13030	*slx4*	*ENSDARG00000061414*	*SLX4 structure-specific endonuclease subunit homolog (S. cerevisiae) [Source:ZFIN;Acc:ZDB-GENE-050208-359]*
g_9907	*smap1*	*ENSDARG00000031302*	*small ArfGAP 1 [Source:ZFIN;Acc:ZDB-GENE-060920-2]*
g_2363	*cenpk*	*ENSDARG00000039616*	*centromere protein K [Source:ZFIN;Acc:ZDB-GENE-090313-204]*
g_13036	*ctnnd1*	*ENSDARG00000078233*	*catenin (cadherin-associated protein), delta 1 [Source:ZFIN;Acc:ZDB-GENE-110208-9]*
g_9123	*aspm*	*ENSDARG00000103754*	*abnormal spindle microtubule assembly [Source:ZFIN;Acc:ZDB-GENE-050208-620]*
g_15614	*NA*	*Not available*	*Not available*
g_19254	*NA*	*Not available*	*Not available*
g_17832	*adgrl4*	*ENSDARG00000013653*	*adhesion G protein-coupled receptor L4 [Source:ZFIN;Acc:ZDB-GENE-040426-2689]*
g_2091	*cpox*	*ENSDARG00000062025*	*coproporphyrinogen oxidase [Source:ZFIN;Acc:ZDB-GENE-030131-9884]*
g_13148	*zgc:153738*	*ENSDARG00000069230*	*zgc:153738 [Source:ZFIN;Acc:ZDB-GENE-061013-622]*
g_9828	*clocka*	*ENSDARG00000011703*	*clock circadian regulator a [Source:ZFIN;Acc:ZDB-GENE-990630-14]*
g_9295	*NA*	*Not available*	*Not available*
g_24492	*arhgap45b*	*ENSDARG00000062049*	*Rho GTPase activating protein 45b [Source:ZFIN;Acc:ZDB-GENE-071213-2]*
g_15284	*hapln4*	*ENSDARG00000018542*	*hyaluronan and proteoglycan link protein 4 [Source:ZFIN;Acc:ZDB-GENE-060503-243]*

Table A.10 – Continued

g_12923	elovl8b	ENSDARG00000057365	ELOVL fatty acid elongase 8b [Source:ZFIN;Acc:ZDB-GENE-050522-453]
g_21735	nsun4	ENSDARG00000021324	NOP2/Sun RNA methyltransferase 4 [Source:ZFIN;Acc:ZDB-GENE-041212-77]
g_1993	pip5k1cb	ENSDARG00000100313	phosphatidylinositol-4-phosphate 5-kinase, type I, gamma b [Source:ZFIN;Acc:ZDB-GENE-110408-21]
g_22857	aire	ENSDARG00000056784	autoimmune regulator [Source:ZFIN;Acc:ZDB-GENE-071008-4]
g_21232	ccdc24	ENSDARG00000038793	coiled-coil domain containing 24 [Source:ZFIN;Acc:ZDB-GENE-050327-18]
g_16610	twsg1a	ENSDARG00000104244	twisted gastrulation BMP signaling modulator 1a [Source:ZFIN;Acc:ZDB-GENE-010509-2]
g_6311	or101-1	ENSDARG00000013014	odorant receptor, family B, subfamily 101, member 1 [Source:ZFIN;Acc:ZDB-GENE-990415-190]
g_30547	NA	Not available	Not available
g_16712	NA	Not available	Not available
g_26253	angptl5	ENSDARG00000056630	angiopoietin-like 5 [Source:ZFIN;Acc:ZDB-GENE-030131-5054]
g_3274	zgc:163098	ENSDARG00000078911	zgc:163098 [Source:ZFIN;Acc:ZDB-GENE-070410-141]
g_13108	ephb4a	ENSDARG00000100725	eph receptor B4a [Source:NCBI gene;Acc:30688]
g_12533	txndc15	ENSDARG00000110357	thioredoxin domain containing 15 [Source:ZFIN;Acc:ZDB-GENE-070615-36]
g_4047	rimbp2	ENSDARG00000001154	RIMS binding protein 2 [Source:ZFIN;Acc:ZDB-GENE-040724-96]
g_26769	mtmr12	ENSDARG00000059817	myotubularin related protein 12 [Source:ZFIN;Acc:ZDB-GENE-050401-1]
g_19984	tmlhe	ENSDARG00000077547	trimethyllysine hydroxylase, epsilon [Source:ZFIN;Acc:ZDB-GENE-091204-144]
g_10062	robo4	ENSDARG00000009387	roundabout, axon guidance receptor, homolog 4 (Drosophila) [Source:ZFIN;Acc:ZDB-GENE-020809-1]
g_26086	NA	Not available	Not available
g_11288	si:dkey-16i5.8	ENSDARG00000096722	si:dkey-16i5.8 [Source:ZFIN;Acc:ZDB-GENE-030131-1207]

Table A.10 – Continued

g_6309	NA	Not available	Not available
g_13628	si:ch211-234p6.5	ENSDARG00000071460	si:ch211-234p6.5 [Source:ZFIN;Acc:ZDB-GENE-060503-692]
g_11245	atxn2l	ENSDARG00000011597	ataxin 2-like [Source:ZFIN;Acc:ZDB-GENE-030131-3246]
g_12931	znf281b	ENSDARG00000035910	zinc finger protein 281b [Source:ZFIN;Acc:ZDB-GENE-050220-1]
g_17473	NA	Not available	Not available
g_2739	tcf7l2	ENSDARG00000004415	transcription factor 7 like 2 [Source:ZFIN;Acc:ZDB-GENE-991110-8]
g_4726	dlg5a	ENSDARG00000074059	discs, large homolog 5a (Drosophila) [Source:ZFIN;Acc:ZDB-GENE-030131-3149]
g_22511	cd79b	ENSDARG00000104691	CD79b molecule, immunoglobulin-associated beta [Source:ZFIN;Acc:ZDB-GENE-121219-1]
g_7163	plpp1a	ENSDARG00000053381	phospholipid phosphatase 1a [Source:ZFIN;Acc:ZDB-GENE-080225-26]
g_25033	dennd1a	ENSDARG00000014592	DENN/MADD domain containing 1A [Source:ZFIN;Acc:ZDB-GENE-060404-6]
g_27606	NA	Not available	Not available
g_21651	entpd2a.1	ENSDARG00000035506	ectonucleoside triphosphate diphosphohydrolase 2a, tandem duplicate 1 [Source:ZFIN;Acc:ZDB-GENE-040724-187]
g_14578	trabd2a	ENSDARG00000089701	TraB domain containing 2A Source:ZFIN;Acc:ZDB-GENE-030131-4053]
g_17466	NA	Not available	Not available
g_26258	surf2	ENSDARG00000112476	surfeit 2 [Source:ZFIN;Acc:ZDB-GENE-040801-86]
g_14187	NA	Not available	Not available
g_14166	ccdc62	ENSDARG00000111759	coiled-coil domain containing 62 [Source:ZFIN;Acc:ZDB-GENE-040718-71]
g_375	kyat1	ENSDARG00000023645	kynurenine aminotransferase 1 [Source:ZFIN;Acc:ZDB-GENE-040426-2676]
g_21419	uap1l1	ENSDARG00000013082	UDP-N-acetylglucosamine pyrophosphorylase 1, like 1 [Source:NCBI gene;Acc:393264]
g_26113	dpp7	ENSDARG00000027750	dipeptidyl-peptidase 7 [Source:ZFIN;Acc:ZDB-GENE-050306-16]

Table A.10 – Continued

g_11926	nos1	ENSDARG00000068910	nitric oxide synthase 1 (neuronal) [Source:ZFIN;Acc:ZDB-GENE-001101-1]
g_5422	adamts12	ENSDARG00000067549	ADAM metallopeptidase with thrombospondin type 1 motif, 12 [Source:ZFIN;Acc:ZDB-GENE-070705-471]
g_24745	agpat9l	ENSDARG00000006491	1-acylglycerol-3-phosphate O-acyltransferase 9, like [Source:ZFIN;Acc:ZDB-GENE-060531-19]
g_28297	snap29	ENSDARG00000038518	synaptosome associated protein 29 [Source:ZFIN;Acc:ZDB-GENE-041111-226]
g_19332	SLC25A1	ENSDARG00000080000	si:dkey-178e17.1 [Source:ZFIN;Acc:ZDB-GENE-081104-41]
g_5559	plcxd3	ENSDARG00000054794	phosphatidylinositol-specific phospholipase C, X domain containing 3 [Source:ZFIN;Acc:ZDB-GENE-050327-10]
g_10250	hnrnpk	ENSDARG00000018914	heterogeneous nuclear ribonucleoprotein K [Source:ZFIN;Acc:ZDB-GENE-040426-1926]
g_13187	ppp2r2aa	ENSDARG00000021996	protein phosphatase 2, regulatory subunit B, alpha a [Source:ZFIN;Acc:ZDB-GENE-130530-565]
g_3793	zgc:110626	ENSDARG00000053159	zgc:110626 [Source:ZFIN;Acc:ZDB-GENE-050417-447]
g_3510	riok2	ENSDARG00000035264	RIO kinase 2 (yeast) [Source:ZFIN;Acc:ZDB-GENE-040426-2913]
g_19377	aifm3	ENSDARG00000062780	apoptosis inducing factor mitochondria associated 3 [Source:ZFIN;Acc:ZDB-GENE-140619-2]
g_22260	ela3l	ENSDARG00000007276	elastase 3 like [Source:ZFIN;Acc:ZDB-GENE-060710-2]
g_25506	rasgrp3	ENSDARG00000077864	RAS guanyl releasing protein 3 (calcium and DAG-regulated) [Source:ZFIN;Acc:ZDB-GENE-070424-82]
g_747	ppm1ba	ENSDARG00000001888	protein phosphatase, Mg2+/Mn2+ dependent, 1Ba [Source:ZFIN;Acc:ZDB-GENE-991102-16]
g_26946	NA	Not available	Not available
g_28147	cryzl1	ENSDARG00000026902	crystallin, zeta (quinone reductase)-like 1 [Source:ZFIN;Acc:ZDB-GENE-040718-378]
g_28922	NA	Not available	Not available
g_1207	gdi2	ENSDARG00000005451	GDP dissociation inhibitor 2 [Source:ZFIN;Acc:ZDB-GENE-030131-2485]

Table A.10 – Continued

g_8434	*cnot4b*	*ENSDARG00000007639*	*CCR4-NOT transcription complex, subunit 4b [Source:ZFIN;Acc:ZDB-GENE-040426-1164]*
g_30839	*NA*	*Not available*	*Not available*
g_28729	*myf5*	*ENSDARG00000007277*	*myogenic factor 5 [Source:ZFIN;Acc:ZDB-GENE-000616-6]*
g_24117	*napepld*	*ENSDARG00000009252*	*N-acyl phosphatidylethanolamine phospholipase D [Source:ZFIN;Acc:ZDB-GENE-030131-3856]*
g_14849	*NA*	*Not available*	*Not available*
g_1560	*slc9a3.1*	*ENSDARG00000058498*	*solute carrier family 9 member A3, tandem duplicate 1 [Source:ZFIN;Acc:ZDB-GENE-060503-545]*
g_20396	*NA*	*Not available*	*Not available*
g_27497	*CABZ01101996.1*	*ENSDARG00000109996*	*Not available*
g_10668	*thrap3b*	*ENSDARG00000098228*	*thyroid hormone receptor associated protein 3b [Source:ZFIN;Acc:ZDB-GENE-040516-9]*
g_25418	*NA*	*Not available*	*Not available*
g_12375	*dync1li1*	*ENSDARG00000098317*	*dynein, cytoplasmic 1, light intermediate chain 1 [Source:ZFIN;Acc:ZDB-GENE-030131-4108]*
g_8949	*calcr*	*ENSDARG00000028845*	*calcitonin receptor [Source:ZFIN;Acc:ZDB-GENE-060503-420]*
g_17229	*nsun2*	*ENSDARG00000056665*	*NOP2/Sun RNA methyltransferase 2 [Source:ZFIN;Acc:ZDB-GENE-030131-4017]*
g_14602	*cited4b*	*ENSDARG00000101009*	*Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 4b [Source:ZFIN;Acc:ZDB-GENE-030425-5]*
g_18822	*si:dkey-106g10.7*	*ENSDARG00000088036*	*si:dkey-106g10.7 [Source:ZFIN;Acc:ZDB-GENE-160728-46]*
g_4306	*spata6l*	*ENSDARG00000004874*	*spermatogenesis associated 6-like [Source:ZFIN;Acc:ZDB-GENE-040426-1369]*
g_22025	*ino80b*	*ENSDARG00000062749*	*INO80 complex subunit B [Source:ZFIN;Acc:ZDB-GENE-061013-69]*
g_6247	*si:cabz01074946.1*	*ENSDARG00000090396*	*si:cabz01074946.1 [Source:ZFIN;Acc:ZDB-GENE-160113-134]*
g_7328	*b4galt7*	*ENSDARG00000021899*	*xylosylprotein beta 1,4-galactosyltransferase, polypeptide 7 (galactosyltransferase I) [Source:ZFIN;Acc:ZDB-GENE-040727-3]*

Table A.10 – Continued

g_9009	sec24b	ENSDARG00000071906	SEC24 homolog B, COPII coat complex component [Source:ZFIN;Acc:ZDB-GENE-030131-6565]
g_19800	NA	Not available	Not available
g_74	ift172	ENSDARG00000041870	intraflagellar transport 172 [Source:NCBI gene;Acc:432389]
g_16204	enpp1	ENSDARG00000005789	ectonucleotide pyrophosphatase/phosphodiesterase 1 [Source:ZFIN;Acc:ZDB-GENE-040724-172]
g_21326	mrps10	ENSDARG00000045913	mitochondrial ribosomal protein S10 [Source:ZFIN;Acc:ZDB-GENE-040914-39]
g_26350	NA	Not available	Not available
g_8191	cenpe	ENSDARG00000063385	centromere protein E [Source:ZFIN;Acc:ZDB-GENE-060929-860]
g_16070	yipf2	ENSDARG00000021399	Yip1 domain family, member 2 [Source:ZFIN;Acc:ZDB-GENE-040724-124]
g_14843	eef2kmt	ENSDARG00000054950	eukaryotic elongation factor 2 lysine methyltransferase [Source:ZFIN;Acc:ZDB-GENE-041010-160]
g_1225	stard3	ENSDARG00000017809	StAR-related lipid transfer (START) domain containing 3 [Source:ZFIN;Acc:ZDB-GENE-001120-2]
g_22480	qtrt1	ENSDARG00000043105	queuine tRNA-ribosyltransferase 1 [Source:ZFIN;Acc:ZDB-GENE-040426-1625]
g_18415	CU138547.1	ENSDARG00000074231	Not available
g_30618	mrps34	ENSDARG00000057910	mitochondrial ribosomal protein S34 [Source:ZFIN;Acc:ZDB-GENE-041114-71]
g_3276	NA	Not available	Not available
g_21252	uba5	ENSDARG00000063588	ubiquitin-like modifier activating enzyme 5 Source:ZFIN;Acc:ZDB-GENE-031112-2]
g_7374	spice1	ENSDARG00000004647	spindle and centriole associated protein 1 [Source:ZFIN;Acc:ZDB-GENE-041212-64]
g_5438	NA	Not available	Not available
g_2848	map7d2b	ENSDARG00000045316	MAP7 domain containing 2b [Source:ZFIN;Acc:ZDB-GENE-091118-82]

Table A.10 – Continued

g_7878	*NA*	*Not available*	*Not available*
g_27514	*pex2*	*ENSDARG00000062421*	*peroxisomal biogenesis factor 2 [Source:ZFIN;Acc:ZDB-GENE-070530-2]*
g_15022	*NA*	*Not available*	*Not available*

Table A.11 Two hundred and ten *"dN/dS"* candidates based on branch model and their ID, zebrafish orthologs Ensemble ID, chromosome location, as well as start and end position in the genome

Gene ID	Gene name	Ensemble ID	Chromosome	Start	End
g_17780	sdcbp2	ENSDARG00000012513	1	10850420	10860495
g_32331	pard6b	ENSDARG00000003865	1	11333532	11334389
g_2285	NA	Not available	1	18924231	18946748
g_30769	ccdc114	ENSDARG00000015010	1	22787740	22791981
g_25897	sypl2b	ENSDARG00000000690	1	26435099	26437851
g_16307	apobec2b	ENSDARG00000113992	1	29204950	29207283
g_9141	etv7	ENSDARG00000089434	1	29506653	29510107
g_13402	NA	Not available	1	33345573	33371674
g_12217	ptpdc1b	ENSDARG00000058873	1	33528889	33532670
g_4229	uhrf1bp1	ENSDARG00000077011	1	35308580	35329010
g_2079	plekhg5b	ENSDARG00000101752	1	39709750	39778580
g_27323	zbtb48	ENSDARG00000039263	1	39803839	39808353
g_13829	itga5	ENSDARG00000006353	1	43983863	44039491
g_4376	NA	Not available	1	56096087	56104045
g_25302	zgc:101731	ENSDARG00000040965	1	56543441	56545616
g_9073	si:ch211-137a8.4	ENSDARG00000078748	1	65278538	65285802
g_25237	NA	Not available	1	70538123	70556075
g_8223	NA	Not available	1	74466747	74469178
g_19278	yif1b	ENSDARG00000040505	1	78854505	78858653
g_27423	inppl1a	ENSDARG00000104222	1	79988652	80015111
g_23555	meis3	ENSDARG00000002795	1	87486037	87500772
g_10008	lpar5b	ENSDARG00000068638	1	88083911	88084774
g_10350	cntn2	ENSDARG00000000472	2	4964514	4994125
g_2381	atp2b4	ENSDARG00000044902	2	14694494	14762310
g_10766	fer1l4	ENSDARG00000076952	2	15224895	15243799
g_6097	skib	ENSDARG00000008034	2	29319011	29345663
g_5149	suclg2	ENSDARG00000044914	2	35337183	35438089
g_27671	NA	Not available	2	37190250	37192374
g_7401	camkvb	ENSDARG00000005141	2	37525144	37533428

Table A.11 - Continued

g_30422	*cntn4*	*ENSDARG00000098161*	2	39895656	39908914
g_15303	*si:dkey-156n14.3*	*ENSDARG00000052351*	2	42254106	42265426
g_3513	*cand2*	*ENSDARG00000005749*	2	43087080	43111262
g_30555	*NA*	*Not available*	2	49179758	49181828
g_11692	*tnnt3b*	*ENSDARG00000068457*	2	49607623	49621247
g_16386	*NA*	*Not available*	2	49903270	49910813
g_1586	*sox6*	*ENSDARG00000015536*	2	55280787	55366032
g_4370	*scamp2*	*ENSDARG00000010279*	2	56238476	56256919
g_8976	*sin3aa*	*ENSDARG00000079716*	2	56741396	56769842
g_7519	*snupn*	*ENSDARG00000008395*	2	56802864	56812552
g_23708	*sigirr*	*ENSDARG00000062204*	2	57934670	57938420
g_22086	*NA*	*Not available*	2	58933539	58988085
g_8854	*taf3*	*ENSDARG00000045513*	2	60637595	60648435
g_18423	*kcnc1b*	*ENSDARG00000032959*	2	62677850	62686447
g_22865	*lactb*	*ENSDARG00000040803*	2	64622087	64626973
g_30963	*kti12*	*ENSDARG00000054301*	2	64768740	64771404
g_19186	*cd9a*	*ENSDARG00000005842*	2	66148559	66151880
g_31300	*NA*	*Not available*	2	67231604	67234531
g_8751	*ush1c*	*ENSDARG00000051876*	2	71880220	71895084
g_5125	*bicd1a*	*ENSDARG00000079496*	2	74021591	74036260
g_3070	*aars1*	*ENSDARG00000069142*	2	76744135	76772171
g_5203	*cmip*	*ENSDARG00000062933*	2	87942332	87956656
g_31324	*NA*	*Not available*	2	89538725	89568493
g_21318	*ZNF276*	*ENSDARG00000110991*	2	92808730	92813531
g_5105	*bmper*	*ENSDARG00000101980*	3	7844495	7866533
g_10035	*dhdds*	*ENSDARG00000039851*	3	15579515	15592651
g_3152	*pbx2*	*ENSDARG00000019717*	3	17830772	17839338
g_22274	*si:dkey-17m8.1*	*ENSDARG00000079530*	3	19585371	19596148
g_13715	*NA*	*Not available*	3	21608250	21609353
g_14904	*tnxba*	*ENSDARG00000001760*	3	23757351	23763527

Table A.11 - Continued

g_10648	lyplal1	ENSDARG00000088764	3	37372350	37395181
g_8654	NA	Not available	3	46223361	46327721
g_10131	ppie	ENSDARG00000103234	3	66884052	66889053
g_17915	NA	Not available	3	69627632	69634352
g_5455	atg2b	ENSDARG00000097650	3	70636201	70645753
g_5016	fgfrl1a	ENSDARG00000032617	3	71555626	71580400
g_21129	prlh2r	ENSDARG00000054700	3	72051427	72053110
g_7392	TTC9	ENSDARG00000074363	3	83271545	83276842
g_27153	wdr32	ENSDARG00000029600	3	84543869	84554210
g_4655	numb	ENSDARG00000027279	3	84911273	84936706
g_22608	guca1g	ENSDARG00000045737	4	7853858	7862781
g_31014	brd7	ENSDARG00000008380	4	9648622	9665545
g_28942	chrna3	ENSDARG00000100991	4	11901577	11909512
g_20923	ppfibp2b	ENSDARG00000029168	4	17576476	17596603
g_5635	tead1b	ENSDARG00000059483	4	20534591	20551400
g_11371	rasa3	ENSDARG00000063371	4	24326722	24360978
g_1316	scml2	ENSDARG00000012949	4	26969500	27008775
g_91	lmnl3	ENSDARG00000007751	4	45499475	45519284
g_7582	dok4	ENSDARG00000073731	4	45722430	45736427
g_8116	il16	ENSDARG00000102908	4	50413232	50479596
g_22445	GTPBP8	ENSDARG00000075033	4	58149048	58170013
g_25957	znf142	ENSDARG00000061373	4	61307995	61315858
g_21492	lrrc3	ENSDARG00000078415	4	64747836	64748636
g_9290	pofut2	ENSDARG00000045175	4	75829214	75839711
g_7993	si:dkey-11f4.16	ENSDARG00000099799	4	79798285	79807148
g_23837	NA	Not available	4	89482765	89485852
g_2129	NA	Not available	4	89557528	89577540
g_9432	rftn2	ENSDARG00000056078	4	91531537	91557082
g_2217	efhc2	ENSDARG00000004204	4	92923188	92927120
g_3352	ifngr1	ENSDARG00000074771	5	7485001	7501858
g_17536	snrpb2	ENSDARG00000039424	5	9047893	9051543

Table A.11 - Continued

g_14774	*pex6*	*ENSDARG00000070958*	5	9797776	9812785
g_10167	*cyp2u1*	*ENSDARG00000026548*	5	12046258	12062220
g_34736	*casp6a*	*ENSDARG00000093405*	5	12189335	12195254
g_10757	*psip1a*	*ENSDARG00000104710*	5	12244066	12248282
g_17232	*socs1b*	*ENSDARG00000089873*	5	16228489	16230906
g_19811	*primpol*	*ENSDARG00000033273*	5	17622571	17626636
g_10706	*NA*	*Not available*	5	18672134	18673618
g_17667	*pdcd4b*	*ENSDARG00000041022*	5	26695980	26703345
g_13030	*slx4*	*ENSDARG00000061414*	5	27514960	27531111
g_9907	*smap1*	*ENSDARG00000031302*	5	29980123	29986905
g_2363	*cenpk*	*ENSDARG00000039616*	5	29987825	29992659
g_13036	*ctnnd1*	*ENSDARG00000078233*	5	36128885	36161345
g_9123	*aspm*	*ENSDARG00000103754*	5	70610342	70612485
g_15614	*NA*	*Not available*	5	71610136	71615985
g_19254	*NA*	*Not available*	5	72566748	72622332
g_17832	*adgrl4*	*ENSDARG00000013653*	5	80475319	80481571
g_2091	*cpox*	*ENSDARG00000062025*	5	81627975	81631977
g_13148	*zgc:153738*	*ENSDARG00000069230*	5	81697294	81709239
g_9828	*clocka*	*ENSDARG00000011703*	5	82137887	82146343
g_9295	*NA*	*Not available*	5	83834464	83841685
g_24492	*arhgap45b*	*ENSDARG00000062049*	5	83985652	84001769
g_15284	*hapln4*	*ENSDARG00000018542*	5	84128984	84132591
g_12923	*elovl8b*	*ENSDARG00000057365*	5	86644912	86648080
g_21735	*nsun4*	*ENSDARG00000021324*	6	19538741	19542819
g_1993	*pip5k1cb*	*ENSDARG00000100313*	6	21106970	21121252
g_22857	*aire*	*ENSDARG00000056784*	6	24364739	24369835
g_21232	*ccdc24*	*ENSDARG00000038793*	6	28382902	28393631
g_16610	*twsg1a*	*ENSDARG00000104244*	6	31550348	31565332
g_6311	*or101-1*	*ENSDARG00000013014*	6	44137912	44140817
g_30547	*NA*	*Not available*	6	45026389	45027305
g_16712	*NA*	*Not available*	6	46733231	46737499

Table A.11 - Continued

g_26253	angptl5	ENSDARG00000056630	6	53159220	53163517
g_3274	zgc:163098	ENSDARG00000078911	6	53579666	53586724
g_13108	ephb4a	ENSDARG00000100725	6	53850049	53883473
g_12533	txndc15	ENSDARG00000110357	6	59804857	59809310
g_4047	rimbp2	ENSDARG00000001154	6	66117501	66146539
g_26769	mtmr12	ENSDARG00000059817	6	66307020	66325910
g_19984	tmlhe	ENSDARG00000077547	6	67663287	67678842
g_10062	robo4	ENSDARG00000009387	6	76678843	76693473
g_26086	NA	Not available	8	2466424	2478151
g_11288	si:dkey-16i5.8	ENSDARG00000096722	8	6097161	6097625
g_6309	NA	Not available	8	6936716	6956678
g_13628	si:ch211-234p6.5	ENSDARG00000071460	8	10083374	10092673
g_11245	atxn2l	ENSDARG00000011597	8	10689139	10700891
g_12931	znf281b	ENSDARG00000035910	8	11201336	11202222
g_17473	NA	Not available	8	12794859	12797128
g_2739	tcf7l2	ENSDARG00000004415	8	13072377	13155703
g_4726	dlg5a	ENSDARG00000074059	8	13747591	13776701
g_22511	cd79b	ENSDARG00000104691	8	18336390	18340741
g_7163	plpp1a	ENSDARG00000053381	8	30277700	30302897
g_25033	dennd1a	ENSDARG00000014592	8	31397980	31436335
g_27606	NA	Not available	8	33765235	33769992
g_21651	entpd2a.1	ENSDARG00000035506	8	35797651	35801123
g_14578	trabd2a	ENSDARG00000089701	8	40528104	40530091
g_17466	NA	Not available	8	41220642	41241576
g_26258	surf2	ENSDARG00000112476	8	44428548	44430347
g_14187	NA	Not available	8	47942428	47946236
g_14166	ccdc62	ENSDARG00000111759	8	48529329	48539535
g_375	kyat1	ENSDARG00000023645	8	53909056	53914043
g_21419	uap1l1	ENSDARG00000013082	8	56349414	56387363
g_26113	dpp7	ENSDARG00000027750	12	3805337	3811669

Table A.11 - Continued

g_11926	*nos1*	*ENSDARG00000068910*	12	19136566	19137258
g_5422	*adamts12*	*ENSDARG00000067549*	12	23679470	23681078
g_24745	*agpat9l*	*ENSDARG00000006491*	12	23701878	23706828
g_28297	*snap29*	*ENSDARG00000038518*	12	23908812	23912125
g_19332	*SLC25A1*	*ENSDARG00000080000*	12	23934079	23944372
g_5559	*plcxd3*	*ENSDARG00000054794*	12	26597693	26605732
g_10250	*hnrnpk*	*ENSDARG00000018914*	12	28730314	28735425
g_13187	*ppp2r2aa*	*ENSDARG00000021996*	12	30717370	30726526
g_3793	*zgc:110626*	*ENSDARG00000053159*	12	35565969	35570274
g_3510	*riok2*	*ENSDARG00000035264*	12	36321334	36326705
g_19377	*aifm3*	*ENSDARG00000062780*	12	36619494	36629847
g_22260	*ela3l*	*ENSDARG00000007276*	12	39437387	39447531
g_25506	*rasgrp3*	*ENSDARG00000077864*	12	55499649	55519704
g_747	*ppm1ba*	*ENSDARG00000001888*	12	79330684	79337424
g_26946	*NA*	*Not available*	12	83565382	83566419
g_28147	*cryzl1*	*ENSDARG00000026902*	12	85118461	85128840
g_28922	*NA*	*Not available*	13	1598685	1618679
g_1207	*gdi2*	*ENSDARG00000005451*	13	4268921	4283154
g_24544	*NA*	*Not available*	13	7640927	7665198
g_8434	*cnot4b*	*ENSDARG00000007639*	13	11651318	11664248
g_30839	*NA*	*Not available*	13	14482715	14497072
g_28729	*myf5*	*ENSDARG00000007277*	13	15408597	15410848
g_24117	*napepld*	*ENSDARG00000009252*	13	20458318	20468874
g_14849	*NA*	*Not available*	13	35284736	35286982
g_1560	*slc9a3.1*	*ENSDARG00000058498*	13	38130114	38161467
g_20396	*NA*	*Not available*	13	38505550	38519887
g_27497	*CABZ01101996.1*	*ENSDARG00000109996*	13	39806799	39809089
g_10668	*thrap3b*	*ENSDARG00000098228*	13	43864750	43874170
g_25418	*NA*	*Not available*	13	45765039	45768320
g_12375	*dync1li1*	*ENSDARG00000098317*	13	52737678	52739650
g_8949	*calcr*	*ENSDARG00000028845*	13	53459188	53487418

Table A.11 - Continued

g_17229	*nsun2*	*ENSDARG00000056665*	13	54876791	54889966
g_14602	*cited4b*	*ENSDARG00000101009*	13	59480386	59481153
g_18822	*si:dkey-106g10.7*	*ENSDARG00000088036*	14	1337033	1340119
g_4306	*spata6l*	*ENSDARG00000004874*	14	2581558	2591619
g_22025	*ino80b*	*ENSDARG00000062749*	14	17097273	17116148
g_6247	*si:cabz01074946.1*	*ENSDARG00000090396*	14	17715666	17721096
g_7328	*b4galt7*	*ENSDARG00000021899*	14	35311231	35317558
g_9009	*sec24b*	*ENSDARG00000071906*	14	36029910	36075187
g_19800	*NA*	*Not available*	14	53565591	53569187
g_74	*ift172*	*ENSDARG00000041870*	15	6000119	6059193
g_16204	*enpp1*	*ENSDARG00000005789*	15	11952092	11982611
g_21326	*mrps10*	*ENSDARG00000045913*	15	14098270	14100158
g_26350	*NA*	*Not available*	15	16582193	16587535
g_8191	*cenpe*	*ENSDARG00000063385*	15	32036278	32140881
g_16070	*yipf2*	*ENSDARG00000021399*	15	49172590	49187538
g_14843	*eef2kmt*	*ENSDARG00000054950*	15	56143356	56147905
g_1225	*stard3*	*ENSDARG00000017809*	15	59586088	59594914
g_22480	*qtrt1*	*ENSDARG00000043105*	15	63505685	63508963
g_18415	*CU138547.1*	*ENSDARG00000074231*	15	68092374	68122951
g_30618	*mrps34*	*ENSDARG00000057910*	15	69759746	69761761
g_3276	*NA*	*Not available*	15	70375255	70378745
g_21252	*uba5*	*ENSDARG00000063588*	21	3239319	3242261
g_7374	*spice1*	*ENSDARG00000004647*	21	5025445	5031487
g_5438	*NA*	*Not available*	21	5140522	5145162
g_2848	*map7d2b*	*ENSDARG00000045316*	21	7978884	7997723
g_7878	*NA*	*Not available*	21	12327126	12328926
g_27514	*pex2*	*ENSDARG00000062421*	21	21062173	21068580
g_15022	*NA*	*Not available*	24	2780656	2813844

Table A.12 Thirty-one *"dN/dS"* candidates obtained from both branch and branch-site model-based analyses, and their IDs, names, location in chromosomes, as well as start and end position in the genome

Gene ID	Gene name	Chromosome	Start	End
g_2285	*NA*	1	18924231	18946748
g_16307	*apobec2b*	1	29204950	29207283
g_8223	*NA*	1	74466747	74469178
g_19278	*yif1b*	1	78854505	78858653
g_27671	*NA*	2	37190250	37192374
g_16386	*NA*	2	49903270	49910813
g_18423	*kcnc1b*	2	62677850	62686447
g_21318	*ZNF276*	2	92808730	92813531
g_5105	*bmper*	3	7844495	7866533
g_10648	*lyplal1*	3	37372350	37395181
g_7392	*TTC9*	3	83271545	83276842
g_27153	*wdr32*	3	84543869	84554210
g_8116	*il16*	4	50413232	50479596
g_17536	*snrpb2*	5	9047893	9051543
g_17232	*socs1b*	5	16228489	16230906
g_13030	*slx4*	5	27514960	27531111
g_2363	*cenpk*	5	29987825	29992659
g_15614	*NA*	5	71610136	71615985
g_26086	*NA*	8	2466424	2478151
g_21651	*entpd2a.1*	8	35797651	35801123
g_26258	*surf2*	8	44428548	44430347
g_14166	*ccdc62*	8	48529329	48539535
g_375	*kyat1*	8	53909056	53914043
g_5422	*adamts12*	12	23679470	23681078
g_10250	*hnrnpk*	12	28730314	28735425

Table A.12 – Continued

g_3793	*zgc:110626*	12	35565969	35570274
g_3510	*riok2*	12	36321334	36326705
g_28922	*NA*	13	1598685	1618679
g_4306	*spata6l*	14	2581558	2591619
g_16204	*enpp1*	15	11952092	11982611
g_27514	*pex2*	21	21062173	21068580