Transposable Elements in the Salmonid Genome

by

David Richard Minkley
B.Sc., University of Victoria, 2011

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Biology

© David Richard Minkley, 2018
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.

Supervisory Committee

Transposable Elements in the Salmonid Genome

by

David Richard Minkley
B.Sc., University of Victoria, 2011

Supervisory Committee

Dr. Ben F. Koop (Department of Biology)
Supervisor

Dr. Jürgen Ehlting (Department of Biology)
Departmental Member

Dr. John Taylor (Department of Biology)
Departmental Member

Abstract

Supervisory Committee
Dr. Ben F. Koop (Department of Biology)
Supervisor

Dr. Jürgen Ehlting (Department of Biology)
Departmental Member

Dr. John Taylor (Department of Biology)
Departmental Member

Salmonids are a diverse group of fishes whose common ancestor experienced an evolutionarily important whole genome duplication (WGD) event approximately 90 MYA. This event has shaped the evolutionary trajectory of salmonids, and may have contributed to a proliferation of the repeated DNA sequences known as transposable elements (TEs). In this work I characterized repeated DNA in five salmonid genomes. I found that over half of the DNA within each of these genomes was derived from repeats, a value which is amongst the highest of all vertebrates. I investigated repeats of the most abundant TE superfamily, Tc1-Mariner, and found that large proliferative bursts of this element occurred shortly after the WGD and continued during salmonid speciation, where they have produced dramatic differences in TE content among extant salmonid lineages. This work provides important resources for future studies of salmonids, and advances the understanding of two important evolutionary forces: TEs and WGDs.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BLAST	Basic local alignment search tool
cDNA	Complimentary DNA
DSB	Double-stranded break
ERV	Endogenous retrovirus
HC	High confidence
HR	Homologous recombination
HSP	High-scoring segment pair
HT	Horizontal transfer
HTT	Horizontal transposon transfer
ICSASG	International Consortium to Sequence the Atlantic Salmon Genome
LARD	Large retrotransposon derivative
LC	Low confidence
LINE	Long interspersed nuclear element
lncRNA	Long noncoding RNA
LORe	Lineage-specific ohnolog resolution
LTR	Long terminal repeat
MITE	Miniature inverted-repeat transposable element
MYA	Million years ago
NHEJ	Non-homologous end joining
NJ	Neighbour-joining
ORF	Open reading frame
piRNA	PIWI-interacting RNA
RBH	Reciprocal best hit
RNAi	RNA interference
rRNA	Ribosomal RNA
RT	Reverse transcriptase
SDR	Split direct repeats
SINE	Short interspersed nuclear element
TCE	Tc1-Mariner-like element
TE	Transposable element
TIR	Terminal inverted repeat
TRIM	Terminal repeats in miniature
tRNA	Transfer RNA
TSD	Target site duplication
UTR	Untranslated region
WGD	Whole genome duplication

Acknowledgments

An enormous number of people have helped me over the last five years, without the support of whom this thesis would have never been finished. My most sincere gratitude and thanks to…

… my supervisor, Dr. Ben Koop, for giving me incredible opportunities, for being compassionate and understanding, for his guidance and insight, for keeping his door always open, and for supporting me in my development as a scientist.

… my lab-mates past and present. Thanks Katy, Hollie, Kris, Eric, Cody, Stuart, Jong, Laura, Eric, Amber, Marj, Johanna, Kim, Nathan, Graeme, Jordan, Amy, Ben, Steph, Kris von S. You're all wonderful and have made the lab great!

… my fellow grad students – you've let me know that I'm not alone in this craziness.

… the entire biology community at UVic, but especially my committee members John Taylor and Jürgen Ehlting, as well as Steve Perlman and Michelle Chen, who helped me keep my head above the water.

… the staff, researchers and students of the Bamfield Marine Sciences Centre, for inspiring and centering me.

… Roger Aubin and the crew of Annie, for challenge and comradery whether the seas were stormy or calm.

… my friends, who have supported me in too many ways to count.

… my funders and institutional supporters. Thank you Compute Canada, NSERC, The Province of British Columbia, The Government of Canada, and the University of Victoria. Without you science just doesn't happen.

… my family. You have lifted me up more times than I can count, loved me all along the way, and inspired in me my love of life and science. Mom, Dad, Michael and John – I couldn't have done it without you.

Dedication

For my loving family.

Chapter 1 - Introduction

1.1 Thesis overview

In this work I will examine the repeated DNA sequences known as transposable elements (TEs) in five salmonid species and describe historical patterns of the proliferation of one TE group – Tc1-Mariner elements. In Chapter I establish the economic and cultural importance of salmon, and introduce an important evolutionary event – a whole genome duplication (WGD) – which occurred in the common ancestor of all salmonids and which has contributed to the evolution of this lineage over the past 90 million years. I describe WGDs, their importance as an evolutionary event, the resulting state of polyploidy, and their potential effects on the development of new traits and lineage diversification. Further, I summarize the ways in which the WGD is thought to have influenced the evolutionary trajectory of salmon. In addition to WGDs, I describe another evolutionary force that has shaped salmonid genomes – TEs. TEs are repeated DNA sequences which occupy significant portions of the genomes of most eukaryote species. I outline the many different TE taxa, discuss the methods by which they are regulated in a host cell, and describe the multitude of ways in which TE-derived sequences can facilitate genomic change and evolutionary novelty. In a final section I introduce the field of computational biology and discuss the many ways in which its techniques are being applied to investigate biological questions.

In my second chapter, I review the importance of annotating TEs within the genome in order to facilitate both the study of TEs themselves as well as other biological tasks such as comparative genome research or genome assembly. I then outline the methods and challenges of constructing a database of TE sequences with which genome annotation can be performed and describe in detail the methodology I have used to construct such a database for each of my five salmonid species of interest: Atlantic salmon (*Salmo salar*), arctic char (*Salvelinus alpinus*), rainbow trout (*Oncorhynchus mykiss*), coho salmon (*O. kisutch*) and chinook salmon (*O. tshawytscha*). Using these databases, I identify repeat-derived DNA within salmonid genomes and determine that at least ~55% of each genome I interrogated is composed of such sequence. I describe patterns in the abundances of individual TE superfamilies which inhabit salmonid genomes and compare these findings

with other vertebrates, which generally exhibit markedly less repeat-derived DNA. Finally, I propose that the increase in repeat-derived DNA is a result of the WGD and closely reflects a general vertebrate trend in which TEs are observed to occupy a larger portion of DNA in species with larger genomes.

In my third and final chapter I further investigate TEs of the Tc1-Mariner superfamily, which is by far the most abundant TE superfamily within the genomes of my five salmonid species of interest. I review relevant aspects of Tc1-Mariner biology, and then describe an intensive manual curation process through which I identified representative sequences for 60 distinct Tc1-Mariner families from Atlantic salmon and rainbow trout. Using these sequences I obtain TE copies from the genomes of all five of my salmonid species and, by comparing the sequence similarity between elements of the same family, I construct a timeline of historical Tc1-Mariner activity in the salmonid lineage. Further, I identify intron pairs from gene duplicates which originated at the salmonid-specific WGD, and use the sequence divergence between these duplicates to estimate when the WGD occurred in relation to historical Tc1-Mariner activity. With some caveats, this timeline reveals a series of bursts of TE proliferation which occurred coincidentally with or shortly after the salmonid-specific WGD, and which has continued throughout salmonid speciation to the present day. Further, certain TE families have been much more active in some salmonid lineages than others. Finally, I discuss the ways in which this prolonged intense TE activity could have been facilitated by the WGD and its aftermath, and investigate the potential impacts of massive lineage-specific TE proliferation on the evolutionary trajectories of salmonid species.

The work I outline in this thesis represents a significant step forward in the contemporary understanding of salmonid genome biology, as well as useful insights into the relationship between TEs and WGDs. The resources developed over the course of these projects, particularly five salmonid repeat libraries, provide essential tools for further research in areas such as genome assembly, genetic marker discovery, and comparative genomics. By describing the genetic elements that make up more than half of modern salmonid genomes, my work advances the current understanding of evolution, and provides an important stepping stone into developing further understanding of the economically, culturally and scientifically important salmonid group.

1.2 Salmonids

The salmonids are a diverse group of bony fish consisting of at least 70 species that make up the monophyletic group Salmonidae, the only family within the order Salmoniformes (Nelson et al. 2016). Since a split from the ancestor of their sister taxa, Esociformes (mudminnows and pikes) 100-130 million years ago (MYA), salmonids have diverged into three major clades (Coregoninae, Thymallinae and Salmoninae) which are in turn composed of fish from 11 genera including greyling, ciscoes, whitefish, trout, char and salmon (Near et al. 2012, Betancur-R et al. 2013, Nelson et al. 2016). Salmonid species are natively found around the world in the fresh and salt waters of the Northern hemisphere, and have also been successfully introduced in many other regions around the world (Berra 1981). Individuals from the Salmoninae subfamily, which includes the majority of intensively-researched salmonid species, are the focus of this work and are indicated in Figure 1.

Salmon possess a great variety of morphological and life history traits. Principle among these is anadromy, which sees salmon born in freshwater, migrate to the ocean

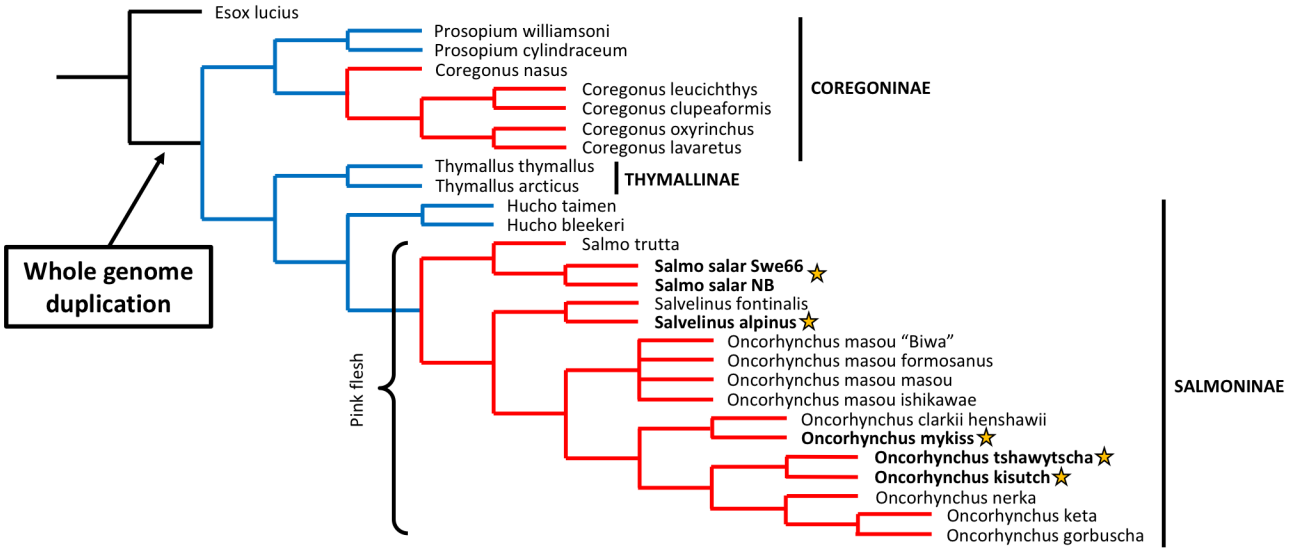

Figure 1 Salmonid taxa and the WGD based on Davidson, 2013. Blue and red branches correspond to strictly freshwater and anadramous life history strategies, respectively. Stars indicate species which are the focus of the present work: Atlantic salmon (*Salmo salar*), arctic char (*Salvelinus alpinus*), rainbow trout (*Oncorhynchus mykiss*), coho (*Oncorhynchus kisutch*) and chinook (*Oncorhynchus tshawytscha*)

(where they remain for the majority of their adult life), and finally swim back to freshwater to spawn. Anadromy, which requires substantial morphological, osmoregulatory and immunological plasticity during the transition between fresh and saltwater (Björnsson et al. 2011) has developed independently in at least two salmonid lineages and been lost many times, even within different populations of a single species (Berg 1985, Davidson 2013). Salmonids also differ in the number of times they spawn; iteroparous species spawn many times, while semelparous species spawn only once before they die. Some salmonids, such as arctic char, evince physiological adaptations to freezing temperatures that have seen them called the most cold-adapted freshwater fish (Budy and Luecke 2014, Vindas et al. 2017), while all species belonging to the *Oncorhynchus*, *Salvelinus*, *Salmo* and *Parahucho* genera possess a charismatic red-pigmented flesh which is the result of carotenoid sequestration within body tissues. Because many carotenoids possess antioxidant properties, this sequestration may have first arisen as a response to damage caused by reactive oxygen species that result from physical deterioration during extended migrations and mating (Rajasingh et al. 2007).

1.3 Culturally and economically important species

Along with their diverse physical and life history traits, the cultural and economic importance of salmonids has led to them being perhaps the most intensively-researched fish group of the last half-century; the past 60 years has seen the publication of well over 70,000 scientific reports on salmonids (Davidson et al. 2010). Salmonids have long been of great cultural relevance, and have captured the imaginations of human societies across the world for thousands of years. Testament to this is the distinct relief carving of an Atlantic salmon found in a cave near the Vézère River in France, created by humans over 22,000 years ago. Salmon have been similarly important in the Americas, where their presence has helped to facilitate the establishment of permanent human settlements over the past 7,000 years and elevated them to a prominent position in the art and culture of many Pacific-coast indigenous societies (Cannon and Yang 2006). Today, the importance of salmon in some cultures is perhaps nowhere better exemplified than in the furious debate over the conservation of salmonids in North America's Pacific Northwest. The declines of the titanic Pacific salmon spawning runs and the potential involvement of Atlantic salmon aquaculture and other human activity in this trend has given rise to

protest, extensive media coverage, intensive research and a commission by the Government of Canada (Commission of Inquiry into the Decline of Sockeye Salmon in the Fraser River (Canada) and Cohen 2012)

The economic importance of salmonid species is substantial. Globally, the combined share of salmonid fisheries and aquaculture has increased over recent decades, and in 2013 made up 16.6% of world trade by value (FAO 2016). This increase is in large part driven by the growing demand for products such as farmed Atlantic salmon from the middle class of both developed and emerging economies. In Norway, the world's largest exporter of salmonid products, the value of sold farmed salmonids was 63.3 billion NOK (~10.2 billion CAD) in 2015-2016 (Statistics Norway 2017). In British Columbia, Canada, farmed Atlantic salmon were the province's largest agrifoods export in 2016, totalling $524.2 million CAD (British Columbia Ministry of Agriculture 2016), while a recent report determined that from 2012-2015 Canadian commercial and recreational salmon fisheries had a combined output of $1.4 billion USD and produced 12,400 full-time equivalent jobs for the national economy (Gislason et al. 2017). Through aquaculture, tourism, and both commercial and recreational fisheries, salmon species are important contributors to the economy at the local and global level.

1.4 The salmonid genome and WGD

The evolutionary history and character of salmonid genomes is complex. The most prominent event in the evolution of the salmonid genome was a WGD which took place approximately ~90 MYA (Allendorf and Thorgaard 1984, Berthelot et al. 2014, Macqueen and Johnston 2014) in the progenitor of all salmonids, shortly after its divergence from the ancestor of the Esociformes. WGDs are monumental mutations that occur when all of the chromosomes within a genome are doubled, and result in individuals that are polyploid (have more than two sets of chromosomes within adult somatic cells). Two rounds of WGD (1R and 2R) are believed to have occurred in the common ancestor of all vertebrates, and contributed to the complexity and evolutionary success of the lineage (Dehal and Boore 2005, Smith et al. 2013). A third WGD (3R) also occurred in the ancestor of all teleost fish, a group whose 26,840 members make up half of all vertebrate species and includes salmon (Amores et al. 1998, Taylor et al. 2001a, Jaillon et al. 2004). Including the most recent salmonid-specific WGD (4R), the

genomes of modern salmonids are the products of at least four of these significant evolutionary events.

Since the 4R WGD, salmon have been reverting from their post-WGD tetraploid state back to one of diploidy (Ohno 1970, Wright et al. 1983, Allendorf and Thorgaard 1984). This process, rediploidization, is common in the aftermath of a WGD and proceeds over time as duplicate chromosome pairs (homeologs; also refers to duplicate gene pairs resulting from a WGD) accumulate mutations and diverge from each other to such an extent that they no longer pair with each other during meiosis (Wolfe 2001). This differs from polyploid meiotic pairing, in which all four corresponding chromosomes (two homologous chromosomes for each homeolog in a pair) can generally pair with each other in a bivalent fashion, or in some cases combine with each other to form single tetravalent structures (Otto and Whitton 2000). Extant salmon are intriguing because they are only part of the way through the process of rediploidization; the majority of chromosome sequences pair in a diploid fashion but some sequences, particularly in the sub-telomeric regions near the ends of chromosomes, exhibit tetraploid meiotic character. For this reason, salmon are termed 'pseudo-tetraploid', and possess both genomic loci that have a maximum of two alleles, and loci that effectively have four.

Rediploidization has occurred differently in different salmonid lineages and has contributed to the wide array of karyotypes present in this group. The ancestral karyotype of teleosts most likely consisted of either 48 or 50 chromosomes in somatic cells (Mank and Avise 2006), a number which has remained remarkably consistent in the majority of daughter lineages. The diploid ancestor of salmon is similarly believed to have possessed 50 chromosomes in somatic cells (Phillips et al. 2009, Lien et al. 2016). In the time since the 4R WGD, this number has varied widely, however; extant salmonids exhibit diploid chromosome numbers between 52 and 102 (Phillips and Ráb 2001). Even more remarkably, karyotypes vary even within some salmonid species. Different Atlantic salmon populations, for example, have been identified with between 54 and 58 chromosomes (Phillips and Ráb 2001). Varying paths of rediploidization, involving different large-scale chromosomal events such as fissions and fusions have probably contributed to reproductive isolation between nascent salmonid lineages and the resultant variation in karyotypes therein (Lien et al. 2016, Robertson et al. 2017).

1.5 Polyploidy

A WGD can result from two types of polyploidy: allopolyploidy and autopolyploidy. Allopolyploidy describes the scenario in which half of the chromosomes in a duplicated genome originate from each of two different but closely-related species. Allopolyploidization thus occurs as the result of hybridization between divergent genomes, and results in homeolog pairs which differ to some extent at the nucleotide level. Because of these divergent homeologs, the chromosomes of allopolyploid species for the most part pair only bivalently during meiosis, and individuals with recently-duplicated genomes are likely to possess more than two alleles for a subset of loci immediately following their WGD event. By contrast, autopolyploidy occurs within a single species following the duplication of one chromosome set. As a result, homeologs in autopolyploid individuals are nearly or completely identical upon duplication, facilitating both bivalent and tetravalent pairing between homeologs and their homologous chromosomes. Both auto- and allo-polyploidy can result from a number of mechanisms. These include nonreduction, in which unreduced diploid gametes are produced (very rarely) by both parents during gametogenesis and result in a tetraploid embryo following fertilization, and errors in cell division such as when an early germ cell undergoes DNA replication but does not subsequently divide (Ramsey and Schemske 1998, Van de Peer et al. 2017). The salmonid WGD is currently believed to have been the result of an autopolyploidization event (Wright et al. 1983, Allendorf and Thorgaard 1984, Hartley 1987), as are other ancient events such as the 1R and 2R vertebrate WGDs (Furlong and Holland 2002).

Polyploidy occurs in species across the entire eukaryotic domain, with notable examples in plants, vertebrates and fungi (Albertin and Marullo 2012, Van de Peer et al. 2017). It is much more common in plants than in animals; many angiosperm lineages, for example, exhibit evidence of both recent and ancestral WGDs (Cui et al. 2006, Soltis et al. 2008) while in animals it is comparatively rare (Otto and Whitton 2000). Interestingly, among vertebrates polyploidy is notably more common in amphibian and fish lineages, perhaps because many of their constituent species do not regulate their internal temperature and are susceptible to polyploidy-inducing temperature shocks - traits that they share with plants (Mable et al. 2011). Fish groups evincing polyploid

character are predominantly from less-derived lineages such as Acipenseriformes, Siluriformes, Cypriniformes and Salmoniformes; there is, for example, only one documented genus containing polyploid species amongst the highly-derived Perciformes group, which is the most numerous order of vertebrates and contains over 10,000 species (Mable et al. 2011). As in plants, the vast majority of polyploid fish species show evidence of past hybridization and are thus considered to be allopolyploid. Salmonids are therefore in the small minority of autopolyploid species.

1.6 Costs and advantages of polyploidy

The vast majority of WGD events have been observed relatively recently in the evolutionary tree (Arrigo and Barker 2012, Soltis et al. 2015). This fact suggests that polyploidy is generally an evolutionary dead end, with the preponderance of ancient duplication events having occurred within lineages which subsequently became extinct. There are many immediate disadvantages associated with WGDs. The fitness of the triploid offspring of a nascent polyploid and a diploid of the same species is low, and a newly formed polyploid is often in direct competition with its diploid cousins (Otto 2007). Polyploid genomes, particularly those of allopolyploid species, can be notably unstable and are susceptible to disruptive changes in gene regulation as well as ectopic genome recombination events such as deletions and translocations (Gaeta and Chris Pires 2010, Song and Chen 2015). Furthermore, the presence of more than two alleles at genomic loci can mask the effects of deleterious mutations from natural selection, thereby allowing them to persist within a population and over time to reach higher frequencies than they would otherwise be able to achieve, which in an equilibrium state is expected to decrease the mean fitness of tetraploids compared to a similar diploid population (Otto 2007). Interestingly, polyploidy can also result in disruptive changes in regulatory processes, which can be detrimental to species which cannot tolerate regulatory divergence, and it can cause increases in nuclear volume that are often associated with increases in cell size. The body size increases that are associated with polyploidy in insects and plants are not observed in fish and amphibians, at least in part due to a decrease in the number of (larger) cells (Mable et al. 2011). Given the dearth of ancient WGDs that have been identified, in most cases these disadvantages and disruptions evidently outweigh any benefits provided by a duplicated genome.

Both neutral and selective evolutionary processes could contribute to the perpetuation of polyploid lineages in the few cases when they are successful. In the short term, a somewhat-contentious theory posits that the increased genetic variation present in polyploids allows them to adapt more rapidly to a broader range of ecological and environmental conditions (Van De Peer et al. 2009, Te Beest et al. 2012, Van de Peer et al. 2017). Such variation is hypothesized to grant increased robustness during periods of substantial environmental upheaval and stress, as well as the ability to exploit niches that would otherwise be unavailable to a polyploid species' diploid progenitors. A newly-formed polyploid population may also gain a brief respite from deleterious mutations, as they are less likely to occur in a homozygous fashion and, in the case of loss-of-function mutations, can have their negative effects at least somewhat ameliorated by the presence of multiple functional alleles (Otto and Whitton 2000). This observation implies an initial relief from inbreeding depression, and so may also help polyploid populations persist following their initial severe bottleneck.

Provided that polyploids are able to persist long enough to overcome their initially severe challenges, longer-term evolutionary processes can take effect that provide the opportunity for substantial and novel changes. Initially following a WGD, one copy from many duplicate pairs can be lost (Lynch and Conery 2000, Brunet et al. 2006, Lien et al. 2016). Those homeologs that retain both copies are often sensitive to dosage changes; it is believed that for some pairs both duplicates must be retained in order to maintain specific stoichiometric relationships with the products of other genes which themselves have been duplicated (Schnable et al. 2011). Over time, homeolog pairs in which both copies have remained functional can begin to diverge, a process which offers the opportunity for novel functional developments. Retained gene duplicates have two primary fates: subfunctionalization and neofunctionalization (Ohno 1970, Force et al. 1999). In cases of subfunctionalization, any tasks which were performed by the single ancestor of an homeologous pair are partitioned between the two duplicates. This outcome is hypothesized to allow for each duplicate to further refine their roles, where previously this may not have been possible due to pleiotropic constraints. Alternatively, neofunctionalization results in one copy retaining the ancestral role of the progenitor gene, while the other copy is freed to develop novel functions. When these processes

occur on a genome-wide scale as in the aftermath of a WGD, they offer substantial opportunities for evolutionary novelty and the potential for increasing complexity. Indeed, the functional categories of gene duplicate pairs that are disproportionately retained following WGD events in animals are typically associated with signalling, development, transcriptional regulation and form (Van de Peer et al. 2017). Theoretical models in yeast have indicated that in the aftermath of a WGD the concerted increase in dosage of an entire pathway's gene complement can lead to increases in fitness that would not occur were genes individually duplicated, while experimental work in the polyploid plant *Arabidopsis thaliana* has identified cases where groups of WGD-duplicated genes have diverged in concert to form distinct networks (Blanc and Wolfe 2004, Van Hoek and Hogeweg 2009, De Smet and Van de Peer 2012). By duplicating every gene in the genome, WGDs provide evolution with ample raw material and the crucial opportunity to innovate.

1.7 WGD and lineage diversification

Polyploidy has long been speculated to play a role in speciation. Recent research has found however that in the shorter term, diploids form new species faster and go extinct more slowly than those populations that have undergone WGD. As a result, recent polyploids seem to have lower diversification rates than do diploids (Mayrose et al. 2011, 2015). Despite this observation, and given the extensive species radiations and development of regulatory novelty in many lineages that descend from an ancient WGD (such as vertebrates, teleosts and angiosperms), polyploidization is suspected to play a role in long-term speciation.

A framework that has been presented to explain longer-term species diversification in lineages that have experienced a WGD is called the 'radiation lag-time model' (Schranz et al. 2012). Under this model, lineage diversification is proposed to occur in only a subset of polyploid daughter lineages millions of years after a WGD. The WGD is hypothesized to imbue its descendants with an 'evolutionary potential' that can subsequently interact with lineage-specific ecological factors and promote diversification. The question of how this potential is maintained over periods that sometimes exceed one hundred million years has recently been addressed through the study of the salmonid WGD, which occurred long enough ago that rediploidization has begun to take place, but

recently enough that genetic signatures of evolution have not become overly obscured. In a report by Robertson et al. (2017), the authors note that a state of tetraploidy can be maintained by genetic recombination between homeologous chromosomes, preventing them from diverging enough to revert to diploid pairing during meiosis. If this state of concerted evolution continues as speciation takes place, the eventual rediploidization of the genome can occur in dramatically different ways and at different times in each daughter lineage. Because rediploidization implies the divergence of homeologs and the associated development of novel gene function, each lineage is able to use the WGD's 'evolutionary potential' to adapt to their unique ecological conditions in their own way. This elements of this model, which is termed 'Lineage-specific Ohnologue Resolution' (LORe), are clearly in evidence in the evolutionary history of salmon, where different lineages have undergone rediploidization at different times as the homeologs in different genomic regions begin to diverge. The LORe model helps explain why the vast majority of salmonid speciation occurred long after the WGD and why it is correlated with a period of climactic cooling and strongly associated with the development of anadromy, a trait which may offer a selective advantage in modern (cooler) temperate latitudes. (Macqueen and Johnston 2014). Notably, the 'delayed rediploidization' that is required for LORe generally only occurs in cases of autopolyploidy; allopolyploids have two distinct subgenomes (one from each parent species) and so in most cases instantly accomplish rediploidization upon their formation.

1.8 Repeats in the salmonid genome

The complexity of the salmonid genome is not limited to the occurrence and aftereffects of a significant ancestral WGD - it also plays host to a diverse collection of repeated sequences known as TEs. Isolated sequences from a variety of TE taxa were first characterized in the genomes of numerous salmonid species during the late eighties and early nineties (Moir and Dixon 1988, Winkfein et al. 1988, Kido et al. 1991, Stuart et al. 1992, Goodier and Davidson 1993). Early on, it became clear that some TEs were notably abundant in many of these species (Goodier and Davidson 1994, Radice et al. 1994). This observation was exploited in order to identify the conserved sequence of a hyperactive salmonid TE family called Sleeping Beauty, which has subsequently been used as a vector in genetic modification experiments (Ivics et al. 1997, Ivics and Izsvák

2015). Further research into the nature of TEs in the salmonid genome has identified many additional types of these sequences, and has also suggested that bursts of TE duplication and proliferation within the genome were ongoing during the process of speciation (de Boer et al. 2007, Matveev and Okada 2009). The presence and abundance of TEs in the salmonid genome imply a potential role for these repeated sequences in the evolution of salmon.

1.9 An introduction to TEs

TEs are DNA sequences that facilitate their own change of position and/or duplication within a host cell's genome using a variety of TE-encoded proteins and signaling motifs. They are found in virtually all eukaryotes, and through a wide array of molecular mechanisms most TEs are capable not only of moving to new locations, but also of replicating themselves (Wicker et al. 2007). The ability of TEs to proliferate within a genome contributes to their natural selection as a discrete evolutionary entity separate from their host, and has led to their frequent characterization as 'genomic parasites'. TE insertions directly into protein-coding or regulatory regions of DNA can have both subtle and extreme effects on the survival and fitness of their hosts. Insertion directly into the exons of a gene, for example, will almost certainly cause a loss-of-function mutation; they are rarely observed presumably due to strong negative selection against their effects (Stewart et al. 2011). Integration into important non-coding regions of the genome such as promoters, enhancers or introns can also wreak havoc on regulatory processes within the cell through either direct disruption of important regulatory motifs or through the injection of TE sequence that itself may contain signals that alter the genetic neighborhood. Some TEs, for example, contain splicing signals that can change the intron/exon boundaries of surrounding genes, or transcription factor binding sites that encourage nearby gene expression or repression (Polak and Domany 2006, Solyom and Kazazian 2012). The mere presence of TEs can be sufficient to encourage major genomic recombination events including chromosome-level translocations and inversions (Lim and Simmons 1994, Hedges and Deininger 2007); such events can dramatically contribute to evolutionary processes such as speciation.

As a result of the disruptions they can cause, TEs must balance the need for survival through continued replication with that of minimizing their effect on a host. At the same

time, the host experiences pressure to suppress TE activity and in some cases to benefit from it. The resulting evolutionary arms race has left the genomes of most species littered with fragments of TE remnants that are no longer mobile, representing whole groups of once-functional elements that accumulated inactivating mutations faster than they could replicate. TEs frequently make up a large proportion of their host genome and in many species the majority of DNA is derived from this these elements. In humans, for example, estimates of the repeat-derived proportion of the genome (which is predominantly composed of TEs) differ depending on the repeat-identification process and range from 45% (Lander et al. 2001) to over two-thirds (de Koning et al. 2011). TEs are one of the most significant forces affecting the structure and function of the genome.

1.10 TE taxonomy

TEs, which vary in size from less than 80 bp to more than 25 Kbp, come in many forms and the mechanisms by which they achieve their dispersal within a genome are diverse. At the broadest level TEs are divided into Class I and Class II elements based on whether or not an RNA intermediate is required to facilitate their transposition (Wicker et al. 2007). Prototypical Class I and Class II mobilization mechanisms are outlined in Figure 2. All Class I elements rely on the creation of a RNA transcript which is processed by a TE-encoded reverse transcriptase (RT) enzyme. This enzyme generates a complimentary DNA (cDNA) molecule that is inserted into the genome. Class II TEs do not utilize a reverse-transcribed RNA transcript and use a mobile DNA intermediate instead. For some Class II TEs this intermediate takes the form of an excised element from the genome that is simply moved from one location to another, while in others a new TE molecule is generated directly through the use of a template DNA molecule. Further classification levels within the TE taxonomic hierarchy divide elements based on differences in their specific replication strategy, constituent protein sequences, structural characteristics and internal signaling motifs. TE taxonomic categories in the classification regime established by Wicker et al. (2007) are, in order from the least to the most specific: Class, Order, Superfamily, Family and Subfamily. Because a comprehensive TE taxonomy has only recently been established and categories have evolved dynamically over the past 50 years as new groups of elements are discovered,

there still remain cases (particularly in older literature) where these groupings are not strictly adhered to.

1.11 Autonomous vs non-autonomous elements

TEs can be either autonomous or non-autonomous. Autonomous TEs are prototypical elements that possess all of the characteristics of a given TE taxa and which do not rely on other TEs in order to replicate and integrate into a new locus. With the exception of certain non-TE host factors that some TEs require, autonomous elements encode all of the proteins and signaling motifs required for transposition. Non-autonomous elements are lacking some of these features and require assistance from proteins encoded by other TEs in order to replicate. If a non-autonomous member of a TE family has lost a critical protein through a random mutation, for example, a separate TE copy from the same family with a functional protein-coding gene may be able to generate a protein that can mobilize the non-autonomous element. Non-autonomous elements account for a significant amount of TE activity in many genomes (Eickbush and Malik 2002).

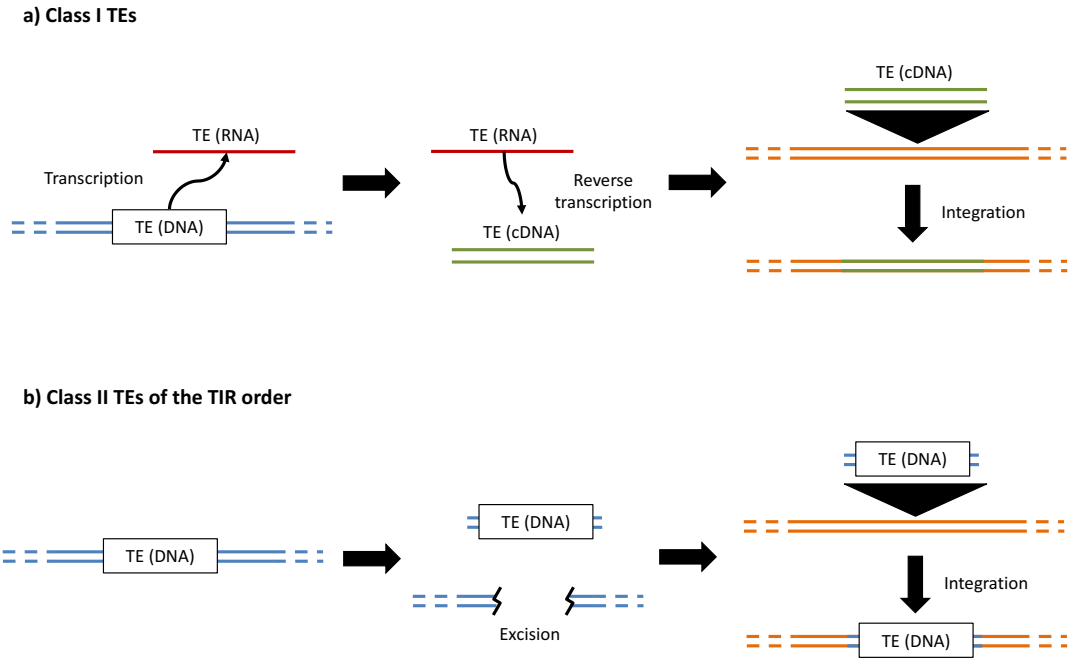

Figure 2 Transposition mechanisms of Class I TEs and Class II TEs of the TIR order. **a)** All Class I TEs replicate through the use of an RNA intermediate which is reverse-transcribed into cDNA and inserted into the target site. **b)** Class II TEs of the TIR order represent the preponderance of described Class II TEs, and mobilize through excision at one locus and reintegration at another.

1.12 Class I TEs

The Class I elements are known as retrotransposons and utilize an RNA intermediate for propagation. Following transcription of the entire TE to an RNA transcript, these elements rely on a TE-encoded RT enzyme to generate the cDNA which is subsequently inserted elsewhere within the genome. Because a new copy is generated during transposition and the original template TE remains intact, all Class I elements are colloquially described as being 'copy-and-paste' elements. With few exceptions, the most well-characterized retrotransposons are broadly divided into two major groups: long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons (which are also occasionally termed retroposons). More recently discovered and less well-studied groups include elements of the DIRS and Penelope orders.

LTR elements retrotranspose by a mechanism similar to that of retroviruses and are characterized by large tracts of nearly-identical sequence that flank the ends of each element and contain regulatory signals important for replication (the 'long terminal repeats'). Major LTR retrotransposon superfamiles include Copia, Gypsy, and Bel-Pao. The LTR order also includes retroviruses and their endogenous derivatives (endogenous retroviruses - ERVs), which are remnants of past retroviral infections that have lost the ability to produce a complete envelope protein and so can no longer escape the cell by conventional mechanisms. Although they are closely related to other LTR retrotransposons (indeed some vertebrate retroviruses are thought to be derived from ancestral LTR retrotransposons) and included in some major TE classification schemes, retroviruses encompass a field of their own within virology and are not further explored within this work (Deininger and Roy-Engel 2002, Wicker et al. 2007, Piégu et al. 2015).

As with many TEs, all LTR retrotransposons produce short duplications of the sequence present at their target insertion locus when they integrate into the genome. In LTR retroelements these target-site duplications (TSDs) vary in length from 4-6 bp and ultimately end up flanking the two ends of the newly-inserted element. LTR elements also possess certain characteristic open reading frames (ORFs) that encode proteins important for transposition; encoded proteins include the structural protein GAG, an aspartic proteinase, an RT enzyme, RNase H, and an integrase.

Non-autonomous derivatives of LTR retrotransposons are present in many genomes and are termed either large retrotransposon derivatives (LARDs), for elements that are longer than 4 kbp, or terminal repeats in miniature (TRIMs), for elements that are shorter than 4 kbp. These elements generally contain no ORF sequences that are reminiscent of the characteristic LTR retrotransposon genes, however they are still flanked by LTRs believed to contain the signals necessary for transposition. In many cases, the autonomous elements that are responsible for mobilizing LARDs and TRIMs are not known – like many TEs, evidence for their activity is only found in polymorphisms between individuals of a single species (Kalendar et al. 2004).

The non-LTR retrotransposon group includes both the autonomous long interspersed nuclear elements (LINEs) and the non-autonomous short interspersed nuclear elements (SINEs). Like LTR retrotransposons, LINEs contain a gene encoding an RT enzyme but unlike them they do not exhibit flanking LTR sequences. In addition to their RT gene, LINEs also code for an endonuclease. This endonuclease is responsible for creating a 'nick' in a target DNA strand, producing a free 3' DNA end that is used by a LINE RT enzyme to generate a cDNA in proximity to the target DNA sequence in a process called target-primed reverse transcription (Cost et al. 2002). This process is distinct from that used by LTR elements, in which the reverse transcription reaction itself takes place independently of the target DNA strand and generally outside of the nucleus within viral-like particles encoded by the LTR element GAG ORF (Finnegan 2012). At their 5' end LINEs possess an untranslated region (UTR) that contains a promoter sequence necessary for transcription, while their 3' tails are generally composed of single-base adenosine tracts (poly(A) tracts), tandem repeats or A-rich regions. Curiously, LINEs are frequently present in the genome with a random-length portion of their 5' end missing. This state of 5' truncation is believed to result from premature termination of reverse transcription and can often make it difficult to detect the variable-length TSDs produced by all LINE elements (Leeton and Smyth 1993, Eickbush and Malik 2002). Major LINE superfamilies include, but are not limited to: R2, RandI, L1, RTE, I and Jockey.

Whereas LINEs are generally at least a few thousand bases in length SINEs are much shorter, almost always less than 500 bp. In order to proliferate, SINEs rely on the reverse transcription machinery of one or more autonomous 'partner' LINE families. SINEs are

unique from most other TEs in that their 5' head contains a promoter sequence which encourages transcription by the RNA polymerase III (Pol III) enzyme instead of the more common RNA Pol II promoter used by most TEs and protein-coding genes (Okada 1991). Pol III-expressed genes generally encode short, innately-functional RNA transcripts such as 5S ribosomal RNA (5S ribosomal RNA – 5S rRNA), transfer RNA (tRNA) and signal recognition particle RNA (7SL RNA). SINE superfamiles are thus defined by the RNA type from which their Pol-III promoter sequence is derived; the major superfamilies are 5S, tRNA and 7SL. Like LINEs, SINEs may exhibit 3' tails that are A-rich or contain tandem repeats, and will occasionally resemble the 3' tail of the LINE partner which facilitates their mobilization. In many cases, however, the SINE 3' end will instead consist of a poly-thymine tract that serves as a Pol III termination signal and/or show no discernable similarity to any known LINEs. Because they rely on the LINE biomolecular machinery to integrate into a new genomic locus, SINEs produce variable-length TSDs.

Apart from the classic LTR and non-LTR retrotransposons, two groups of Class I TEs have recently been described that are sufficiently distinct for them to have been placed within their own orders: DIRS and Penelope elements. Like LTR retrotransposons, DIRS elements possess a GAG ORF, and genes encoding an aspartic proteinase, an RT enzyme and RNase H, however they differ in that instead of an endonuclease gene they encode a tyrosine recombinase (Cappello et al. 1985, Goodwin and Poulter 2004). This tyrosine recombinase is responsible for the integration of a DIRS cDNA molecule into a genomic target locus, and is distinct from other Class I elements because its mechanism of action produces no TSDs. DIRS elements are not flanked by LTRs and the ends are instead bounded by either terminal inverted repeats (TIRs), which occur where an element is flanked on one end by a sequence motif and on the other end by that motif's reverse complement, or by split direct repeats (SDRs), a structure in which a sequence occurs twice in tandem, with some amount of interleaving non-duplicated sequence. Elements from the final Class I order, Penelope, contain genes encoding RT and endonuclease proteins that are sufficiently distinct from those of other orders that these elements were placed within their own clade (Evgen'ev and Arkhipova 2005). They possess flanking

LTR-like sequences which can either be in a direct or inverse orientation, and produce variable-length TSDs.

1.13 Class II TEs

The most well-characterized group of Class II elements - DNA transposons - replicate through a 'cut-and-paste' mechanism in which a TE-encoded transposase protein recognizes a source TE sequence (DNA), excises it, and then integrates it elsewhere in the genome. Because they do not directly replicate their own sequence, DNA transposons increase their number by 'jumping' during DNA synthesis from a locus that was previously replicated during the normal course of DNA replication to one which has not yet been duplicated, or by taking advantage of DNA gap repair machinery that will occasionally replace an excised TE if an identical insertion is present at the same location on the homologous chromosome that guides repair (Feschotte and Pritham 2007).

With the exception of elements of the Crypton and Helitron superfamilies, DNA transposons utilizing a cut-and-paste mechanism are flanked by characteristic TIR sequences. These sequences define the boundaries for excision that are acted on by a transposase enzyme. For most Class II superfamilies, the transposase itself relies on a catalytic domain containing two aspartic acid residues followed at some point by either a glutamic acid or another aspartic acid residue, which as a result of these amino acid's single-letter IUPAC codes is either termed a DDE or DDD transposase (Wicker et al. 2007). TIR elements with this form of transposase include TEs of the superfamilies Tc1-Mariner, hAT, Mutator, Merlin, Transib, and PIF-Harbinger. Each of these superfamilies differs in the exact sequence of the DDE/DDD transposase, as well as in the specific motif sought for target-site integration. Class II TIR elements that contain an otherwise similar transposase that does not appear to utilize a DDE/DDD catalytic domain include P, PiggyBac and CACTA elements; the exact catalytic mechanism of these TE groups are poorly understood (Wicker et al. 2007). In addition to their transposase gene, elements of some superfamilies (PIF-Harbinger and CACTA) also include an additional ORF that encodes a protein of unclear function. As a by-product of insertion all of these elements exhibit varying-length TSDs. Distinct from other cut-and-paste transposons, elements of the Crypton superfamily do not possess TIRs and rely on a tyrosine recombinase instead of a transposase for target site integration (Goodwin et al. 2003).

Two types of DNA transposons exist which do not exclusively make use of a cut-and-paste mechanism: Helitron and Maverick elements. Based on in silico analyses that established certain similarities to catalytic motifs and replication initiators in plasmids and single-stranded DNA viruses, Helitrons have generally been considered to replicate using a 'copy-and-paste' rolling circle mechanism (Kapitonov and Jurka 2001), however recent research implies that these elements may also make use of an excision-based cut-and-paste strategy (Li and Dooner 2009, Borgognone et al. 2017). Helitrons possess no TIRs and do not produce TSDs upon insertion. TEs of the Maverick superfamily, also known as Polintons, are thought to replicate in a purely copy-and-paste fashion through the use of a self-encoded DNA polymerase. Elements of this superfamily, which possess TIRs and produce 6 bp TSDs, are particularly remarkable as they range in size from 15kbp to 40kbp and can encode up to 10 proteins including the aforementioned DNA polymerase, a retroviral-like integrase, a protease and a putative ATPase (Gao and Voytas 2005, Feschotte and Pritham 2005, Kapitonov and Jurka 2006, Pritham et al. 2007).

For TIR-flanked DNA transposons, non-autonomous elements take the form of an unspecified sequence that is still flanked by the TIRs characteristic for a given TE family. The internal sequence in such elements can consist of a transposase gene that has suffered a loss-of-function mutation or (potentially very short) random sequence that was shuffled between two TIRs through genomic recombination. In cases where these non-autonomous TEs have been so reduced in size that they are essentially composed only of two abutting TIRs, they are termed miniature inverted-repeat transposable elements (MITEs). As long as there remains a TE copy within the genome that encodes a functional transposase recognizing its TIRs, a non-autonomous DNA element can be excised and integrated elsewhere in the same way as an autonomous element.

1.14 Host defences against TEs

The mutagenic nature of TEs and the wide variety of ways in which they can negatively impact the proper function of host cellular processes (see Section 1.15 below) is balanced in eukaryotes by a number of defense mechanisms which reduce their activity and potential to do damage. These host defenses operate on both the epigenetic level, where they can regulate the transcriptional environment of TEs, as well as on the post-

transcriptional level. The primary defense against TE activity is provided by an RNA interference (RNAi) pathway that is centered on short RNA molecules known as PIWI-interacting RNA (piRNA). Primary piRNAs are generated from intergenic regions of the genome known as piRNA clusters, which attract TE insertions (they act as 'transposon traps') and over time grow to contain a diverse collection of TE fragments (Iwasaki et al. 2015). Long RNA molecules containing TE fragments are bidirectionally transcribed from piRNA clusters before they are cleaved in the cytoplasm into short active piRNAs. Once created, piRNAs are directly loaded into Argonaute protein effectors which enter the nucleus and, relying on the complementarity of their piRNA partner to TE motifs, methylate both histone proteins and DNA itself in a manner which silences TE activity (Haase 2016).

In addition to their direct silencing of TE expression, piRNAs can also be used by protein effectors to target other transcripts in the cytoplasm containing complementary TE motifs. In a process called 'ping-pong', these target transcripts are themselves cleaved and converted into secondary piRNA molecules, which like primary piRNAs can participate in the silencing of TE expression or guide an attack on TE transcripts (Brennecke et al. 2007). This process has the effect of refining the whole pool of defensive piRNA molecules towards active TE elements. The piRNA RNAi system acts as a form of intracellular immune system – it possesses both an 'innate' immune capacity, facilitated by the genetic memory of past TE insertions sequestered in piRNA clusters, and an 'adaptive' capacity, as the pool of piRNAs is refined to more specifically target active elements. In some cases, this defense system is so essential to maintaining genomic integrity that pools of piRNA will be built up in the female parent and transferred to her offspring, thereby conferring an increased level of defense to the germ cells that are critical to evolutionary success (Brennecke et al. 2008).

1.15 The disruptive effects of TEs

The vast majority of TE insertions have relatively minor deleterious effects on host fitness, the severity of which can vary between species (Lynch 2007). However, the effects that TEs exert on their resident genomes can be diverse and may contribute positively or negatively to the adaptation of their host species. The most conceptually straightforward way in which TEs affect their host is through insertion into a protein-

coding exon of a gene (Stewart et al. 2011). When this occurs, the resultant protein product is more often than not dysfunctional, given that the most prolific eukaryotic TEs are hundreds or thousands of bases long. This disruption can take the form of a frame shift or the addition of a multitude of new amino acids. It is also possible for a novel TE insertion to alter the splicing environment of a gene through the addition or disruption of splice sites themselves (Sorek et al. 2002). Through this mechanism, even insertions into non-coding introns or UTRs of a gene can affect its final protein product.

Beyond the modification of splicing regulation, TE insertions are also capable of significantly altering the general regulatory environment of a gene or entire region of the genome. This can be accomplished in a direct way by insertion and disruption of regulatory elements; TE insertions into both promoters and enhancers have been demonstrated to affect the expression of nearby and distant genes (Hollister and Gaut 2009). Apart from direct disruption, the observation that many TEs contain host-recognizable regulatory motifs evinces an intriguing situation in which novel regulatory elements can be distributed throughout the genome by TEs (Shankar et al. 2004, Polak and Domany 2006, Lynch et al. 2011). This situation, together with that in which trans-acting regulatory RNA or proteins are mutated to suddenly recognize a novel motif contained within a TE family spread throughout the genome, presents a very powerful evolutionary potential in which regulatory networks can be born *de novo* or extensively modified over very short periods of time (Feschotte 2008, Kunarso et al. 2010, Rebollo et al. 2011, Jacques et al. 2013).

The host suppression mechanisms that operate on TEs present another vector through which a TE insertion can affect the surrounding genomic region. Regions that are rich in TEs are heavily repressed through a number of epigenetic mechanisms including both DNA and histone methylation. In many cases, sequence within a particular TE family is actively sought out by epigenetic factors whose imprecision can result in the suppression of not only the TE but also of the surrounding DNA. In this way, a nearby TE can reduce the transcription of both protein-coding and non-protein coding transcripts (Hollister and Gaut 2009). When TEs insert into transcribed regions of the genome host suppression strategies operating on the RNA level can similarly have off-target effects. In these cases, RNA degradation pathways such as RNAi that are targeted to TE motifs may seek

out and destroy non-TE transcripts in which a TE insertion has occurred (Elbarbary et al. 2016).

TEs can also encourage structural changes in a host's DNA. The presence of a large number of highly similar sequences within a genome can wreak havoc with a variety of processes involved in the repair of DNA, as well as in normal chromosomal crossover (Konkel and Batzer 2010). This disruption is particularly prominent with the homologous recombination (HR) DNA double-stranded break (DSB) repair pathway, which in fixing a break relies on the existence of a 'template' molecule located at the same locus on the homologous chromosome. When multiple nearly-identical sequences are present in the region near the DSB (as is the case with recently-active TE families), the wrong chromosome locus can be selected to guide repair, resulting in non-allelic (ectopic) recombination events such as deletions, inversions, duplications and even inter-chromosomal translocations (Lim and Simmons 1994, Gray 2000, Hedges and Deininger 2007, Grabundzija et al. 2016). Interestingly, other DNA repair processes such as the non-homologous end joining (NHEJ) pathway which do not require a template homologous chromosome can also be discombobulated by the presence of TEs (Elliott et al. 2005). Beyond TE-induced recombination, gene duplications can be also directly facilitated by the biomolecular machinery encoded by Class I retrotransposon elements (Kaessmann et al. 2009). This latter method of duplication occurs when the transcribed mRNA of a non-TE gene acts as a substrate for TE-encoded RT enzyme, forming a cDNA that is then inserted randomly into the genome as if it were a retrotransposon. When compared to their homologs, genes that have undergone this process can often be identified by their missing 5' regions (a result of the 5' truncation that is characteristic of non-LTR retrotransposon insertions) or by the absence of introns (when reverse transcription and insertion occurs on previously-spliced mRNA).

1.16 TEs as an evolutionary toolkit

While TEs can be enormously disruptive to protein-coding genes, regulatory processes and genome architecture, they also contribute sequences that can be co-opted ('exapted') by the host for other useful purposes. There are many domains within TE proteins that provide utilities and functions that are also required by a multitude of non-TE proteins. These include DNA- and RNA-interaction domains, nuclear localization signals that

allow for a translated protein to be transported into the nucleus, dimerization domains, and protein-interaction domains (Feschotte 2008). Following insertion and given a sufficient amount of time, these TE components can be shuffled around the genome and end up providing novel capabilities to previously-existing non-TE proteins. Similarly, entire TE proteins can be modified over time and evolve to contribute to essential host functions, as is the case with the RAG1 and RAG2 proteins central to the vertebrate adaptive immune system and the telomerase-like activity of retrotransposon derivatives in Drosophila (Levis et al. 1993, Kapitonov and Jurka 2005).

The propensity for TEs to donate motifs to other host factors is not limited to protein-coding sequences. TEs have recently been recognized to contribute a substantial amount of sequence to the important long non-coding RNA (lncRNA) class of regulatory molecules (Kapusta et al. 2013). lncRNA transcripts are notably abundant and participate in processes as diverse as transcription regulation, mRNA processing, post-transcriptional control, protein regulation and higher-order RNA-protein complex formation (Geisler and Coller 2013). TE-derived sequence within these molecules has been repeatedly shown to be critical to their function (Elisaphenko et al. 2008, Gong and Maquat 2011, Carrieri et al. 2012). Because lncRNA sequence has a relatively short evolutionary turnover rate (functionally important lncRNAs are frequently present in only a single taxa and poorly conserved between major groups), ongoing novel TE insertions represent a potentially major substrate for the formation of new functional molecules (Kapusta and Feschotte 2014). Between their contributions to protein-coding genes, regulatory sequences and non-coding RNAs, TEs provide a diverse toolkit of functional components that serves as important fodder for the evolutionary process.

1.17 TE dynamics

The factors that affect the prevalence of TEs within a genome include traditional evolutionary determinants such as natural selection and drift, as well as more specific forces that are reflective of the unique relationship between selfish TEs and their hosts. TEs predominantly exhibit a negative correlation with recombination rate, possibly as a result of selection acting to eliminate higher incidences of TE-induced ectopic recombination in regions with higher recombination rates, or because of the increased power of selection in such areas (Kent et al. 2017). Genetic drift is hypothesized to play

a prominent role in the abundance of TEs between different populations and species, with greater drift implying a decreased ability for natural selection to remove deleterious TE insertions; this effect is driven in large part by changes in effective population size (Hua-Van et al. 2011, Szitenberg et al. 2016), and could also explain the positive correlation between genome size and TE abundance (Kidwell 2002, Lynch and Conery 2003, Lynch 2007, Touchon and Rocha 2007). The presence or absence of sex is also important in determining the success of TEs. Asexual populations will tend to be resilient to initial TE invasion because TE elements cannot spread to other genomes during zygote formation, and neutral and/or negative fitness effects would be expected to drive the initially-invaded genome to extinction. For similar reasons, a transition from a sexual to an asexual life history strategy in a species which already possesses TEs will over time result in TE reduction (Arkhipova 2005).

The effects of the previous factors are all in some part dependent on the selective disadvantages imposed by TE proliferation itself, which can be attenuated by the development of an insertion preference for intergenic regions or the adoption of lower transposition rates. The ability of the host to suppress TE activity can also affect TE abundance - changes in the RNAi system can alter TE dynamics, while the epigenetic disturbances that can result from events such as WGDs are also able to unharness TE activity (Obbard et al. 2009, Parisod and Senerchia 2012, Vicient and Casacuberta 2017). Ultimately, the proliferation and evolution of TEs is defined by a balance between a TE's own transposition and a variety of evolutionary forces.

Across all eukaryotes the abundance and diversity of TEs present within a species vary substantially, even between closely-related lineages. For vertebrates few general patterns exist, although larger and more deeply-branching clades tend to evince a greater variety of TE diversity and TE content (Sotero-Caio et al. 2017). Of all vertebrate taxa, Actinopterygiian fishes display the greatest amount of TE diversity with an average of 24 TE superfamiles per species, and a large variety in the abundance of TE-derived DNA within the genome, which varies from 6% in the small-genomed green spotted puffer (*Tetraodon nigroviridis*) to nearly 60% in some salmonids, as I describe in Chapter 2 (Aparicio et al. 2002, Volff et al. 2003, Lien et al. 2016). TE abundance, in fact, is thought to be the major determinant of genome size across Actinopterygii (Chalopin et al.

2015, Gao et al. 2016). A sample of fish genomes from across the vertebrate phylogeny are displayed along with their repeat content in Table 1. These fishes have also often been host to 'bursts' of TE amplification that introduce substantial diversity between related species and even between individual populations.

Such bursts have been associated with many different superfamilies and include those identified in merry widows (*Phallichtys amates*), platyfish (*Xiphophorus maculatus*), medaka (*Oryzias latipes*), zebrafish (*Danio rerio*) and salmonids (Volff et al. 2000, de Boer et al. 2007, Koga et al. 2009, Gao et al. 2016). The genome sequencing and analysis of further fish genomes is required to begin to fully understand TE dynamics within these species, however it is clear that TEs have had varied and important impacts on Actinopterygiian genomic landscapes.

Table 1 Genome size and repeat content in published vertebrate fish genomes Fish species are ordered from most basal to most derived according to the phylogeny of Near et al. 2012.

Common name	Latin name	Major groups	Genome size (Mbp)	Repeat content (%)	Reference
Sea Lamprey	*Petromyzon marinus*	Cyclostomata; Petromyzontiformes	1,130	~60*	Smith et al. 2018
Elephant shark	*Callorhinchus milii*	Chondrichthyes; Chimaeriformes	937	28.2	Venkatesh et al. 2014
Coelacanth	*Latimeria chalumnae*	Sarcopterygii; Coelacanthiformes	2,736	~60*	Nikaido et al. 2013
Spotted Gar	*Lepisosteus oculatus*	Holostei; Semionotiformes	945	20.1	Braasch et al. 2016
Electric fish	*Paramormyrops kingsleyae*	Osteoglossomorphs; Osteoglossiformes	880	26.0	Gallant et al. 2017
Zebrafish	*Danio rerio*	Ostariophysi; Cypriniformes	1,412	52.2	Howe et al. 2013
Common carp	*Cyprinus carpio*	Ostariophysi; Cypriniformes	1,690	31.2	Xu et al. 2014
Atlantic salmon	*Salmo salar*	Protacanthopterygii; Salmoniformes	2,966	59.9	Lien et al. 2016
Northern pike**	*Esox lucius*	Protacanthopterygii; Esociformes	904**	40.78**	Rondeau et al. 2014**
Atlantic cod	*Gadus morhua*	Neoteleost; Gadiformes	643	31.3	Tørresen et al. 2017
Nile tilapia	*Oreochromis niloticus*	Neoteleost; Percomorph; Cichliformes	1,009	29.5	Conte et al. 2017
Medaka	*Oryzias latipes*	Neoteleost; Percomorph; Beloniformes	764	17.5	Kasahara et al. 2007
Platyfish	*Xiphophorus maculatus*	Neoteleost; Percomorph; Cyprinodontiformes	669	~16*	Schartl et al. 2013
Stickleback	*Gasterosteus aculeatus*	Neoteleost; Percomorph; Perciformes	463	13.5	Jones et al. 2012, Xu et al. 2014
Tiger Puffer	*Takifugu rubripes*	Neoteleost; Percomorph; Tetraodontiformes	390	7.1	Aparicio et al. 2002, Xu et al. 2014
Green Spotted Puffer	*Tetraodon nigroviridis*	Neoteleost; Percomorph; Tetraodontiformes	342	5.7	Jaillon et al. 2004, Xu et al. 2014

* Approximations are the most specific information present in the corresponding publication

** Values for Northern pike correspond to as-yet unpublished work on the most recent genome assembly (NCBI RefSeq accession GCF_000721915.3), using a repeat library creation process similar to that described for salmon in Chapter 2.

1.18 The rise of computational biology

With the rise in power of both personal and high-performance computers over the past decade, the use of computing systems to address problems at all levels of biology has become commonplace. The simultaneous dramatic reduction in the price of high-throughput sequencing has allowed for the elucidation of genome and transcriptomic datasets for everything from single cells to entire evolutionary clades. Recent ambitious large-scale projects include BGI-Shenzhen's commitment to sequence and assemble genome sequences for every extant bird species by 2020 (Zhang 2015), the ENCODE project which sought to characterize all of the functional elements within the human genome (Dunham et al. 2012), an initiative to map all of the cells in the human body (Rozenblatt-Rosen et al. 2017), and multiple independent endeavours to sequence millions of human genomes (Cyranoski, 2016; Ledford, 2016; PMI Working Group, 2015). Beyond highly-publicized and well-funded collaborations, the decreasing cost of computational and sequencing platforms as well as the availability of publicly-funded resources such as the Genbank and EMBL-EBI databases puts these tools within reach of individual researchers (Benson et al. 2005, Kanz et al. 2005).

The analysis of the staggering amounts of data being generated by sequencing projects has driven the progression of computational biology and bioinformatics, two fields of study with often murky and overlapping definitions. These fields are concerned with the development and application of computational methods to study biological phenomena, and frequently involve hardware-tuning, algorithm design, software development, database management, statistical analysis and data visualization. Some applications of these fields are reference genome assembly, genome annotation, transcriptome assembly, epigenetic analysis, comparative genomics, genome-wide association studies, population genetic analyses, molecular sequence comparison, phylogeny work and, importantly, TE analysis. The applications of bioinformatics and computational biology are already enormous and still-growing, and help make incomprehensibly-large datasets manageable for human researchers.

1.19 Objectives and the analysis of TEs within the genomes of five salmonids

Over the course of this thesis I seek to address three principle questions: 1) What proportion of salmonid genomes are derived from repetitive DNA, including TEs? 2) What types of TEs are present in salmonid genomes, and what is their abundance? 3) How have TEs shaped Salmonid evolutionary history? In this work I use bioinformatic techniques to address these questions by first identifying libraries of TE sequences from the genomes of five salmonid species: Atlantic salmon, arctic char, rainbow trout, coho salmon, and chinook salmon. These databases will provide essential resources for the entire salmonid research community, as they help facilitate many downstream analyses that must account for the presence of confounding and highly-repetitive TEs. I investigate the patterns of abundance of these elements within their genomes, and speculate about the implications of these patterns for the current understanding of salmonid evolution, particularly in the context of the salmonid-specific WGD and rediploidization. For elements of one superfamily, Tc1-Mariner, I further elucidate historical patterns of transposition activity as they relate to the WGD and salmonid speciation, and identify differences in proliferation and copy number that occur between different salmonid lineages. By establishing a robust foundation of salmonid TE resources, my work creates valuable datasets for future researchers and provides insights into the relationship between two enormously transformative evolutionary forces: TEs and WGD.

Chapter 2 - Repeats in five salmonid genomes

2.1 Introduction

2.1.1 Building references resources

The burgeoning field of genomics is heavily reliant on the creation of high-quality reference resources. These resources include exemplary genome sequences as well as databases of genes, regulatory elements, TEs and other important features. The development of publicly accessible resources provides a common point of reference for all experiments that utilize them, and saves researchers the effort of having to develop their own frequently expensive tools. After a reference resource has been developed, it can be submitted to international repositories such as NCBI's Genbank, the UCSC Genome Browser, or EMBL-EBI (Kent et al. 2002, Benson et al. 2005, Kanz et al. 2005). Alternatively, it can be made available online through a privately- or institutionally-hosted web server. Once distributed, a resource such as a reference genome sequence can be combined with the annotation information or experimental investigations of many other researchers, thereby producing a central hub for a particular research subject. For example, there are currently 5,262 eukaryote genome sequences included in NCBI's Genomes database and 104 species listed in the UCSC genome browser, all of which include detailed annotation information (accessed Feb 27, 2018).

Reference resources form the basis of a large variety of research endeavours and encourage reproducibility. For example, analysis of orthologous protein sequences from a variety of species can allow for the identification of conserved and functionally important protein regions, which can in turn help inform an assessment of the impact of novel mutations of that protein within an individual (Bendl et al. 2014). Reference resources are similarly important for the annotation of novel genome sequences; existing gene, transcript, protein, and repeat databases are used extensively to identify orthologous or similar features in a new genome.

2.1.2 The importance of repeat libraries

A particularly critical type of reference resource is a repeat library. A repeat library is a set of sequences in which each sequence corresponds to a whole group (or family) of similar (repeated) sequences that are currently present, or were historically present,

within a genome. Once it has been developed, each constituent sequence can be used to identify its corresponding family members within a genome, and so annotate the repeat-derived genomic fraction. Because the vast majority of repeated DNA in most genomes is composed of TEs (rather than other other types of repeats such as satellites, tandem repeats or microsatellites), the reliable identification and classification of repeats within a genome is obviously crucial for any study of TE biology or of the impact of TEs on evolutionary processes, however their identification is also important because their presence in a genome can create complex problems for many analyses if they are not properly accounted for. The complications created by TEs are due to a combination of their frequently similar and highly repeated natures. These complications can manifest themselves over the course of a diverse array of experimental tasks. Two processes for which the presence of repeat sequences can be particularly problematic are described below: comparative genomic analysis and genome assembly.

2.1.3 The importance of repeat libraries: comparative genomics

Repeat-related issues frequently arise in rudimentary tasks when comparing the genomes of different species (Frith 2011, Platt et al. 2016). A first step in these analyses is often the identification of sequences that represent corresponding copies of the same ancestral sequence (ie orthologous DNA). Often, the same TE families are present and active in both of the compared species, however due to continued TE sequence proliferation following the species' divergence they may have inserted into different loci in the two lineages. In approaches where sequence similarity is used as a proxy for orthology, tools such as the Basic Local Alignment Search Tool (BLAST) will generate high-scoring alignments between truly orthologous sequences as well as between TE insertions from the same family that don't actually represent ancestral 'orthologous' insertions (Altschul et al. 1990). In many cases, the large number of these confounding TE insertions will overwhelm true orthologous sequence pairs and lead to such an enormous number of false positives that orthologous pairs of interest cannot be identified. A process called genome masking typically makes use of the RepeatMasker program to identify (or 'mask') repeat-derived sequences within the genome, thereby removing the potential for such spurious matches (Smit, Hubley & Green, 2015).

2.1.4 The importance of repeat libraries: genome assembly

Repeated DNA sequences also present issues for tasks commonly performed during genome assembly (Pop 2009). Modern genome sequence creation frequently relies on the generation of enormous libraries of short (approximately 125 bp - 250 bp long) overlapping genomic 'reads'. Once obtained, these reads are assembled into longer contiguous genomic sequences ('contigs') using their overlapping regions. In situations where highly similar repeats occur in multiple locations across the genome, many different seemingly legitimate overlaps may occur between short reads; this situation is analogous to being unable to place a given repeat in any one location in the genome. The end result of this ambiguity on the genome assembly process is that contigs are broken up into pieces that cannot be linked together due to repeats. One way to at least partially overcome this limitation is through the use of paired reads. Paired reads are common in modern sequencing libraries and are linked in such a way that each pair of reads originates from either end of a genome fragment with a known length. A repeat library can help determine the length-profile of the most potentially-problematic repeats in a to-be-assembled genome, and thus offer insight into the ideal internal fragment length for paired reads that will be able to straddle and overcome repeat-induced issues. Thus, the creation of a high-quality repeat library can aid in the process of genome assembly.

2.1.5 Creating a repeat library

The construction of a reliable and complete repeat library is a formidable task. Libraries are typically built from a collection of exemplar and/or consensus sequences; exemplars are single repeat copies obtained from the genome and are each representative of an entire repeat family, while consensus sequences are created from an 'average' of all of the genomic instances from a particular family. The ultimate goal for both of these types of sequences is to facilitate the identification and masking of their repeat family members in the genome. Creating a reference database is difficult due to the nature of repeats, and in particular of TEs. Different TEs vary substantially in their structure and sequence; examples of heterogeneous TEs within the genome include LINEs, which are frequently fragmented due to 5' truncation, Class II TIR elements, which exhibit prominent TIR structures, and LTR retrotransposons, which are flanked by long repeated motifs that are frequently fragmented by confounding ectopic recombination (Leeton and

Smyth 1993, Eickbush and Malik 2002, Sanchez et al. 2017). As with all DNA that is subject to limited or no selection, TEs experience a regular mutational force that drives the divergence of sequences that may once have been identical. For this reason recently active TEs are comparatively easy to find, while more ancient families may be fragmented and possess copies that are nearly impossible to identify or reconstruct. Further complicating the identification of repeat copies and the assignment of family membership, some TEs may have been exceptionally prolific, while others are rare. All of these characteristics make TE identification problematic, and increase the difficulty of creating reliable consensus sequences for a reference repeat database (Sotero-Caio et al. 2017).

Automated tools for repeat identification generally utilize techniques from at least one of three different approaches: similarity-based, repetitiveness-based and structure-based (Lerat 2010, Hoen et al. 2015). Similarity-based approaches make use of a database of previously identified TE sequences, often from other closely-related species. Using this preliminary database, local sequence alignment software tools such as HMMER (Eddy 1998) or BLAST identify TE fragments in the genome that are subsequently processed and added to a species-specific repeat library. Importantly, this technique relies on the existence of a reliable TE database in a species that is closely related to the species of interest, and requires that TEs within the two species have not diverged to such an extent that they are no longer detectable with common sequence search tools. Because of this reliance on previously-identified sequences, similarity-based approaches cannot identify previously unreported TEs.

Repetitiveness-based tools are commonly used for *de novo* repeat library creation in organisms which have not already been the focus of TE research. These approaches rely on the existence of multiple highly similar repeat copies in order to identify repeat instances which originate from the same family. Tools that make use of this approach, such as REPET (Flutre et al. 2011), RECON (Bao and Eddy 2003) and RepeatScout (Price et al. 2005), generally first perform an ‘all-by-all’ BLAST (or other initial search) step in which local alignments are found between fragments of a whole genome sequence. Any genomic sequences that possess multiple alignments to many other regions of the genome are classified as probable repeats and are subsequently clustered,

with each cluster ideally representing a single TE family. In a final step, sequences from each cluster are aligned, and consensi are generated from each alignment. Repetitiveness-based tools are ideal for identifying recently-active repeat families, as these families have had little time to diverge within the genome and are easily identified by local alignment search tools. Such families are often lineage-specific, and as a result this approach is particularly complimentary to similarity-based methods.

The final category of repeat-identification tools follow a structure-based approach. This strategy relies on the presence of diagnostic sequence motifs or structural features to identify TEs. Examples of software that implements these methods include LTRharvest (Ellinghaus et al. 2008), LTR_FINDER (Xu and Wang 2007) and LTR_STRUC (McCarthy and McDonald 2003), which identify LTR retrotransposons wherever their characteristic flanking long terminal repeats are detected, and MITE-Hunter (Han and Wessler 2010), which identifies Class II TEs on the basis of flanking TIRs. These tools unveil individual TE copies in the genome but are especially susceptible to false positives, so the results must often be validated before they are clustered into individual families (Campbell et al. 2014, Jiang et al. 2016). For those cases in which targeted TE insertions have maintained the characteristic features on which a structural search is based, these methods are able to identify older TEs, previously undiscovered elements, and those with very low copy numbers.

2.1.6 Repeat resources

Previously discovered repeats are often made publically available through large online repositories, private websites or within scientific reports. A major online repository is RepBase, a resource maintained by the Genetic Information Research Institute which promotes itself as "a database of representative repetitive sequences from eukaryotic species" (Bao et al. 2015). Rather than unselectively incorporating all potential repeats, RepBase employs a rigorous manual and automatic curation process that has allowed it to grow to include over 44,000 very high quality reference sequences. This process utilizes a custom pipeline which identifies TEs using both the RECON and LTR_FINDER tools, as well as an iterative refinement procedure and further manual processing to extend incomplete TE sequences to their true termini (Bao et al. 2015). Other sources of repeats include both Genbank and, for TE protein sequences, UniProtKB (Apweiler et al. 2004).

Both of these databases store many other sequences apart from repeats and therefore do not enforce a consistent repeat taxonomy or annotation system, however they are still useful when exploring sequences which have yet to make their way into RepBase. More esoteric repeat information is frequently available from domain-specific websites or individual publications. A very recent example is FishTEdb (www.fishtedb.org) which was announced by Shao et al in January 2018 and does not yet contain any Protacanthopterygian species (Shao et al. 2018). Salmonid repeats, particularly TEs, are available in previous reports such as the rainbow trout genome analysis of Berthelot et al. (2014), a survey of retrotransposons by Matveev and Okada (2009), an analysis of Tc1-Mariner activity by de Boer et al. (2007) and others (Moir and Dixon 1988, Kido et al. 1991, Stuart et al. 1992, Goodier and Davidson 1994).

2.1.7 Repeat libraries for salmonids

Multiple reference genome sequences have now been reported for salmonid species (Berthelot et al. 2014, Lien et al. 2016) and many more have been recently released to genome repositories in anticipation of publication (Benson et al. 2005). The existence of these sequences demands the development of a set of robust, high-quality reference repeat libraries that can enable future work in this important taxa. The identification of repeat sequences within the salmonids will facilitate not only unrelated genomic analysis, but also an increased understanding of TE biology and the role of these important sequences in the evolution and rediploidization of the salmonid genome.

In this chapter I describe the process of creating reference repeat libraries for each of five salmonid species: Atlantic salmon, arctic char, rainbow trout, coho salmon and chinook salmon. Using these libraries I interrogate each species' genome and find that the repeat-derived fraction of DNA is greater than 50% in salmonid genomes, a value which is amongst the largest for all vertebrate sequences that have been analyzed to date. Finally, I explore the repeated fraction of these five genomes and assess the relative abundance of different TE taxa both within salmonids and between salmonids and other vertebrate species. I find that individual salmonid species differ only a little in the relative abundance of their constituent TEs, and that high level of TE diversity within the salmonid genome is consistent with patterns that have been previously reported in other fish species.

2.2 Methods

2.2.1 Repeat library construction

In order to investigate the repeat-derived portion of salmonid genomes and create a set of reference sequences that could be used to facilitate further genomic studies, I constructed repeat libraries for each of five salmonid species: Atlantic salmon, arctic char, rainbow trout, coho and chinook. The exact approaches I took to library construction varied to some extent between species due both to time constraints and the maturation of my methodology. Below I first describe the process used to create the Atlantic salmon repeat database, and then discuss the few changes I made during library construction for the four other species.

2.2.2 Atlantic salmon repeat library

The approach I used to create repeat libraries is divided into four general steps: 1) identify putative repeat sequences from existing sources or using *de novo* methods; 2) verify the repetitive character of the putative repeats and split any potentially chimeric sequences; 3) merge libraries from different repeat sources and remove redundant sequences; and, 4) classify repeats and remove non-TE host genes. Together, these steps ensure that any sequences in the final library are truly repetitive (few if any false positives), that they are non-redundant, and that no repeated non-TE genes are present in the library. This latter criterion is important as it ensures that non-TE genes of interest are not inappropriately discarded in future studies when the present repeat libraries are used to mask a genome.

2.2.3 A note on BLAST alignments

At many points in the ensuing process I utilized the BLASTN software tool to create local alignments between different sequences. In all of these cases I used the 'word_size 7' parameter in order to identify more-divergent alignments between sequences of interest; this is the lowest permitted word_size value for BLASTN. This parameter reduces the default size of the exact match ('seed') required between two sequences during BLASTN's initial search step. It has the effect of allowing for the creation of alignments between more-dissimilar sequences that have at least seven sequential bases that are identical between them, but not 11 (the default).

2.2.4 A note on computational analysis

Many different bioinformatics software tools and pipelines were used over the course of this work, much of which was performed on resources provided by Compute Canada (www.computecanada.ca). Apart from those cases in which I explicitly reference a software tool or program, any computational tasks I describe were performed using scripts created with the Python, Bash or R programming languages. Three external Python libraries, BioPython (Cock et al. 2009), NetworkX (Hagberg et al. 2008) and NumPy (Van Der Walt et al. 2011), were particularly helpful over the course of my research.

2.2.5 Step 1: Identify putative repeat sequences

2.2.5.1 Existing sources

I obtained previously-identified preliminary repeats from three different sources: i) nine salmonid non-LTR retrotransposons reviewed by Matveev and Okada (2009); ii) 445 repeats in the RepBase database originating from species within the family Salmonidae as of February 2015 (Bao et al., 2015); and, iii) 634 repeats identified in the rainbow trout genome by Berthelot et al. (2014).

2.2.5.2 *De novo* methods – REPET and RepeatModeler

Three *de novo* repeat-finding programs were used to identify repetitive sequences in the genome: REPET v1.3.9 (Flutre et al. 2011), RepeatModeler v1.0.8 (Smit and Hubley 2008) and LTRharvest, which is included within GenomeTools v1.5.1 (Gremme et al. 2005, Ellinghaus et al. 2008). REPET is a software pipeline which combines a number of repeat-finding programs including RECON (Bao and Eddy 2003), GROUPER (Quesneville et al. 2003) and PILER (Edgar and Myers 2005) with filtering steps to produce a final library. REPET follows three major steps: i) all-by-all self-alignment of genome sequences using BLAST; ii) clustering of BLAST pairwise alignments (using the clustering approaches of GROUPER, PILER and RECON); and iii) multiple sequence alignment followed by consensus sequence construction. Similarly to REPET, RepeatModeler is a wrapper program which encapsulates RECON and another repeat-finding program, RepeatScout (Price et al. 2005). The standalone RECON component of RepeatModeler operates similarly to REPET, utilizing an all-by-all BLAST and clustering method. RepeatScout follows a different approach by efficiently identifying

short, highly repetitive nucleotide sequences ('seed' kmers) scattered through the genome, the flanking sequences of which are iteratively extended to create a consensus sequence representative of each seed group. Two REPET libraries containing 581 and 919 sequences respectively were produced using contigs longer than 10 kb from two preliminary draft Atlantic salmon genome assemblies that were made available through the International Consortium to Sequence the Atlantic Salmon Genome (ICSASG; see Davidson et al., 2010) and constructed primarily using Sanger sequencing data and the Celera assembler. A single library of 927 sequences was produced by RepeatModeler using all of the contigs in Atlantic salmon genome assembly v3.6 (Lien et al. 2016), which was curated and submitted to GenBank (accession number GCF_000233375.1) and incorporated Illumina HiSeq, GAIIx, PacBio and Sanger data.

2.2.5.3 Manual examination and extension of REPET sequences

A large selection of sequences from the two REPET libraries were manually examined and processed in an effort to create longer, more complete consensus sequences. This process initially entailed the separate alignment of 475 REPET sequences against the genome using BLASTN and, for each REPET query, the extraction of a subset of the genomic sequences corresponding to high-quality high-scoring segment pairs (HSPs). Any HSPs which were at least 80% of the length of the query sequence with at least 80% similarity were considered to be high-quality; the associated genomic sequences correspond to repeat instances within the genome. Up to 2,000 bp of 5' and 3' flanking sequence was extracted along with each genomic region. The extracted samples were then aligned using MUSCLE 3.8.31 (Edgar 2004) in the MEGA5 software (Tamura et al. 2011), any obviously non-repeat flanking sequence was manually removed, and new consensus sequences were created from the alignments. The initial REPET sequences were frequently only fragments of longer repeats, and so only identified partial repeat HSPs during the initial BLASTN search. In many cases, the additional flanking bases that were extracted from the genome allowed for the capture of additional valid repeat sequence, which following alignment and trimming had the effect of extending the consensus sequence. A similar procedure is described in more detail for the creation of a curated Tc1-Mariner repeat library in Chapter 3.

2.2.5.4 *De novo* methods – LTRharvest

The third *de novo* program used to identify preliminary repeats, LTRharvest, identifies LTR structures within a genome sequence and infers the existence of LTR retrotransposons (or their remnants) from these flanking sequences. Because of LTRharvest's propensity to identify false-positives, a series of filtering and validation steps were performed using a procedure outlined in the MAKER genome annotation tool documentation (Campbell et al. 2014, Jiang et al. 2016). A similar filtering procedure was also recently used in developing a repeat library for Atlantic cod (Tørresen et al. 2017). Two libraries were initially created using LTRharvest – one corresponding to recently active full-length LTR retroelements and one composed of more ancient LTR retroelements. A putative LTR element identified by LTRharvest was included in the "recent" library if its characteristic terminal repeats were at least 99% similar to each other (LTR99 library) and otherwise included in an 'ancient' library if its terminal repeats were at least 85% similar (LTR85 library). All LTR elements were required to possess either a poly-purine tract or primer binding site identified using the LTRdigest tool packaged with GenomeTools v1.5.1 (Steinbiss et al. 2009). Primer binding site detection was assisted by a library of eukaryotic tRNA sequences obtained from the Genomic tRNA Database (Chan and Lowe 2009). In order to identify and exclude common false positives caused by tandem local repeats, local gene clusters or adjacently-inserted TEs, the 50 bases upstream of each constituent LTR were aligned using MUSCLE v3.8.31 and, separately, the downstream 50 bases were similarly aligned. If the aligned upstream or downstream sequences had at least 25 identical bases and were at least 60% similar to each other, the entire element was excluded.

Other steps were taken to ensure that the chosen LTR retroelements were in fact 'exemplars' suitable for retention in the final Atlantic salmon library. For both the LTR99 and LTR85 collections, all of the flanking LTR regions of the elements in a given library were used with RepeatMasker v4.0.5 to mask the internal regions of the library (maximum 20% sequence divergence); if any bases in a given internal sequence were masked, the element was removed. Subsequently, retroelements containing nested insertions of other non-LTR TEs were detected and removed based on the presence of a good (E-value $\leq 1e^{-10}$) TBLASTN HSP when compared to a TE protein library (REPET-

formatted RepBase v4.0.5) from which all LTR retrotransposon-like proteins were removed.

In an effort to obtain only a single representative sequence for each LTR TE family, a redundancy removal step was performed. First, an all-by-all BLASTN search was undertaken separately on the LTR99 and LTR85 libraries using a minimum percent identity threshold of 80%. Redundant sequence pairs were then identified. A pair of sequences was deemed redundant if there existed a single HSP between the two sequences that occupied at least 90% of the length of one of them, or if there existed two HSPs, anchored within 50 bp of the 5' and 3' ends of the query sequence and in an intuitive orientation, which together occupied at least 90% of the query sequence. This second 'two-HSP' redundancy check was implemented after observing a large number of LTR retroelement family members which differed only in the length of an internal tandem repeat tract that frequently disrupted a single HSP. Once redundant pairs were identified, the sequence that was a member of the largest number of such pairs was retained and all other sequences to which it was paired were removed. This process was repeated again for the sequence that was a member of the second-largest number of pairs, and continued in this iterative fashion until there were no pairs remaining. This process had the effect of 'greedily' identifying sequences which were more likely to be longer and more complete elements, and which could act as representatives for a large group of individual retrotransposon insertion sequences. In a final step, the LTR99 and LTR85 libraries were merged. RepeatMasker was used to mask the LTR85 library with LTR99 sequences (maximum sequence divergence of 20% between LTR99 library query sequences and LTR85 targets), and any sequences that were more than 80% masked were removed as redundant. The remaining sequences from both libraries were combined into a single LTRharvest library containing 407 sequences.

2.2.6 Step 2: Verification of repetitiveness

In order to address false-positives, sequences from all four of the *de novo* libraries as well as from the rainbow trout library of Berthelot et al. (2014) were subjected to varying degrees of validation to verify that the sequences were truly repetitive. The rainbow trout library sequences were included in this step to ensure that sequences which were repetitive in the rainbow trout genome were also present and repetitive in Atlantic

salmon; rainbow trout sequences which failed the repetitiveness check might be those that were lost in the Atlantic salmon lineage, or gained *de novo* in the rainbow trout lineage, since the divergence of these species' ancestors.

Putative repeats from the two REPET libraries (including the manually curated sequences), the RepeatModeler library and the rainbow trout genome library were assigned a confidence level based on the length and number of BLASTN hits on contigs in the Atlantic salmon genome. Any sequences that generated three or more HSPs at least 80% of their length were designated as high confidence (HC). Sequences not designated as HC were classified as lower confidence (LC) if they produced 10 or more hits of at least 100 bp; otherwise they were eliminated. Because of the intensive and manual lengths taken to initially characterize them, sequences from Matveev & Okada's TE library and RepBase were assumed to be well-curated and were automatically included in the HC library.

Many LC sequences failed to generate long (80%) HSPs due to a stretch of ambiguous ('N') bases or the inappropriate concatenation of separate repeats by the source *de novo* repeat-finding program. Such chimeric sequences were generally composed of two or more distinct repeats joined by non-repeated sequences. To reduce the number of chimeras the number of long HSPs (80+ bp) overlapping each LC sequence base was determined using BLASTN and the repeat sequence was split wherever the HSP coverage dropped below 10 over 10 consecutive bases, with low-coverage sequence being removed.

LTRharvest-derived sequences, as part of a different analysis pipeline, were validated separately from the other libraries. In order to be retained, a sequence in the LTRharvest library was required to generate 10 or more 500 bp HSPs when aligned to the genome using BLASTN.

2.2.7 Step 3: Library merging and redundancy removal

Merging of the validated repeat libraries was guided by the LTRharvest procedure in the MAKER documentation, as well as by the 80-80-80 rule of Wicker et al. (2007) which proposes that two TE sequences should be considered to be within the same taxonomic family if, when considering BLASTN HSPs over 80 bp between the two

sequences, they overlap over 80% of their length with at least 80% identity. Multiple merge steps were undertaken to produce a single repeat library.

First, redundancy was reduced separately within both the HC and LC libraries. This was accomplished by removing a sequence A if there existed an equal-length or longer sequence B within the same library such that A was at least 80% covered by long (80+ bp) non-overlapping HSPs from B possessing at least 80% identity. Following within-library removal, the remaining sequences in the LC library were removed using the same criteria if they possessed a match to a sequence in the (reduced) HC library.

Sequences from the LTRharvest library were merged with the HC+LC sequences using a RepeatMasker-based approach. Specifically, the sequences in the LTRharvest library were used to mask the HC and LC sequences (maximum sequence divergence of 20%), and any HC or LC sequences that were more than 90% masked were removed. The libraries were then combined.

2.2.8 Step 4: Non-TE host gene identification and repeat classification

2.2.8.1 Non-TE host gene detection and removal

In order to identify and remove non-TE host genes (such as duplicated non-TE genes) the merged library was compared to both the SwissProt UniProtKB database and a TE protein database made of REPET-formatted RepBase v19.06 sequences. Non-TE host genes were identified as those repeats having a strong (E-value $\leq 1e^{-10}$) BLASTX hit to a non-TE UniProtKB protein while not possessing a hit to a RepBase TE protein with a better bit score. During the subsequent repeat classification step, any sequences categorized as rRNA genes by the PASTE Classifier tool included with REPET (Hoede et al. 2014) were similarly removed. In total, 40 repeat sequences were classified as putative non-TE host genes and removed from the library.

2.2.8.2 Repeat classification

The classification of repeat sequences was based on the guidelines established by Wicker et al. (2007). The PASTE Classifier tool included with REPET v2.0 was used in combination with BLASTN and BLASTX to identify structural motifs and to establish similarity to reference sequences for classification. Reference databases consisted of: i) a REPET-formatted set of PFAM HMM Gypsy profiles (v26.0) available on the REPET

website (https://urgi.versailles.inra.fr/Tools/REPET); ii) nucleotide and protein sequences from the REPET-formatted RepBase v19.06 library; iii) all proteins from RepBase TEs found in Actinopyerygii species as of May 2015; v) eukaryotic rRNA sequences from release 115 of the SILVA rRNA database (Quast et al. 2013); and, vi) salmonid SINE and LINE retroelement sequences reviewed by Matveev and Okada (2009).

Repeat library sequences were first classified to the superfamily level using BLASTN if, when compared to a nucleotide reference sequence, at least 80% of their sequence was covered by long (80+ bp) non-overlapping HSPs with greater than 80% similarity. If no classification was performed with BLASTN, sequences were similarly classified at the superfamily level based on their best BLASTX hit to a reference TE protein (E-value ≤ $1e^{-10}$). LTRharvest elements without a superfamily-level categorization were classified as being in the LTR order and sequences identified as MITEs by the PASTE Classifier tool were assigned to the TIR order. Using Geneious (Kearse et al. 2012), all sequences were analyzed by dot plot to identify and visually classify those consisting solely of a tandem repeat (satellite) or of a simple repeat motif. The annotation information of library sequences with conflicting classifications and those labeled as 'PotentialChimeric' by PASTE Classifier was manually examined and, if any obvious category could not be established, were labelled as 'Unknown', along with all other sequences possessing no annotations.

2.2.9 Library creation for other salmonid species

The repeat library creation process for arctic char, rainbow trout, coho and chinook was guided by the same four-step method as was developed for Atlantic salmon. The preliminary repeat libraries created and retrieved in the initial phase of the process were different, however. Notably, LTRharvest and REPET libraries were not included (including the manually curated library), and instead were replaced by the finished Atlantic salmon library. Additionally, the repeat library created by Berthelot et al. (2014) was only included during the creation of the Atlantic salmon and rainbow trout libraries, but no others. I made these choices due primarily to time and resource constraints. These exclusions were theoretically ameliorated by the novel inclusion of the Atlantic salmon library, which was likely to contain any older repeat families (active before the divergence of the Salmoninae lineages) that would have been otherwise identified by

LTRharvest or REPET. More recent lineage-specific families, having experienced less genomic degradation, should still be readily detectable using the repetitiveness-based process of RepeatModeler. A further change made to the repeat library creation process was to the redundancy removal step. Specifically, when attempting to identify sets of high-scoring HSPs which covered up to 80% of each of two redundant sequence pairs, individual HSPs were allowed to overlap by up to 15 bp on either the query or the subject sequence. This represented a change from the Atlantic salmon library approach in which they were permitted no overlap, and was implemented after observing a number of HSP sets which would have eliminated a redundant sequence had they not overlapped by a few bases.

As they were created at different times over the course of four years, the exact versions and content of the input and reference databases differed between the five salmonid species. For both arctic char and rainbow trout, the RepBase preliminary library consisted of all Salmoniformes sequences available as of February 2016 (rather than February 2015, as in Atlantic salmon). For coho and chinook, the RepBase input sequences used were those obtained in January 2017. All four non-Atlantic salmon species used version 20.05 of the REPET-formatted RepBase database as a reference database during the non-TE host gene identification and repeat classification steps, rather than version 19.06.

Different input genome sequences were used by RepeatModeler for *de novo* element detection in each of the four non-Atlantic salmon species. For arctic char, an in-house preliminary ALLPATHS-LG (Gnerre et al. 2011) genome assembly was used; this genome formed the basis of the more-refined reference genome that has since been submitted to NCBI (Genbank Accession: GCF_002910315.2). The input for RepeatModeler runs for rainbow trout, coho, and chinook took the form of final or near-final assemblies that have since been made available through Genbank with Accessions: GCF_002163495.1 (rainbow trout), GCF_002021735.1 (coho), and GCA_002872995.1 (chinook). It is worth emphasizing that the rainbow trout genome sequence used during the course of my research is different from that which formed the basis of the first whole-genome survey of repeats in a salmonid (Berthelot et al. 2014). In the ensuing years since this first rainbow trout sequence was released, a much more contiguous and mature

genome sequence has been produced through a collaboration led by researchers at the United States Department of Agriculture.

2.2.10 Repeat identification and assessment

The total amount of repeat-derived content within the genome and the abundance of individual TE superfamilies were assessed for all five salmonid species using RepeatMasker v4.0.8. All masked genomes were the most recent versions available from Genbank as of February 2018 (Table 2), and possessed the following accessions: GCF_000233375.1 (Atlantic salmon), GCF_002910315.2 (arctic char), GCF_002163495.1 (rainbow trout), GCF_002021735.1 (coho), and GCA_002872995.1 (chinook). For calculations of repeat percentage, the total genome size was that reported by RepeatMasker, a value which does not include runs of ambiguous ('N' or 'X') bases greater than 19 bp.

Table 2 Summary statistics for five repeat-masked salmonid genomes

Species	Contig N50 (Kbp)	Scaffold N50 (Mbp)	Sequences masked	Sequence length (Gbp)	Genbank accession	Library types used
Atlantic salmon	57.62	1.37	Chromosomes + Unmapped scaffolds	2.97	GCF_000233375.1	Illumina HiSeq PE+MP Illumina GAIIx PE+MP Illumina MiSeq PacBio Sanger
Arctic char	55.62	1.02	Chromosomes only	1.52	GCF_002910315.2	Illumina HiSeq PE+MP PacBio
Rainbow trout	13.83	1.67	Chromosomes + Unmapped scaffolds	2.18	GCF_002163495.1	Dovetail HiRise
Coho salmon	58.12	1.27	Chromosomes + Unmapped scaffolds	2.37	GCF_002021735.1	Illumina HiSeq PE+MP PacBio
Chinook salmon	133.17	1.73	Chromosomes + Unmapped scaffolds	2.43	GCA_002872995.1	Illumina HiSeq PacBio

Following genome masking by RepeatMasker, repeat abundance figures were retrieved in two ways. The estimate of the total amount of repeat-derived DNA within each genome was obtained directly from the RepeatMasker *.tbl file, as were estimates of the abundance of unclassified sequences, satellites, simple repeats and low-complexity DNA. These values correspond directly to the number of bases masked of each feature type in the genome of interest. Abundance figures for individual TE taxa are not explicitly reported, however, and so must be indirectly obtained from the RepeatMasker *.out file, which individually lists each repeat fragment detected in the genome. These individual annotations occasionally overlap with each other, in rare cases by as much as 80%. Because of this overlap, the abundance totals created by summing individual annotations are likely to represent slight overestimates. Nevertheless, this approach is common (see for example the 2013 zebrafish genome report by Howe et al.), and in practice does not seem to dramatically inflate the estimated abundance values.

2.3 Results and Discussion

2.3.1 Repeat libraries for five salmonid species

I created repeat libraries for each of five salmonid species in order to investigate the abundance and character of repeat-derived DNA (the majority of which generally consists of TEs) within their genomes, as well as to facilitate the repeat-masking of DNA in furtherance of future studies. These databases are the result of a process which incorporated TEs from existing sources as well as repeats found using *de novo* repeat-identification programs. After combining the repeats from different sources, a process of repetitiveness validation and redundancy removal was performed which resulted in libraries containing between roughly two and three thousand individual sequences. Specifically, 1,997 repeats were identified in Atlantic salmon, 2,854 in arctic char, 2,940 in rainbow trout, 2,372 in coho and 2,419 in chinook (Table 3). The repeats within each library include consensus sequences that reflect an average of the nucleotide sequences for the members of an individual TE family, as well as 'exemplar' repeats, the sequences of which were obtained directly from individual TE copies in the genome and which, like consensus sequences, are representative of an entire TE family.

The repeats identified in all five species were put through a categorization process which classified sequences using the method and taxonomy established by Wicker et al. (2007). The approach I employed performed classification based primarily on the similarity of repeat sequences to previously-classified TEs in databases such as RepBase, but also incorporated both automated and manual examination of diagnostic patterns such as TIRs or the consistently repeating motifs which define satellites. Consistently across all five salmonid libraries, approximately half of the identified representative repeat sequences were classified to some degree (Table 3). Those repeats which remain unknown possess no similarity to previously described TEs, and don't evince structures which can easily be used to facilitate their categorization. Such repeats may correspond to novel TEs present only in the genomes of salmonids, or to difficult-to-identify tandem repeat sequences.

Table 3 Repeat libraries for five salmonids

Species	Number of representative repeats in library	Number of sequences classified
Atlantic salmon	1,997	1,091 (54.6%)
Arctic char	2,854	1,535 (53.8%)
Rainbow trout	2,599	1,578 (53.7%)
Coho salmon	2,372	1,148 (48.4%)
Chinook salmon	2,419	1,165 (48.2%)

2.3.2 The need for manual curation

Creating a high-quality repeat library and annotating such sequences within a genome are particularly difficult tasks due to the extensive divergence, fragmentation and diversity of TEs that exist within most species. The presence of these obstacles is the result of both the evolutionary imperative of TEs to proliferate (even at the expense of the host) and of the fact that most TE insertions produce only minor fitness disadvantages. The combination of these factors allows TEs to accumulate in a genome and exposes them to very few forces (such as selection) that would prevent the gradual divergence and breakup of their constituent sequence after a successful genomic insertion. Because of these characteristics of repeat-derived DNA, all of the automated approaches to repeat discovery - similarity-based, repetitiveness-based and structure-based - have shortcomings that often prevent the retrieval of full-length sequences, particularly for older, more-fragmented families. The process of creating truly comprehensive repeat libraries thus requires an extensive and highly labour-intensive manual curation step that is often not undertaken, particularly in genomic projects where repeats are analyzed only as a minor component (Hoen et al. 2015, Sotero-Caio et al. 2017). Over the course of such a manual process fragments of TE copies from single families are often sought in the genome, extracted, trimmed and analyzed in a batch, with the final goal of reconstructing full-length representative sequences.

In this work I pursued a compromise method in which I performed a manual curation step during the creation of the repeat library for Atlantic salmon, but not during that of the other four species. In total I analyzed 475 repeats obtained from REPET preliminary libraries and manually attempted to extend the sequences beyond their often-fragmented state. Once longer sequences were determined, I observed that many of these repeats were redundant; in such cases the initial REPET library consensus sequences originated from different sections of repeats that belonged to the same family. The identification of

incomplete repeat sequences by automated methods would be expected for repeat families that are older, as they are more likely to exist only in fragmented pieces throughout the genome which may themselves vary in copy number and which are not amenable to full-length consensus sequence reconstruction. Although I didn't explicitly manually examine TEs from the other salmonid species, the fact that the Atlantic salmon repeat library was used as an input during the construction of the other species' libraries suggests that in many cases the Atlantic salmon sequences would have made their way into other libraries. Thus, all repeat libraries likely benefited in some way from this manual curation effort, either directly or indirectly.

2.3.3 The repeat-derived component of salmonid genomes

The results of my repeat annotation effort are described in Table 3. At least ~55% of the DNA content in all five salmonid genomes is derived from repeated DNA. At 58.7%, Atlantic salmon evinced the greatest repeat-derived genome fraction, quite possibly as a result of the more comprehensive repeat-discovery and curation processes that were used for this species. The slight deviation of less than 3% between the non-Atlantic salmon species could be attributed to true biological variation or to technical factors such as differences between the input reference database versions used for repeat library creation, or variation in the quality of the underlying assemblies (see Table 2); repeats are one of the primary obstacles to genome assembly and are often 'collapsed' or missing from even intensively-researched modern genomes (Pop 2009, Treangen and Salzberg 2012). The high percentage of the genome composed of repeated DNA is notable, as it is amongst the highest ever reported in a vertebrate; with the exception of zebrafish and coelacanth (*Latimeria chalumnae*), fish species generally evince a repetitive genome content of between 5% and 40% (Howe et al., 2013; Nikaido et al., 2013; Rondeau et al., 2014; Schartl et al., 2013; P. Xu et al., 2014).

Table 4 Repeat abundance in salmonid genomes. Percentage values are based on the number of bases in five salmonid genome assemblies, ignoring tracts of >= 20 ambiguous (N/X) bases: 2.62 Gbp (Atlantic salmon), 1.45 Gbp (arctic char), 1.93 Gbp (rainbow trout), 2.26 Gbp (coho) and 2.36 Gbp (chinook).

			Genome abundance (%)				
Repeat Type	**Order**	**Superfamily**	**Atlantic salmon**	**Arctic char**	**Rainbow trout**	**Coho**	**Chinook**
Class I TEs	**All**	**All**	**20.86**	**18.48**	**17.79**	**17.24**	**17.11**
	LTR	**All**	**5.84**	**3.46**	**2.75**	**2.06**	**2.43**
		Gypsy	2.67	2.00	2.12	1.54	1.77
		ERV	0.81	0.39	0.42	0.34	0.36
		Copia	0.28	0.14	0.11	0.11	0.20
		Bel-Pao	0.10	0.09	0.09	0.07	0.09
	DIRS	**DIRS**	**0.10**	**0.14**	**0.11**	**0.10**	**0.11**
	PLE	**Penelope**	**0.23**	**0.17**	**0.17**	**0.16**	**0.16**
	LINE	**All**	**13.96**	**14.41**	**14.49**	**14.67**	**14.17**
		CR1	0.01	< 0.01	0.03	0.03	0.03
		Crack	3.94	4.24	4.49	5.40	4.64
		Hero	0.00*	0.00*	0.07	0.03	0.02
		L1	0.47	0.67	0.64	0.48	0.52
		L2	3.19	1.96	1.73	1.59	1.45
		Nimb	0.19	0.24	0.21	0.19	0.21
		R2	0.02	< 0.01	< 0.01	< 0.01	< 0.01
		Rex1	4.53	5.65	5.40	5.08	5.45
		RTE	0.01	0.01	0.01	0.01	0.01
		RTEX	0.90	0.73	1.12	1.06	1.00
		Tx1	0.69	0.78	0.77	0.78	0.83
	SINE	**All**	**0.75**	**0.30**	**0.27**	**0.26**	**0.25**
		tRNA	0.68	0.19	0.17	0.17	0.15
		Deu	0.06	0.11	0.10	0.09	0.10
Class II TEs	**All**	**All**	**20.45**	**21.82**	**23.41**	**19.60**	**20.05**
	TIR	**All**	**17.11**	**19.61**	**20.38**	**16.71**	**17.04**
		CMC-EnSpm	0.19	0.18	0.21	0.14	0.16
		Ginger	0.11	0.02	0.04	0.08	0.07
		Harbinger	0.09	0.15	0.09	0.10	0.08
		hAT	2.80	3.38	2.43	2.31	2.25
		IS3EU	0.06	0.08	0.07	0.08	0.06
		ISL2EU	0.01	0.01	0.01	0.01	0.01
		Kolobok	0.02	0.04	0.04	0.04	0.04
		PiggyBac	0.26	0.36	0.40	0.31	0.32
		Sola	0.08	0.06	0.04	0.04	0.04
		Tc1-Mariner	13.49	15.31	17.03	13.58	14.00
	Dada	**Dada**	**0.01**	**< 0.01**	**< 0.01**	**< 0.01**	**< 0.01**
	Crypton	**Crypton**	**0.19**	**0.18**	**0.17**	**0.15**	**0.16**
	Maverick	**Maverick**	**0.05**	**0.04**	**0.06**	**0.09**	**0.09**
Unclassified			**10.65**	**15.03**	**14.09**	**14.62**	**15.44**
Satellites			**5.70**	**0.69**	**0.95**	**1.40**	**0.95**
Simple repeats			**2.39**	**2.07**	**1.96**	**3.03**	**2.65**
Low-complexity			**0.43**	**0.44**	**0.36**	**0.81**	**0.71**
Total Masked			**58.66**	**56.38**	**56.85**	**54.72**	**55.08**

* Although absent in their respective repeat libraries, TEs of the Hero superfamily were subsequently found in the genomes of Atlantic salmon and arctic char

While roughly half of the sequences in the salmonid repeat libraries were not classified, less than 16% of each genome was ultimately masked by those unknown repeats. This indicates that the unclassified repeat library sequences are on average less repetitive than their classified fellows. The Atlantic salmon genome had the lowest number of unidentified repeats (10.65% of the genome), a possible result of a larger percent of its library being classified (54.6%), as well as the presence of longer classified repeats originating from the manual curation step. It's also likely that LTRharvest discovery process in Atlantic salmon, the results of which were classified as being from the LTR order if no more precise annotation was found, would have led to the annotation of additional unknown sequences in the other species were they to have had such a benefit. If one makes the tenuous assumption that there is roughly the same number of LTR repeat copies within all five salmonid genomes, this assertion is supported by the observation that 5.8% of the (LTRharvest-assisted) Atlantic salmon genome was found to be derived from LTR TEs, a value which was only 1.5-3.5% in the other four salmonid genomes. This difference of ~2-3% could help account for the decreased amount of unknown repeat-derived DNA in Atlantic salmon. If not a technical artifact, the larger proportion of LTR retrotransposons could reflect a recent increase in activity that occurred after the divergence of the Atlantic salmon ancestor from the other groups or, alternatively, a generally high level of activity that existed in the Salmoninae ancestor and which experienced a marked decrease in the *Onchorhynchus* lineage and, to a lesser extent, the *Salvelinus* lineage (arctic char has the next highest abundance of LTR retrotransposons, at 3.46%). In all, approximately 14-15% of the genomes of arctic char, rainbow trout, coho and chinook are derived from repeated sequences that could not be assigned to a TE taxa or other repeat class.

Apart from the apparent increase in the prevalence of LTR retrotransposons in Atlantic salmon compared to the other species, the influence of Class I TEs on the genome composition of the five salmonids is markedly similar. This finding implies either that the activity of retroelements occurred for the most part in the common ancestor of the Salmoninae lineage, or that any activity which occurred since lineage diversification has continued at approximately the same rate in all extant species. TEs of the Hero superfamily stand out distinctly from this general trend, as they at first appeared to be

absent from the genomes of Atlantic salmon and arctic char. This result pointed to the potential existence of a horizontal transposon transfer (HTT) event of Hero into the genome of the *Oncorhynchus* ancestor. To investigate this possibility, TBLASTX was used to search the genomes of Atlantic salmon and arctic char for TE copies using a query set of 23 Hero nucleotide sequences obtained from RepBase. Hero TE instances were found in both genomes, a finding which implies a failure to identify and include a legitimate repeat in the corresponding libraries. This result emphasizes the occasional imperfection inherent in automated methods operating on entire genome sequences.

Of all elements present in the five salmonid genomes, Tc1-Mariner is by far the most prolific superfamily. While other Class II elements are present in appreciable quantities (eg hAT), they vary little in their abundance between species. Tc1-Mariner elements occupy at least 13.5% of the genome in all species surveyed, and in two species, arctic char and rainbow trout, they occupy considerably more (up to 17% in rainbow trout). These findings are consistent with previous observations that Tc1-Mariner elements are especially prolific in fish lineages, and that they have been especially active over the course of salmonid evolutionary history. I examine the activity of Tc1-Mariner elements in further detail in Chapter 3.

A number of non-TE repeat types can be identified by RepeatMasker and are also described in Table 4. As with most TEs, the relative abundances of both simple repeats (such as microsatellites) and low-complexity repeats (poly-nucleotide tracts) are generally consistent between genomes. Only minor differences of approximately 1% of the genome exist for simple repeats, with coho evincing a slightly higher percentage than do the other species. Such inconsistency could well be attributable to differences in the underlying sequence assemblies, since simple repeats are a common source of missed or confounding genome sequence (De Bustos et al. 2016). Dissimilarly from the case of simple and low-complexity repeats, there appears to be a substantial difference in satellite (large tandemly-replicated repeat) DNA content between species; 5.7% of the Atlantic salmon genome is composed of these repeats, while the other species all possess less than 1.5%. A few different hypotheses potentially address this variation. Satellite sequences have been shown to be highly variable between even closely-related species, so the possibility exists that the differences observed here reflect the biological reality (Louzada

et al. 2015, Jagannathan et al. 2017). Other, technical, causes include differences in the underlying genome assemblies as well as the fact that REPET seemed to generate a higher number of satellite sequences than did other *de novo* repeat discovery methods. In this latter case, the lack of a species-specific REPET library would potentially reduce the number of satellite sequences discovered and classified as such in the non-Atlantic salmon species.

2.3.4 Previous work in rainbow trout and Atlantic salmon

Previous genome repeat abundance estimates have been made for both Atlantic salmon and rainbow trout. In their 2010 announcement of a coordinated effort to sequence the genome of Atlantic salmon (Davidson et al. 2010), members of the ICSASG reported their initial estimate that 30-35% of the genome was composed of repetitive elements, based on a library created from an amalgamation of previously reported salmonid TEs and a RepeatModeler run performed on an Atlantic salmon BAC library. Similar results were also more-recently reported by Berthelot et al. (2014) in their survey of the rainbow trout genome, in which they estimated that 38% of the genome was derived from repeated DNA. This later work also posited an occasionally different breakdown of the abundance of different TE taxa within the rainbow trout genome than what I have reported here. For example, Berthelot et al. report only 6.67% of the genome as being derived from DNA transposons (vs 20.4% in the present work), with only 5.5% arising from Tc1-Mariner elements (compared to my estimate of 17%).

There are a few potential causes for the differences in repeat abundance estimates between my work and previous studies. The first of these, that the libraries I describe herein are somehow classifying non-repetitive sequences as repetitive and leading to an artificially inflated estimate of genome abundance, can be discounted for at least two reasons: i) I validated that all of the repeats in my libraries are truly repetitive, and ii) if for some reason some sequences in my repeat libraries are not repetitive, they do not contain nearly enough total sequence to mask an additional ~20% of the genome. In this latter case, consider that there are only 3.32 million nucleotides in my rainbow trout repeat library, while 386 million bases would be required for RepeatMasker to identify as repeat-derived an additional 20% of the 1.93 Gbp genome assembly if they were each used only once; clearly the majority of additional DNA masked by my library must arise

from truly repetitive library sequences. Thus, the true repeat-derived DNA fraction is almost certainly much closer to my estimate than those of previous endeavors.

After discounting the theory that my libraries are inappropriately masking non-repetitive sequence, the reason that previous efforts failed to identify a large number of repeat-derived bases is likely due to missing repeats in the underlying libraries. One factor that contributes to this deficiency is the quality of the genome sequence that is provided to *de novo* repeat-finding programs. Both the previous Atlantic salmon and rainbow trout libraries were created at least in part from BAC sequences, which are composed of smaller subsets of a much larger genome sequence. Repeat-discovery processes relying on such sequences may fail to detect repeats, particularly those that are dispersed across the genome, fragmented or present at low copy numbers. Because repeats are often difficult to resolve during genome assembly (or BAC sequence assembly), my use of the latest salmonid genome sequences also likely assisted in the recovery of more repeats. The newer genome sequences for rainbow trout and Atlantic salmon are the result of time-consuming and expensive genome creation processes that incorporate a diverse array of sequencing data and customized assembly methods. The most recent rainbow trout genome, for instance, evinces a scaffold N50 value of 1.6 Mbp which indicates a considerably more contiguous assembly than that used by Berthelot et al., which possessed a scaffold N50 of 384 kbp. A final potential source of repeat library differences is variation in the discovery process itself; all of my libraries benefitted either directly or indirectly from the inclusion of the previously-identified Berthelot et al. sequences as well as from the presence of two distinct TE discovery programs: REPET and RepeatModeler. As has been discussed previously (Hoen et al. 2015, Platt et al. 2016), this work emphasizes the fact that repeat library creation and genome annotation are sensitive to differences in their underlying methodology, and that caution must always be used in interpreting results which are the product of different analysis pipelines.

2.3.5 Comparison of salmonid TE diversity to other species

The patterns of both TE diversity and abundance I observed in salmonids are consistent with those few that have been previously described in surveys of fish genomes. In an analysis of 23 vertebrate genomes which did not include salmonids, Chalopin et al.

(2015) found that Actinopterygian species evinced an average of 24 TE superfamilies per genome, the highest amongst any vertebrate clade (tetrapods only possessed an average of 14 superfamilies). By adopting the taxonomic categories employed by these researchers (which differ slightly from those used here), I found that salmonids possess the average number of 24 TE superfamilies, a value that is much higher than that of tetrapods and lower than that of species such as zebrafish and Atlantic cod, which evince 27 superfamilies and the greatest diversity of all vertebrates examined. Chalopin et al. also found that, of four major TE clades (LTR, LINE, SINE, DNA), the genomes of teleosts were dominated by DNA elements and had particularly large contributions from the Tc1-Mariner, hAT and L2 superfamilies, all of which I also found to have made higher than average contributions to the repeat-derived portion of salmonid genomes. With only a sparse representation from many vertebrate lineages it remains difficult to speculate about the factors that determine the presence and/or absence of individual TE taxa in different lineages, however the diversity of repeats in salmonid genomes is clearly in keeping with those of most other teleost fish species.

Given the relatively large size of pseudotetraploid salmonid genomes, which commonly possess C-values in excess of 3 pg (Gregory et al. 2007), the present work reinforces the previously-observed fish association between genome size and the proportion of DNA that is derived from repeats that can be seen in Figure 3 (Chalopin et al. 2015, Gao et al. 2016, Yuan et al. 2018). With a genomic repeat fraction of at least 55%, salmonids evince one of the highest such fractions of all sequenced vertebrates, comparable only to species such as the axolotl (*Ambystoma mexicanum*) at 66%, coelacanth at ~60% and zebrafish at 52.2% (Howe et al. 2013, Nikaido et al. 2013, Nowoshilow et al. 2018). The vast majority of all other vertebrate species possess repeat fractions that are less than 50% of the genome (Berthelot et al. 2014, Yuan et al. 2018). I explore the potential mechanisms that could lead to an increase in TE-derived genome content in Chapter 3, in the context of the WGD and multiple bursts of Tc1-Mariner proliferation.

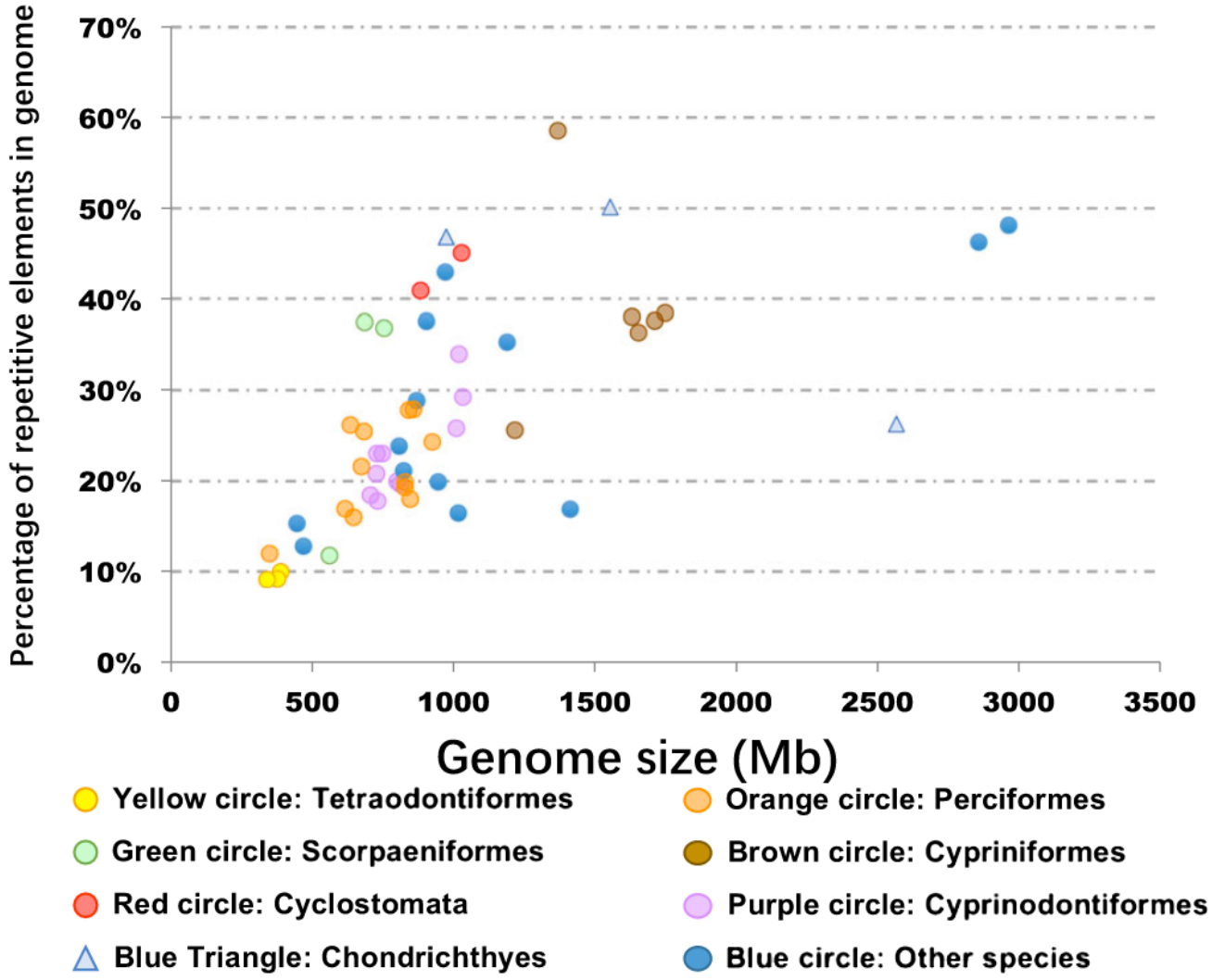

Figure 3 Relationship between genome size and repeat content in 52 fish species from Yuan et al. 2018.

2.4 Conclusions and Future Directions

The repeat libraries I present in this chapter build substantially on those that were previously developed for salmonids, and provide the resources necessary to facilitate future experiments on both salmonid and TE biology. The analyses I perform describe only the broad trends in repeat abundance and diversity which exist within salmonids, and will be greatly enhanced by future, more detailed work. Future investigations could reveal the locations and patterns of TE activity across the genome and could potentially lead to a better understanding of the forces that shape regulatory processes and the structure of genes. Further manual repeat identification and classification efforts could reduce the fraction of unknown repeats in salmonid genomes, identify repeats that have been miscategorised as a result of automated processes, and reveal any families that are unique to the salmonid lineage. With the increasingly rapid development and release of vertebrate genomes and the concomitant filling of phylogenetic gaps, fields such as comparative genomics and TE evolutionary biology promise to remain at the forefront of genomic analysis far into the future. This work provides a firm foundation for such studies, and positions this fascinating species group in a place where it will continue to offer insights into the evolutionary processes which have shaped the modern biosphere.

Chapter 3 - Tc1-Mariner proliferation and the evolution of salmonids

3.1 Introduction

3.1.1 Tc1-Mariner TEs

Repeats belonging to the Tc1-Mariner superfamily (often abbreviated termed Tc1-Mariner-like elements – TCEs) are Class II DNA TEs that proliferate in the genome by means of a 'copy-and-paste' mechanism. Although their size varies between 1 kbp and 5 kbp (Munoz-Lopez and Garcia-Perez 2010), autonomous copies of these TEs are most commonly 1.3 to 2.4 kbp in length (Plasterk and van Luenen 2002). They TIRs that vary in size from 10s to hundreds of bases (Plasterk et al. 1999), as well as characteristic two-base TSDs of their overwhelmingly preferred insertion site within the genome: a thymine followed by an adenine residue ('TA'). Autonomous TCEs possess a single ORF that codes for the transposase protein which facilitates TCE excision and re-integration elsewhere in the genome. Non-autonomous TCEs are formed when the transposase gene of an otherwise autonomous element develops a mutation which prevents functional transposase activity; such a mutation can be a relatively minor change in coding sequence, or the loss of the entire transposase ORF. As long as the TIRs of a TCE are retained, a non-autonomous element can generally continue to be mobilized by the functional transposase of an autonomous member of the same TCE family found elsewhere in the genome. Insertion target-region preference varies between different Tc1-Mariner families, with some evincing an apparently random insertion pattern and other preferring more specific regions such as introns (Zagoraiou et al. 2001, Plasterk and van Luenen 2002, Miskey et al. 2003). Additionally, TCEs are also subject to 'local hopping', in which the reintegration site of an element is more likely to be located in close proximity to its excision site than at some distant genomic location (Munoz-Lopez and Garcia-Perez 2010).

As 'cut-and-paste' transposons, TCEs can only proliferate in a host lineage through indirect mechanisms (Feschotte and Pritham 2007). One way for TCEs to increase their number is to 'jump' from a locus that has already been replicated during DNA synthesis to one which is located ahead of the replication fork (and has not yet been duplicated). By this mechanism, a TCE can increase its copy number by one when its target locus is

subsequently replicated. Another process by which TCEs can increase in copy number is through the action of DNA gap repair machinery. When a TCE exits a source locus, it will generally leave behind a DSB that must be repaired by cellular machinery in a process that is guided by following a 'template' on the homologous chromosome. If a copy of the TCE is found on the template loci (ie, if the individual was homozygous for the TCE insertion before the jump) then it is possible that the DNA repair process will replace the TCE that has jumped to a new locus.

Of all the myriad types of TEs, the structural characteristics of TCEs make them strong candidates for bioinformatics analysis. These elements almost always possess a distinctive 'TA' TSD as well as a family-conserved TIR, and often evince at least a fragment of a transposase ORF. They are shorter than elements of many other superfamilies and so are less-likely to be degraded and fragmented in the genome, and are also not subject to phenomena such as 5' truncation which can make LINEs difficult to work with. Together these characteristics imply very well-defined terminal regions that are relatively easy to distinguish from random surrounding flanking DNA in multiple-sequence alignments, and ultimately make certain important bioinformatic tasks such as copy identification and manual consensus sequence construction much easier.

3.1.2 TCE life history

Tc1-Mariner elements possess a distinct 'life cycle' within their host genome that is characterized by bursts of proliferation followed by self-induced suppression (Hartl et al. 1992, 1997, Le Rouzic and Capy 2006, Feschotte and Pritham 2007). In their initial period of activity TCEs that have evaded host-cell suppression mechanisms begin to accumulate in the genome. After a sufficient amount of time has passed, simple mutations and ectopic recombination will naturally begin to generate non-autonomous variants of the active TCE family. Once these non-autonomous elements emerge, they begin to act as an alternative target for transposase proteins generated from functional TCEs; a transposase will divide its biochemical effort between the transposition of autonomous sequences, which can in turn generate more transposase protein, and non-autonomous sequences, which cannot. Over time more and more non-autonomous elements are created, thereby increasing the likelihood of non-autonomous proliferation and concomitantly reducing the number of autonomous TEs that are duplicated. In this

cycle, non-autonomous family members are analogous to a molecular sponge that soaks up transposase activity, and eventually lead to a genome with many non-autonomous TE copies and proportionately fewer 'functional' copies. The end result of this course occurs when all copies develop mutations making them non-autonomous. This process, combined with host-cell suppression mechanisms such as piRNA and epigenetic modification, ensure that after their initial burst of proliferative activity Tc1-Mariner elements are rapidly silenced within the genome. It is also responsible for the characteristic 'starburst' pattern observed in unrooted phylogenies created using Tc1-Mariner sequences from the same family (Hartl et al. 1997, Feschotte and Pritham 2007, Pace et al. 2008); this pattern is exemplified by copies of a prolific TCE family from the genome of rainbow trout (see Figure 4a).

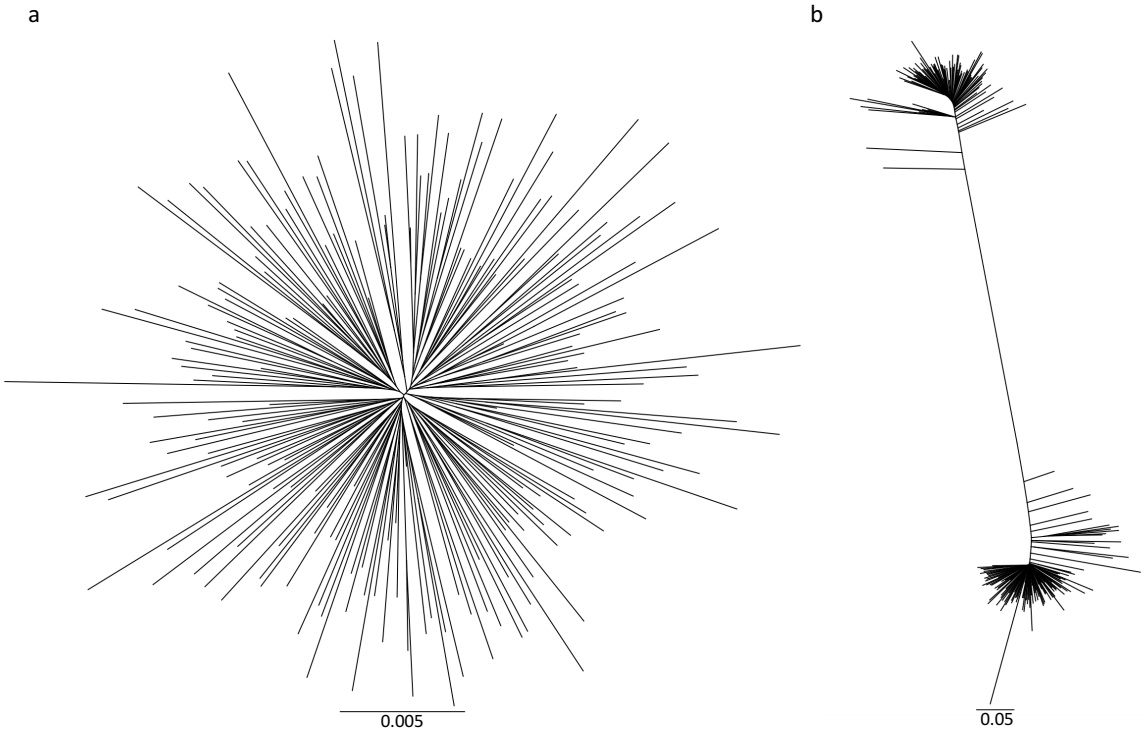

Figure 4 Unrooted NJ trees of TCE copies identified in salmon (see Methods section 3.2.4). **a**, TCE elements from a single family (omyk_TCE_37) from rainbow trout. **b**, TCE elements from two different families retrieved from the genome of arctic char using a single TCE consensus sequence derived from rainbow trout (omyk_TCE_10).

3.1.3 TCE phylogenetics and HTT

Although the first TE from the Tc1-Mariner superfamily was initially discovered in the genome of *Caenorhabditis elegans*, TCEs have now been identified in many major eukaryotic lineages including protozoa, fungi, nematodes, arthropods and chordates (Plasterk and van Luenen 2002) and are probably the most widespread DNA transposons in nature (Plasterk et al. 1999, Ivics and Izsvák 2015). In many taxa - including most mammalian lineages - Tc1-Mariner transposition has entirely ceased, however even in these groups it is clear that historical TCE activity has helped shape the architecture of the genome (Pace and Feschotte 2007, Ray et al. 2008). Curiously, phylogenetic cladograms constructed on the basis of sequence similarity between TCEs frequently indicate that elements from very distantly related host species are themselves closely related (Pace et al. 2008, de Carvalho and Loreto 2012). Such a pattern is indicative either of a selective force driving the retention of the same functional TE sequence in different lineages, or of horizontal transfer (HT) between disparate species. HT is the process by which DNA sequence from one lineage makes its way into the genome of another by a mechanism independent of the standard germ line transmission of genetic HTT is facilitated by viruses or other parasites which infect multiple hosts (Gilbert et al. 2014, Panaud 2016). Under the proposed viral mechanism, a TE would insert itself non-offensively into functional viral DNA which is then packaged into viral particles before infecting a member of a different species. Following this infection, the virus integrates into a new cell while at the same time introducing its TE passenger to a naïve genome, thereby accomplishing HTT. Many recent studies have demonstrated that, like most TEs, the vast majority of Tc1-Mariner elements are under very little selection at the host level and thus that the many examples of phylogenetic anomalies involving TCEs are almost certainly a result of HTT (Ivics et al. 1996, Schaack et al. 2010, Peccoud et al. 2017). In salmon species specifically, a number of cases of HTT involving TCEs have been proposed (Goodier and Davidson 1994, de Boer et al. 2007).

3.1.4 Sleeping beauty

Tc1-Mariner elements within fish, and salmonids in particular, have historically been of particular note as a result of their contribution to the Sleeping Beauty transfection system. The central component of the Sleeping Beauty system is a functional TCE that

was ‘resurrected’ by aligning and generating a consensus sequence from non-functional ancient copies of the element obtained from a number of different salmonid species (Ivics et al. 1997, Ivics and Izsvák 2015). In the laboratory transfection system, the transposase gene from the resultant Sleeping Beauty transposon is included on a plasmid, separate from a gene of interest that is flanked by the Sleeping Beauty TIRs. Once inside a target cell the transposase protein will excise the target gene and insert it randomly into the cell’s genome. Salmonid-based Sleeping Beauty was the first Tc1-Mariner superfamily member shown to be active in vertebrates, as well as the first gene reconstructed using inactive DNA with an archaic origin. It has been used for many tasks including the generation of transgenic cells in tissue culture, the production of germline-transgenic animals, and therapy of genetic disorders such as lymphoma in humans (Ivics and Izsvák 2015, Kebriaei et al. 2016). Since its development, similar transposon transfection systems have been developed around other TCE families from a variety of species – examples of these include *Frog Prince*, *Minos* and *Himar1* (Zagoraiou et al. 2001, Miskey et al. 2003, Maier et al. 2006).

3.1.5 TCEs in the salmonid genome

Since the first descriptions of individual Tc1-Mariner elements (Goodier and Davidson 1994, Radice et al. 1994, Ivics et al. 1996) in salmonids, the number of TCE families identified in this group has slowly grown (de Boer et al. 2007). In the previous chapter I observed that TCEs are by far the most prolific TE superfamily within the genome of five salmonid species: Atlantic salmon, arctic char, rainbow trout, coho salmon and chinook salmon. Given the marked abundance of these elements in salmon genomes, their intriguing life history characteristics and the previous speculation that they may have played a role in salmonid speciation (de Boer et al. 2007), I performed a detailed analysis of their historical activity. In this chapter I describe the identification of 60 Tc1-Mariner families and the construction of a timeline of their historical activity in the salmonid genome. I find that TCEs have proliferated in major bursts, which occurred coincidentally with the salmonid WGD and continued during speciation. I examine the more recent TCE activity that continued throughout salmonid speciation and resulted in occasionally enormous within-family variation in TCE copy number between different salmonid lineages, a trend which is exemplified by the particularly notable expansion of a

TCE family which occurred recently in the rainbow trout lineage. Finally, I discuss the potential causes and impacts of the many waves of Tc1-Mariner expansion that have shaped salmonid genomes and have undoubtedly affected the evolutionary trajectory of this important taxa.

3.2 Methods

3.2.1 Creating a Tc1-Mariner curated library

In order to investigate patterns of Tc1-Mariner element activity in salmonid genomes, I first required manually-curated repeat libraries containing consensus sequences for each TCE family. I constructed such libraries independently for both Atlantic salmon and rainbow trout; TCEs identified in these two genomes were then used to interrogate the genomes of these as well as other salmonids.

For each of Atlantic salmon and rainbow trout, a set of 'seed' repeats was first identified by extracting Tc1-Mariner sequences from the REPET (Atlantic salmon) and/or RepeatModeler (Atlantic salmon, rainbow trout) repeat libraries described in the previous chapter. Although those libraries facilitate a good preliminary picture of the repeats within their respective genomes, they in many cases contain incomplete sequences, nested multi-superfamily chimeric TEs, and at least partially redundant repeats. For more detailed analyses of a specific repeat superfamily, such as Tc1-Mariner, a more exacting curation process must first be performed. The Tc1-Mariner seed sequences formed the starting point for this curation.

All TCE seed sequences were initially aligned to their respective species' reference genome using BLASTN (no dust filter; wordsize 7). For Atlantic salmon, a preliminary draft genome assembly was used that was made available through the ICSASG (Davidson et al., 2010). The rainbow trout sequence was a close-to-final version of the sequence that is now available in Genbank with accession GCF_002163495.1. For each seed repeat, the sequence corresponding to those HSPs which occupied 60% or 80% of the query sequence (for Atlantic salmon) or 70% of the query (for rainbow trout) were extracted. For rainbow trout, HSPs were also required to possess a percent identity of at least 60% to the query sequence. Once these sequences were identified a subset of them were randomly chosen; these correspond to a collection of genomic instances of a given TCE family. Subsequently, the corresponding genome sequences plus both their upstream and downstream flanking regions were extracted, and all instances associated with a particular seed TCE were aligned using MUSCLE v3.8.31 (Edgar 2004). The flanking sequences were included in the alignments in the hope of extracting full-length element copies from partial fragment seed TCEs. Following these preliminary steps, I

manually examined the alignments corresponding to each TCE and visually identified the regions where sequences ceased to be similar and became a collection of apparently random bases (see Figure 5) – such boundary regions corresponded to the 5' and 3' flanking regions of full-length TCE copies. Once identified, I removed the non-repetitive DNA portion of the sequences, leaving a collection of aligned full-length TCE instances which could be processed to create a single representative consensus.

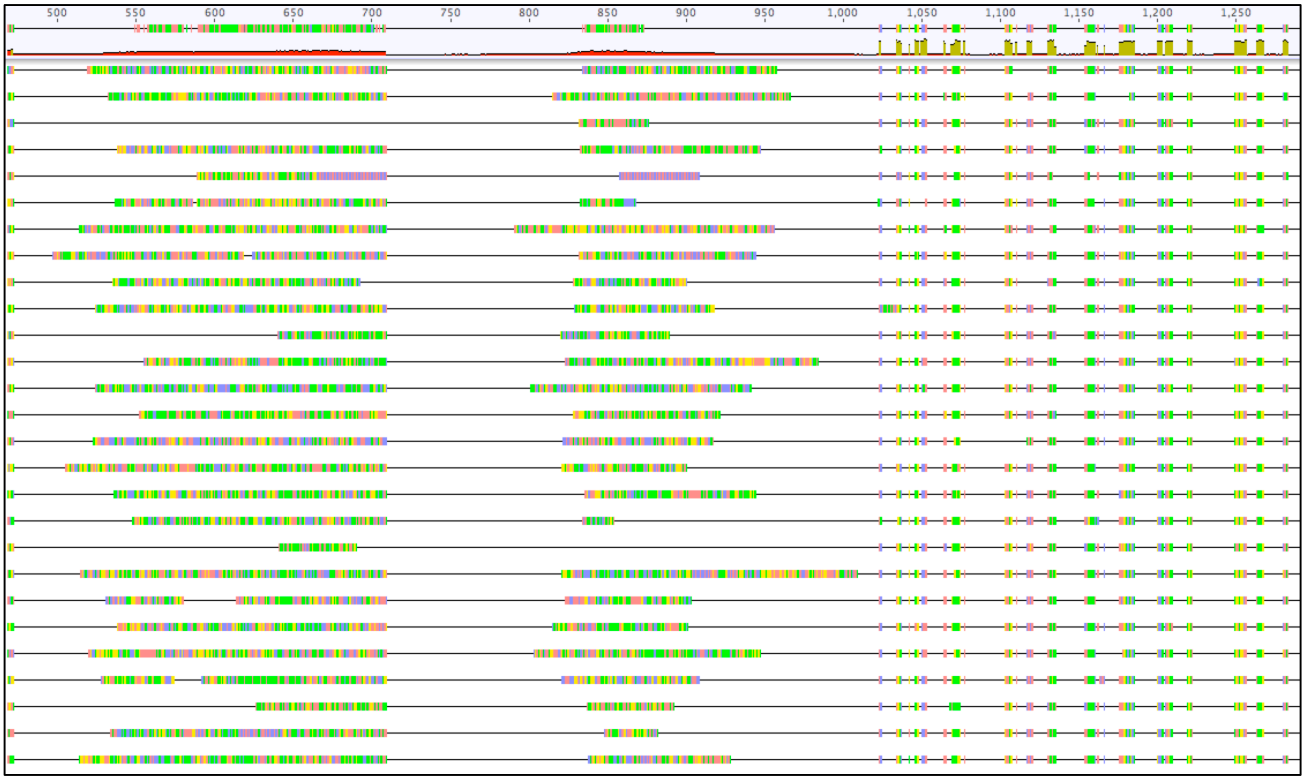

Figure 5 Geneious multiple sequence alignment of members of a single TCE family plus flanking regions. Different rows correspond to individual TCE instances, colours correspond to different nucleotide bases and each column represents an alignment position. The transition between presumably non-TCE flanking bases and TCE sequence is readily apparent; in this alignment, flanking sequence occurring before base ~1,020 would have been removed (Kearse et al. 2012).

During this labour-intensive process, obvious alignment subgroups would occasionally emerge that corresponded to TCE instances from two or more distinct families. Such sequences generally shared some region that was common to all of the families (contained within the original seed query sequence) even though they diverged to varying degrees outside of this conserved region. In the cases where only a minority of sequences were notably different from the others, I removed the offending sequences. In other situations (such as when there were two large distinct groups), I typically trimmed a conservative amount of random flanking sequence and then performed a re-alignment from which I constructed an unrooted Neighbour-Joining (NJ) tree using the Jukes-Cantor genetic distance model in Geneious v7 (Kearse et al. 2012). These trees revealed obvious groupings of TCE instances (see Figure 4b) which were then extracted into separate sub-alignments. From then on, each of these sub-alignments were all treated as if they originated from their own seed Tc1-Mariner repeat sequence.

A significant difficulty occurred in some cases in which the TE/random sequence boundary for an instance was exceptionally unclear, appearing ‘staggered’ from that of other instances in the alignment. Different factors could contribute to this state: instances from more anciently-active TCE families could be highly fragmented, or the instances under consideration could have originated from a non-Tc1-Mariner group that is characterized by random truncation of its elements, such as the LINE superfamily. In the cases where this obfuscation occurred, efforts were made to identify even a small number of instances which shared common 5’ or 3’ end sequence. These longer instances then formed a guide for the other sequences in the alignment, which were each individually trimmed wherever their random (flanking) DNA began to resemble that of the guide sequences. In cases where establishing a firm boundary was not feasible, a point was identified at which most instances exhibited a relatively conserved motif and all alignment bases past this point were removed. This latter, last-resort approach inevitably and unfortunately gave rise to only partial consensus sequences.

Once high-quality alignments were created with well-defined TE boundaries, consensus sequences were generated. Consensus construction followed the ‘50% - Strict’ threshold in Geneious v7: if at least 50% of the bases in any alignment column were the same symbol (A, T, C, G, ‘-‘), the corresponding consensus base was chosen to be that

symbol; otherwise it was marked as ambiguous ('N'). Subsequently, all consensi were aligned to a high-quality reference database of TE protein sequences (REPET-formatted RepBase version 20.05; W. Bao et al., 2015) using BLASTX in order to orient all of sequences in the same direction, and to remove those which were even partly composed of non-Tc1-Mariner sequence. First, the BLASTX HSP with the lowest E-value was identified for each repeat; this corresponded to TCE transposase sequence. If the best BLASTX HSP was located on the negative strand of the TCE consensus, the reverse complement of the consensus was used going forward. Any sequences generating a non-Tc1-Mariner HSP with an E-value less than $1e^{-5}$ were removed as possible chimeras.

The final step of TCE library construction was the identification and removal of redundant repeats. Many of the initial seed library sequences were fragments from different regions of the same element and as a result when they were each individually extended to full-length TCEs, many consensus sequences would emerge for a single family. To identify such consensus sequences belonging to the same family, all sequences were aligned using MUSCLE and a phylogenetic tree was constructed using the NJ method with a Jukes-Cantor distance model. Sequences from the same family were exceptionally similar (at least 97%) and grouped into very apparent nodes within the resulting tree. The longer sequence in each group was retained; the others were discarded.

Once the TCE consensus curation process was completed for Atlantic salmon, the resulting sequences were used as input seeds for the subsequent creation of the rainbow trout TCE library.

3.2.2 Creating a combined salmonid Tc1-Mariner library

To create a consistent standard by which to assess the relative impacts of TCE activity on genome composition in different salmon species, I combined the Atlantic salmon and rainbow trout TCE consensus libraries. This amalgamation was performed by: i) aligning all consensus sequences using MUSCLE and creating a NJ tree using Geneious (Jukes-Cantor distance model); ii) visually identifying the obvious very closely-related sequence pairs; iii) creating a two-sequence alignment for each such pair, and; iv) randomly removing one sequence from any {Atlantic salmon, rainbow trout} pair with a similarity higher than 97% to avoid redundancy. It is important to note in downstream analyses that

this library may be missing consensus sequences for TCE families that have been very recently active in the genomes of arctic char, coho and chinook; a full TCE identification process was not performed for these species.

3.2.3 Reconstructing Tc1-Mariner activity in the Atlantic salmon and rainbow trout lineages

In order to establish a timeline of historical activity for those Tc1-Mariner elements that are present in modern salmonid genomes, I investigated the pairwise similarity between TCE elements within each family. Pairwise similarity or p-distance, which is the proportion of nucleotide bases that are the same between a pair of aligned sequences, can be used as a useful proxy for the relative time that has passed since a duplication event occurred that resulted in two identical homologous sequences. This method relies on the observation that, absent natural selection and certain types of recombination, the random accumulation of single-base mutations over time will drive two once-identical sequences apart. As a result, sequence duplication events can be placed on a relative timeline in which more recently-arising homologues have a higher percent similarity than those that are more ancient. Using the combined salmonid TCE library, I performed a percent similarity analysis for TCE families in both the rainbow trout and Atlantic salmon genomes. This analysis was bolstered by the incorporation of genomic abundance information for each TCE family, which helped to paint a complete picture of the magnitude and timeline of historical TCE activity in salmonid genomes.

For both Atlantic salmon and rainbow trout, this analysis followed three principle steps: 1) obtain genome abundance estimates for each family; 2) obtain an aligned sample of representative instances for each TCE family; and, 3) combine abundance and pairwise similarity data to construct a complete timeline of historical activity. All steps in this process utilized the most recently Atlantic salmon and rainbow trout assemblies in Genbank, which have accessions GCF_000233375.1 and GCF_002163495.1, respectively.

3.2.3.1 Obtaining genome abundance estimates for each family

Using the combined salmonid TCE library, RepeatMasker v4.0.8 (Smit, AFA, Hubley, R & Green 2015) was used to mask the genomes of either Atlantic salmon or rainbow trout and to identify individual TCE copies. Subsequently, the number of bases in the

genome attributed to each family was summed by parsing the RepeatMasker *.out file. As mentioned in Chapter 2, individual RepeatMasker annotations can infrequently overlap by as much as 80%; accordingly, TCE abundance estimates are not necessarily additive.

3.2.3.2 Obtaining an aligned sample of representative instances for each TCE family

TCE instances for each family were identified by first aligning curated Tc1-Mariner consensus sequences to the genome of either rainbow trout or Atlantic salmon using BLASTN. All HSPs were retained that were at least 60% of the query consensus sequence length, and the corresponding genomic sequences were extracted; in order to obtain older TCE copies that may have diverged significantly from the consensus sequence, no E-value or percent identity thresholds were used. Because it was possible that HSPs for multiple families overlapped the same genomic region, an additional reciprocal best hit (RBH) quality control step was performed. In this step, all extracted genomic instances were aligned to the original TCE consensus library with BLASTN, and the only genomic instances retained were those whose highest-scoring HSP was to the original sequence that first identified it (ie the reciprocal best BLAST - RBH - hit).

Once genomic instances were obtained, up to 200 were randomly selected for each family and aligned using MUSCLE. This sample of 200 (or fewer) elements was used because of the expenses incurred in both accuracy and computational resource use when very large sequence alignments are performed. From the MUSCLE alignments, the percent similarity was determined between each pair of sequences in the alignment. When calculating this pairwise similarity, both gaps and ambiguous bases in a particular alignment position caused that alignment position to be ignored.

3.2.3.3 Combine abundance and pairwise similarity data to construct a timeline of historical activity

For both Atlantic salmon and rainbow trout, visualized historical Tc1-Mariner activity in the genome using a stacked density plot created using the ggplot2 package in R (Wickham 2009). In this visualization method, individual density plots were first constructed for each family using all of their constituent pairwise similarity values, and then scaled by the abundance of their particular family in the genome as inferred from

RepeatMasker. The resulting plots were then 'stacked' in order to reveal overall TCE trends.

3.2.4 Comparing TCEs between species

In order to characterize patterns of TCE activity across the salmonid lineage, I analyzed TCE pairwise similarity and abundance in the genomes of Atlantic salmon, arctic char, rainbow trout, coho and chinook. To this end, the TCE consensus sequences within the Atlantic salmon/rainbow trout combined library were separately aligned against the genomes of all five salmonid species using BLASTN. The genome sequences used for this step were obtained from Genbank in early 2018 and possess the following accessions: GCF_000233375.1 (Atlantic salmon), GCF_002910315.2 (arctic char), GCF_002163495.1 (rainbow trout), GCF_002021735.1 (coho), and GCA_002872995.1 (chinook). HSPs were filtered, with those overlapping at least 60% of the query TCE sequence being retained. The genomic sequence of each remaining HSP was then extracted; these sequences correspond to TCE copies within the genome. A RBH procedure was subsequently performed in order to ensure the correct TCE family assignment for each TCE copy. This was accomplished by aligning the extracted genomic sequences back to the original TCE consensus library using a second BLASTN search, and only retaining those genome sequences with a highest-scoring hit back to the TCE consensus that allowed its identification in the first BLASTN search.

Once identified, 200 random TCE copies for each family in each salmonid genome were aligned to each other using MUSCLE and the percent similarity between each pair was calculated. The abundance of each TCE family in each genome was obtained by using RepeatMasker.

3.2.5 Creating a reference point for TCE activity – the salmonid WGD

Although a relative timeline of Tc1-Mariner activity is valuable on its own to understand the historic forces shaping salmonid genomes, I also sought to add a 'fixed landmark' to the timeline in the form of the salmonid-specific WGD. To integrate both timing estimates, I had to create an estimate of when the WGD occurred using a similar metric as that used for the TCEs. I accomplished this by comparing sequence that was unlikely to be under significant selection, in the form of introns, between gene duplicates

('homeologues') that originated at the WGD. In a similar way to TCEs, such duplicate sequences would be expected to randomly diverge from each other with each successive generation, thereby facilitating the use of pairwise similarity as a proxy for time. This analysis was performed on the Atlantic salmon genome only, and was guided by previous approaches (Hoffman and Birney 2007, Resch et al. 2007).

The first step in this approach was to identify a set of homeolog pairs suitable for pairwise similarity-based dating. Using NCBI RefSeq annotation information, the longest protein isoform for each gene in the Atlantic salmon genome was identified. These sequences were aligned against themselves using the BLASTP tool in accordance with an all-by-all BLAST approach. The resulting HSPs were then filtered; hits were removed if they possessed a E-value greater than $1e^{-10}$, if their similarity was less than 60%, if the query or subject coverage was less than 50%, or if the query and subject didn't originate from Atlantic salmon homeologous paired regions identified previously (see Lien et al., 2016). This initial work was largely performed by Simen R. Sandve, Gareth Gillard and Torfinn Nome of the Norwegian University of Life Sciences, and produced a set of putatively homeologous proteins (and associated genes) for further work.

Next, I endeavored to identify intron pairs that I could confidently assess as being homeologous. This was made challenging by the fact that intron-exon boundaries, as well as the existence of specific introns at all, could have changed in the time since the WGD occurred. My approach relied on the identification of homeologous exon sequences that flanked both ends of a putative homeologous intron. Because exons are composed of coding sequence that is presumed to be constrained by natural selection, two homeologous exons are much more likely to remain similar to each other than are nearby introns, which are expected to exhibit confounding base substitutions and indels which might prevent the easy identification of homeologous regions. In this strategy exons therefore act as 'anchors', flanking intronic sequence which can then be aligned with a large degree of confidence in its homeologous nature.

For each homeologous protein identified in the previous step, I obtained the exon sequences from the corresponding transcript (in cDNA form) using the NCBI RefSeq genome annotation (O'Leary et al. 2016). The library of these exon sequences was

aligned to itself using BLASTN in an all-by-all manner and any HSPs with a query or subject coverage less than 80% were discarded, as were any query-subject pairs that were not between the previously-identified homeologous pairs. Following this, pairs of HSPs were identified such that their corresponding exons in each homeolog followed each other in the genome. An example of such a pair would be an HSP in which exon 1 of some protein A matches with exon 3 of a homeologous protein B, and another HSP in which protein A exon 2 matches with protein B exon 4. These pairs correspond to anchor exons in the two homeologous genes, both of which flank putatively homeologous introns. Where homeologous flanking anchor exons were discovered, the corresponding intron pairs were identified. The introns were subsequently extracted from a genome that had been previously masked with RepeatMasker, using the Atlantic salmon repeat library described in the previous chapter.

Extracted introns had the terminal 30 bp removed from both their 5' and 3' ends – such regions are more likely to contain splice signaling motifs under selection (Hoffman and Birney 2007) which could potentially interfere with my objective of obtaining neutrally-evolving homeologous sequence. Intron pairs which included intron 1 were excluded for similar reasons (Park et al. 2014). Following end removal, the repeat-masked homeologous introns were aligned against each other using the LastZ tool v. 1.02 (Harris 2007). LastZ is similar to BLAST in that it computes local alignments, however it is better able to create single alignment blocks that span large insertions or deletions between two sequences. The LastZ-aligned sequence blocks were subsequently processed and the average percent similarity for each block was determined. In the final step, I removed the small minority of intron alignments over 10 kbp (which are more likely to engender alignments between non-homeologous sequence such as tandem duplicates) and created an alignment length-weighted percent-similarity histogram using the ggplot2 package in R.

3.3 Results and Discussion

3.3.1 Properties of TCEs

A detailed investigation of historical trends in Tc1-Mariner activity within salmonid genomes required a library of higher quality than that created for the initial genome-wide TE analyses described in the previous chapter. To this end, I created separate manually-curated Tc1-Mariner consensus sequence libraries for both Atlantic salmon and rainbow trout, as well as a non-redundant combined library containing sequences from both species (Table 5). The final Atlantic salmon library contained 40 consensus sequences which ranged in size from 523 bp to 2,944 bp with a median length of 1,628 bp, while the 48 sequences from the rainbow trout library had a median length of 1,607 bp and were between 535 bp and 2,653 bp long. Consistent with its two component libraries, the final non-redundant omyk+ssal combined library contained 60 repeat sequences with a median length of 1,610 bp (min: 523 bp; max: 2,944 bp). The vast majority of identified Tc1-Mariner consensus sequences were found to be within the 1.3 kbp to 2.4 kbp range (Plasterk and van Luenen 2002) that is typical for repeats of this superfamily; 31 (78%) Atlantic salmon TCEs and 35 (73%) rainbow trout TCEs were within this interval. Those sequences with more-extreme lengths are potential candidates to represent incomplete TCE consensi, or non-autonomous families of TCEs that have either lost important transposase ORF sequence or acquired extraneous tracts of bases which may interrupt their autonomous function.

Table 5 Properties of manually-curated TCE libraries

TCE Library	No. of consensus sequences	Min Length (bp)	Max Length (bp)	Median Length (bp)
Atlantic salmon only	40	523	2,944	1,628
Rainbow trout only	48	535	2,653	1,607
Combined	60	523	2,944	1,610

Of the 60 sequences present in the combined TCE library, I found that 50 (83%) showed evidence of TIRs following a visual examination of dotplots created using the Geneious software package (for example, see Figure 6). The large number of sequences possessing intact TIRs lends confidence to the methods I used to identify full-length sequences. The length of the TIRs themselves vary from 22 bp to an extreme of 764 bp, with the vast majority falling into one of two groups: those that are approximately 28 bp and that are approximately 213 bp. These lengths are consistent with the previously observed length range of 17 bp to 1,110 bp (Munoz-Lopez and Garcia-Perez 2010). The 10 sequences in which no TIRs were observed are almost certainly fragments for which the whole consensus was not identified; with only one exception they are all notably shorter than the average library length, and possess a median length of 946 bp.

3.3.2 TCE family activity

Tc1-Mariner activity within a host genome is frequently characterized by a star-like topology pattern that appears in unrooted phylogenies created from individual TCE copies (Pace et al. 2008). I observed this pattern for the vast majority of TCE families I identified in Atlantic salmon and rainbow trout (for example, see Figure 4a). This pattern, in which the majority of final branches (leaves) are roughly equal in length and the branching of the tree occurs in close proximity to the sequences' common ancestor, is indicative of a rapid period of sequence duplication followed by a relatively longer phase of mutation accumulation and divergence. The presence of this sequence relationship for TCE families provides confidence that there is truly only a single family represented by each library consensus sequence, a property which is important for the interpretation of my historical activity analysis. In the few cases where my TCE family instances deviate from this pattern, two cases are possible. First, the family may have been active recently-enough that individual branches have not yet had the chance to 'grow' by acquiring neutral mutations. Alternatively, there may be an undescribed family whose members are being grouped with a described family. A very extreme example of this arrangement (which was removed from analyses) can be observed in Figure 4b; the sequences in this figure originate from arctic char, where there are clearly two true families associated with a single consensus sequence.

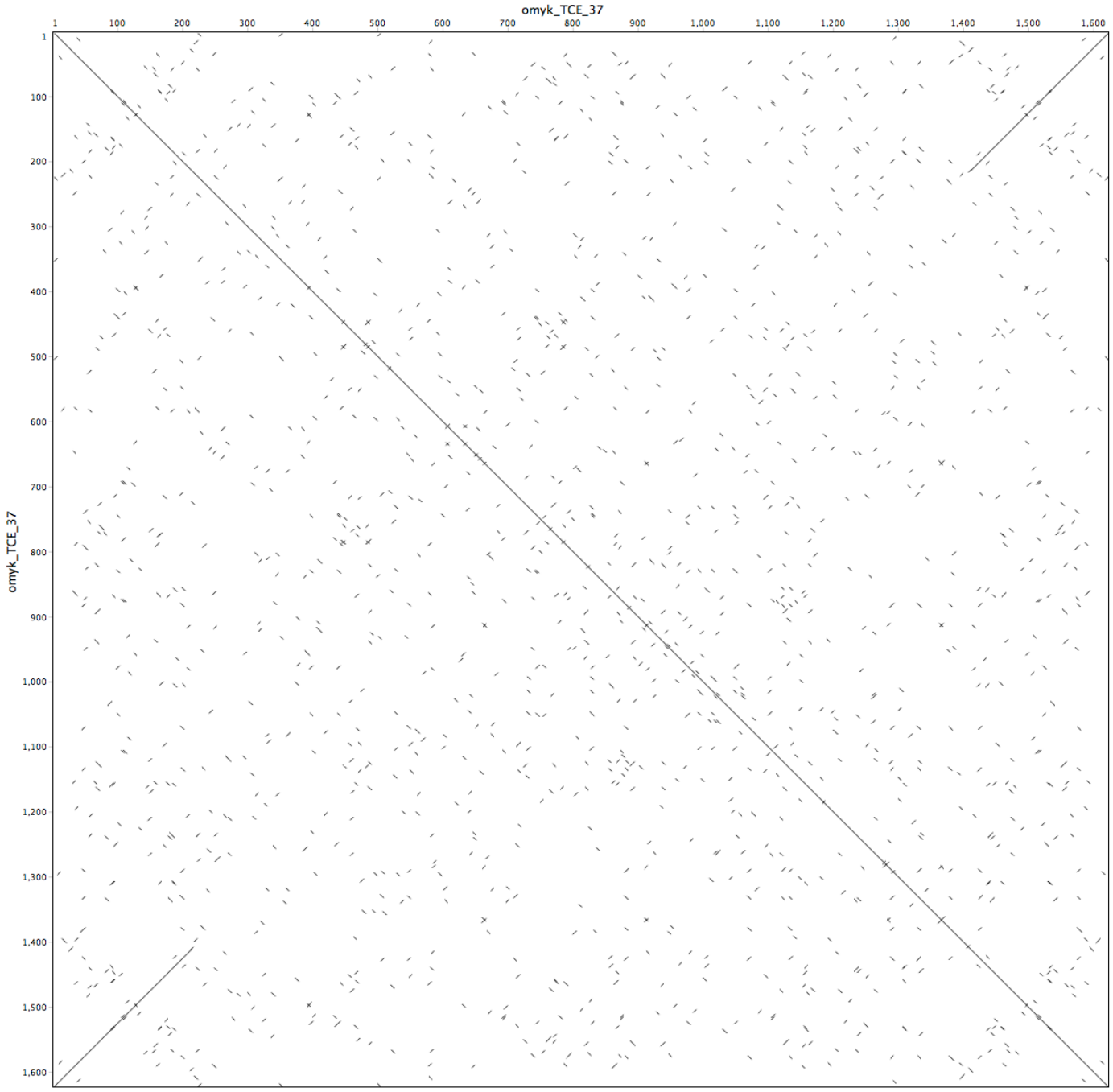

Figure 6 Geneious dotplot of a single TCE consensus sequence (omyk_TCE_37) compared to itself. Regions of similarity between two sequences are marked with a dot. Lines progressing from upper left to lower right correspond to regions of similarity between the forward orientations of two sequences, while lines progressing from lower left to upper right correspond to similarity between a forward-oriented sequence and a reverse-complemented sequence. The TIRs for this TCE are clearly visible and approximately 215 bp in length.

In order to assess historical trends in Tc1-Mariner activity, I identified a subset of genomic copies for each family, aligned them, and estimated their relative age in the genome by using the pairwise similarity between each pair of family members as a proxy for time. By creating a density plot of these pairwise similarity values, and scaling this plot by the abundance of each family in the genome, I created an approximation of the magnitude and timing of duplication events of TCEs in both the Atlantic salmon and rainbow trout lineages (Atlantic salmon: Figure 7a; rainbow trout: Figure A1 in the Appendix). The relationship of these historical trends to the salmonid lineage-specific WGD event was determined by calculating the pairwise similarity values between presumed neutrally-evolving introns from WGD-duplicated genes, and visualizing this information alongside the Atlantic salmon TCE activity plot (Figure 7b).

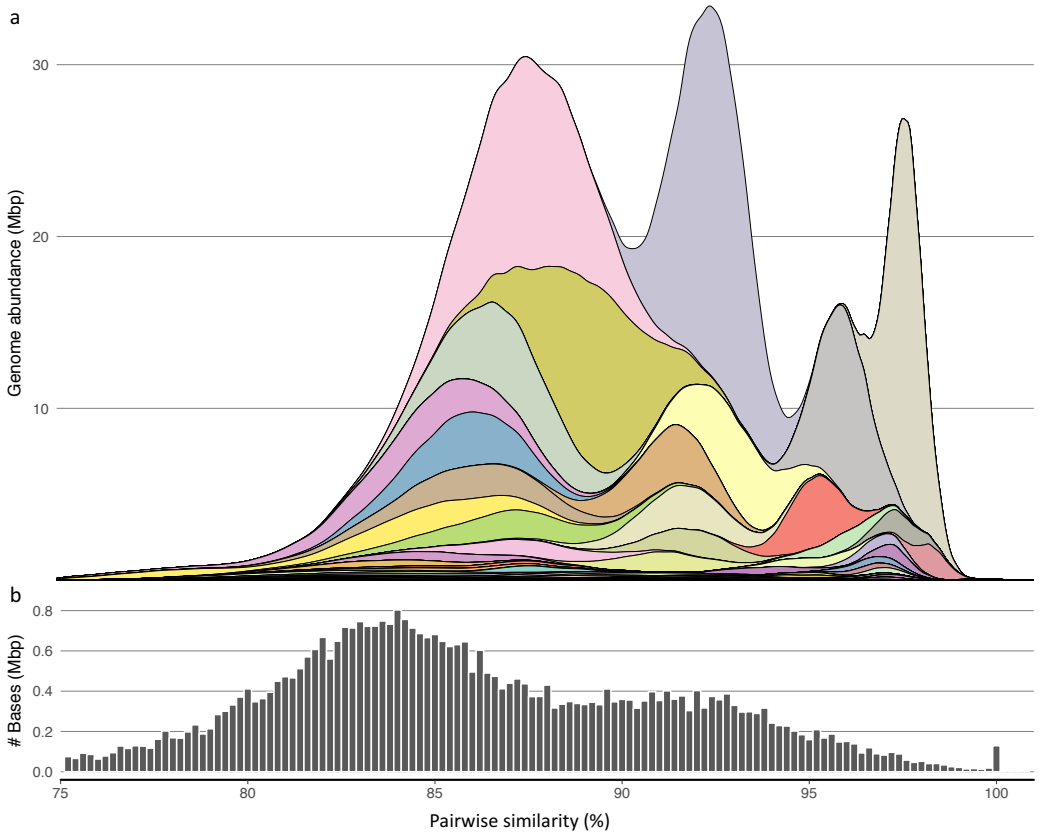

Figure 7 Historical TCE proliferation in the context of the 4R WGD. a, Stacked TCE age plot scaled by genome abundance. Each colour represents an individual TCE family. Historical age of families was estimated by calculating the pairwise similarity between a random subset of family members; if family members were less similar they were created in more ancient duplication events. **b**, Percent similarity between homeologous introns arising from the 4R WGD.

Historical estimates of both the timing of the WGD and ancient TCE activity are present in Figure 7, and reveal a striking apparent correlation: a significant portion of the TCE copies within present-day salmonid genomes were created in enormous bursts of activity, the largest of which occurred either coincidentally with or shortly following the WGD. The WGD-origin component of the figure has two distinct peaks – a major peak at approximately 84% similarity as well as a secondary peak at approximately 92% similarity. This distinct distribution of homeolog sequence similarity has been noted previously and is the result of different homeologous regions of the ancestral salmonid genome undergoing rediploidization at different times; WGD homeologs could not begin to diverge from one another while in a state of autotetraploidy that facilitated recombination between these sequences. The first, larger peak corresponds to the earliest sequences to begin transitioning to a diploid state, and as a result it is the most relevant point for comparisons to patterns of TCE activity. With this understanding, Figure 7 reveals a possible relationship between the WGD and TCEs in the genome. The most conceptually straightforward explanation of this relationship is that, from relatively low levels of activity prior to the WGD, transposition rates increased markedly near to or following the time when the WGD occurred. Assuming this explanation is true, TCE families continued to expand in the genome after the WGD in prolific bursts up until the present day. An alternative mechanism that could also explain the observed trends requires changes in the removal rate of TE sequence from the genome. If TCE family members were removed from the genome at varying rates over time, an apparent peak in TE abundance could reflect a very low level of repeat removal from a particular epoch, rather than a high level of TE proliferation. If this mechanism were the predominant force producing the trends observed in Figure 7, it would imply a very high degree of ancient TCE removal which was greatly reduced in temporal proximity to the WGD, and which continued to be relatively relaxed into the modern era.

3.3.3 Confounding factors

Caution must be applied in interpreting Figure 7, as both technical and biological factors could potentially confound the results. The primary technical concern involves the method by which TCE copies were identified in the genome. In my approach, I used each TCE family consensus sequence in the combined Atlantic salmon/rainbow trout

library as a query for a BLASTN search, and identified as TCE copies any genomic regions which aligned to more than 60% of this query sequence. This 60% threshold was chosen as it ensures that at least some portion of all copies overlap during the subsequent alignment creation step. The threshold could be potentially problematic, however, as it biases selection towards longer TE copies that are less likely to have experienced a fragmentation event. Longer TE copies are disproportionately likely to have been deposited more recently in the genome, as the potential for a given TE copy to have experienced a confounding mutation increases over time. A comprehensive assessment of the impact of this ascertainment bias would require a detailed investigation of historical deletion and recombination rates in salmonids, which does not at present exist. Suffice to say, the number of transposition events is likely to be increasingly underestimated with greater historical age.

The percent similarity estimates for WGD and TCE activity are also likely to be affected by two additional confounding forces: recombination and gene conversion. Both of these processes can involve the exchange of genetic material between homeologous sequences that pair during meiosis, which through different mechanisms has the effect of homogenizing duplicated sequences. Sequences which are constrained together this way are said to be undergoing concerted evolution. Recombination and gene conversion can act both in an allelic fashion, in which DNA exchange occurs between the same loci on homologous (or homeologous) chromosomes, or in an ectopic (non-allelic) way, where exchange occurs between other (highly similar) sequences. Following the WGD, the ancestral salmonid genome entered a tetraploid state during which recombination enforced similarity between homeologous chromosomes. Over time, different regions of the genome began to revert back to a diploid state under which homeologues were free to diverge, however this process was initially delayed and occurred well after the WGD (Macqueen and Johnston 2014, Robertson et al. 2017). As a result of this period of homeologous recombination, the degree of percent divergence between homeologues is not a perfect linear proxy for time; the WGD likely occurred earlier than it appears in Figure 7. In a similar although certainly less pronounced fashion, ectopic gene conversion, which is known to occasionally occur between TE copies, may have caused

an underestimation of how far in the past some TCE activity occurred (Roeder and Fink 1982, Webster et al. 2005, Ellison and Bachtrog 2015).

3.3.4 Patterns of TCE activity across the salmonid lineage

In addition to examining historical TCE trends within individual salmonid species such as Atlantic salmon and rainbow trout, I also investigated both the abundance and timing of TCEs between these and three additional salmonid species: arctic char, chinook salmon and coho salmon. The results of this comparison can be seen in Figure 8; an alternative version of this figure with outliers present is found in Figure A2 of the Appendix.

A trend is immediately apparent when comparing the activity periods of the five salmonids for each TCE family: the *Oncorhynchus* species evince consistently more divergent sequences than do Atlantic salmon and arctic char. For example, in rainbow trout, coho and chinook copies of ssal_TCE_26 are on average 90% similar to each other, while copies in the Atlantic salmon and arctic char genomes are about 91.5% similar. Given the consistency of this pattern even in very old families that were presumably active in the common salmonid ancestor, the probable explanation is that species within the *Oncorhynchus* lineage have experienced an increase in their effective mutation rate since their divergence from the other lineages. With this caveat in mind it is also apparent that before the time corresponding to 94% similarity the ancestral genomes of all five species experienced the same relative rates of TCE activity, at roughly the same times. This observation implies that TCE activity before this point was occurring in the common ancestor of all of these salmonids.

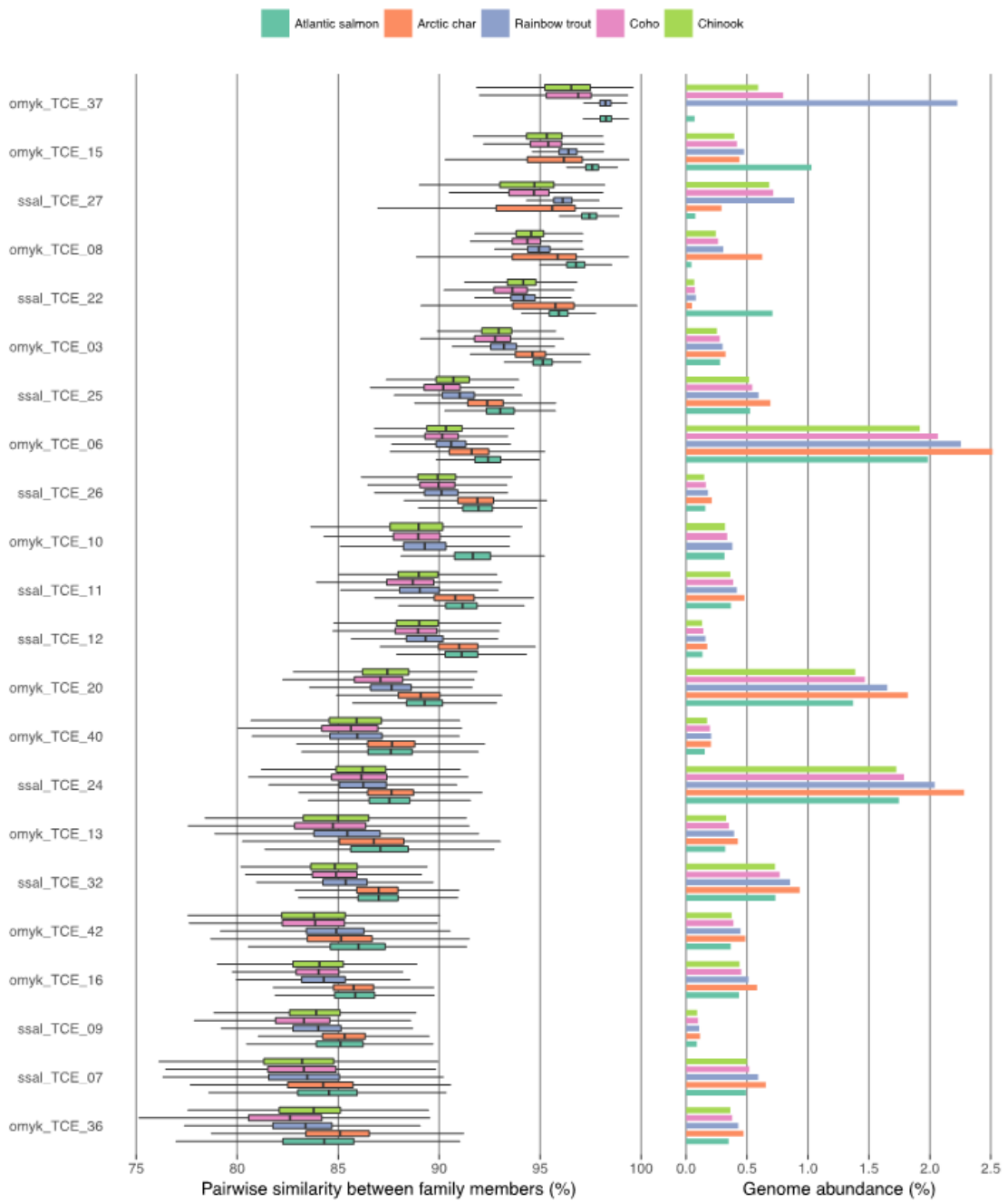

Figure 8 Age and abundance of TCEs in the genomes of five salmonids. Only those families occupying more than 0.1% of any genome are shown. Families with increased pairwise similarity between members have experienced less sequence divergence since they were rendered inactive and reflect more recent additions to the genome. No intact elements from family omyk_TCE_37 were found in the genome of arctic char. Elements assigned to omyk_TCE_10 in arctic char are not shown as they clearly include members from two different families (see Figure 1b). For clarity, outlier pairwise similarity values are not displayed however they are included in Figure A2 of the Appendix.

The time period after 94% sequence similarity, corresponding to lineage diversification in salmon, is rife with examples of TCE activity that differs between individual species for the same TCE family. The variations in genomic abundance are self-evident, and imply large differences in the degree of proliferation of different TCE families in different lineages. The observation that the interquartile range is suddenly and dramatically broader in some species past this time point, however, requires further explanation. This trend, which occurs only in arctic char, coho and chinook, likely appears because consensus sequences are not present in the combined Atlantic salmon/rainbow trout library for all of the TCE families within these lineages. In cases where families have markedly diverged or in which HTT has introduced a new family, that family will be present in the genomes of these species, but not represented in the query TCE library. As a result, when family copies were obtained from the genome to create Figure 8, copies from multiple families could be grouped into a single family/box - but generally only in arctic char, coho and chinook. This in turn led to inflated percent similarity values between members of the two different families. For an extreme visualization of this effect see Figure 4b, which contains the TE copies that were identified in the arctic char genome using the omyk_TCE_10 family consensus from the combined Atlantic salmon/rainbow trout library. Clearly this single consensus sequence is identifying members of two different families. The copies of omyk_TCE_10 from arctic char were not included in Figure 8, as they were so divergent that the dramatically disrupted the visualization.

The dramatic differences in proliferation that have occurred in different salmonid lineages for TCE families omyk_TCE_15, ssal_TCE_27, omyk_TCE_08, and ssal_TCE_22 are notable, however the activity of omyk_TCE_37 is particularly notable. This family closely follows omyk_TCE_06 as the second most abundant Tc1-Mariner superfamily in the rainbow trout genome; on its own it occupies an astonishing 42.9 Mbp, which corresponds to 2.23% of the entire genome. The proliferation of omyk_TCE_37 occurs predominantly in rainbow trout, and is almost certainly responsible for the differences in Tc1-Mariner abundances that were observed between salmonid species in the previous chapter. This element's expansion likely began in the *Oncorhynchus* ancestor, as moderately-high levels of abundance can be observed in both coho and

chinook. Under this hypothesis, activity that began in the *Oncorhynchus* ancestor would have either decreased in the coho/chinook ancestor following its divergence from that of rainbow trout, or the activity in the rainbow trout lineage would have increased following the split.

Importantly, no copies of omyk_TCE_37 were found in the arctic char genome assembly even though small numbers were present in Atlantic salmon. The lack of this family in arctic char could feasibly be a technical artifact resulting from the difficulty of resolving repetitive regions during genome assembly, however this is unlikely due to the ability of assemblers to resolve the repeat in the other salmonid genomes. Two biological scenarios potentially explain the observed pattern. In the first and perhaps most parsimonious, omyk_TCE_37 was present in the common Salmoninae ancestor and active at very low levels. This activity ceased in both the *Salmo* and *Salvelinus* lineages independently, and, importantly, all copies within the *Salvelinus* ancestor were either lost or rendered undetectable. An alternative explanation invokes HTT. In this scenario, omyk_TCE_37 invaded between the *Salmo* and *Oncorhynchus* ancestors, or invaded both lineages independently. While this may be less likely than the 'common ancestor' hypothesis, multiple instances of TE HTT into independent lineages have been previously reported.

3.3.5 Why a burst of TEs?

The factors governing the abundance of TEs within a genome are those that underscore all evolutionary change: natural selection, genetic drift, recombination and mutation. As such, an increase in the apparent number of TCEs that is associated with a particular event or time period could be the result of many interacting influences including decreased selection, increased transposition activity or decreased DNA loss.

A decrease in the ability of selection to remove subtly deleterious TE insertions from the genome could in part explain the patterns of TCE origin and retention that I observed in the salmonid genome. Such changes could be accomplished in two principle ways: i) a reduction in the negative fitness costs of TCE insertions; or ii) a decrease in the power of selection. Interestingly, a decrease in the fitness costs of TE insertions has been noted in polyploids and is thought to be due to the functional redundancy of essential genes,

which are more amenable to mutation when a 'backup' copy is present (Ågren et al. 2016, Vicient and Casacuberta 2017).

The power of selection to act on mildly deleterious mutations is in large part driven by the susceptibility of a population to the effects of random genetic drift, which is capable of fixing detrimental alleles (TE insertions in this case) in a population before selective processes can eliminate them. Species with small effective population sizes are much more exposed to drift than are abundant species (Lynch 2007). It can be reasonably assumed that the salmonid ancestor went through a dramatic bottleneck in the aftermath of the WGD, and the concomitant decrease in population size could have indirectly decreased the power of selection to remove TE insertions with minor (or even moderate) negative fitness effects. Determining the population size of salmonids beyond this initial phase is largely intractable, however, due to the timescales involved and the relative dearth of salmonid fossils (Behnke 2002). Apart from population size, the condition of tetraploidy itself provides a buffer against random drift compared to diploid populations of similar size (Meirmans and Van Tienderen 2013), a benefit which is conferred by the greater number of alleles that can be present when each constituent individual effectively harbours four genomes rather than two. The impact of natural selection on the retention of TCE copies, particularly in the period well after the WGD, is clearly opaque and requires further investigation.

An alternative mechanism by which the number of TCE copies in the ancestral salmonid genome could expand is through a decrease in host TE suppression, a process which has been frequently associated with stress (see review in Horváth, Merenciano, & González, 2017). Importantly, so-called 'genomic stressors' such as interspecific crosses and polyploidization events have been frequently shown to result in reductions in the ability of a host to silence its TE residents. In plants, both auto- and allopolyploidization events can induce substantial changes in the methylation environment of the genome that are in turn responsible for increases in TE activity (Kashkush et al. 2003, Xu et al. 2012, Santos et al. 2015, Piednoel et al. 2015). There are also a number of counterexamples, however, when polyploidization leads to TE hypermethylation, potentially as secondary response to a dramatic initial increase in activity (Kraitshtein et al. 2010, Yaakov and Kashkush 2011, Zhang et al. 2015). Given the catastrophic nature of WGDs, the changes

they bring to genome regulation, and their tendency to occur and be particularly adaptive during periods of increased environmental stress, an increase in raw TCE transposition activity provides a promising explanation for the frequent bursts of these sequences in the salmonid lineage.

An alternative to explanations that invoke the increased retention or insertion of TCE elements to explore historical TCE activity trends is one which supposes a decrease in the rate of DNA loss within the genome of the salmonid ancestor. Such an explanation is feasibly supported by the observation that larger genomes tend to have reduced rates of DNA loss generally, and TE loss specifically (Petrov et al. 2000, Canapa et al. 2016, Kapusta et al. 2017). The exact mechanisms underlying this trend, be they neutral, selective, or mutational forces, remain contentious (Lynch 2007, Kapusta et al. 2017).

In trying to justify a decrease in the rate of DNA loss in the aftermath of the salmonid WGD, mechanisms which involve changes in the underlying recombination rate are intriguing. Specifically, species with larger genomes (such as salmonids following the WGD) tend to evince a lower recombination rate than their smaller relatives (Tiley and Burleigh 2015). Further, recombination rate has also been shown to change following polyploidization, although not always in a negative direction (Bingham 1967, Pecinka et al. 2011, Lambing et al. 2017). Increases in recombination rate have specifically been shown to drive DNA loss and genome shrinkage (Nam and Ellegren 2012); the reverse situation may be true in salmonids. Mechanistically, positive correlations between recombination and DNA loss could be driven by increases in the opportunity for deletion-inducing ectopic recombination, or by a greater ability of natural selection to remove excess DNA that only has a minor decreased fitness effect on its host. Bringing these concepts together, a potential process emerges to explain the trends in Figure 7: first, the sudden increase in genome size following the Ss4R WGD leads to a decrease in recombination rate. At the same time as this genome size increase, the recombination rate is suppressed, resulting in a decrease in the rate at which DNA is removed from the genome. This model accounts for the apparent spike in the retention of TCEs from the period immediately following the WGD, as well as the large number of TCEs originating throughout salmonid speciation that have been retained to the present day.

Determining the exact reason that a large number of TCEs arising in close proximity to the WGD remain in the genome to this day is clearly a complex task. Given the feasibility of all of the above scenarios, the ultimate explanation likely involves aspects of all of them.

3.3.6 Impact of a historical TCE proliferation in the salmonids

The analyses I have presented here show that for at least the last ~95 million years, when the 4R WGD occurred, there have been continual and dramatic bursts of Tc1-Mariner activity and retention in the salmonid lineage that have resulted in the deposition of tens of millions of bases in extant salmonid genomes. The exact particulars of these bursts, especially in the distant past, remain obscured by phenomena including ascertainment bias, recombination, gene conversion and potential changes in the rates of DNA deletion. Nonetheless, it is clear that TCE activity has had a monumental impact on the architecture of salmonid genomes.

The potential impact of Tc1-Mariner activity ranges from the relatively minor to major changes that could have affected the entire evolutionary trajectory of salmonid rediploidization and speciation. At their most basic, single TCE insertions could alter the regulatory environment of a gene, induce splicing changes, or knock it out altogether. TCE sequences could also be directly co-opted and 'domesticated' to serve host functions, as has occurred in other lineages (Levis et al. 1993, Schatz 2004, Sakai et al. 2007). More severe effects include the ability of TCEs to propagate regulatory elements around the genome, which can result in the *de novo* creation of new regulatory networks (Feschotte 2008, Chuong et al. 2017), or their ability to encourage ectopic recombination and its host of related outcomes: inversions, deletions, duplications, and translocations (Lim and Simmons 1994, Gray 2000, Hedges and Deininger 2007, Grabundzija et al. 2016).

The high level of TCE activity following the WGD is particularly notable because it has led to extremely variable expansion of different TCE families in different salmonid lineages. To varying degrees, all of the salmonid species in Figure 8 have experienced increases in TCE genomic abundance that are not shared by all of the other salmonids. The rainbow trout lineage in particular has been host to a very large TCE expansion that

is not shared by other salmonids. Mass random insertions of mutagenic TCEs at different times in different lineages could easily drive lineage-specific innovations and changes.

The role of TCEs in the speciation of the salmonid lineage is of particular interest, especially given the concurrent rediploidization process that has been ongoing since the 4R WGD. De Boer et al. (2007) were the first to postulate the involvement of TCE activity in the speciation process of salmonids, after having observed a large number of TE copies present in the modern genome which originated during this time period. My work has extended de Boer et al.'s theoretical window of prolific TCE burst activity to the period immediately following the WGD, suggesting that it has been a major evolutionary force both before and during speciation. Importantly, a WGD and subsequent rediploidization process are not necessary to suggest a link between TE activity and speciation. Indeed, TCEs have been implicated in a number of species radiations including those of primates (Feschotte and Pritham 2007, Pace et al. 2008), multiple clades of bats (Ray et al. 2008) and *Anolis* lizards (Feiner 2016). By incorporating both TCEs and WGDs into one model, however, many other intriguing opportunities for lineage diversification present themselves.

One proposed mechanism for the development of the hybrid incompatibility in the aftermath of a WGD involves the process of reciprocal gene loss and divergent resolution (Lynch and Conery 2000, Taylor et al. 2001b). In this scenario, a population of newly tetraploid individuals is divided into two subsequently isolated subpopulations which experience very little gene flow between them. In each subpopulation, the opposite copy of a pair of essential homeologous genes is rendered non-functional; in this example such a pseudogenization is not catastrophic, as a copy of the essential gene still exists in each subpopulation on different chromosomes. The opportunity for speciation arises when the gene flow barrier is lifted, and individuals from different subpopulations are once again permitted to breed. In the first (F1) generation of such a hybrid coupling, all offspring will be heterozygous for the first homeolog copy and heterozygous for the second, having received only one functional copy from each of their parents. Dramatic fitness effects occur in the second (F2) generation of the hybrid population, however, because 1/16 of the constituent progeny would have no functional copy of an essential gene. Effectively, 1/16th of all F2 individuals would be non-viable, which would in turn reduce the success

of the hybrid lineage. In the case of a WGD divergent resolution is not limited to just one essential gene; potential substrates for the process are present in the thousands of newly-formed homeologous pairs. Bursts of TCE activity enter into this model by providing a dramatic way to instantly create a non-functional gene copy. The insertion of a TE into a protein-coding exon of a gene (and in some cases, into introns or UTRs) is almost certainly guaranteed to completely disrupt proper function. Thus, if TEs are prolifically expanding in temporarily isolated subpopulations in the aftermath of a WGD, they may feasibly contribute to hybrid infertility and subsequent speciation.

An alternative opportunity for the involvement of TCE bursts in lineage diversification relies on their tendency to encourage ectopic recombination events. Large-scale genome reorganization events such as chromosomal fusions and fissions or intra-chromosomal duplications and deletions have frequently been suggested to play a role in the introduction of barriers to hybridization (Feulner and De-Kayne 2017). Under this model, a population which had experienced successive genome rearrangements would be less likely to produce viable offspring with another population which did not experience the same events, due to incompatibilities during meiotic pairing. By helping to facilitate the ectopic recombination events that result in such dramatic rearrangements, a burst of TCEs could indirectly drive speciation.

The ability of TEs to promote genome rearrangement is granted an added significance in the context of a WGD and rediploidization in salmon. In particular, large-scale genome reorganization events have been proposed to disrupt the meiotic pairing of homeologous chromosomes, thereby pushing a genomic region towards diploidy (Wendel 2000, Gerstein et al. 2006). In the recent Atlantic salmon genome report in which some of the present findings were first reported, a dramatic association was observed between large-scale genome rearrangements and sequences which had clearly completed the rediploidization process (Lien et al. 2016). Regions that were inferred to possess residual tetraploidy, however, were devoid of such events. Given this finding, and the massive waves of TCE proliferation that were bombarding the genome immediately following the WGD and during the entire rediploidization process, it is possible that TE-induced rearrangement events may have helped encourage rediploidization in salmonids.

3.4 Conclusions and Future Directions

In this chapter I have provided evidence for the existence of continuing massive proliferations of Tc1-Mariner TEs within the evolutionary history of salmonids. I have shown that this high level of activity is likely to be associated with the aftermath of the salmonid WGD, and that in continuing throughout salmonid speciation it has led to enormous differences in the genomic content of modern salmonids. Although not directly shown, it is highly probable that this large number of TCE insertions has affected the evolutionary trajectory of this import taxon.

The opportunities for future work are enormous and exciting. Further research could investigate the patterns of TCE insertion and determine whether or not these elements have driven pseudogenization in homeologous gene copies. Alternatively, an extensive survey of the breakpoints of genome rearrangements might identify a direct role of TCEs in the processes that facilitate rediploidization. A particularly intriguing future task would be the search for a HTT event that led directly to the dramatic lineage-specific proliferation of a TCE in salmonids; discovering such a case would offer a fascinating example of an evolutionary force which operated outside the bounds of the traditional paradigm that is centered on vertical inheritance.

The role that TEs play in the aftermath of a WGD event is not fully understood. The work I present in this chapter underscores the contribution of TEs to genome architecture, and in suggesting a relationship between two powerful evolutionary forces helps develop a better understanding of a taxonomic group that has become an important model in the study of evolution.

Final Thoughts

Over the course of this thesis I have presented evidence that repeats and TEs make up a substantial portion of five salmonid genomes. Further, I have demonstrated that a dramatic increase in Tc1-Mariner element abundance occurred following the WGD and well into speciation, with the result that it has differentially affected the genome architecture of extant salmonids. While this work falls short of conclusively demonstrating that TEs have had tangible impacts on lineage diversification or the development of salmonid-specific traits, it is reasonable to assume that such impacts have occurred. This assumption arises from the observation that TEs are a consistent mutagenic force across genomes in which they reside, and that, among vertebrates, salmonids evince an exceptionally large TE-derived genome fraction. Simply put, it is extremely unlikely that hundreds of millions of bases could be randomly deposited throughout a genome without that activity influencing the evolution of its host species in some way.

This work is largely descriptive, and future work has the promising potential of resolving many of the hypotheses which I have left unanswered. In particular, the impact of elements from Tc1-Mariner and other superfamilies on specific genome processes is of great interest. Are TEs disproportionately present at the breakpoints of chromosomal rearrangements that have likely contributed to rediploidization? Did any of the massive Tc1-Mariner proliferations that occurred arise as the result of HTT? Are certain classes of genes disproportionately likely to have retained TE insertions, implying that they may have served a functional purpose? There is also potential for TEs to help answer other longstanding questions about salmonids: Can TE insertions be used to resolve ambiguities in the salmonid phylogeny, for example to help determine whether the Thymallinae lineage is a sister taxa to the Salmoninae lineage (as proposed by Crête-Lafrenière et al. in 2012) or to the Coregoninae lineage (as proposed by Macqueen and Johnston in 2014)? How many TE families are active in the present day, and do their insertions differ dramatically enough that they could be used to easily distinguish different populations?

The intersection of genomics, computer science and evolutionary biology is a fascinating and increasingly important area of research, and I have thoroughly enjoyed exploring its depths. With the work I present in this thesis I have helped to expand the current understanding of two fundamental evolutionary forces: TEs and WGDs. I have also developed a set of salmonid reference repeat libraries which are already helping to advance myriad other biological investigations. This work provides a solid base from which to pursue future studies in both salmonids and evolution, and I look forward to seeing how it and the work of thousands of other researchers contributes to a better understanding of the processes which have given rise to the incredible diversity of life on Earth.

Bibliography

Ågren, J.A., Huang, H.R., and Wright, S.I. 2016. Transposable element evolution in the allotetraploid *Capsella bursa-pastoris*. Am. J. Bot. **103**(7): 1197–1202. doi:10.3732/ajb.1600103.

Albertin, W., and Marullo, P. 2012. Polyploidy in fungi: evolution after whole-genome duplication. Proc. R. Soc. B Biol. Sci. **279**(1738): 2497–2509. doi:10.1098/rspb.2012.0434.

Allendorf, F.W., and Thorgaard, G.H. 1984. Evolutionary Genetics of Fishes. *Edited by* B.J. Turner. Plenum Press. doi:10.1007/978-1-4684-4652-4.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. **215**(3): 403–410. doi:10.1016/S0022-2836(05)80360-2.

Amores, A., Force, A., Yan, Y.-L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.-L., Westerfield, M., Ekker, M., and Postlethwait, J.H. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science **282**(5394): 1711–1714. doi:10.1126/science.282.5394.1711.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. ming, Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Sollewijn Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J.K., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science **297**(5585): 1301–1310. doi:10.1126/science.1072104.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.-S.L. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. **32**(Database issue): D115-9. doi:10.1093/nar/gkh131.

Arkhipova, I.R. 2005. Mobile genetic elements and sexual reproduction. Cytogenet. Genome Res. **110**(1–4): 372–382. doi:10.1159/000084969.

Arrigo, N., and Barker, M.S. 2012. Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. **15**(2): 140–146. Elsevier Ltd. doi:10.1016/j.pbi.2012.03.010.

Bao, W., Kojima, K.K., and Kohany, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6**(1): 11. doi:10.1186/s13100-015-0041-9.

Bao, Z., and Eddy, S.R. 2003. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. **13**(1): 1269–1276. doi:10.1101/gr.88502.

Te Beest, M., Le Roux, J.J., Richardson, D.M., Brysting, A.K., Suda, J., Kubešová, M., and Pyšek, P. 2012. The more the better? The role of polyploidy in facilitating plant invasions. Ann. Bot. **109**(1): 19–45. doi:10.1093/aob/mcr277.

Behnke, R.J. 2002. Trout and Salmon of North America. *In* Trout and salmon of North America. Available from https://books.google.com/books?hl=en&lr=&id=3WlHElmgQVgC&oi=fnd&pg=PR7&dq=Trout+and+Salmon+of+North+America+behnke&ots=pxKVl6BzcX&sig=8pnnI54JSwLSFDEsisvplAEJLJg.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., and Damborsky, J. 2014. PredictSNP: robust and accurate consensus classifier for prediction of Disease-Related Mutations. PLoS Comput. Biol. **10**(1): 1–11. doi:10.1371/journal.pcbi.1003440.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. Nucleic Acids Res. **33**(DATABASE ISS.): D34–D38. doi:10.1093/nar/gki063.

Berg, O. 1985. The formation of non-anadromous populations of Atlantic salmon, *Salmon salar* L., in Europe. J. Fish Biol. **27**: 805–815. doi:10.1111/j.1095-8649.1985.tb03222.x.

Berra, T.M. 1981. An atlas of distribution of the freshwater fish families of the world. University of Nebraska Press.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G.H., Boussaha, M., Quillet, E., Guyomard, R., Galiana, D., Bobe, J., Volff, J.N., Genêt, C., Wincker, P., Jaillon, O., Crollius, H.R., and Guiguen, Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. **5**. doi:10.1038/ncomms4657.

Betancur-R, R., Broughton, R.E., Wiley, E.O., Carpenter, K., López, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton Ii, J.C., Zhang, F., Buser, T., Campbell, M.A., Ballesteros, J.A., Roa-Varon, A., Willis, S., Borden, W.C., Rowley, T., Reneau, P.C., Hough, D.J., Lu, G., Grande, T., Arratia, G., and Ortí, G. 2013. The tree of life and a new classification of bony fishes. PLoS Curr. **5**. Available from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3644299&tool=pmcentrez&rendertype=abstract [accessed 24 July 2013].

Bingham, J. 1967. Breeding cereals for improved yielding capacity. Ann. Appl. Biol.

59(2): 312–315. doi:10.1111/j.1744-7348.1967.tb04442.x.

Björnsson, B.T., Stefansson, S.O., and McCormick, S.D. 2011. Environmental endocrinology of salmon smoltification. Gen. Comp. Endocrinol. **170**(2): 290–298. doi:10.1016/j.ygcen.2010.07.003.

Blanc, G., and Wolfe, K.H. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution author(s): Guillaume Blanc and Kenneth H . Wolfe Source : The Plant Cell, Vol. 16, No. 7 (Jul., 2004), pp. 1679-1691 Published by: American Society of. Plant Cell **16**(7): 1679–1691. doi:10.1105/tpc.021410.tion.

de Boer, J.G., Yazawa, R., Davidson, W.S., and Koop, B.F. 2007. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics **8**(1): 422. doi:10.1186/1471-2164-8-422.

Borgognone, A., Castanera, R., Muguerza, E., Pisabarro, A.G., and Ramírez, L. 2017. Somatic transposition and meiotically driven elimination of an active helitron family in Pleurotus ostreatus. DNA Res. **24**(2): 103–115. doi:10.1093/dnares/dsw060.

Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A.M., Campbell, M.S., Barrell, D., Martin, K.J., Mulley, J.F., Ravi, V., Lee, A.P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F.E.G., Sun, Y., Hertel, J., Beam, M.J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J.H., Litman, G.W., Litman, R.T., Mikami, M., Ota, T., Saha, N.R., Williams, L., Stadler, P.F., Wang, H., Taylor, J.S., Fontenot, Q., Ferrara, A., Searle, S.M.J., Aken, B., Yandell, M., Schneider, I., Yoder, J.A., Volff, J.-N., Meyer, A., Amemiya, C.T., Venkatesh, B., Holland, P.W.H., Guiguen, Y., Bobe, J., Shubin, N.H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J.H. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. **48**(4): 427–437. doi:10.1038/ng.3526.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell **128**(6): 1089–1103. doi:10.1016/j.cell.2007.01.043.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G.J. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. Science (80-.). **322**(5906): 1387–1392. doi:10.1126/science.1165171.

British Columbia Ministry of Agriculture. 2016. 2016 Export Highlights. British Columbia Agrifood & Seafood.

Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. **23**(9): 1808–1816.

doi:10.1093/molbev/msl049.

Budy, P., and Luecke, C. 2014. Understanding how lake populations of arctic char are structured and function with special consideration of the potential effects of climate change: A multi-faceted approach. Oecologia **176**(1): 81–94. doi:10.1007/s00442-014-2993-8.

De Bustos, A., Cuadrado, A., and Jouve, N. 2016. Sequencing of long stretches of repetitive DNA. Sci. Rep. **6**(1): 36665. doi:10.1038/srep36665.

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.-H., Childs, K.L., Sun, Y., Jiang, N., and Yandell, M. 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. **164**(2): 513–524. doi:10.1104/pp.113.230144.

Canapa, A., Barucca, M., Biscotti, M.A., Forconi, M., and Olmo, E. 2016. Transposons, genome size, and evolutionary insights in animals. Cytogenet. Genome Res. **147**(4): 217–239. doi:10.1159/000444429.

Cannon, A., and Yang, D.Y. 2006. Early storage and sedentism on the Pacific Northwest Coast: ancient DNA analysis of salmon remains from Namu, British Columbia. Am. Antiq. **71**(1): 123–140. doi:10.2307/40035324.

Cappello, J., Handelsman, K., and Lodish, H.F. 1985. Sequence of *Dictyostelium* DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell **43**(1): 105–115. doi:10.1016/0092-8674(85)90016-9.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., Forrest, A.R.R., Carninci, P., Biffo, S., Stupka, E., and Gustincich, S. 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature **491**(7424): 454–457. Nature Publishing Group. doi:10.1038/nature11508.

de Carvalho, M.O., and Loreto, E.L.S. 2012. Methods for detection of horizontal transfer of transposable elements in complete genomes. Genet. Mol. Biol. **35**(4 SUPPL.): 1078–1084. doi:10.1590/S1415-47572012000600024.

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.N. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol. Evol. **7**(2): 567–580. doi:10.1093/gbe/evv005.

Chan, P.P., and Lowe, T.M. 2009. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. **37**(SUPPL. 1): D93–D97. doi:10.1093/nar/gkn787.

Chuong, E.B., Elde, N.C., and Feschotte, C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. **18**(2): 71–86. Nature

Publishing Group. doi:10.1038/nrg.2016.139.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M.J.L. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11): 1422–1423. doi:10.1093/bioinformatics/btp163.

Commission of Inquiry into the Decline of Sockeye Salmon in the Fraser River (Canada), and Cohen, B.I. 2012. The uncertain future of Fraser River sockeye. Volume 1, Volume 1,. **3**(October). Available from http://publications.gc.ca/collections/collection_2012/bcp-pco/CP32-93-2012-1-eng.pdf.

Conte, M.A., Gammerdinger, W.J., Bartie, K.L., Penman, D.J., and Kocher, T.D. 2017. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. BMC Genomics **18**(1): 341. BMC Genomics. doi:10.1186/s12864-017-3723-5.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. 2002. Human L1 element target-primed reverse transcription in vitro. EMBO J. **21**(21): 5899–5910. doi:10.1093/emboj/cdf592.

Crête-Lafrenière, A., Weir, L.K., and Bernatchez, L. 2012. Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. PLoS One **7**(10). doi:10.1371/journal.pone.0046662.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and DePamphilis, C.W. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. **16**(6): 738–749. doi:10.1101/gr.4825606.

Cyranoski, D. 2016. China embraces precision medicine on a massive scale. Nature **529**(7584): 9–10. doi:10.1038/529009a.

Davidson, W.S. 2013. Understanding salmonid biology from the Atlantic salmon genome. Genome **56**(10): 548–550. doi:10.1139/gen-2013-0163.

Davidson, W.S., Koop, B.F., Jones, S.J., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S., and Omholt, S.W. 2010. Sequencing the genome of the Atlantic salmon (*Salmo salar)*. Genome Biol. 2010 119 **11**(9): 403. doi:10.1186/GB-2010-11-9-403.

Dehal, P., and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. **3**(10). doi:10.1371/journal.pbio.0030314.

Deininger, P.L., and Roy-Engel, A.M. 2002. Mobile DNA II. *Edited by* N.L. Craig, A.M. Lambowitz, R. Craigie, and M. Gellert. American Society of Microbiology. doi:10.1128/9781555817954.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B.R., Landt, S.G., Lee, B.K., Pauli, F., Rosenbloom, K.R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J.M., Song, L., Trinklein, N.D., Altshuler, R.C., Birney, E., Brown, J.B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T.S., Gerstein, M., Giardine, B., Greven, M., Hardison, R.C., Harris, R.S., Herrero, J., Hoffman, M.M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G.K., Merkel, A., Mortazavi, A., Parker, S.C.J., Reddy, T.E., Rozowsky, J., Schlesinger, F., Thurman, R.E., Wang, J., Ward, L.D., Whitfield, T.W., Wilder, S.P., Wu, W., Xi, H.S., Yip, K.Y., Zhuang, J., Bernstein, B.E., Green, E.D., Gunter, C., Snyder, M., Pazin, M.J., Lowdon, R.F., Dillon, L.A.L., Adams, L.B., Kelly, C.J., Zhang, J., Wexler, J.R., Good, P.J., Feingold, E.A., Crawford, G.E., Dekker, J., Elnitski, L., Farnham, P.J., Giddings, M.C., Gingeras, T.R., Guigó, R., Hubbard, T.J., Kent, W.J., Lieb, J.D., Margulies, E.H., Myers, R.M., Stamatoyannopoulos, J.A., Tenenbaum, S.A., Weng, Z., White, K.P., Wold, B., Yu, Y., Wrobel, J., Risk, B.A., Gunawardena, H.P., Kuiper, H.C., Maier, C.W., Xie, L., Chen, X., Mikkelsen, T.S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M.J., Durham, T., Ku, M., Truong, T., Eaton, M.L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B.A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O.J., Park, E., Preall, J.B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K.S., Schaeffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S.E., Hannon, G.J., Ruan, Y., Carninci, P., Sloan, C.A., Learned, K., Malladi, V.S., Wong, M.C., Barber, G.P., Cline, M.S., Dreszer, T.R., Heitner, S.G., Karolchik, D., Kirkup, V.M., Meyer, L.R., Long, J.C., Maddren, M., Raney, B.J., Grasfeder, L.L., Giresi, P.G., Battenhouse, A., Sheffield, N.C., Showers, K.A., London, D., Bhinge, A.A., Shestak, C., Schaner, M.R., Kim, S.K., Zhang, Z.Z., Mieczkowski, P.A., Mieczkowska, J.O., Liu, Z., McDaniell, R.M., Ni, Y., Rashid, N.U., Kim, M.J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V.R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E.C., Varley, K.E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K.M., Anaya, M., Cross, M.K., Muratet, M.A., Newberry, K.M., McCue, K., Nesmith, A.S., Fisher-Aylor, K.I., Pusey, B., DeSalvo, G., Parker, S.L., Balasubramanian, S., Davis, N.S., Meadows, S.K., Eggleston, T., Newberry, J.S., Levy, S.E., Absher, D.M., Wong, W.H., Blow, M.J., Visel, A., Pennachio, L.A., Petrykowska, H.M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J.M., Griffiths, E., Harte, R., Hendrix, D.A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M.F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J.M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward,

C., Tapanari, E., Tress, M.L., Van Baren, M.J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R.P., Auerbach, R.K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A.P., Cao, A.R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J.D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V.X., Karczewski, K.J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X.J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.K., Yang, X., Struhl, K., Weissman, S.M., Penalva, L.O., Karmakar, S., Bhanvadia, R.R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D.L., Byron, R., Canfield, T.K., Diegel, M.J., Dunn, D., Ebersol, A.K., Frum, T., Garg, K., Gist, E., Hansen, R.S., Boatman, L., Haugen, E., Humbert, R., Johnson, A.K., Johnson, E.M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M.T., Neph, S.J., Neri, F. V., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Rynes, E., Sanchez, M.E., Sandstrom, R.S., Shafer, A.O., Stergachis, A.B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M.A., Yan, Y., Zhang, M., Akey, J.M., Bender, M., Dorschner, M.O., Groudine, M., MacCoss, M.J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N.M., Sobral, D., Vaquerizas, J.M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M.W., Schaub, M.A., Miller, W., Bickel, P.J., Banfai, B., Boley, N.P., Huang, H., Li, J.J., Noble, W.S., Bilmes, J.A., Buske, O.J., Sahu, A.D., Kharchenko, P. V., Park, P.J., Baker, D., Taylor, J., and Lochovsky, L. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414): 57–74. doi:10.1038/nature11247.

Eddy, S.R. 1998. Profile hidden Markov models. Bioinformatics **14**(9): 755–763. doi:10.1093/bioinformatics/14.9.755.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R.C., and Myers, E.W. 2005. PILER: identification and classification of genomic repeats. Bioinformatics **21**(SUPPL. 1): i152–i158. doi:10.1093/bioinformatics/bti1003.

Eickbush, T.H., and Malik, H.S. 2002. Mobile DNA II. *Edited by* N.L. Craig, A.M. Lambowitz, R. Craigie, and M. Gellert. American Society of Microbiology. doi:10.1128/9781555817954.

Elbarbary, R.A., Lucas, B.A., and Maquat, L.E. 2016. Retrotransposons as regulators of gene expression. Science (80-.). **351**(6274). doi:10.1126/science.aac7247.

Elisaphenko, E.A., Kolesnikov, N.N., Shevchenko, A.I., Rogozin, I.B., Nesterova, T.B.,

Brockdorff, N., and Zakian, S.M. 2008. A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS One **3**(6): e2521. Public Library of Science. doi:10.1371/journal.pone.0002521.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics **9**: 18. doi:10.1186/1471-2105-9-18.

Elliott, B., Richardson, C., and Jasin, M. 2005. Chromosomal translocation mechanisms at intronic Alu elements in mammalian cells. Mol. Cell **17**(6): 885–894. doi:10.1016/j.molcel.2005.02.028.

Ellison, C.E., and Bachtrog, D. 2015. Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. Elife **4**. doi:10.7554/eLife.05899.

Evgen'ev, M.B., and Arkhipova, I.R. 2005. Penelope-like elements - A new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet. Genome Res. **110**(1–4): 510–521. doi:10.1159/000084984.

FAO. 2016. The state of world fisheries and aquaculture. *In* The State of World Fisheries and Aquaculture. doi:92-5-105177-1.

Feiner, N. 2016. Accumulation of transposable elements in *Hox* gene clusters during adaptive radiation of *Anolis* lizards. Proc. R. Soc. B Biol. Sci. **283**(1840): 20161555. doi:10.1098/rspb.2016.1555.

Feschotte, C. 2008. Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. **9**(5): 397–405. Nature Publishing Group. doi:10.1038/nrg2337.

Feschotte, C., and Pritham, E.J. 2005. Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet. **21**(10): 551–552. doi:10.1016/j.tig.2005.07.007.

Feschotte, C., and Pritham, E.J. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. **41**(1): 331–368. Annual Reviews. doi:10.1146/annurev.genet.40.110405.090448.

Feulner, P.G.D., and De-Kayne, R. 2017. Genome evolution, structural rearrangements and speciation. J. Evol. Biol. **30**(8): 1488–1490. doi:10.1111/jeb.13101.

Finnegan, D.J. 2012. Retrotransposons. Curr. Biol. **22**(11): 432–437. doi:10.1016/j.cub.2012.04.025.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. 2011. Considering transposable element diversification in *de novo* annotation approaches. PLoS One **6**(1). doi:10.1371/journal.pone.0016526.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**(4): 1531–1545. doi:10101175.

Frith, M.C. 2011. A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. **39**(4): e23–e23. doi:10.1093/nar/gkq1212.

Furlong, R.F., and Holland, P.W.H. 2002. Were vertebrates octoploid? Philos. Trans. R. Soc. B Biol. Sci. **357**(1420): 531–544. doi:10.1098/rstb.2001.1035.

Gaeta, R.T., and Chris Pires, J. 2010. Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol. **186**(1): 18–28. doi:10.1111/j.1469-8137.2009.03089.x.

Gallant, J.R., Losilla, M., Tomlinson, C., and Warren, W.C. 2017. The genome and adult somatic transcriptome of the Mormyrid electric fish *Paramormyrops kingsleyae*. Genome Biol. Evol. **9**(12): 3525–3530. doi:10.1093/gbe/evx265.

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., and Song, C. 2016. The contribution of transposable elements to size variations between four teleost genomes. Mob. DNA **7**(1): 4. Mobile DNA. doi:10.1186/s13100-016-0059-7.

Gao, X., and Voytas, D.F. 2005, March. A eukaryotic gene family related to retroelement integrases. Trends in Genetics **21**(3): 133-137. doi:10.1016/j.tig.2005.01.006.

Geisler, S., and Coller, J. 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat. Rev. Mol. Cell Biol. **14**(11): 699–712. Nature Publishing Group. doi:10.1038/nrm3679.

Gerstein, A.C., Chun, H.J.E., Grant, A., and Otto, S.P. 2006. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. PLoS Genet. **2**(9): 1396–1401. doi:10.1371/journal.pgen.0020145.

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E.A., and Cordaux, R. 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat. Commun. **5**: 3348. doi:10.1038/ncomms4348.

Gislason, G., Lam, E., Gunnar, K., and Guettabi, M. 2017. Economic Impacts of Pacific Salmon Fisheries. (July). Prepared for the Pacific Salmon Commission.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., and Jaffe, D.B. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. **108**(4): 1513–1518. doi:10.1073/pnas.1017351108.

Gong, C., and Maquat, L.E. 2011. LncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 39 UTRs via Alu elements. Nature **470**(7333): 284–290. Nature Publishing Group. doi:10.1038/nature09701.

Goodier, J.L., and Davidson, W.S. 1993. A repetitive element in the genome of Atlantic salmon, *Salmo salar*. Gene **131**(2): 237–42. Available from http://www.ncbi.nlm.nih.gov/pubmed/8406016.

Goodier, J.L., and Davidson, W.S. 1994. Tc1 transposon-like sequences are widely distributed in salmonids. J. Mol. Biol. **241**(1): 26–34. doi:10.1006/jmbi.1994.1470.

Goodwin, T.J.D., Butler, M.I., and Poulter, R.T.M. 2003. Cryptons: A group of tyrosine-recombinase- encoding DNA transposons from pathogenic fungi. Microbiology **149**(11): 3099–3109. doi:10.1099/mic.0.26529-0.

Goodwin, T.J.D., and Poulter, R.T.M. 2004. A new group of tyrosine recombinase-encoding retrotransposons. Mol. Biol. Evol. **21**(4): 746–759. doi:10.1093/molbev/msh072.

Grabundzija, I., Messing, S.A., Thomas, J., Cosby, R.L., Bilic, I., Miskey, C., Gogol-Doring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E.J., Dyda, F., Izsvak, Z., and Ivics, Z. 2016. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat. Commun. **7**. doi:10.1038/ncomms10716.

Gray, Y.H.M. 2000. It takes two transposons to tango: Transposable-element-mediated chromosomal rearrangements. Trends Genet. **16**(10): 461–468. doi:10.1016/S0168-9525(00)02104-1.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J., and Bennett, M.D. 2007. Eukaryotic genome size databases. Nucleic Acids Res. **35**(SUPPL. 1): D332–D338. doi:10.1093/nar/gkl828.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. 2005. Engineering a software tool for gene structure prediction in higher organisms. Inf. Softw. Technol. **47**(15): 965–978. doi:10.1016/j.infsof.2005.09.005.

Haase, A.D. 2016. A small RNA-based immune system defends germ cells against mobile genetic elements. Stem Cells Int. **2016**. doi:10.1155/2016/7595791.

Hagberg, A.A., Schult, D.A., and Swart, P.J. 2008. Exploring network structure, dynamics, and function using NetworkX. *In* Proceedings of the 7th Python in Science Conference (SciPy2008). *Edited by* G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA. pp. 11–15.

Han, Y., and Wessler, S.R. 2010. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res.

38(22): e199–e199. doi:10.1093/nar/gkq862.

Harris, R.S. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University. Available from http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf.

Hartl, D.L., Lohe, A.R., and Lozovskaya, E.R. 1997. Modern thoughts on an ancyent marinere: function, evolution, regulation. Annu. Rev. Genet. **31**(1): 337–358. doi:10.1146/annurev.genet.31.1.337.

Hartl, D.L., Lozovskaya, E.R., and Lawrence, J.G. 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica **86**(1–3): 47–53. Available from http://www.ncbi.nlm.nih.gov/pubmed/1334917.

Hartley, S.E. 1987. The chromosomes of salmonid fishes. Biol. Rev. **62**(3): 197–214. doi:10.1111/j.1469-185X.1987.tb00663.x.

Hedges, D.J., and Deininger, P.L. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat. Res. - Fundam. Mol. Mech. Mutagen. **616**(1–2): 46–59. doi:10.1016/j.mrfmmm.2006.11.021.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. 2014. PASTEC: An automatic transposable element classification tool. PLoS One **9**(5): e91929. doi:10.1371/journal.pone.0091929.

Van Hoek, M.J.A., and Hogeweg, P. 2009. Metabolic adaptation after whole genome duplication. Mol. Biol. Evol. **26**(11): 2441–2453. doi:10.1093/molbev/msp160.

Hoen, D.R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D.D., Quesneville, H., Smit, A., Wheeler, T.J., Bureau, T.E., and Blanchette, M. 2015. A call for benchmarking transposable element annotation methods. Mob. DNA **6**: 13. Mobile DNA. doi:10.1186/s13100-015-0044-6.

Hoffman, M.M., and Birney, E. 2007. Estimating the neutral rate of nucleotide substitution using introns. Mol. Biol. Evol. **24**(2): 522–531. doi:10.1093/molbev/msl179.

Hollister, J.D., and Gaut, B.S. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. **19**(8): 1419–1428. doi:10.1101/gr.091678.109.

Horváth, V., Merenciano, M., and González, J. 2017. Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends Genet. **33**(11): 832–841. doi:10.1016/j.tig.2017.08.007.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins,

J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assunção, J.A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Eliott, D., Threadgold, G., Harden, G., Ware, D., Mortimer, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Ürün, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberländer, M., Rudolph-Geiger, S., Teucke, M., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M.J., Enright, A., Geisler, R., Plasterk, R.H.A., Lee, C., Westerfield, M., De Jong, P.J., Zon, L.I., Postlethwait, J.H., Nüsslein-Volhard, C., Hubbard, T.J.P., Crollius, H.R., Rogers, J., and Stemple, D.L. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature **496**(7446): 498–503. doi:10.1038/nature12111.

Hua-Van, A., Le Rouzic, A., Boutin, T.S., Filée, J., and Capy, P. 2011. The struggle for life of the genome's selfish architects. Biol. Direct **6**(1): 19. doi:10.1186/1745-6150-6-19.

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvák, Z. 1997. Molecular reconstruction of sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell **91**(4): 501–510. doi:10.1016/S0092-8674(00)80436-5.

Ivics, Z., and Izsvák, Z. 2015. Sleeping Beauty transposition. Microbiol. Spectr. **3**(2): 1–21. doi:10.1128/microbiolspec.MDNA3-0042-2014.

Ivics, Z., Izsvak, Z., Minter, A., and Hackett, P.B. 1996. Identification of functional domains and evolution of Tc1-like transposable elements. Proc. Natl. Acad. Sci. U. S. A. **93**(10): 5008–5013. doi:10.1073/pnas.93.10.5008.

Iwasaki, Y.W., Siomi, M.C., and Siomi, H. 2015. PIWI-interacting RNA: its biogenesis and functions. Annu. Rev. Biochem. **84**(1): 405–433. doi:10.1146/annurev-biochem-060614-034258.

Jacques, P.E., Jeyakani, J., and Bourque, G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. **9**(5). doi:10.1371/journal.pgen.1003504.

Jagannathan, M., Warsinger-Pepe, N., Watase, G.J., and Yamashita, Y.M. 2017. Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. G3 **7**(2): 693–704. doi:10.1534/g3.116.035352.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J., and Roest Crollius, H. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature **431**(7011): 946–957. doi:10.1038/nature03025.

Jiang, N., Bowman, M., and Childs, K. 2016. Repeat library construction - advanced. Available from http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/%0ARepeat_Library_Construction-Advanced [accessed 4 March 2018].

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., Di Palma, F., Lindblad-Toh, K., and Kingsley, D.M. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**(7392): 55–61. doi:10.1038/nature10944.

Kaessmann, H., Vinckenbosch, N., and Long, M. 2009. RNA-based gene duplication: Mechanistic and evolutionary insights. Nat. Rev. Genet. **10**(1): 19–31. doi:10.1038/nrg2487.

Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A.H. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics **166**(3): 1437–1450. doi:10.1534/genetics.166.3.1437.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Garcia Diez, F., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr,

P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. 2005. The EMBL nucleotide sequence database. Nucleic Acids Res. **33**(Database issue): D29–D33. doi:10.1093/nar/gki098.

Kapitonov, V. V., and Jurka, J. 2001. Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. **98**(15): 8714–8719. doi:10.1073/pnas.151269298.

Kapitonov, V. V., and Jurka, J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. **3**(6): 0998–1011. doi:10.1371/journal.pbio.0030181.

Kapitonov, V. V., and Jurka, J. 2006. Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. **103**(12): 4540–4545. doi:10.1073/pnas.0600833103.

Kapusta, A., and Feschotte, C. 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. **30**(10): 439–452. doi:10.1016/j.tig.2014.08.004.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L. a., Bourque, G., Yandell, M., and Feschotte, C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. **9**(4). doi:10.1371/journal.pgen.1003470.

Kapusta, A., Suh, A., and Feschotte, C. 2017. Dynamics of genome size evolution in birds and mammals. Proc. Natl. Acad. Sci. **114**(8): E1460–E1469. doi:10.1073/pnas.1616702114.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S.I., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., and Kohara, Y. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature **447**(7145): 714–719. doi:10.1038/nature05846.

Kashkush, K., Feldman, M., and Levy, A.A. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. **33**(1): 102–106. doi:10.1038/ng1063.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**(12): 1647–1649. doi:10.1093/bioinformatics/bts199.

Kebriaei, P., Singh, H., Huls, M.H., Figliola, M.J., Bassett, R., Olivares, S., Jena, B., Dawson, M.J., Kumaresan, P.R., Su, S., Maiti, S., Dai, J., Moriarity, B., Forget, M.,

Senyukov, V., Orozco, A., Liu, T., Mccarty, J., Jackson, R.N., Moyes, J.S., Rondon, G., Qazilbash, M., Ciurea, S., Alousi, A., Nieto, Y., Rezvani, K., Marin, D., Popat, U., Hosing, C., Shpall, E.J., Kantarjian, H., Keating, M., Wierda, W., Do, K.A., Largaespada, D.A., Lee, D.A., Hackett, P.B., Champlin, R.E., and Cooper, L.J.N. 2016. Phase I trials using Sleeping Beauty to generate CD19-specific CAR T cells. J Clin Invest. **126**(9): 3363–3376. doi:10.1172/JCI86721DS1.

Kent, T. V, Uzunović, J., and Wright, S.I. 2017. Coevolution between transposable elements and recombination. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **372**(1736): 20160458. doi:10.1098/rstb.2016.0458.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, a. D. 2002. The human genome browser at UCSC. Genome Res. **12**(6): 996–1006. doi:10.1101/gr.229102.

Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M., and Okada, N. 1991. Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. Proc. Natl. Acad. Sci. U. S. A. **88**(6): 2326–30. Available from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51224&tool=pmcentrez&rendertype=abstract.

Kidwell, M.G. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica **115**(1): 49–63. doi:10.1023/A:1016072014259.

Koga, A., Wakamatsu, Y., Sakaizumi, M., Hamaguchi, S., and Shimada, A. 2009. Distribution of complete and defective copies of the Tol1 transposable element in natural populations of the medaka fish *Oryzias latipes*. Genes Genet. Syst. **84**(5): 345–52. doi:10.1266/ggs.84.345.

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. 2011. Repetitive elements may comprise over two thirds of the human genome. PLoS Genet. **7**(12): e1002384. Public Library of Science. doi:10.1371/journal.pgen.1002384.

Konkel, M.K., and Batzer, M.A. 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. Semin. Cancer Biol. **20**(4): 211–221. Elsevier Ltd. doi:10.1016/j.semcancer.2010.03.001.

Kraitshtein, Z., Yaakov, B., Khasdan, V., and Kashkush, K. 2010. Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. Genetics **186**(3): 801–812. doi:10.1534/genetics.110.120790.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. **42**(7): 631–634. doi:10.1038/ng.600.

Lambing, C., Franklin, F.C.H., and Wang, C.-J.R. 2017. Understanding and Manipulating

Meiotic Recombination in Plants. Plant Physiol. **173**(3): 1530–1542. doi:10.1104/pp.16.01530.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M.L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M.J. 2001. Initial sequencing and analysis of the human genome. Nature **409**(6822): 860–921. Macmillian Magazines Ltd. doi:10.1038/35057062.

Ledford, H. 2016. AstraZeneca launches project to sequence 2 million genomes. Nature **532**(7600): 427. doi:10.1038/nature.2016.19797.

Leeton, P.R., and Smyth, D.R. 1993. An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Mol. Gen. Genet. **237**(1–2): 97–104. doi:10.1007/BF00282789.

Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. Heredity (Edinb). **104**(6): 520–533. doi:10.1038/hdy.2009.165.

Levis, R.W., Ganesan, R., Houtchens, K., Tolar, L.A., and Sheen, F. miin. 1993. Transposons in place of telomeric repeats at a *Drosophila* telomere. Cell **75**(6): 1083–1093. doi:10.1016/0092-8674(93)90318-K.

Li, Y., and Dooner, H.K. 2009. Excision of Helitron transposons in maize. Genetics **182**(1): 399–402. doi:10.1534/genetics.109.101527.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W., and Davidson, W.S. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature **advance on**. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. Available from http://dx.doi.org/10.1038/nature17164.

Lim, J.K., and Simmons, M.J. 1994. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. BioEssays **16**(4): 269–275. doi:10.1002/bies.950160410.

Louzada, S., Vieira-da-Silva, A., Mendes-da-Silva, A., Kubickova, S., Rubes, J., Adega, F., and Chaves, R. 2015. A novel satellite DNA sequence in the *Peromyscus* genome (PMSat): evolution via copy number fluctuation. Mol. Phylogenet. Evol. **92**: 193–203. doi:10.1016/j.ympev.2015.06.008.

Lynch, M. 2007. The Origins of Genome Architecture. 1st edition. Sinauer Assotiates Inc.

Lynch, M., and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science (80-.). **290**(5494): 1151–1155. doi:10.1126/science.290.5494.1151.

Lynch, M., and Conery, J.S. 2003. The origins of genome complexity. Science **302**(5649): 1401–1404. doi:10.1126/science.1089370.

Lynch, V.J., Leclerc, R.D., May, G., and Wagner, G.P. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat. Genet. **43**(11): 1154–1159. Nature Publishing Group. doi:10.1038/ng.917.

Mable, B.K., Alexandrou, M.A., and Taylor, M.I. 2011. Genome duplication in amphibians and fish: An extended synthesis. J. Zool. **284**(3): 151–182. doi:10.1111/j.1469-7998.2011.00829.x.

Macqueen, D.J., and Johnston, I.A. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc. R. Soc. B Biol. Sci. **281**(1778): 20132881–20132881. doi:10.1098/rspb.2013.2881.

Maier, T.M., Pechous, R., Casey, M., Zahrt, T.C., and Frank, D.W. 2006. In vivo Himar1-based transposon mutagenesis of *Francisella tularensis*. Appl. Environ. Microbiol. **72**(3): 1878–1885. doi:10.1128/AEM.72.3.1878-1885.2006.

Mank, J.E., and Avise, J.C. 2006. Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes. Genetica **127**(1–3): 321–327. doi:10.1007/s10709-005-5248-0.

Matveev, V., and Okada, N. 2009. Retroposons of salmonoid fishes (Actinopterygii: Salmonoidei) and their evolution. Gene **434**(1–2): 16–28. doi:10.1016/j.gene.2008.04.022.

Mayrose, I., Zhan, S.H., Rothfels, C.J., Arrigo, N., Barker, M.S., Rieseberg, L.H., and Otto, S.P. 2015. Methods for studying polyploid diversification and the dead end hypothesis: A reply to Soltis et al. (2014). New Phytol. **206**(1): 27–35. doi:10.1111/nph.13192.

Mayrose, I., Zhan, S.H., Rothfels, C.J., Magnuson-Ford, K., Barker, M.S., Rieseberg, L.H., and Otto, S.P. 2011. Recently formed polyploid plants diversify at lower rates. Science (80-.). **333**(6047): 1257. doi:10.1126/science.1207205.

McCarthy, E.M., and McDonald, J.F. 2003. LTR STRUC: A novel search and identification program for LTR retrotransposons. Bioinformatics **19**(3): 362–367. doi:10.1093/bioinformatics/btf878.

McGaugh, S.E., Gross, J.B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., Hinaux, H., Jeffery, W.R., Keene, A., Ma, L., Minx, P., Murphy, D., O'Quin, K.E., Rétaux, S., Rohner, N., Searle, S.M.J., Stahl, B.A., Tabin, C., Volff, J.-N., Yoshizawa, M., and Warren, W.C. 2014. The cavefish genome reveals candidate genes for eye loss. Nat. Commun. **5**: 5307. doi:10.1038/ncomms6307.

Meirmans, P.G., and Van Tienderen, P.H. 2013. The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity (Edinb). **110**(2): 131–137. Nature Publishing Group. doi:10.1038/hdy.2012.80.

Miskey, C., Izsvák, Z., Plasterk, R.H., and Ivics, Z. 2003. The Frog Prince: A reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells. Nucleic Acids Res. **31**(23): 6873–6881. doi:10.1093/nar/gkg910.

Moir, R.D., and Dixon, G.H. 1988. A repetitive DNA sequence in the salmonid fishes similar to a retroviral long terminal repeat. J. Mol. Evol. **27**(1): 1–7. Available from http://www.ncbi.nlm.nih.gov/pubmed/3133484.

Munoz-Lopez, M., and Garcia-Perez, J. 2010. DNA transposons: nature and applications in genomics. Curr. Genomics **11**(2): 115–128. doi:10.2174/138920210790886871.

Nam, K., and Ellegren, H. 2012. Recombination drives vertebrate genome contraction. PLoS Genet. **8**(5). doi:10.1371/journal.pgen.1002680.

Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., Friedman, M., and Smith, W.L. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. **109**(34): 13698–13703. doi:10.1073/pnas.1206625109.

Nelson, J.S., Grande, T.C., and Wilson, M.V.H. 2016. Fishes of the World. *In* 5th edition. Wiley. doi:10.1002/9781119174844.

Nikaido, M., Noguchi, H., Nishihara, H., Toyoda, A., Suzuki, Y., Kajitani, R., Suzuki, H., Okuno, M., Aibara, M., Ngatunga, B.P., Mzighani, S.I., Kalombo, H.W.J., Masengi, K.W.A., Tuda, J., Nogami, S., Maeda, R., Iwata, M., Abe, Y., Fujimura, K., Okabe, M., Amano, T., Maeno, A., Shiroishi, T., Itoh, T., Sugano, S., Kohara, Y., Fujiyama, A., and Okada, N. 2013. Coelacanth genomes reveal signatures for evolutionary transition from water to land. Genome Res. **23**(10): 1740–1748. doi:10.1101/gr.158105.113.

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., Falcon, F., Knapp, D., Powell, S., Cruz, A., Cao, H., Habermann, B., Hiller, M., Tanaka, E.M., and Myers, E.W. 2018. The axolotl genome and the evolution of key tissue formation regulators. Nature **554**(7690): 50–55. doi:10.1038/nature25458.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., and Pruitt, K.D. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. **44**(D1): D733–D745. doi:10.1093/nar/gkv1189.

Obbard, D.J., Gordon, K.H.., Buck, A.H., and Jiggins, F.M. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. B Biol. Sci. **364**(1513): 99–115. doi:10.1098/rstb.2008.0168.

Ohno, S. 1970. Evolution by gene duplication. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-86659-3.

Okada, N. 1991. SINEs: Short interspersed repeated elements of the eukaryotic genome. Trends Ecol. Evol. **6**(11): 358–361. doi:10.1016/0169-5347(91)90226-N.

Otto, S.P. 2007. The evolutionary consequences of polyploidy. Cell **131**(3): 452–462. doi:10.1016/j.cell.2007.10.022.

Otto, S.P., and Whitton, J. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. **34**(1): 401–437. doi:10.1146/annurev.genet.34.1.401.

Pace, J.K., and Feschotte, C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. **17**(4): 422–432. doi:10.1101/gr.5826307.

Pace, J.K., Gilbert, C., Clark, M.S., and Feschotte, C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc. Natl. Acad. Sci. **105**(44): 17023–17028. doi:10.1073/pnas.0806548105.

Panaud, O. 2016. Horizontal transfers of transposable elements in eukaryotes: the flying genes. Comptes Rendus - Biol. **339**(7–8): 296–299. doi:10.1016/j.crvi.2016.04.013.

Parisod, C., and Senerchia, N. 2012. Responses of transposable elements to polyploidy. *In* Topics in Current Genetics. pp. 147–168. doi:10.1007/978-3-642-31842-9-9.

Park, S.G., Hannenhalli, S., and Choi, S.S. 2014. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics **15**(1): 526. doi:10.1186/1471-2164-15-526.

Peccoud, J., Loiseau, V., Cordaux, R., and Gilbert, C. 2017. Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. **114**(18): 4721–4726. doi:10.1073/pnas.1621178114.

Pecinka, A., Fang, W., Rehmsmeier, M., Levy, A.A., and Mittelsten Scheid, O. 2011. Polyploidization increases meiotic recombination frequency in *Arabidopsis*. BMC Biol. **9**. doi:10.1186/1741-7007-9-24.

Van De Peer, Y., Maere, S., and Meyer, A. 2009. The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10**(10): 725–732. doi:10.1038/nrg2600.

Van de Peer, Y., Mizrachi, E., and Marchal, K. 2017. The evolutionary significance of

polyploidy. Nat. Rev. Genet. **8**(3): 206–16. Nature Publishing Group. doi:10.1038/nrg.2017.26.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. 2000. Evidence for DNA loss as a determinant of genome size. Science (80-.). **287**(5455): 1060–1062. doi:10.1126/science.287.5455.1060.

Phillips, R., and Ráb, P. 2001. Chromosome evolution in the Salmonidae (Pisces): an update. Biol. Rev. Camb. Philos. Soc. **76**(1): 1–25. doi:10.1111/j.1469-185X.2000.tb00057.x.

Phillips, R.B., Keatley, K. a, Morasch, M.R., Ventura, A.B., Lubieniecki, K.P., Koop, B.F., Danzmann, R.G., and Davidson, W.S. 2009. Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). BMC Genet. **10**: 46. doi:10.1186/1471-2156-10-46.

Piednoel, M., Sousa, A., and Renner, S.S. 2015. Transposable elements in a clade of three tetraploids and a diploid relative, focusing on Gypsy amplification. Mob. DNA **6**(1): 5. doi:10.1186/s13100-015-0034-8.

Piégu, B., Bire, S., Arensburger, P., and Bigot, Y. 2015. A survey of transposable element classification systems - a call for a fundamental update to meet the challenge of their diversity and complexity. Mol. Phylogenet. Evol. **86**: 90–109. doi:10.1016/j.ympev.2015.03.009.

Plasterk, R.H.A., Izsvák, Z., and Ivics, Z. 1999. Resident aliens the Tc1/mariner superfamily of transposable elements. Trends Genet. **15**(8): 326–332. doi:10.1016/S0168-9525(99)01777-1.

Plasterk, R.H.A., and van Luenen, H.G.A.M. 2002. Mobile DNA II. *Edited by* N.L. Craig, A.M. Lambowitz, R. Craigie, and M. Gellert. American Society of Microbiology. doi:10.1128/9781555817954.

Platt, R.N., Blanco-Berdugo, L., and Ray, D.A. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol. Evol. **8**(2): 403–410. doi:10.1093/gbe/evw009.

Polak, P., and Domany, E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics **7**(1): 133. doi:10.1186/1471-2164-7-133.

Pop, M. 2009. Genome assembly reborn: Recent computational challenges. Brief. Bioinformatics. **10**(4): 354–366. doi:10.1093/bib/bbp026.

Precision Medicine Initiative (PMI) Working Group. 2015. The precision medicine initiative cohort program – building a research foundation for 21st century medicine. Precis. Med. Initiat. Work. Gr. Rep. to Advis. Comm. to Dir. NIH **Sept 17**: 1–108.

Available from http://www.nih.gov/precisionmedicine/.

Price, A.L., Jones, N.C., and Pevzner, P.A. 2005. *De novo* identification of repeat families in large genomes. Bioinformatics **21**(Supp. 1): 351–358. doi:10.1093/bioinformatics/bti1018.

Pritham, E.J., Putliwala, T., and Feschotte, C. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene **390**(1–2): 3–17. doi:10.1016/j.gene.2006.08.008.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. **41**(D1): D590–D596. doi:10.1093/nar/gks1219.

Quesneville, H., Nouaud, D., and Anxolabéhère, D. 2003. Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. J. Mol. Evol. **57**(Supp. 1): S50-9. doi:10.1007/s00239-003-0007-2.

Radice, a D., Bugaj, B., Fitch, D.H., and Emmons, S.W. 1994. Widespread occurrence of the Tc1 transposon family: Tc1-like transposons from teleost fish. Mol. Gen. Genet. **244**(6): 606–12. Available from http://www.ncbi.nlm.nih.gov/pubmed/7969029.

Rajasingh, H., Våge, D.I., Pavey, S. a, and Omholt, S.W. 2007. Why are salmonids pink? Can. J. Fish. Aquat. Sci. **64**(11): 1614–1627. doi:10.1139/f07-119.

Ramsey, J., and Schemske, D.W. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. **29**(1): 467–501. doi:10.1146/annurev.ecolsys.29.1.467.

Ray, D.A., Feschotte, C., Pagan, H.J.T., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W., and Craig, N.L. 2008. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. **18**(5): 717–728. doi:10.1101/gr.071886.107.

Rebollo, R., Romanish, M.T., and Mager, D.L. 2011. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet. **46**(1): 120913153128008. doi:10.1146/annurev-genet-110711-155621.

Resch, A.M., Carmel, L., Mariño-Ramírez, L., Ogurtsov, A.Y., Shabalina, S.A., Rogozin, I.B., and Koonin, E. V. 2007. Widespread positive selection in synonymous sites of mammalian genes. Mol. Biol. Evol. **24**(8): 1821–1831. doi:10.1093/molbev/msm100.

Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R., Redmond, A.K., Martin, S.A.M., Holland, P.W.H., Sandve, S.R., Macqueen, D.J., Lien, S., Martin, S.A.M., Holland, P.W.H., Sandve, S.R., and Macqueen, D.J. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication

and evolutionary diversification. Genome Biol. **18**(1): 111. Genome Biology. doi:doi:10.1101/098582.

Roeder, G.S., and Fink, G.R. 1982. Movement of yeast transposable elements by gene conversion. Proc. Natl. Acad. Sci. U. S. A. **79**(18): 5621–5625. doi:10.1073/pnas.79.18.5621.

Rondeau, E.B., Minkley, D.R., Leong, J.S., Messmer, A.M., Jantzen, J.R., Von Schalburg, K.R., Lemon, C., Bird, N.H., and Koop, B.F. 2014. The genome and linkage map of the northern pike (*Esox lucius*): Conserved synteny revealed between the salmonid sister group and the neoteleostei. PLoS One **9**(7). doi:10.1371/journal.pone.0102089.

Le Rouzic, A., and Capy, P. 2006. Population genetics models of competition between transposable element subfamilies. Genetics **174**(2): 785–793. doi:10.1534/genetics.105.052241.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. 2017. The Human Cell Atlas: from vision to reality. Nature **550**(7677): 451–453. doi:10.1038/550451a.

Sakai, H., Tanaka, T., and Itoh, T. 2007. Birth and death of genes promoted by transposable elements in *Oryza sativa*. Gene **392**(1–2): 59–63. doi:10.1016/j.gene.2006.11.010.

Sanchez, D.H., Gaubert, H., Drost, H.G., Zabet, N.R., and Paszkowski, J. 2017. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat. Commun. **8**(1): 1283. doi:10.1038/s41467-017-01374-x.

Santos, F.C., Guyot, R., do Valle, C.B., Chiari, L., Techio, V.H., Heslop-Harrison, P., and Vanzela, A.L.L. 2015. Chromosomal distribution and evolution of abundant retrotransposons in plants: gypsy elements in diploid and polyploid *Brachiaria* forage grasses. Chromosom. Res. **23**(3): 571–582. doi:10.1007/s10577-015-9492-6.

Schaack, S., Gilbert, C., and Feschotte, C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. **25**(9): 537–546. Elsevier Science London. doi:10.1016/j.tree.2010.06.001.

Schartl, M., Walter, R.B., Shen, Y., Garcia, T., Catchen, J., Amores, A., Braasch, I., Chalopin, D., Volff, J.-N., Lesch, K.-P., Bisazza, A., Minx, P., Hillier, L., Wilson, R.K., Fuerstenberg, S., Boore, J., Searle, S., Postlethwait, J.H., and Warren, W.C. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nat. Genet. **45**(5): 567–72. Nature Publishing Group. doi:10.1038/ng.2604.

Schatz, D.G. 2004. Antigen receptor genes and the evolution of a recombinase. Semin. Immunol. **16**(4): 245–256. doi:10.1016/j.smim.2004.08.004.

Schnable, J.C., Pedersen, B.S., Subramaniam, S., and Freeling, M. 2011. Dose–sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses. Front. Plant Sci. **2**(March): 1–7. doi:10.3389/fpls.2011.00002.

Schranz, E.M., Mohammadin, S., and Edger, P.P. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. Curr. Opin. Plant Biol. **15**(2): 147–153. Elsevier Ltd. doi:10.1016/j.pbi.2012.03.011.

Shankar, R., Grover, D., Brahmachari, S.K., and Mukerji, M. 2004. Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. BMC Evol. Biol. **4**(1): 37. doi:10.1186/1471-2148-4-37.

Shao, F., Wang, J., Xu, H., and Peng, Z. 2018. FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. Database **2018**. doi:10.1093/database/bax106.

De Smet, R., and Van de Peer, Y. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr. Opin. Plant Biol. **15**(2): 168–176. Elsevier Ltd. doi:10.1016/j.pbi.2012.01.003.

Smit, AFA, Hubley, R & Green, P. 2015. RepeatMasker Open-4.0. Available from www.repeatmasker.org.

Smit, A.F.A., and Hubley, R. 2008. RepeatModeler Open-1.0. Available from http://www.repeatmasker.org.

Smith, J.J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M.S., Yandell, M.D., Manousaki, T., Meyer, A., Bloom, O.E., Morgan, J.R., Buxbaum, J.D., Sachidanandam, R., Sims, C., Garruss, A.S., Cook, M., Krumlauf, R., Wiedemann, L.M., Sower, S.A., Decatur, W.A., Hall, J.A., Amemiya, C.T., Saha, N.R., Buckley, K.M., Rast, J.P., Das, S., Hirano, M., McCurley, N., Guo, P., Rohner, N., Tabin, C.J., Piccinelli, P., Elgar, G., Ruffier, M., Aken, B.L., Searle, S.M.J., Muffato, M., Pignatelli, M., Herrero, J., Jones, M., Brown, C.T., Chung-Davidson, Y.W., Nanlohy, K.G., Libants, S. V., Yeh, C.Y., McCauley, D.W., Langeland, J.A., Pancer, Z., Fritzsch, B., De Jong, P.J., Zhu, B., Fulton, L.L., Theising, B., Flicek, P., Bronner, M.E., Warren, W.C., Clifton, S.W., Wilson, R.K., and Li, W. 2013. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat. Genet. **45**(4): 415–421. Nature Publishing Group. doi:10.1038/ng.2568.

Smith, J.J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M.C., Parker, H.J., Cook, M.E., Hess, J.E., Narum, S.R., Lamanna, F., Kaessmann, H., Timoshevskiy, V.A., Waterbury, C.K.M., Saraceno, C., Wiedemann, L.M., Robb, S.M.C., Baker, C., Eichler, E.E., Hockman, D., Sauka-Spengler, T., Yandell, M., Krumlauf, R., Elgar, G., and Amemiya, C.T. 2018. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat. Genet.

50(2): 270–277. Springer US. doi:10.1038/s41588-017-0036-1.

Soltis, D.E., Bell, C.D., Kim, S., and Soltis, P.S. 2008. Origin and early evolution of angiosperms. Ann. N. Y. Acad. Sci. **1133**: 3–25. doi:10.1196/annals.1438.005.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., and Soltis, D.E. 2015. Polyploidy and genome evolution in plants. Curr. Opin. Genet. Dev. **35**: 119–125. Elsevier Ltd. doi:10.1016/j.gde.2015.11.003.

Solyom, S., and Kazazian, H.H. 2012. Mobile elements in the human genome: implications for disease. Genome Med. **4**(2): 1–8. doi:10.1186/gm311.

Song, Q., and Chen, J.Z. 2015. Epigenetic and developmental regulation in plant polyploids. Curr. Opin. Plant Biol. **24**: 101–109. Elsevier Ltd. doi:10.1016/j.pbi.2015.02.007.

Sorek, R., Ast, G., and Graur, D. 2002. Alu-containing exons are alternatively spliced. Genome Res. **12**(7): 1060–1067. doi:10.1101/gr.229302.

Sotero-Caio, C.G., Platt, R.N., Suh, A., and Ray, D.A. 2017. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol. Evol. **9**(1): 161–177. doi:10.1093/gbe/evw264.

Statistics Norway. 2017. Norwegian aquaculture annual preliminary figures. Available from https://www.ssb.no/en/jord-skog-jakt-og-fiskeri/statistikker/fiskeoppdrett/aar-forelopige [accessed 4 March 2018].

Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. 2009. Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. Nucleic Acids Res. **37**(21): 7002–7013. doi:10.1093/nar/gkp759.

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.P., Busby, M., Indap, A.R., Garrison, E., Huff, C., Xing, J., Snyder, M.P., Jorde, L.B., Batzer, M.A., Korbel, J.O., and Marth, G.T. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. **7**(8). doi:10.1371/journal.pgen.1002236.

Stuart, G.R., Dixon, B., and Pohajdak, B. 1992. Isolation of a putative retrovirus pol gene fragment from trout. Comp. Biochem. Physiol. B. **102**(1): 137–42. Available from http://www.ncbi.nlm.nih.gov/pubmed/1526119.

Szitenberg, A., Cha, S., Opperman, C.H., Bird, D.M., Blaxter, M.L., and Lunt, D.H. 2016. Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. Genome Biol. Evol. **8**(9): 2964–2978. doi:10.1093/gbe/evw208.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood,

evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**(10): 2731–2739. doi:10.1093/molbev/msr121.

Taylor, J.S., Van de Peer, Y., Braasch, I., and Meyer, A. 2001a. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos. Trans. R. Soc. B Biol. Sci. **356**(1414): 1661–1679. doi:10.1098/rstb.2001.0975.

Taylor, J.S., Van De Peer, Y., and Meyer, A. 2001b. Genome duplication, divergent resolution and speciation. Trends Genet. **17**(6): 299–301. doi:10.1016/S0168-9525(01)02318-6.

Tiley, G.P., and Burleigh, G. 2015. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. BMC Evol. Biol. **15**(1): 1–14. BMC Evolutionary Biology. doi:10.1186/s12862-015-0473-3.

Tørresen, O.K., Star, B., Jentoft, S., Reinar, W.B., Grove, H., Miller, J.R., Walenz, B.P., Knight, J., Ekholm, J.M., Peluso, P., Edvardsen, R.B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K.S., and Nederbragt, A.J. 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics **18**(1): 95. doi:10.1186/s12864-016-3448-x.

Touchon, M., and Rocha, E.P.C. 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. **24**(4): 969–981. doi:10.1093/molbev/msm014.

Treangen, T.J., and Salzberg, S.L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. **13**(1): 36–46. doi:10.1038/nrg3117.

Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B., Ohta, Y., Flajnik, M.F., Sutoh, Y., Kasahara, M., Hoon, S., Gangu, V., Roy, S.W., Irimia, M., Korzh, V., Kondrychyn, I., Lim, Z.W., Tay, B.H., Tohari, S., Kong, K.W., Ho, S., Lorente-Galdos, B., Quilez, J., Marques-Bonet, T., Raney, B.J., Ingham, P.W., Tay, A., Hillier, L.W., Minx, P., Boehm, T., Wilson, R.K., Brenner, S., and Warren, W.C. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature **505**(7482): 174–179. doi:10.1038/nature12826.

Vicient, C.M., and Casacuberta, J.M. 2017. Impact of transposable elements on polyploid plant genomes. Ann. Bot. **120**(2): 195–207. doi:10.1093/aob/mcx078.

Vindas, M.A., Magnhagen, C., Brännäs, E., Øverli, Ø., Winberg, S., Nilsson, J., and Backström, T. 2017. Brain cortisol receptor expression differs in Arctic charr displaying opposite coping styles. Physiol. Behav. **177**(January): 161–168. Elsevier. doi:10.1016/j.physbeh.2017.04.024.

Volff, J.N., Bouneau, L., Ozouf-Costaz, C., and Fischer, C. 2003. Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet. **19**(12): 674–678. doi:10.1016/j.tig.2003.10.006.

Volff, J.N., Körting, C., and Schartl, M. 2000. Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol. Biol. Evol. **17**(11): 1673–1684. doi:10.1093/oxfordjournals.molbev.a026266.

Van Der Walt, S., Colbert, S.C., and Varoquaux, G. 2011. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2): 22–30. doi:10.1109/MCSE.2011.37.

Webster, M.T., Smith, N.G.C., Hultin-Rosenberg, L., Arndt, P.F., and Ellegren, H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human Alu repeats. Mol. Biol. Evol. **22**(6): 1468–1474. doi:10.1093/molbev/msi136.

Wendel, J.F. 2000. Genome evolution in polyploids. Plant Mol. Biol. **42**(1): 225–249. doi:10.1023/A:1006392424384.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A.H. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**(12): 973–982. doi:10.1038/nrg2165.

Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York, New York, NY. doi:10.1007/978-0-387-98141-3.

Winkfein, R.J., Moir, R.D., Krawetz, S. a, Blanco, J., States, J.C., and Dixon, G.H. 1988. A new family of repetitive, retroposon-like sequences in the genome of the rainbow trout. Eur. J. Biochem. **176**(2): 255–64. Available from http://www.ncbi.nlm.nih.gov/pubmed/2843369.

Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2**(5): 333–41. doi:10.1038/35072009.

Wright, J.E., Johnson, K., Hollister, A., and May, B. 1983. Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. Isozymes Curr. Top. Biol. Med. Res. **10**: 239–60. Available from http://www.ncbi.nlm.nih.gov/pubmed/6354984.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Huo, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., Shan, G., Li, X., Wu, S., Song, L., Hou, G., Jiang, Y., Jeney, Z., Yu, D., Wang, L., Shao, C., Song, L., Sun, J., Ji, P., Wang, J., Li, Q., Xu, L., Sun, F., Feng, J., Wang, C., Wang, S., Wang, B., Li, Y., Zhu, Y., Xue, W., Zhao, L., Wang, J., Gu, Y., Lv, W., Wu, K., Xiao, J., Wu, J., Zhang, Z., Yu, J., and Sun, X. 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nat. Genet. **46**(11): 1212–1219. Nature Publishing Group. doi:10.1038/ng.3098.

Xu, Y., Zhao, Q., Mei, S., and Wang, J. 2012. Genomic and transcriptomic alterations following hybridisation and genome doubling in trigenomic allohexaploid *Brassica*

carinata×*Brassica rapa*. Plant Biol. **14**(5): 734–744. doi:10.1111/j.1438-8677.2011.00553.x.

Xu, Z., and Wang, H. 2007. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**(Supp. 2): W265-8. doi:10.1093/nar/gkm286.

Yaakov, B., and Kashkush, K. 2011. Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. Genome **54**(1): 42–49. doi:10.1139/G10-091.

Yuan, Z., Liu, S., Zhou, T., Tian, C., Bao, L., Dunham, R., and Liu, Z. 2018. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics **19**(1): 141. doi:10.1186/s12864-018-4516-1.

Zagoraiou, L., Drabek, D., Alexaki, S., Guy, J. a, Klinakis, a G., Langeveld, a, Skavdis, G., Mamalaki, C., Grosveld, F., and Savakis, C. 2001. *In vivo* transposition of Minos, a *Drosophila* mobile element, in mammalian tissues. Proc. Natl. Acad. Sci. U. S. A. **98**(20): 11474–8. doi:10.1073/pnas.201392398.

Zhang, G. 2015. Genomics: Bird sequencing project takes off. Nature **522**(7554): 34. doi:10.1038/522034d.

Zhang, J., Liu, Y., Xia, E.-H., Yao, Q.-Y., Liu, X.-D., and Gao, L.-Z. 2015. Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. Proc. Natl. Acad. Sci. **112**(50): E7022–E7029. doi:10.1073/pnas.1515170112.

Appendix

Historical activity of TCEs in rainbow trout

The below figure was created with the same approach as was used for Figure 7a, using the combined Atlantic salmon/rainbow trout TCE library to identify TCE families in the rainbow trout genome. The patterns which emerge are initially very similar to those observed in Atlantic salmon, however they begin to deviate from each other after the time corresponding to 95% similarity, in concert with the divergence of the *Salmo* and *Oncorhynchus* lineages.

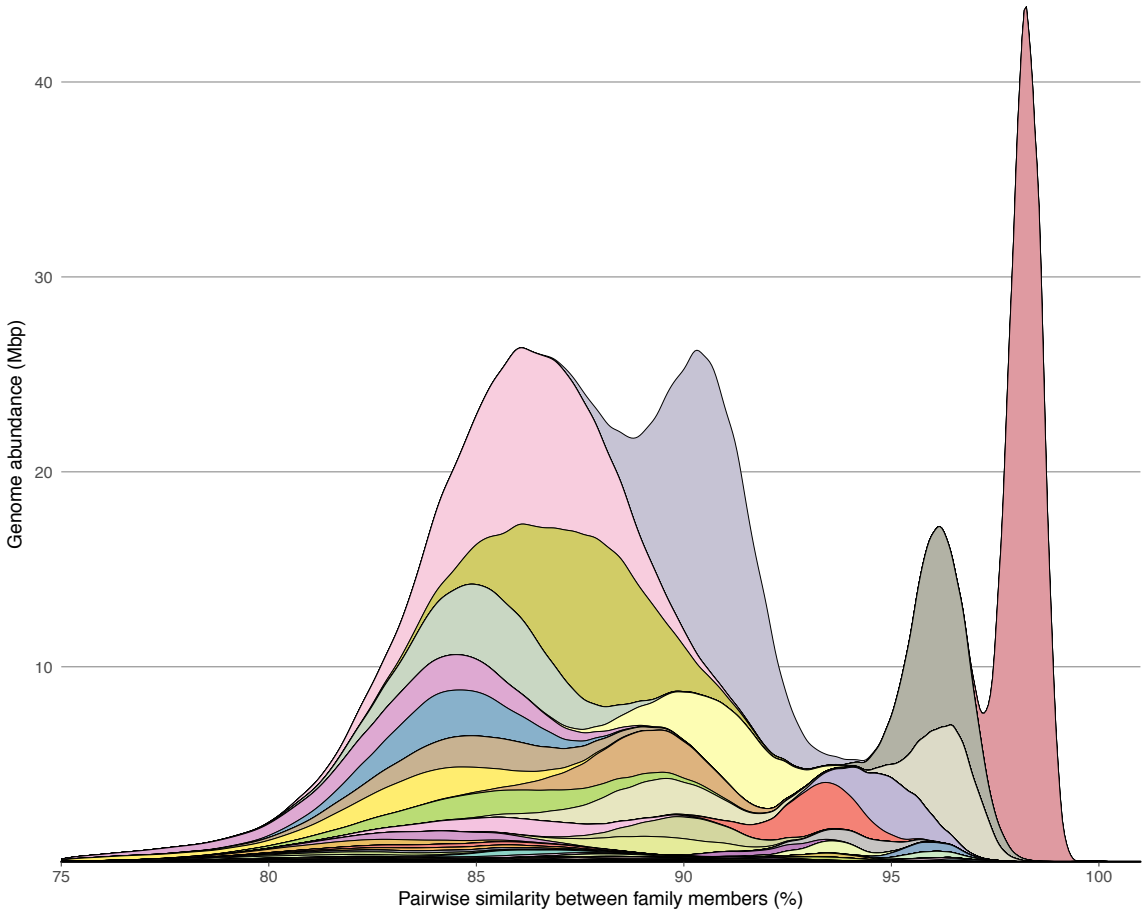

Figure A1 Historical TCE activity in rainbow trout. Stacked TCE age plot scaled by genome abundance. The colours used to represent each individual TCE family are the same as those used for the families in Figure 7a. Historical age of families was estimated by calculating the pairwise similarity between a random subset of family members; if family members were less similar they were created in more ancient duplication events.

TCE activity in five salmonid genomes, with outliers

When determining the average pairwise similarity between family members in order to date TCE activity (as in Figure 8), a data point is generated for each comparison; with a maximum of 200 randomly-chosen TCE sequences per family, up to 19,900 data points can be created. If even a small number of the 200 randomly chosen sequences are not true family members, or are otherwise substantially different from their fellows, a large number of 'outlier' percent similarity values will be generated and obscure the trends illustrated by a boxplot visualization. As such, outliers were removed from Figure 8. In the interest of transparency, the same figure is included below with outliers present.

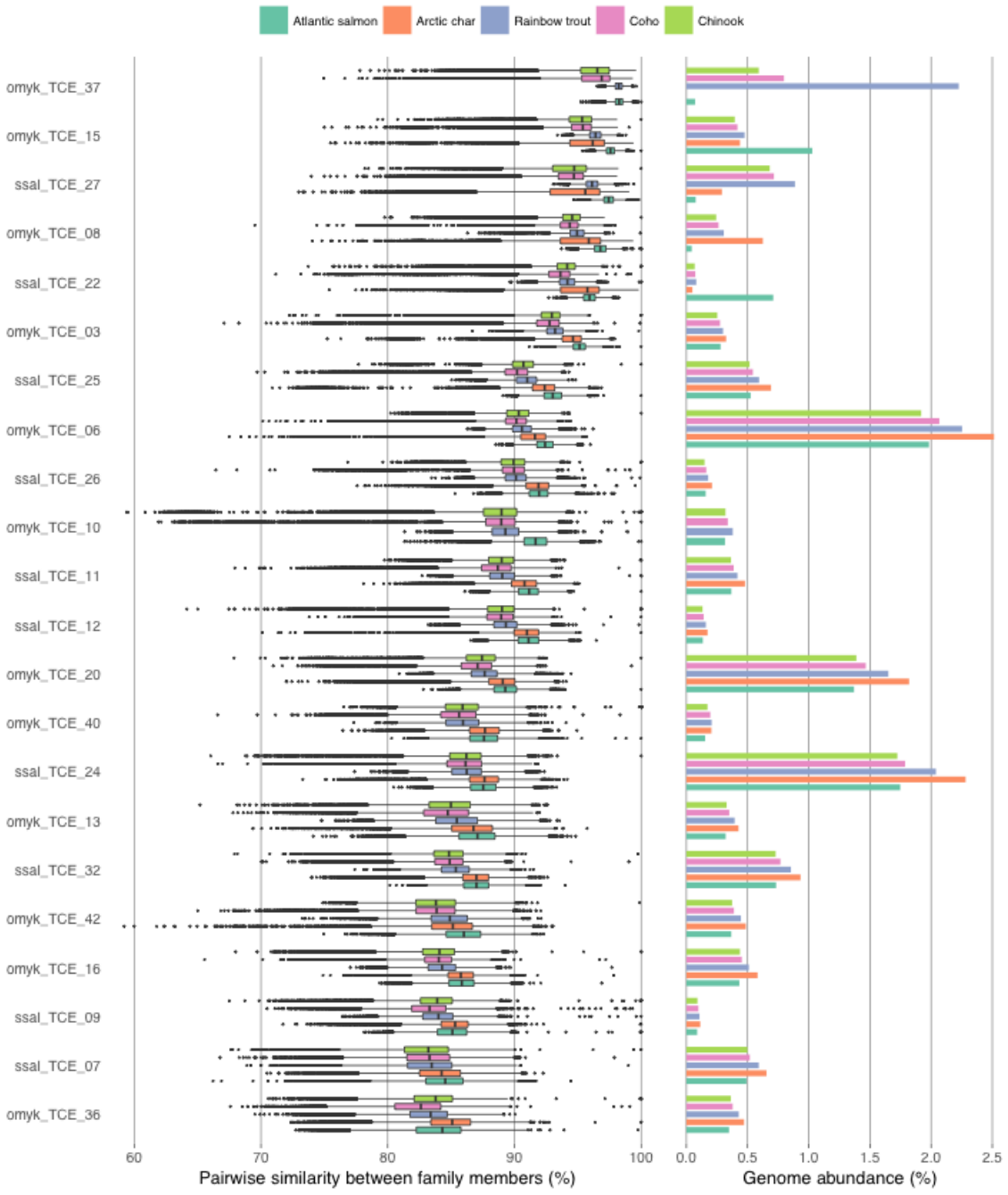

Figure A2 Age and abundance of TCEs in the genomes of five salmonids, with outliers.

Families with increased pairwise similarity between members have experienced less sequence divergence since they were rendered inactive and reflect more recent additions to the genome. No intact elements from family omyk_TCE_37 were found in the genome of arctic char. Elements assigned to omyk_TCE_10 in arctic char are not shown as they clearly include members from two different families (see Figure 4b).

Publications during Masters degree period

† Authors contributed equally

Pearse DF, Barson N, Nome T, Gao G, Campbell M, Abadía-Cardoso A, Anderson E, Rundio DE, Williams TH, Naish KA, Baranski M, Moen T, Liu S, Kent M, **Minkley DR**, [14 others]. Double inversion mediates selection on sex- and temperature-dependent migration in rainbow trout. *In preparation.*

Christensen KA†, Leong JS†, **Minkley DR**†, Rondeau EB, Nugent CM, Danzmann RG, Ferguson MM, Stadnik A, Devlin RH, Davidson WS, Koop BF. The Arctic Charr (*Salvelinus alpinus*) Genome. *In preparation.*

Messmer AM, Leong JS, Rondeau EB, Mueller AM, Despins CA, **Minkley DR**, Kent MP, Lien S, Boyce B, Morrison D, Fast MD, Norman JD, Danzmann RG, Koop BF. A 200K SNP chip reveals a novel Pacific salmon louse genotype linked to differential efficacy of Emamectin benzoate (2018). *Marine Genomics*. doi.org/10.1016/j.margen.2018.03.005

Christensen KA, Leong JS, Sakhrani D, Biagi CA, **Minkley DR**, Withler RE, Rondeau EB, Koop BF, Devlin RH (2018). Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PLoS One* **13**(4): e0195461.

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong J**, Minkley DR**, … [37 others] (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**:200-205.

Minkley DR, Whitney MJ, Lin S, Barsky MG, Kelly C, Upton C (2014). Suffix tree searcher: exploration of common substrings in large DNA sequence sets. *BMC Research Notes* **7**:466.

Rondeau EB, **Minkley DR**, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, Lemon C, Bird NH, Koop BF (2014). The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. *PLoS One* **9**: e102089.

Rondeau EB, Messmer AM, Sanderson DS, Jantzen SG, von Schalburg KR, **Minkley DR**, Leong JS, Macdonald GM, Davidsen AE, Parker WA, Mazzola RSA, Campbell B, Koop BF (2013). Genomics of sablefish (*Anoplopoma fimbria*): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. *BMC Genomics* **14**:452.

von Schalburg KR, Gowen BE, Rondeau EB, Johnson NW, **Minkley DR**, Leong JS, Davidson WS, Koop BF (2013). Sex-specific expression, synthesis and localization of aromatase regulators in one-year-old Atlantic salmon ovaries and testes. *CBP Part B* **164**: 236-246.

Presentations during Masters degree period

Expansion of repeated DNA has differentially shaped the genomes of salmonids. Oral presentation at the 2017 UVic Biology Department Graduate Student Symposium. Victoria, BC, Canada

Salmon bioinformatics: Whole-genome duplications and transposable elements. Invited oral presentation as part of the Bamfield Marine Science Centre 2016 Fall Seminar Series. Bamfield, BC, Canada

Travelling DNA: Horizontal transfer of transposable elements and the evolution of salmon. Oral presentation at the 2014 UVic Biology Department Graduate Student Symposium. Victoria, BC, Canada.

Travelling DNA: Horizontal transfer of transposable elements and the evolution of salmon. Oral presentation at the 2014 Brackendale Ecology and Evolution Retreat. Brackendale, BC, Canada.

Massive proliferation of transposable elements following the salmonid whole-genome duplication. Poster presented in 2014 at the 2nd International Conference on Integrative Salmonid Biology. Vancouver, BC, Canada.

Massive proliferation of Tc1-Mariner transposable elements following the salmonid whole-genome duplication. Poster presented at the 2014 Keystone Symposium on Mobile Genetic Elements and Genome Evolution. Santa Fe, USA.

Evolution of the Salmonids: transposable element proliferation and the Salmonid whole-genome duplication. Oral presentation at the 2014 Pacific Ecology and Evolution Conference (PEEC). Bamfield, BC, Canada.

Proliferation of transposable elements following Salmonid genome duplication. Oral presentation at the 2013 Annual UVic Biology Department Graduate Student Symposium. Victoria, BC, Canada.