Population genomics of Salish Sea chum salmon: The legacy of the salmonid whole genome duplication

Ryan K. Waples

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2015

Reading Committee:

Lisa Seeb, Co-chair

James Seeb, Co-chair

Steven Roberts

Paul Hohenlohe

Program Authorized to Offer Degree:

School of Aquatic and Fishery Sciences

University of Washington

Abstract

Population genomics of Salish Sea chum salmon:

The legacy of the salmonid whole genome duplication

Ryan Kele Waples

Co-chairs of the Supervisory Committee:
Lisa Seeb, Research Professor
School of Aquatic and Fishery Sciences
James Seeb, Research Professor
School of Aquatic and Fishery Sciences

The common ancestor of salmonids underwent a whole genome duplication approximately 88 million years ago. This duplication event still has a lasting impact on the form and structure of salmonid genomes today and is evident in many duplicated genes and ongoing residual tetrasomic inheritance. This duplication also serves to complicate genetic analyses, as paralogous genes and sequences are difficult to distinguish, and often fully excluded prior to study. The goal of this thesis is to demonstrate how to incorporate duplicated loci into genetic studies of salmonids using high-throughput sequencing of chum salmon from the Salish Sea. In the first chapter, I develop a method to resolve paralogous loci within a pedigree and include them on a high-density linkage map. I show that paralogous loci are concentrated in 16 regions near the

ends of linkage groups. These regions are inferred to have ongoing residual tetrasomic inheritance and we find that they have a lower incidence of transposable elements than the rest of the genome, a possible explanation for their stability since the whole genome duplication. In the second chapter, I use the discovered paralogous loci in a population genetic study of 10 collections of chum salmon from the Salish Sea. I compare genetic diversity and population structure at paralogous and non-paralogous loci and conduct a genome scan for association with run timing. I demonstrate that it is possible to characterize paralogous loci in wild populations and that they show similar patterns of population structure as the rest of the genome. The genome scan reveals genomic regions of elevated association with run timing, highlighting the potential downside of excluding paralogous loci in studies looking for genetic signals of adaptation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my graduate advisors Lisa and Jim Seeb. They have given me both the freedom and the direction necessary to succeed, all the while showing me what it is to be a professional scientist. I am also indebted to my other committee members Paul Hohenlohe and Steven Roberts for their insight and guidance.

I would also like to thank my lab mates, particularly Morten Limborg, who has been a thoughtful and engaging friend, Wes Larson for showing me how to get stuff done, and Carita Pascal for keeping me out of the laboratory. I would be remiss if I didn't acknowledge the support of many others along the way: Garret McKinney, Carolyn Tarpey, Daniel Gomez Uchida, Fred Utter, Tyler Dann, Bill Templin, Ken Warheit, Marine Brieuc, Marissa Jones, Meredith Everett, Sewall Young, Linda Park, thank you all.

Last, and certainly not least, I thank my family. They have always supported me completely, and without hesitation.

DEDICATION

For my father, my role model in science and in life

LINKAGE MAPPING WITH PARALOGS EXPOSES REGIONS OF RESIDUAL TETRASOMIC INHERITANCE IN CHUM SALMON (*ONCORHYNCHUS KETA*)[1]

1.1 ABSTRACT

Gene sequence similarity due to shared ancestry after a duplication event, i.e. paralogy, complicates the assessment of genetic variation, as sequences originating from paralogs can be difficult to distinguish. These confounded sequences are often removed prior to further analyses, leaving the underlying loci uncharacterized. Salmonids have only partially re-diploidized subsequent to a whole-genome duplication; residual tetrasomic inheritance has been observed in males. We present a maximum likelihood-based method to resolve confounded paralogous loci by observing the segregation of alleles in gynogenetic haploid offspring and demonstrate its effectiveness by constructing two linkage maps for chum salmon (*Oncorhynchus keta*): with and without these newly resolved loci. We find that the resolved paralogous loci are not randomly distributed across the genome. A majority are clustered in expanded sub-telomeric regions of 14 linkage groups, suggesting a significant fraction of the chum salmon genome may be missed by the exclusion of paralogous loci. Transposable elements have been proposed as drivers of genome evolution and, in salmonids, may have an important role in the re-diploidization process by driving differentiation between homeologous chromosomes. Consistent with that hypothesis, we find a reduced fraction of transposable element annotations among paralogous loci, and these loci predominately occur in the genomic regions that lag in the re-diploidization process.

[1] Citation: Waples, R. K., Seeb, L. W. and Seeb, J. E. (2015), Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). Molecular Ecology Resources. doi: 10.1111/1755-0998.12394

1.2 INTRODUCTION

Gene and chromosome duplications appear in the evolutionary history of all species. These duplications create two paralogous sequences from a single ancestral sequence. Paralogs can be identified by sequence alignment, but sequence similarity complicates genetic analysis. The genetic variation observed within and between paralogs is often confounded, leaving them uncharacterized. Failure to differentiate paralogs and correctly resolve loci confuses the assessment of genetic variation and complicates assemblies of genomes and transcriptomes (Davidson *et al.* 2010; Seeb *et al.* 2011a; Wang *et al.* 2011). Paralogs are especially difficult to identify and resolve in non-model species that lack a high-quality reference genome.

A common strategy when faced with confounded paralogs is to identify and exclude them. For example, paralogous sequence variants (PSVs), i.e. variant calls resulting from the alignment of paralogous sequences, can be distinguished from SNPs by assessing measures of heterozygosity and Hardy-Weinberg equilibrium (e.g. Davidson *et al.* 2010; Keller *et al.* 2013). Excluding paralogous loci impoverishes our genetic understanding by discarding all genealogical information they contain, but is often necessary given our inability to resolve and genotype them. The cumulative effect of this exclusion on genetic inferences is not clear, but the potential for bias is real, especially if the excluded loci experience different rates of evolutionary forces such as genetic drift and selection than the rest of the genome.

Rather than exclusion, another approach to dealing with paralogs in non-model species is to use a distance-based metric to separate paralogs based on the underlying sequences (e.g. Seeb *et al.* 2011a; Catchen *et al.* 2013). This approach works well if the genetic distances between sequence haplotypes form a hierarchical pattern with larger differences between haplotypes originating from paralogous loci than from the haplotypes within each locus. Conversely, this approach will fail if haplotypes are not locus-specific and are present at both paralogous loci, a situation akin to incomplete lineage sorting. Thus, this distance-based approach depends on the ability to distinguish alleles and the underlying pattern of divergence.

Paralogy is an acute issue in species with a recent whole genome duplication (WGD). WGDs are a duplication of each ancestral chromosome and occur in two forms: auto- and allopolyploid duplications. In autopolyploid duplications, existing chromosomes are duplicated in-place, creating two complete sets of homeologous chromosomes. In allopolyploid

duplications, hybridization assembles two sets of orthologous chromosomes from related species into a single genome. Polysomic inheritance occurs directly after an autopolyploid WGD due to the complete identity of homeologous chromosome pairs (Soltis & Soltis 1999). Subsequently, homeologous chromosomes tend to diverge and the frequency of polysomic inheritance drops during a process of re-diploidization (Makino & McLysaght 2012). The re-diploidization process is not well understood and has been characterized as 'rapid' in yeast (Scannell *et al.* 2006) and 'slow and stepwise' in salmonids (Berthelot *et al.* 2014)

Salmonids are particularly well suited for studying vertebrate genome evolution subsequent to a WGD because they have experienced at least four WGD events. Two (1R and 2R) occurred in the ancestral vertebrate lineage (Dehal & Boore 2005), one (3R) in the ancestral teleost lineage (Jaillon *et al.* 2004) and, most recently, the common ancestor of salmonids underwent an autopolyploid WGD (4R) approximately 100 MYA (Ohno 1970; Macqueen & Johnston 2014). Since their most recent WGD, salmonids have only partially re-diploidized and have genes with both disomic and tetrasomic patterns, but tetrasomic inheritance has only observed in males and is understood not to occur in females (Allendorf 1978; Wright *et al.* 1983; Allendorf & Danzmann 1997). It is not known if the rate of re-diploidization is constant over time, but the 'extreme stability' of the retained salmonid chromosomes subsequent to the WGD (Berthelot et al. 2014) suggest that tetrasomic inheritance is being conserved. The majority of genetic studies in salmonids take steps to identify and remove paralogs (e.g., Hohenlohe *et al.* 2013; Larson *et al.* 2014; many others) Exclusion of paralogous loci creates a potential for bias because modes of inheritance affect evolutionary forces. In particular, tetrasomic inheritance increases effective population size relative to disomic inheritance (Charlesworth 2009), leading to uneven rates of genetic drift across the genome.

Genotype phase and allele dosage are important aspects of genetic data that can be hard to infer from sequence data, especially in polyploids (Dufresne *et al.* 2014). Codominant genotyping methods rely on the assumption that a diploid individual has either one or two distinct alleles at each locus. If two distinct alleles are observed, a heterozygote genotype is inferred; if only a single allele is observed, a homozygous genotype is called. Codominant genotyping breaks down for confounded paralogous loci and in polyploid taxa as the observed allelic presence/absence signals are often consistent with multiple underlying genotypes. Haploids are relatively easy to produce in salmonids (Spruell *et al.* 1999) and provide an

opportunity to sort paralogs (Brieuc *et al.* 2014; Limborg *et al.* 2014). Haploids have genetic material from only one parent which makes them ideal for constructing linkage maps; genotypes are completely phased, making recombination events easier to locate (Young *et al.* 1998), and PSVs appear as heterozygous genotypes where complete homozygosity is expected (e.g., Palti *et al.* 2014). Here we use gynogenetic haploids, which have their paternal genetic contribution disrupted by UV radiation and thus contain only maternal DNA.

Our three primary objectives are to: 1) develop a method to resolve confounded paralogous loci, 2) build a chum salmon linkage map that includes the resolved loci, and 3) identify and genetically characterize homeologous regions of the chum salmon genome.

We present a novel method to resolve paralogous loci and use it to genotype the maternal parent of a gynogenetic haploid family of chum salmon. We apply a maximum-likelihood approach that extends the work described in Brieuc *et al.* (2014) by formally testing alternative parental genotypes. By following segregation patterns in offspring, we are able to resolve cases where two loci share an allelic sequence (isoloci) and also resolve loci where the sequence similarity between alleles at paralogs and homologs is of the same magnitude. Both of these cases are frequent in salmonids where residual tetrasomic inheritance constrains the divergence of homeologous chromosomes through ongoing gene exchange.

1.3 MATERIALS AND METHODS

1.3.1 *Haploid families*

For this project we required a single family of haploid individuals. We obtained eggs from 12 chum salmon females from the Hoodsport Hatchery, Hoodsport, Washington, USA, for use in this project and other SNP discovery projects (data not shown). Success rate of induced haploidy can vary, and redundancy insured the availability of adequate numbers of families of validated haploids. All animal handling procedures and animal care followed University of Washington IACUC protocol #4229-01. Fin clips were taken from all adults used in the matings (12 chum salmon females and one coho salmon male) and stored in ethanol.

Haploids were produced by fertilizing chum salmon eggs with UV-inactivated sperm from coho salmon as in Seeb & Seeb (1986). Embryos were incubated at a constant 11°C; the date of

hatch was estimated with the software IncubWin. After 42 days, just prior to hatch, the putative haploid embryos were euthanized and removed to ethanol.

DNA was extracted from adult and embryo tissues using DNEasy-96 kits from Qiagen (Venlo, Netherlands). Haploidy was confirmed by screening the parents and embryos with 5'-nuclease assays developed in chum and coho salmon (Smith *et al.* 2006; Elfstrom *et al.* 2007; Petrou *et al.* 2013). Only embryos expressing no paternal (coho salmon) alleles were retained for RAD sequencing; the family with the largest number of haploid offspring was selected for use in this study. Genotypes obtained during haploidy screening were included with the RAD-derived genotypes (see below) and were subject to the same downstream filters and analysis.

1.3.2 *Sequencing*

Haploid and diploid tissues were sequenced on 8 lanes of a HiqSeq2000 (Illumina, San Diego, CA). A total of 192 haploid offspring were prepared for RAD sequencing with the *SbfI* restriction enzyme as per Baird *et al.* (2008) and Etter *et al.* (2011a). Genotyping by sequencing protocols using *SbfI*-based loci have been well optimized for salmonid genomes; use of *SbfI* further enables direct comparisons across populations and species analyzed in a similar fashion (for example see comparisons made by Larson et al. 2014 to data from Everett & Seeb (2014). The *SbfI* recognition sequence is GC-rich (CCTGCA/GG); this may result in an overrepresentation of *SbfI* cut sites in gene-rich regions of the genome. GC-rich sequences have small, but known, biases during PCR (Davey *et al.* 2011) that are not expected to contribute significantly to subsequent analyses. DNA from each haploid was uniquely barcoded (6bp) and multiplexed into seven libraries for single-end sequencing. RAD libraries for the female parent and the 11 other diploid adult chum salmon were prepared as above and multiplexed into one library for paired-end sequencing. Sequencing was conducted for 101 cycles at the Genomics Core Facility at the University of Oregon, with one library per lane.

1.3.3 *Sequence Analysis*

We quality-filtered the raw sequence reads and analyzed the remaining high-quality sequences to discover and genotype SNP loci. Sequence data were received as Phred33 FASTQ files as produced by Cassava (v1.8.x) (Illumina, San Diego, CA). The single-end sequences from the haploids and the P1 sequences from the paired-end sequencing of diploids were de-

multiplexed, stripped of the barcode, trimmed to 84bp, and filtered for chastity and quality with the process_radtags program within the Stacks software suite (v1.05) (Catchen et al. 2013). Stacks pipeline component ustacks was used to discover and assign variant allelic haplotypes to each individual *de novo*. SNP ascertainment (catalog construction) proceeded on three diploid females including the parent using cstacks; three individuals were used to facilitate the relation to other genomic resources. Genotyping proceeded in the haploid offspring by matching the sequences from each individual with the ascertained variation within the catalog using sstacks.

Within the Stacks analysis pipeline, *catalog entries* are groups of aligned sequences. Individual catalog entries nominally represent a unique locus, but they can also contain sequences originating from multiple loci. In these latter cases, we term the catalog entry confounded as it no longer represents a distinct genetic locus. Genetic variation observed within confounded catalog entries may be an artifact of aligning sequences from multiple loci, i.e. paralogous sequence variants. The Stacks method for constructing catalog entries is designed with the goal of grouping genomic locations that exchange alleles (i.e. loci) and splitting those that do not. Our approach seeks to classify variation observed within each catalog entry as intra- or inter-locus and establish parental genotypes at all constituent loci. For simplicity of communication, we refer to both sequence clusters determined by Stacks and the 5' nuclease assays used for haploidy confirmation as "catalog entries"; they were treated identically in downstream analyses.

We used Stacks parameter values similar to those that have been successfully applied to salmonids and that are generally consistent with published protocols (Everett & Seeb 2014; Mastretta-Yanes *et al.* 2015). The –M parameter, determining the maximum number of nucleotide differences used for grouping alleles within a catalog entry, was set to 4. The –m parameter, the minimum depth to observe an allele, was set to 3, and the bounded-error model was applied with an upper bound of 0.05. In one important departure from established methods, we disabled the deleveraging algorithm present in ustacks. The deleveraging algorithm attempts to resolve loci from confounded catalog entries using differences between allelic haplotypes (Catchen *et al.* 2013). Given that residual tetrasomic inheritance provides an avenue for gene flow between paralogs, we do not expect allelic haplotypes to be unique to a single locus and so a distance-based metric was unsuitable. In its place, we leverage the segregation patterns of alleles in gynogenetic haploid offspring (see below). We used all the allelic haplotypes observed

in each haploid individual at each catalog entry because confounded loci may produce genotypes with more than one or two alleles. Offspring with a no-call rate > 0.25 were excluded during a preliminary analysis. Catalog entries with a no-call rate > 0.25, with > 4 alleles, or without variation, were also excluded.

1.3.4 *Assignment of parental genotypes based on segregation patterns*

Segregation patterns of alleles can be used to infer the genotype of the parent, even if two loci are confounded by alignment, analogous to the use of parent-offspring trios as checks against genotyping errors (e.g.; Geller & Ziegler 2003). We apply a maximum-likelihood approach that classifies each catalog entry based on the underlying parental genotype(s). Each parental genotype is expected to produce offspring with genotypes in a particular ratio (**Figure 1.1**). We calculate the likelihood of the observed offspring genotype data given each of the parental genotypes using a multinomial sampling distribution (see Appendix A), a method similar to calculating genotype likelihoods from sequencing data using allele depths (Hohenlohe *et al.* 2010). Each catalog entry is classified with the parental genotype of maximum likelihood. This is a powerful method for assigning parental genotypes, as it is able to leverage genotype information derived from all offspring. Notably, we do not rely on offspring genotypes that include allele dosage; instead, we assume a codominant model where each allele is scored for presence/absence, an important distinction when genotyping polyploid taxa.

As any number of loci can become confounded by alignment, testing all possible maternal genotypes is not feasible. In light of this, we considered a limited set of 18 genotype categories that cover all the relevant cases. A restricted list of the genotype categories considered is presented in **Figure 1.1**; the full list is available in Supplemental File S1. Of the 18 maternal genotype categories, only five have the possibility of recovering segregating loci suitable for inclusion on a linkage map, and an additional nine maternal genotypes do not allow a resolution of the constituent loci. The remaining four categories predict alleles appearing at random in the offspring (one-four allele doses, sampled with replacement) and serve as dummies, meant to attract nonsense allelic segregation patterns.

1.3.5 *Genotyping errors*

We accounted for errors in haploid genotype assignments by including estimates of the genotypic error rate into the likelihood calculations for parental genotypes (see Appendix B). A separate error rate was estimated within each catalog entry for each of the parental genotype categories. We estimated an error rate that is a maximum-likelihood estimate of the rate at which a haploid's genotype call is replaced by a random genotype. The error rate is a function of the number of impossible offspring genotype assignments given the parental genotype under consideration. This approach is similar to that of Stacks (Catchen *et al.* 2013) and Hohenlohe *et al.* (2010), which accounts for sequencing errors by estimating an error rate for each possible genotype. When calculating likelihoods of parental genotypes, we place upper (0.1) and lower bounds (0.01) on the error rate. Bounding the error rate estimates has the desirable effect of penalizing the likelihood of parental allele distributions that result in very high or low estimates of the error rate.

1.3.6 *Resolving confounded loci*

Some confounded catalog entries can be resolved into one or more constituent loci, while others remain unusable. We convert each catalog entry into zero, one, or two distinct loci by observing the segregation of locus-specific alleles (**Figure 1.1**). The resolved loci are segregating in the parent and suitable for linkage mapping. Loci were tested for segregation distortion using the binomial test within the Python package SciPy (Oliphant 2007) and corrected for false-discovery rate (FDR) using the Python package MNE (Gramfort *et al.* 2013). Loci with significant segregation distortion at $a \leq 0.05$ after FDR correction were excluded from further analysis. Loci with > 0.25 missing data, or with a genotyping error rate > 0.2, were also excluded.

Python code used in the analysis is available at the GitHub repository rwaples/ml-psv.

1.3.7 *Linkage Map Construction*

Phase of the diploid female parent was initially unknown and was inferred from offspring genotypes as the first step of linkage map construction. A preliminary linkage map was constructed using arbitrarily phased data (see below for map construction methods). With

arbitrarily phased data we expect to produce twice the final number of linkage groups (LGs), one for each true LG in each phase. Using the linkage group assignment criteria of Wu *et al.* (2008), we identified pairs of loci that would be co-located in alternate phase. We then identified pairs of LGs that contained many loci that would be co-located in alternate phase. Next, we switched the phase of loci in non-overlapping pairs of LGs, and rebuilt the linkage map. Parental phase was visually confirmed in R/qtl (Broman *et al.* 2003) (data not shown).

Linkage maps were constructed with MSTMap (Wu *et al.* 2008); loci were ordered by minimizing the number of inferred of recombination events (COUNT objective function). Loci were spaced with the Kosambi mapping function (Kosambi 1943) due to strong recombination interference within salmonid chromosome arms (Thorgaard *et al.* 1983). Two separate linkage maps were constructed: *Map1* from the 5221 loci resolved from non-confounded catalog entries, and *Map2* including an additional 1015 loci resulting from confounded catalog entries (Supplemental File S1). LGs containing only a single locus were excluded from the final maps. Kendall's rank correlation (τ) was calculated for each corresponding pair of LGs in order to compare the consistency of locus orders between *Map1* and *Map2*.

1.3.8 *Paired-end Assembly*

Paired-end sequence reads from the 12 diploid individuals were quality-filtered with process_radtags using default settings. Paired-end reads were assigned to catalog entries by alignment to the allelic sequences associated with all catalog entries. Only alignments with full identity were accepted. Paired-end sequences assigned to catalog entries present on *Map1* or *Map2* were locally assembled with CAP3 (Huang & Madan 1999) in a process derived from Etter *et al.* (2011b); all reported contigs were retained for annotation (Supplemental File S4). Many confounded catalog entries were comprised of two or more loci that shared alleles (isoloci), preventing the assembly of locus-specific contigs. For this reason, a separate assembly occurred for each catalog rather than each locus. The set of loci comprising a catalog entry received at most a single annotation shared across them.

1.3.9 *BLAST annotation and GO analysis*

Contigs were compared to the Swissprot annotated protein database (version date 12/13/2013) (Magrane & Consortium 2011) with the blastx algorithm (Altschul *et al.* 1990). For

each catalog entry, the lowest e-value match ($< 10^{-4}$) was taken as the protein annotation. In cases of a tie, an annotation was selected at random from among the highest-scoring matches. Catalog entries associated with low-complexity were identified using RepeatMasker (Smit *et al.* 2010). Gene Ontology (GO) terms were assigned to each protein annotation using AmiGO's generic GO slim (Carbon *et al.* 2009). A chi-squared test was applied to the counts for each GO term to test if the term was equally prevalent in the resolved paralogous loci as in the rest of the loci. A false-discovery rate (FDR) procedure was applied to the resulting *p*-values using the fdrtool package (Strimmer 2013); resulting q values <= 0.05 were taken as significant.

1.4 RESULTS

1.4.1 *Screening and sequencing*

Sequencing was successful on the 192 haploid offspring (single-end) and 12 diploids (paired-end). Species-specific genotypes establishing haploidy are available on Dryad (Waples *et al.* 2015a). We excluded 17 offspring with a no-call rate > 0.25, resulting in a final set of 175 haploid offspring averaging 1,681,625 (SD 819,952) retained sequence reads. The female parent had 1,501,228 retained reads, and a total of 69,543,466 sequence reads were retained across all diploid individuals (Supplemental File S2).

We removed 82,023 (83% of 98,913) catalog entries with no variation observed within the haploid family; we also removed 104 catalog entries with more than four allelic sequences observed in the offspring, leaving 16,786.

1.4.2 *Assignment of maternal genotype based on segregation patterns*

The segregation patterns of alleles within the 16,786 polymorphic catalog entries were used to assign maternal genotypes. We identified 6,603 catalog entries consisting of loci with observable segregation patterns (**Table 1.1**). Of these, 5,321 consisted of a single segregating locus; the remaining 1,282 consisted of confounded loci. In 154 of the confounded catalog entries, both loci had observable segregation patterns, allowing the recovery of two loci. In all other cases (1,128) only one of the confounded loci had observable segregation. Of the 6,757 resolved loci, 262 loci failed the test of segregation distortion, 372 loci had estimated genotyping

error rates > 0.2, and 153 loci had > 25% missing data. These sets are not exclusive; 6,236 loci were found eligible for linkage map construction.

Many polymorphic catalog entries could not be resolved into distinct, segregating loci. The female parent was determined to be homozygous for 8,252 (49%) catalog entries; the observed variation was likely due to genotyping errors in one of the offspring. A total of 669 catalog entries were best explained by maternal genotypes that confounded more than two loci, suggesting that the sequences comprising them appear >2 times across the genome, and 1006 catalog entries were best explained by two confounded loci homozygous for alternate alleles. Our dummy models with random allelic presence/absence were assigned to only 15 catalog entries (Supplemental File S3).

1.4.3 *Linkage map construction*

A preliminary linkage map was constructed to determine parental phase. After setting aside LGs comprised of a single locus, the preliminary map included 74 LGs containing 6,159 loci. A total of 3,075 loci on 37 LGs were phase-switched and the linkage map was subsequently rebuilt to demonstrate successful phasing. Two final linkage maps were constructed from the phased data: *Map1* included only loci resulting from non-confounded catalog entries; *Map2* also included paralogous loci resolved from confounded catalog entries. Within *Map1*, 5,221 loci were placed at 3401 unique locations on 37 LGs. In *Map2*, 37 LGs contained a total of 6,162 loci in 4,006 unique locations (**Table 1.2**), with an additional 74 unplaced loci not considered further. The 37 LGs likely correspond to the 37 chromosomes present in chum salmon karyotypes (Sasaki *et al.* 1968; Phillips *et al.* 2007). The 175 offspring provide a potential map resolution of ~0.57 cM (1 Morgan/175), very close to our observed mean marker spacing of ~ 0.6 cM.

1.4.4 *Comparison of Map1 and Map2*

Map2 is longer than *Map1* (3,728 vs 3,246 cM) and has a slightly increased marker density (1.65 vs 1.60 loci per cM) (**Table 1.2**). LG assignment was consistent for all loci between the two maps (**Figure 1.2**). There were some changes in the order and spacing of loci within LGs, shown as crossed blue lines. The few discrepancies in order that do occur are small in scale; mean Kendall's τ rank correlation coefficient of the locus orders was 0.971 (range 0.889 – 0.999) (Supplemental File S3). MSTmap and other linkage mapping algorithms do not produce

confidence/credible intervals for mapping results, complicating the interpretation of between-map comparisons.

1.4.5 *Identification of paralogous loci and homeologous chromosomes*

The distribution of paralogous loci is not random; the majority of these loci (821, 87%) are located on just 14 LGs, each composed of up to 51% paralogs (range 18-51%) (Brieuc *et al.* 2014). Within these 14 LGs, paralogs are concentrated near chromosome ends (**Figure 1.3**). The remaining paralogous loci are scattered, with no apparent pattern, across the remainder of the genome.

The striking pattern of paired paralogous loci provided insight into their origins. Multiple pairs of paralogs were identified on regions of the 14 LGs noted above, forming eight matched sets. The LG pairs are [36,2], [2,14], [30,37], [5,32], [32,10], [13,33], [16,29], and [31,34] (**Figure 1.3**). These paired regions likely connect homeologous portions of LGs that have not diverged due to residual tetrasomic inheritance in salmonid males (e.g., Lien *et al.* 2011). Two of the 14 LGs are present in two distinct pairs (2 and 32), with a discrete association specific to each end of the LG, representing distinct homeologous relationships. In five other cases, we placed pairs of paralogs onto the same LG, always separated by no more than 2 cM; these likely represent segmental duplications and or regions of low sequence complexity. Notably, there are many LGs without identified homeologous relationships with paralogs placed at or near the end of LGs (**Figure 1.2**, **Figure 1.3**) which is consistent with the known concentration of segmental duplications (Riethman 2009; del Carmen Calderón *et al.* 2014).

1.4.6 *Paired-end assembly*

A total of 710,775 (~1% of 69.5M) paired-end sequences were assigned to a catalog entry included on *Map2*, leading to an average of 117.8 (SD 44.3) sequence pairs available for assembly at each catalog entry. Local assembly was successful, with > 99% of catalog entries producing contigs. These assemblies resulted in 14,157 total contigs representing 6,034 loci on *Map2* and ranged in length from 93 to 595 bp.

1.4.7 *BLAST annotation and GO Enrichment*

We successfully annotated 1,049 catalog entries with proteins found in the SwissProt database (Supplemental File S4). Tc1/mariner transposable element-associated (TE) sequences were the most common annotation, constituting 98 (9.3%) of the annotations. A total of 104 catalog entries had more than 50% of their length masked as low-complexity by RepeatMasker. There were 243 distinct GO terms associated with the annotations, occurring from 1 to 562 times each. After FDR correction, 17 GO terms (7.4%) were significantly differently represented between paralogous and non-paralogous loci (< 0.05), with 10 being enriched in paralogous loci, whereas 7 were less frequent. The most enriched GO term was nucleoplasm (0005654), largely due to an abundance of RNA polymerase annotations. Under-represented GO terms included DNA binding (0003677), transposition (0032196), and DNA metabolic process (0006259). We are more interested in general patterns of functional differences between paralogous and non-paralogous categories than specific GO terms. Taken together, the pattern of significant GO terms under-represented in the paralogous loci suggests that fewer TE-associated sequences are found in paralogous loci.

1.5 DISCUSSION

We produced the first dense linkage map of chum salmon, and our method of resolving confounded paralogs allowed the inclusion of an additional ~20% loci. These paralogs were concentrated in sub-telomeric regions of 14 linkage groups. Our method does not distinguish paralogs derived from a whole genome duplication event from paralogs derived from other types of duplication, but the clustered distribution of paralogs on the linkage map is striking and well-explained by residual effects of the ancient WGD. The pairwise association of paralogous regions (black lines, **Figure 1.3**), and the presence of identical alleles at paralogous loci (**Table 1.1**), suggest regions of 14 chromosomes are still undergoing residual tetrasomic inheritance due to incomplete re-diploidization (May *et al.* 1979; Allendorf & Danzmann 1997).

Sub-telomeres are known to harbor segmental duplicates in the most distal 500 kb (cf., Riethman 2009); however, the eight matched regions of paralogs identified here are orders of magnitude larger, given an estimate of the salmon genome of 3 gb (Davidson *et al.* 2010). We infer these to be eight regions of homeology (Lien *et al.* 2011; Brieuc *et al.* 2014). These pairs

likely represent 16 (2*8) ancestral chromatids that joined into 14 chromosomes thorough Robertsonian translocation (Robertson 1916), as in Atlantic salmon (Brenna-Hansen *et al.* 2012), Chinook salmon (Brieuc *et al.* 2014), and coho salmon (Kodama *et al.* 2014). Noticeably, LGs 2 and 32 form two pairwise associations each, one on each end. These LGs are likely metacentric chromosomes formed by the fusion of at least two ancestral chromosome arms.

1.5.1 *Inclusion of paralogous loci on the linkage map*

Linkage maps provide an important resource for the assembly of complicated genomes such as those in species with a recent WGD (e.g., Davidson *et al.* 2010; Felcher *et al.* 2012; Zhao *et al.* 2012). The additional 941 loci included in *Map2* expand the coverage of the genome represented by the linkage map and allow additional recombination events to be observed. The increased map (cM) length of *Map2* reflects these additional recombination events; the increased cM length also reflects increased genotyping error rates among the paralogous loci. Many high-density linkage maps of salmonids are already published (e.g., Miller *et al.* 2012; Everett & Seeb 2014), but most of these exclude paralogous loci, producing an incomplete picture of the genome (but see: Lien *et al.* 2011; Brieuc *et al.* 2014; Kodama *et al.* 2014).

1.5.2 *Transposable elements and genome evolution*

GO terms relating to DNA-binding and transposition are under-represented among the paralogous loci. This is largely due to a reduced fraction of the paralogous loci annotating to transposable element-associated sequences. TEs make up a large fraction of many eukaryotic genomes and have been a driving force in eukaryotic genome evolution (Fedoroff 2012). Following polyploidy, the differential accumulation of TEs between homeologous chromosomes facilitates differentiation and, ultimately, re-diploidization (Parisod et al. 2010). It is not clear if the residual tetrasomic inheritance in salmonids is stable, or if re-diploidization is still ongoing.

Salmonidae is a species-rich family, and TEs can be important in generating Dobzhansky-Muller incompatibilities (DMI) where negative epistatic interactions between paralogs generate reproductive isolation and speciation (de Boer et al. 2007; Brown & O'Neill 2010). Small and isolated populations are ideal conditions for a rapid build-up of DMI, making this model of speciation especially compatible with the populations of anadromous salmonids (Dittman & Quinn 1996). Recent work by Macqueen & Johnston (2014) found little support for the tight

coupling in time of the salmonid 4R WGD and speciation rates. However, their results are consistent with a scenario in which the ancestral WGD provided the raw genomic material that, when coupled with an anadromous life-history and isolated populations, provides ideal conditions for DMI to promote speciation. A dearth of TE annotations in regions of tetrasomic inheritance has not been previously reported in salmonids and merits further investigation in Salmonidae and other partially re-diploidized taxa.

1.5.3 *Evolutionary significance of tetrasomic inheritance*

The lack of tetrasomic inheritance within female salmonids removes the necessity to account for the possibility of tetrasomic inheritance within gynogenetic haploid offspring, but this lack also prevents direct estimates of tetrasomic inheritance rates (Lien *et al.* 2011). The approach of Wu *et al.* (2004) simultaneously estimates rates of homeologous pairing and recombination fractions; this approach could be applied to investigate patterns of tetrasomic inheritance in male meioses. As presented, the method of Wu *et al.* (2004) assumes fully informative loci such that a segregating parent has four distinct alleles. But in the chum salmon linkage map presented here, we find only 8 catalog entries (16 loci) with four distinct alleles in the female parent (i.e., fully informative), so some adaptation would be necessary. In males, rates of homeologous pairing can vary between individuals and populations and are sensitive to outbreeding and hybridization (Allendorf & Danzmann 1997). Chromosome pairing during meiosis is mediated, at least in part, by sequence similarity, which is maintained by gene flow between homeologs (Scannell *et al.* 2006).

Loci undergoing tetrasomic inheritance have larger effective population sizes than the rest of the diploid genome, raising the effectiveness of selection and lessening the effects of drift (Charlesworth 2009). Tetrasomic inheritance can also reduce inbreeding depression by increasing heterozygosity (Tomekpe & Lumaret 1991), particularly relevant for anadromous salmonids with many small, reproductively-isolated populations. The co-occurrence of disomic and tetrasomic regions within chromosomes, as in salmonids, results in loci co-located on a chromosome experiencing different levels of genetic drift and other evolutionary forces.

The common approach of identifying and excluding duplicated loci in genetic studies provides a restricted view of genetic variation and can introduce bias into genetic estimates of population parameters (Meirmans & Van Tienderen 2013). Recent studies have shown elevated

levels of genetic divergence near telomeres during speciation-with-gene-flow (Ellegren *et al.* 2012; Gagnaire *et al.* 2013). In many salmonid chromosomes, these regions are dominated by paralogous loci and excluding them from genomic analyses such as genome scans will return an incomplete account of genomic divergence patterns.

Comparisons across salmonid taxa would facilitate analysis of post-WGD genome structure in a phylogenetic context, providing better dating of significant genome restructuring events and better estimates of the rate of re-diploidization. The approach presented here is directly applicable to other polyploid taxa but cannot be applied to wild populations without a pedigree. In wild populations the inability to observe allele dosage makes estimating basic population genetic parameters such as allele frequency much more difficult (Dufresne *et al.* 2014).

1.6 TABLES

Table 1.1. Catalog entries are classified into four categories.
Non-confounded catalog correspond to individual loci; confounded catalog entries can be resolved into one or two constituent loci. Some catalog entries could not be resolved into loci and are listed as unresolved. Notice that in many confounded genotypes the same allele is present at both constituent loci (e.g. AB/AA).

Category	Parental Genotype	Count
resolved		**6,603**
non-confounded	AB	5,321
confounded, recover one	AB/AA, AA/BC	1,128
confounded, recover two	AB/AC, AB/CD	154
unresolved	**all others**	**10,183**

Table 1.2. The scope, length, and density of Map1 (paralogs removed) and Map2 (including paralogs).

	Loci	Unique map positions	# of LGs	Mean loci per LG	Total length (cM)	LG size range (cM)	Loci per cM
Map1	5,221	3,401	37	141.1	3,246	51 - 145	1.60
Map2	6,162	4,006	37	166.5	3,728	54 - 174	1.65

1.7 FIGURES

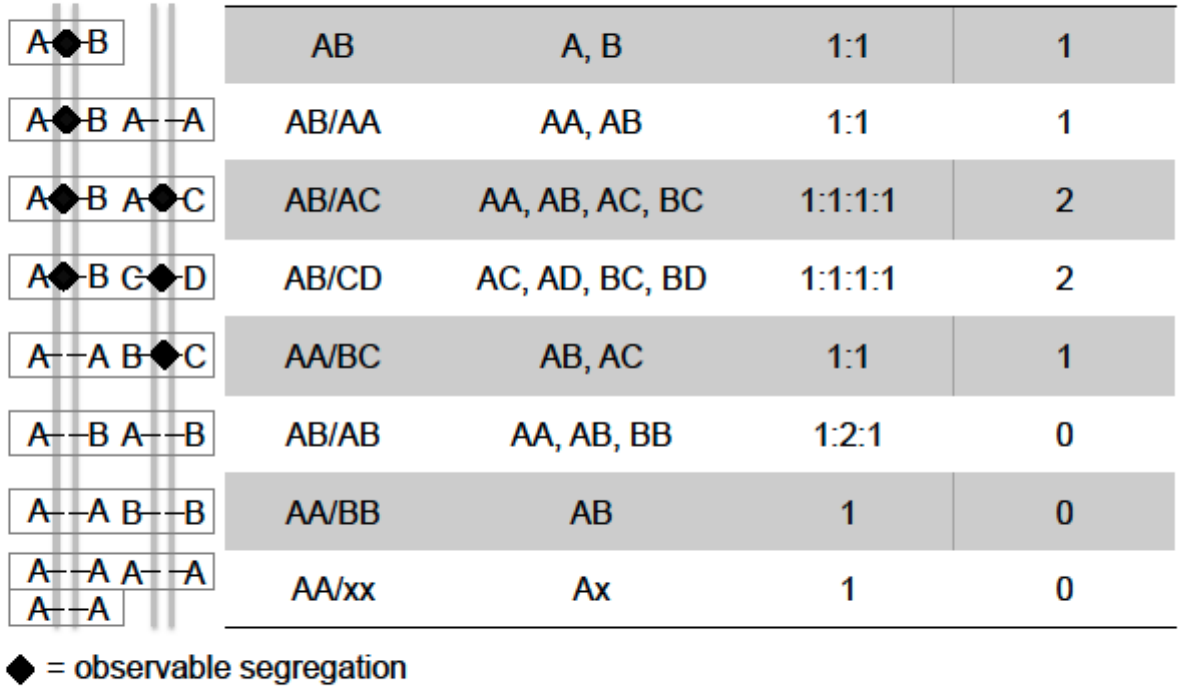

Figure 1.1. Model relating parental and offspring genotypes.
Confounded catalog entries can be resolved into their underlying loci by observing the segregation of up to four alleles. Maternal chromosomes diagram the location of maternal alleles across one or more loci, here assumed to be on separate chromosomes for simplicity. Maternal chromosomes are assumed to segregate strictly disomically (see Methods). A, B, C, and D are alleles; AA/xx represents our inability to distinguish any number of confounded homozygous loci; boxes connect maternal loci that align in a *de novo* analysis. The maternal alleles segregate disomically, forming the offspring genotypes shown in column three in the ratio seen in column four. Observed haploid genotypes are matched to the segregation patterns expected for each maternal genotype with a maximum-likelihood model (see Appendix A). This figure is incomplete, see Supplementary File S1 for a list of considered parental genotypes.

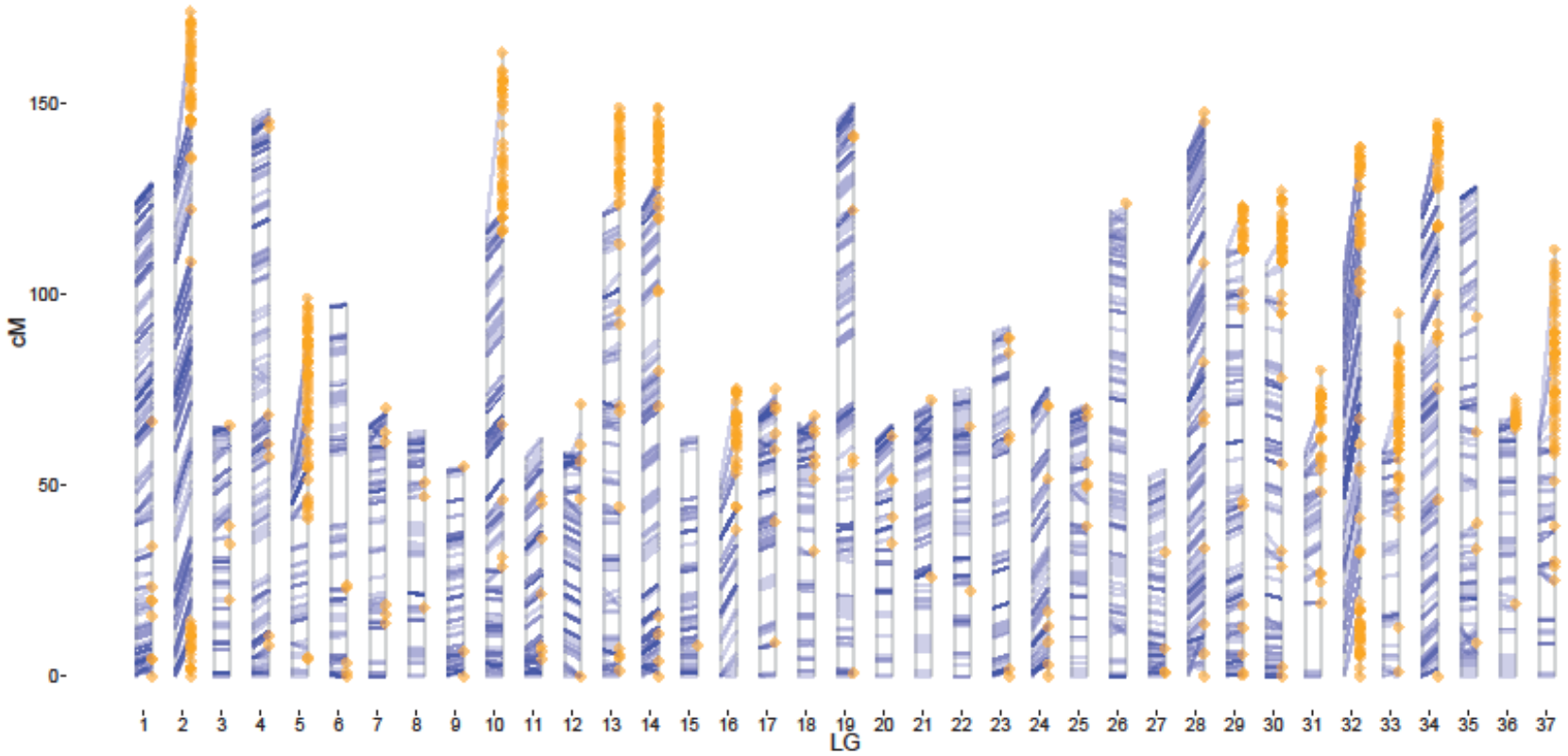

Figure 1.2. Linkage maps constructed with and without the inclusion of paralogs.

Each of 37 linkage groups is represented by two adjacent parallel vertical lines; Map1 (no paralogs) on the left and Map2 (with paralogs) on the right. Blue lines connect the positions of loci appearing on both maps. Yellow diamonds are loci resolved from confounded catalog entries (paralogs) and appear only on Map2. LGs are numbered 1-37 according to the number of loci present on each LG in Map2.

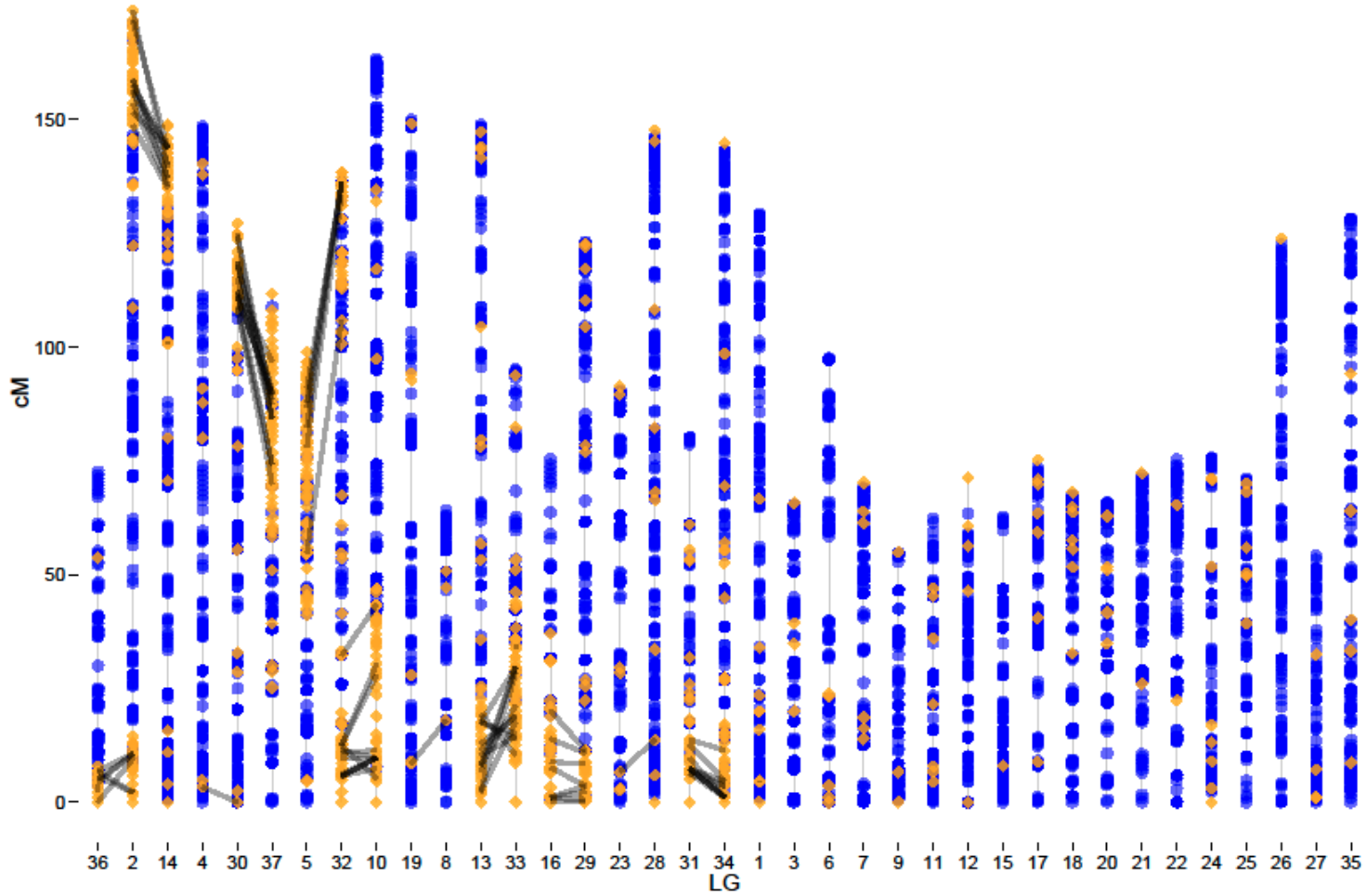

Figure 1.3. Identification of homeologous chromosomes and regions of residual tetrasomic inheritance in *Map2*, which includes resolved paralogous loci.

Non-duplicated loci are shown as blue circles and duplicated loci presented as yellow diamonds. Black lines connect confounded paralogs that have been resolved into two loci. The 16 sub-telomeric concentrations of paralogs form 8 pairs; notice LGs 2 and 32 form distinct pairings on each end. LGs have been reordered from Figure 1.2 to facilitate illustration.

1.8 ACKNOWLEDGMENTS

The authors thank Meredith Carita Pascal for RADseq library prep and primer-based genotyping. We would also like to thank Morten Limborg, Garrett McKinney, and three anonymous reviewers for comments on the manuscript, and Paul Hohenlohe and Steven Roberts for constructive discussion. We would like to thank Ken Warheit and Sewall Young from the Washington Dept. of Fish and Wildlife for biological samples and stimulating conversation. Funding contributing to this research was from NOAA Saltonstall-Kennedy Award NA10NMF4270310, Pacific Salmon Commission Southern Boundary Restoration and Enhancement Fund, and the Gordon and Betty Moore Foundation.

1.9 DATA ACCESSIBILITY

Genotype data are available on Dryad (doi:10.5061/dryad.5b64r).

Python scripts developed for performing the segregation analyses are available on the Github repository "rwaples/ml-psv".

1.10 SUPPLEMENTAL INFORMATION

Supplemental information for this chapter is available online associated with the publication:

Waples, R. K., Seeb, L. W. and Seeb, J. E. (2015), Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). Molecular Ecology Resources. doi: 10.1111/1755-0998.12394

List of supplemental files:

S1.docx Parental allele distribution models (word document)

S2.docx Sequencing read counts and quality filtering (word document)

S3.zip Linkage Mapping (zip archive)

- catalog.stats.txt –Model results for each catalog entry
- kendall_tau.txt – Kendall's tau statistic comparing locus orders
- Map1.txt – Linkage Map 1 (Map1)
- Map2.txt – Linkage Map 2 (Map2)
- Map1_MSTmap.txt – MSTmap input file for Map1
- Map2_MSTmap.txt – MSTMap input file Map2

S4.zip Paired-end assembly and annotation (zip archive)

- P1_consensus.fasta –consensus sequence for each mapped catalog entry (FASTA)
- assembled_contigs.fasta – all contigs assembled for each mapped catalog entry (FASTA)
- swissprot_annotations.txt – BLAST results for the assembled contigs, in BLAST tabular format: (queryId, subjectId, percIdentity, alnLength, mismatchCount, gapOpenCount, queryStart, queryEnd, subjectStart, subjectEnd, eVal, bitScore)
- GO_enrichment.txt – GO enrichment tests (txt file)

Chapter 2. POPULATION GENOMICS OF SALISH SEA CHUM SALMON: POPULATION STRUCTURE AND GENETIC BASIS FOR RUN TIMING INVESTIGATED AT PARALOGOUS LOCI.

2.1 ABSTRACT

Whole genome duplications are major evolutionary event with a lasting impact on genome structure and evolution. The effects of an ancient whole genome duplication approximately 88MYA are still evident in salmonids, in extensive gene duplicates and partial tetrasomic inheritance. While studies have started to reconstruct the complex history of chromosome evolution within salmonids, the continued impact of the whole genome duplication on current patterns of genetic diversity is less well understood. Here were use high-throughput sequencing on ten collections of chum salmon from the Salish Sea in the USA and Canada to investigate genetic diversity and population structure in both tetrasomic and re-diploidized regions of the genome. We start by improving a dense linkage map to identify paralogous loci and use this map to investigate genetic variation across the chum salmon genome. By applying multivariate statistical methods, we show that paralogous genetic loci can be characterized using high-throughput sequencing data and that they display similar patterns of population structure as the rest of the genome. We also investigate genetic associations with run timing, an adaptively important trait that can facilitate reproductive isolation among salmon populations. We show that by including paralogous loci in genome scans we can observe evolutionary signals in genomic regions that have routinely been excluded from population genetic studies in salmon, opening up new avenues for discovery.

2.2 INTRODUCTION

Whole genome duplications are present in the genetic lineages of most eukaryotic species (Crow *et al.* 2006). Despite this, the implications of historical whole genome duplications on current patterns of genetic diversity and genome structure are not fully understood (Comai 2005; Otto 2007). Duplication events are often followed by a period of elevated genome instability, or 'rediploidization', during which many duplicate genes are lost or inactivated. This process has

been characterized as ‘slow and stepwise’ or ‘massive and rapid’ depending on the species and timescale in question (Berthelot *et al.* 2014). The speed and extent of the rediploidization process has implications for speciation (Rieseberg & Willis 2007), adaptation (Selmecki *et al.* 2015), and genetic diversity (Arnold *et al.* 2012). However, many tools of genetic analysis are not well suited to handle paralogous loci.

Polyploid and polyploid-derived genomes complicate many aspects of genetic inference, including genotyping and the assessment of genetic diversity and divergence (Meirmans & Van Tienderen 2013; Dufresne *et al.* 2014). Genotype calling, i.e. determining the genotype for an individual at a particular locus, is more difficult in polyploids and in polyploid-derived genomes. For example, in tetrasomic loci, four chromosomes contribute alleles, and there is increased uncertainty in allele dosage compared to disomic loci. With disomic bi-allelic loci there are three possible genotypes, the heterozygote and the two alternate homozygotes, but each genotype is only consistent with a single pattern of allelic presence/absence. However, with four chromosomes, a unique genotype cannot be determined until a dosage is assigned to each allele, and a simple determination based on presence/absence is insufficient. Additionally genetic analyses methods are often based on assumptions that may not hold, or are difficult to test for paralogous loci. Examples of such assumptions include Hardy-Weinberg equilibrium (HWE), full knowledge of allele dosage and population allele frequencies, and a uniformity of evolutionary forces, such as drift, across the genome.

Salmonids, including whitefish, char and graylings, are an ideal vertebrate system for studying genome evolution subsequent to a whole genome duplication. The common ancestor of all salmonids underwent an autopolyploid whole genome duplication approximately 88 million years ago (Macqueen & Johnston 2014), and the genomic footprint of the ancient duplication event is still evident. Salmonid genomes have many more paralogous sequences than species more removed from WGDs (Wright *et al.* 1983). In Pacific salmon (*Oncorhynchus sp.*) these retained paralogous sequences are not dispersed across the genome, but are concentrated on the distal ends of eight conserved chromosome arms (chapter one, Waples *et al.* (2015b)). In these regions of retained paralogy, rediploidization is not complete, and more than two chromosomes can pair at meiosis, resulting in a multivalent group with tetrasomic inheritance (May *et al.* 1979; Wright *et al.* 1983). The complex pattern of recombination between homologous and homeologous chromosomes results in an atypical arrangement of sequence similarity between

chromosomes that is not always well handled by standard bioinformatic analyses (Aguiar & Istrail 2013). As a result, tetrasomically-inherited regions are often excluded from reference genome assemblies and linkage maps (Davidson *et al.* 2010; Allendorf *et al.* 2015).

Despite a long history of research in salmon genetics and the publication of the first salmonid reference genome (Berthelot *et al.* 2014), the population genetic implications of their ongoing residual tetrasomic inheritance and the importance of the tetrasomically-inherited regions have not been explored in detail. Sequences derived from paralogous loci are often removed from population genetic studies of salmon and other species due to difficulties in genotyping (Dufresne *et al.* 2014; Allendorf *et al.* 2015). In salmon, genomic regions undergoing residual tetrasomic inheritance are rendered invisible to genetic analyses. Studies looking at population structure, genetic diversity, and ecological association in salmon rarely if ever integrate paralogous loci into these analyses (Miller *et al.* 2012; Gagnaire *et al.* 2013; Hohenlohe *et al.* 2013). Here we will first tackle these questions using standard methods and demonstrate what is to be gained through the inclusion of paralogous loci.

We also investigate the genetic associations with life history variation in both paralogous and non-paralogous loci. Previous studies of the genetic basis for salmonid life-history variation have uncovered some promising results. In the Pacific Salmonid *O. mykiss*, Miller *et al.* (2012) uncovered a haplotype associated with rapid growth rate and Pearse *et al.* (2014) found a genomic region with elevated LD associated with smoltification. In Atlantic salmon, Barson *et al.* (2015) found a single locus strongly associated with age at maturity, *VGLL3*. However, none of these studies have extended their analyses to include paralogous loci.

Our analysis focuses on chum salmon (*O. keta*) from that portion of the Salish Sea in Northwest Washington; these populations are characterized by distinct life history variation in run timing including summer, fall, and winter runs (Small *et al.* 2014; Small *et al.* 2015). Run timing is an ecologically important trait in salmon that contributes to reproductive isolation. Neutral population structure between populations in the Salish Sea is well characterized by existing studies utilizing tens of microsatellites, (Small *et al.* 2014), and hundreds of single nucleotide polymorphisms (SNPs) (Seeb *et al.* 2011b; Small *et al.* 2015). While these studies are sufficient to describe relationships between populations defined *a priori*, they are not ideal for describing either continuous clines of genetic variation, identifying migrants, or discovering genetic associations between genotype and phenotype (Rousset 1997). Individual-based

analyses, coupled with large amounts of data provided by Restriction-site associated DNA (RAD) sequencing (Miller *et al.* 2007) provide an opportunity to address these issues and incorporate information from both paralogous and non-paralogous regions of the genome.

The primary goals of this study are to 1) demonstrate that paralogous loci, identified by linkage mapping, can be characterized in wild populations, 2) describe population structure within the Salish Sea chum salmon and compare the ancestry information of paralogous and non-paralogous loci, 3) conduct a genome scan for signals of genetic association with run timing at both paralogous and non-paralogous loci, and 4) improve the utility of the linkage map as a genetic resource. We explore methods to characterize genetic diversity at duplicated loci, first by identifying paralogous loci by observing allelic segregation patterns within a pedigree, and second, by assigning genotypes by the presence/absence of each allele. This is one of the first studies to integrate and explicitly compare salmon population genetic inferences from paralogs to inferences from the rest of the genome.

2.3 METHODS

2.3.1 *Salish Sea chum salmon*

Spawning fish were sampled at ten locations across the Salish Sea, on the western coast of North America on the border of Canada and the USA (**Figure 2.1**). Locations and seasons were designed to capture the existing diversity in genetic structure and life history of chum salmon in this region. The USA National Oceanic and Atmospheric Administration (NOAA) Fisheries recognizes two evolutionarily significant units (ESUs) of chum salmon in the Salish Sea; both are represented within our collections. Most of our collections return to spawn in the fall, as do they the majority of chum salmon in this region, but we also include some of the earliest (summer) and latest (winter) spawning populations in the Americans. The Hood Canal summer-run ESU, represented by the Hamma Hamma River collection, is listed as threatened by NOAA (Johnson *et al.* 1997). Collections were made in four different years (1993, 2003, 2010, and 2011), but each site was sampled only during a single year (**Table 2.1**). For genetic analyses, a fin clip from each fish was preserved in ethanol.

Run timing of each collection was taken from Johnson *et al.* (1997), and classified into summer, fall, or winter. Generally, eggs are deposited in November through December. Embryos

develop and hatch after ~4 months and migrate to sea, with survival and growth rates vary dependent on favorable estuarine and marine conditions (Quinn 2005).

2.3.2 *Sequence analysis*

RAD sequencing was conducted on a total of 200 fish; DNA was extracted from tissues using DNeasy-96 kits from Qiagen (Venlo, Netherlands). Samples were prepared for RAD sequencing with the *SbfI* restriction enzyme as per Baird *et al.* (2008) and Etter *et al.* (2011a). DNA from each sample was uniquely barcoded (6bp), pooled into libraries of 10-30 samples, and sequenced to 101 bp on an Illumina HiSeq2000 at the Genomics Core Facility at the University of Oregon. Hoodsport adults (n = 10) were sequenced using a paired-end protocol to facilitate annotation; all other collections were sequenced using a single-end protocol. Data from 240 gynogenetic haploid offspring were sequenced in a similar manner and used in construction of the consensus linkage map. Sequences from these three mapping families were originally reported in (Waples *et al.* 2015b), but only the largest family was used to construct the map reported there.

Genetic variation was quantified with a reference-based approach using the *Stacks* software pipeline (Catchen *et al.* 2013). A reference FASTA formatted file was constructed by retaining a subset of catalog loci identified in Waples *et al.* (2015b). We conducted an all-by-all self-alignment of catalog loci using *Bowtie2* (Langmead & Salzberg 2012) and formed clusters of loci by grouping those with less than four mismatches. A single locus from each cluster was retained to form the reference; loci present on the linkage map of Waples *et al.* (2015b) were preferentially retained in all cases. Reads from each individual were de-multiplexed and quality-filtered with *process-radtags* and then aligned to the reference with *BWA-mem* (v0.7.5a-r405) (Li 2013). Alignments containing indels or with poor mapping quality (< 20) were removed. Prior to alignment, sequences inferred to be PCR clones were identified and removed from the paired-end reads from Hoodsport adults with the *clone_filter* program that removes identical read pairs. *Stacks* components *pstacks*, *cstacks*, *sstacks*, and *populations* were used to identify and score genetic variants for each individual. A set of 20 individuals, two from each collection, were used to ascertain genetic variation, and all samples were scored at these loci. Paralogous loci were identified based on their segregation pattern within the gynogenetic mapping families and inclusion on the linkage map (see below).

The initial set of genotypes and individuals was assessed and filtered for completeness and quality. This procedure differed slightly for paralogous and non-paralogous loci. For both, loci and individuals with more than 25% missing data were removed. SNP loci with very low allele frequencies are common (Marth *et al.* 2004) but are difficult to distinguish from sequencing errors; loci with a minor allele frequency (MAF) of < 5% in each collection were excluded. Disomic loci were tested for deviations from Hardy-Weinberg equilibrium (HWE). The goal of these HWE tests is to remove loci with erroneously-aligned sequences, rather than to identify small deviations from HWE due to biological processes such as population admixture or selection. HWE was tested within each collection using the mid p-value statistic (Graffelman & Moreno 2013) and loci with $p < 0.05$ in more than 5 collections were removed. Paralogous loci were not subject to HWE tests. Finally, to reduce pseudo-replication caused by physical linkage only the single SNP with the largest minor allele frequency was retained within each locus when calculating diversity statistics. In contrast, for principal components analyses (PCA) and some test of genetic association with run timing (see below), allelic haplotypes comprising all SNPs within each locus were utilized instead of singleton SNPs. All filters were applied in *PLINK* (v1.90beta) (Chang *et al.* 2014).

2.3.3 *Linkage map*

We constructed a consensus linkage map from three families of gynogenetic haploid offspring from Hoodsport Hatchery (family sizes 175, 34, 31). The program *LEPmap* (Rastas *et al.* 2013) was used to assign loci to linkage groups and to determine the order and spacing of loci within each linkage group. This linkage map adds to the map presented in Waples *et al.* (2015b) with the addition of two new families and by the identification metacentric linkage groups and prediction of the centromere location within these metacentric linkage groups. Centromere locations were predicted from the distribution of observed recombination events observed along each linkage group as per Limborg *et al.* (2015). Paralogous loci were identified, resolved, and included in the linkage map using their segregation pattern within the gynogenetic offspring as in Waples *et al.* (2015b). Briefly, for all loci that showed variation in at least one offspring, the observed allelic segregation pattern was fit to the segregation patterns predicted assuming different possible parental genotypes, while accounting for genotyping error. The parental genotype most likely to produce the observed segregation pattern was assigned to the parent.

This parental genotype was used to identify segregating loci suitable for inclusion on the linkage map.

While individuals from each collection were used in the ascertainment of genetic variation, loci could only be included on the linkage map if they segregated in at least one of the Hoodsport Hatchery families. Consequently, the loci on the linkage map have experienced a different ascertainment procedure than the rest of the loci. We examined the effect of this ascertainment by comparing allele frequency histograms and summary statistics for all loci and only those loci present on the linkage map.

Synteny to Chinook salmon (*O. tshawytscha*) and other salmonids was assessed by alignment to the linkage map of McKinney *et al.* (2015). Sequence alignments were conducted with *Bowtie2*; alignments spanning the full template length and with a mapping quality > 30 were inferred to show orthologous relationships between the species. To simplify the interpretation of orthologous relationships, paralogous loci were removed prior to display.

2.3.4 *Population structure and genetic diversity*

Basic population genetic statics were calculated for paralogous and non-paralogous loci. Allele frequencies, heterozygosity, and F_{ST} (Weir & Cockerham 1984) were calculated using *PLINK* for each non-paralogous locus within each collection. F-statistics were not calculated for loci identified as paralogs because we could not fully resolve allele dosage or allele frequencies. A neighbor-joining tree was constructed from non-paralogous loci with *POPTREE2* (Takezaki *et al.* 2014), using F_{ST} between collection locations as the distance metric. Loci were resampled with replacement in 1000 bootstrap replicates to assess support for tree topology (Felsenstein 1985).

Effective population size (N_e) was estimated for each collection using the linkage disequilibrium (LD) method implemented in the *LDNe* software package (Waples & Do 2010). The LD method estimates the average correlation (r^2) of alleles at pairs of loci. The mean pairwise r^2 value across independently-assorting loci provides an estimate of contemporary N_e (Hill 1981). Physically-linked loci can downwardly bias N_e estimates, by confounding LD due to physical linkage and LD due to drift. To avoid this potential bias, only loci placed on the linkage map were included and r^2 measurements between loci co-located on a chromosome were excluded as per Larson *et al.* (2014). Full genotypes are required for this analysis, so only non-

paralogous loci were used. Confidence intervals around the N_e estimates were calculated with the bootstrap procedure implemented in *LDNe*.

Population structure was assessed with two multivariate methods: principal component analysis (PCA) and sparse non-negative matrix factorization (sNMF). Multivariate methods are ideal for analysis of population structure including paralogs because they make fewer assumptions about how genetic variation is partitioned among individuals, and therefore are less tied to specific models of population genetics than alternative methods such as *Structure* (Pritchard et al. 2000). Both multivariate methods (PCA and sNMF) summarize genetic relationships between individuals along distinct axes, often interpreted as representing the contribution of distinct gene pools (Patterson *et al.* 2006).

PCA was applied to genotype matrices that scored the presence/absence of each allelic haplotype in each individual. Three distinct sets of loci were constructed to compare the patterns of population structure evidenced in paralogous loci and non-paralogous loci: a) All non-paralogous loci, b) all paralogous loci, and c) a subset of the non-paralogous loci randomly down-sampled to match the number of paralogous loci. As there are many more non-paralogous than paralogous loci, this down-sampled set allowed a comparison that isolated the type of locus from the number of loci. All PCAs were conducted with *EIGENSOFT* (v6.0.1) (Patterson *et al.* 2006). Similarity between projections of the three PCAs was compared with a pairwise Procrustes analysis. When supplied with two PC projections, this method attempts to find an optimal superimposition, achieved by translation, rotation, reflection and scaling. After this transformation, the remaining difference in shape is a measure of the Procrustes distance between PC projections (Peres-Neto & Jackson 2001).

Within the PC projections, one objective is to isolate the variation associated with allele frequency differences between populations (population structure) from individual-to-individual variation present within a population. A formal test for this type of population structure was developed by Patterson *et al.* (2006) based on the expectation that the eigenvalues of a PCA of genotype matrix from a panmictic population is well-represented by the Tracy-Widom (TW) distribution (Tracy & Widom 1994). Eigenvalues that are outliers to the TW distribution are interpreted to represent allele frequency differences between populations. This leads to the expectation that *K* distinct populations are distinguished by *J*-1 significant PC axes, with the remaining PCs reflecting within-population variation in allele frequencies and sampling noise

(Patterson *et al.* 2006). The largest eigenvalues were sequentially tested for deviation from TW expectation ($p <= .05$) until an axis was found non-significant, at which point all subsequent axes were considered non-significant. This analysis was conducted with the *twstats* program within the *EIGENSOFT* software.

The sparse non-negative matrix factorization (sNMF) method of Frichot *et al.* (2013) was applied to summarize the ancestry coefficients of each individual. These ancestry coefficients serve to control for 'nuisance' population structure in the run-timing association analysis (see below). This method is conceptually similar to PCA but with additional constraints that aid the biological interpretation of the result (see discussion in: Engelhardt & Stephens 2010; Frichot *et al.* 2013). In sNMF, each individual's genetic ancestry is modeled as a linear combination of factors, each reflecting the contribution of a gene pool with specific allele frequencies. The goal is to succinctly summarize the complex set of genetic relationships between individuals represented by the genotype data. This method selects the number of ancestral gene pools (K) through a cross-entropy criterion that minimizes error when a small subset of the genotypes (5%) are hidden and re-predicted by the inferred ancestry coefficients. sNMF was applied to two sets of genotype data: 1) non-paralogous loci with co-dominant genotypes, and 2) all loci (paralogous and non-paralogous) scored as the presence/absence of each allele (same as PCA set b).

2.3.5 *Genome scan*

Genome scans were used to investigate patterns of differentiation across the chum salmon genome and to investigate genetic associations with run timing. Genome scans assess the relative support for the influence of neutral and non-neutral processes in driving patterns of genetic differentiation. The Hoodsport Hatchery collection was excluded from all genome scan analyses because three of the eight individuals within that collection served as the ascertainment individuals for the linkage map. This exclusion did not appreciably change the results (data not shown).

Genetic correlation with run timing was investigated using a latent factor mixed model (LFMM) (Frichot *et al.* 2013) on two sets of loci: non-paralogs and all loci. In this model, individual genotypes are modeled as the response variable in a factor regression that assumes population structure due to shared ancestry, as estimated by sNMF (see above), to be the main driver of variation in allele frequencies. The contributions of two additional evolutionary forces

on allele frequencies are estimated for each locus: 1) association with specified ecological variables and 2) locus-specific effects that are independent of the environment. The weight placed on these two other factors can be interpreted as the strength of support for ecological-driven and ecologically-independent selection, respectively. Run timing, the ecological variable in question, was coded as follows: summer = 1, fall = 2, winter = 3. In each scenario, three replicate runs of the LFMM model were conducted. Reported *z*-scores were converted to *p*-values assuming a normal distribution and combined by taking the median (Fisher's method) and false-discovery rate (FDR) correction was applied to the combined *p*-values. The LFMM analysis was conducted in the R package *LEA* (Frichot & Francois 2015).

In addition to the environmental correlation-based model implemented in LFMM, we calculated five differentiation-based statistics, four for each non-paralogus locus and one for all loci including paralogs. At non-paralogs we calculated F_{ST} (W&C) (Weir & Cockerham 1984), d_{xy} (Nei & Li 1979), *BayeScan* F_{ST}, and *BayeScan* q-value (Foll & Gaggiotti 2008). At all loci, including paralogs, we calculated the informativeness for assignment allele ($I_{n)}$ (Rosenberg et al. 2003), based on the presence absence of each allele within each individual. I_n is a measure of potential information for assignment to population of origin, relative to a hypothetical average population. F_{ST} (W&C) is a relative measure of genetic differentiation between collection locations, while d_{xy} is an absolute measure of genetic differentiation between collections that is independent of the degree of diversity present within each collection. *BayeScan* F_{ST} is an average of the population-specific F_{ST} values estimated by *BayeScan*. The *BayeScan* q-value is the FDR at which the null model of neutral drift is rejected, and is a statistic used to identify loci with allele frequency patterns incongruent with neutral population structure. *BayeScan* (v2.1) was used to calculate the F_{ST} (*BayeScan*) *and q*-value of each locus, Employed *BayeScan* settings were as follows: iterations: 10,000, thinning interval: 10, pilot runs: 20, pilot length: 5K, burn in iterations: 50K. F_{ST} (W&C) was calculated with *PLINK*, as noted above, and d_{xy} was calculated as the average Jaccard distance between allelic haplotypes in separate collections.

Rolling-averages of the above statistics were calculated at each centimorgan (cM) of the linkage map using a sliding-window analysis. At each focal cM, a weighted-mean value was calculated across all loci within 5cM on either side, with weights for each locus within the window inversely proportional to the exponential distance in cM to the focal point (Loh *et al.* 2013). Sliding windows with only a single locus were excluded. At each focal cM, bootstrapped

upper 99% intervals for each statistic were calculated by permuting loci into random positions 1000 times. Because each paralogous locus was split into multiple alleles prior to the genome scan, no rolling-averages were taken for statistics derived from paralogous loci.

2.4 RESULTS

2.4.1 *Sequencing and genotyping*

A total of 174 individuals were retained after exclusion of 26 individuals due to insufficient genotype rates. Individuals per collection varied from a low of n=8 (Squakum Creek and Hoodsport Hatchery) to a high of 32 (Sherwood River Fall). Individuals generally had more than one million aligned sequences available for genotyping (**Table 2.1**). Hoodsport Hatchery had markedly lower sequencing depth, about half of most other collections; this was mainly due to the paired-end sequencing protocol employed on these individual enabling the removal of PCR duplicates. Lower sequencing depth in the Hoodsport Hatchery collection resulted in a reduced genotyping rate (0.85), but this rate was not markedly lower than other populations which ranged between 0.86 - 0.98. After removing loci due to missing data, MAF, and rejection of HWE, 12,399 non-paralogous loci were scored in the retained individuals.

2.4.2 *Linkage map*

We placed 7795 loci onto 37 linkage groups on the consensus linkage map; total map length was 3027 cM (**Figure 2.2**). These 37 linkage groups corresponded 1:1 with those reported in chapter one (Waples *et al.* 2015b) and likely have exact correspondence with the 37 chromosomes in the most common chum salmon karyotype (Phillips & Rab 2001). Included on the linkage map were 1215 confounded catalog entries, representing 1504 paralogous loci in distinct genomic locations. Of the 12,399 non-paralogous loci, 6251 (50%) were present on the linkage map. The 1504 paralogous loci on the linkage map were broken into 4384 separate alleles which were scored for presence/absence in the wild collections. In total, 93% (7259/7795) of the loci on the linkage map were scored in the field collections.

Centromeric regions were placed onto 13 linkage groups inferred to be metacentric (green shading, **Figure 2.2**), consistent with the chum salmon karyotype (Phillips & Rab 2001). The remaining 24 chromosomes are likely a combination of telocentric and acrocentric

chromosomes, but the centromeric location was not predicted. Plotting the recombination fraction from each end of each linkage group revealed some possible errors in map order, but they were restricted to small regions near the end of three linkage groups (linkage groups: 3, 7, and 21, supplemental figure S2.1). Comparison to the Chinook salmon linkage map revealed orthologous relationships for all LGs. For most linkage groups, there was a 1:1 correspondence to Chinook salmon, but at least 10 fission and fusion events were evident (Supplemental Figure S2.2).

2.4.3 *Genetic diversity*

Genetic diversity within each collection was characterized by observed heterozygosity and contemporary N_e. Observed heterozygosity at non-paralogous loci was similar across all collections, ranging from 0.30 to 0.34 (**Table 2.2**). Estimates of N_e were more variable: the finite estimates ranged from a low of 161 in Nisqually River Kalama Creek, to a high of 5,959 in Lilliwaup Creek (**Table 2.2**). Two collections had infinite point estimates of N_e: Squakum Creek and Hoodsport Hatchery. However, the sample sizes of these two populations were very small, both below ten. Hamma Hamma River, the ESA-listed collection, did not fall outside the range of genetic diversity exhibited by the other collections.

2.4.4 *Population structure*

Broad patterns of population structure aligned with geographic proximity, with a few notable exceptions. Population genetic structure between collection locations was assessed with pairwise F_{ST} (W&C) and a neighbor-joining tree from non-paralogous loci (**Figure 2.3**, a + b). Mean global F_{ST} across all non-paralogous loci was 0.027. The subset of loci on the linkage map showed slightly lower levels of genetic differentiation, the Mean global F_{ST} of mapped loci was 0.025. The phylogenetic tree groups Sherwood River Summer and Fall collections with Skookum Creek. The tree also groups Hoodsport Hatchery and Lilliwaup Creek, with the northern collections forming the last major branch. Hamma Hamma River and Nisqually Kalama Creek collections were the most genetically distinct. Notably these are also collections with some of the most extreme run timing. Within the Sherwood River collections, geography, rather than run-timing, was the best predictor of shared genetic ancestry.

Two multivariate methods of assessing population structure were employed that allowed the incorporation of paralogous loci and do not rely on prior population assignment of individuals: PCA and sNMF. Both methods showed that a similar pattern of population structure was present across paralogous and non-paralogous loci.

PCA analyses were run on three different sets of loci: **(a)** only paralogs, **(b)** all loci, and **(c)** all loci with non-paralogs down-sampled to match the number of paralogs. The overall pattern of population structure is highly similar across all three sets of loci (**Figure 2.4**). Pairwise Procrustes similarity of the PCA projections was high, ranging between 0.70 - 0.80, and highly significant as assessed by permutation ($p<<0.01$). The first 12 axes from locus set **b** (all loci) were found to be significant outliers to the Tracy-Widom distribution (Supplemental Figure S2.4); together these axes account for 11.77% of the total genetic variance, with the first four axes accounting for more than 1% each. The 12 significant axes can be interpreted to be 13 distinct gene pools represented among the collections (Patterson *et al.* 2006). Fewer significant axes were found significant when using the paralogs (6) or the subsampled non-paralogs (6).

The sNMF ancestry analysis of all loci and only non-paralogous loci both showed the greatest support for three ancestral factors (**Figure 2.5**), as assess by cross-validation. In both analyses, one factor was nearly exclusive to Hamma Hamma River and the other two factors broadly align with the northern and southern boundaries of the geographical range (Supplemental Figures S2.5 and S2.6).

The Hamma Hamma River collection was well distinguished in our analyses of population structure: pairwise F_{ST} was above 0.06 to all other populations, compared to an average pairwise F_{ST} of 0.027. Hamma Hamma River individuals were separated from all others along one of the primary axes in all PCAs (**Figure 2.4**), with subsequent PC axes representing regional population structure between other collections (Supplemental Figure S2.3). As noted above, the sNMF ancestry analysis consistently assigned a nearly-unique ancestry factor to individuals from the Hamma Hamma collection (Supplemental Figures S2.5 and S2.6).

2.4.5 *Genome scan*

We focus on results from loci on the linkage map for the genome scan. The allele frequency distribution of non-paralogous loci present on the linkage map was flatter than the set of all non-paralogous loci (**Figure 2.6**). This is likely due to the different ascertainment schemes

for these two sets of loci. Loci on the linkage map were ascertained in three individuals from Hoodsport Hatchery while the remaining loci were ascertained in a set of 20 individuals with 2 from each collection.

Population differentiation, as assessed by locus-specific F_{ST} (W&C) at non-paralogous loci, varied across all linkage groups (**Figure 2.7**, Supplemental Table S2.2). Mean *Bayescan* F_{ST} was 0.030. Mean d_{xy} was 0.0056. Across mapped loci, Weir F_{ST} had a positive rank-correlation with *Bayescan* F_{ST} (rho = 0.76) and d_{xy} had a weak positive rank-correlation with Weir F_{ST} (rho = .06) and *BayeScan* F_{ST} (rho = .01). However, this relationship appeared to reverse in the upper tail of each distribution and loci with the highest d_{xy} had low values for both two differentiation measures and loci with high values in d_{xy} were only average differentiation measures (Supplemental Figures S2.7 – S2.11). A total of 219 loci (1.7% of 12,399) had *BayeScan* q-values < 0.05. These loci were located on 25 of the 37 linkage groups.

Kernel-smoothed statistics were measured at each cM position for non-paralogous loci. Local increases in genetic differentiation, as measured by the 99th percentile of kernel-smoothed F_{ST}, were present on 20 out of the 37 linkage groups (54%) (**Figure 2.7**). Positions in the 99% percentile for *BayeScan* q-value were present on eight linkage groups. Three regions, located on linkage groups 1, 21, and 28, were above the 99th percentile for both differentiation statistics. In the association test with non-paralogous loci, the locus with the, A total of 11 linkage groups contained at least one locus with FDR-corrected p-values < 0.05, and 11 linkage groups contained genomic regions where the kernel-smoothed LFMM –log10(*p*-value) was above the 99th percentile (**Figure 2.7**).

In analyses that included paralogous loci, we measured informativeness for assignment (I_n), and association with run timing based on the presence/absence of each allele at each locus. Mean informativeness for assignment (I_n), was lower for alleles at paralogous loci than non-paralogs: 0.0735 vs 0.0864 (Supplemental Figure S2.12). We found statistical support for a genetic association with run timing for multiple alleles from both paralogous and non-paralogous loci (**Figure 2.8**). There were loci with alleles that had $-\log_{10}$(p-values) > 5 on three linkage groups: 2, 28, and 34, including a single paralogous locus, c*48610,* on linkage group 2. Notably, the linkage groups selected by the two association analysis (with and without paralogs) did not overlap.

2.5 DISCUSSION

The aim of this study was to investigate population structure in chum salmon and specifically how the inclusion of duplicated loci affected population genetic inferences. Duplicated loci are difficult to integrate into traditional genetic analyses; therefore we assessed population structure and environmental association with multivariate methods that are more robust to departures from the assumptions of traditional population genetic models (Engelhardt & Stephens 2010).

2.5.1 *Genetic diversity and population structure*

Duplicated loci showed similar patterns of population structure to the rest of the genome in both the PCA and matrix factorization analyses. These results suggest that, in a broad scale, paralogous loci are subject to the same evolutionary forces as the rest of the genome and serves as validation that our genotyping methods were able to capture patterns of shared genetic ancestry.

Observed population structure was consistent with Small *et al.* (2015) and highlights that both geographical proximity and run timing contribute to reproductive isolation between chum salmon populations. The large genetic difference apparent on the first two PC axes (**Figure 2.4**) of the Hamma Hamma River collection is compatible with the summer-run chum from Hood Canal listed within their own ESU (Johnson *et al.* 1997) that is distinct from the other local chum salmon populations.

The PCA provides a clear division of individuals along geographical axes. Only a single individual appears incongruous; an individual collected in South Puget Sound that is projected along with the North Puget Sound individuals. The discrepancy between collection location and genetics suggests this individual is likely a migrant. This situation demonstrates the utility of an individual-based approach to genetic analysis; with a population-based approach this individual would likely *a priori* be assigned to a population based on collection location, and additional steps would be necessary to identify any incongruence.

Chum salmon stray at relatively high rates compared to other Pacific salmon (Quinn 1997); their population structure has often been described as isolation-by-distance of spawning populations connected by local straying (Petrou *et al.* 2014; Small *et al.* 2015). Choosing the

most suitable number of distinct gene pools (*K*), contributing to the observed genetic variation is a long-standing question in population genetics; this question is not always the most relevant question. There was disagreement in the 'best' value for *K* across different methods and data sets. The TW analysis of the PC eigenvalues for the largest non-paralogous data set suggested that there were up to 13 distinct populations. However, Shriner (2011) demonstrated that the TW statistic employed in the PCA analyses tends to overestimate *K*. There was likely an effect of statistical power acting here as well; both TW analyses on smaller data sets (paralogs and non-paralogs) suggested *K*=6. The sNMF method suggested *K* = 3, a significantly smaller number. This difference likely occurs because within the sNMF analysis, higher values of *K* are penalized in the cross-validation procedure. There are also more practical reasons to select a smaller value for K as well; Frichot & Francois (2015) note the negative impact that high values of *K* have on algorithm speed and statistical power when detecting genetic associations with environmental variables.

The estimated effective population sizes ($\widehat{N_e}$) were all above 100 and generally larger than previous studies (e.g., Small *et al.* 2014) that assess overlapping sites. We note that Small *et al.* (2014) estimate N_e over different time periods and also did not utilize a linkage map to exclude loci co-located on chromosomes, a potential source of downward bias in estimates of N_e (Waples & Do 2008). We found two collections with infinite estimates of N_e, (Hoodsport Hatchery and Squakum Creek). These two collections had sample sizes of less than ten. Small sample sizes hinder accurate estimates of N_e based on the LD method. Sampling only a few individuals creates correlations between alleles in a similar manner to genetic drift, thereby increasing r^2. While a statistical correction for this effect is possible, this will often overcorrect and result in infinite estimates of N_e (Waples 2006). The two collections with samples size below 10, Squakum Creek and Hoodsport Hatchery, had infinite estimates of N_e.

Residual tetrasomic inheritance affects our expectations of genetic diversity. For fully tetraploid inheritance, genetic drift occurs at one-half of the rate of disomic inheritance, leading to a doubling of N_e (Charlesworth 2009). The rate of tetrasomic inheritance is variable across the salmonid genome (Ostberg 2015) suggesting that we should expect neutral genetic diversity to vary as well. We were not able to test for differences between paralogous and non-paralogous loci directly without genotypes that have allele dosage information, but mean I_n was lower at paralogous loci. This is consistent with less genetic drift and larger N_e at paralogous loci, but

there are other differences as well. Paralogous loci can have more than two alleles, and in some cases, one allele is present in nearly every individual, with the effect of reducing I_n.

There was slightly lower genetic differentiation evident in loci on the linkage map, showing the effects of ascertainment bias. Lower differentiation occurred because rare variants from other populations were excluded if they were invariable in the three females from Hoodsport Hatchery used to construct the linkage map. To address this, future population genetic studies that rely on linkage maps could take steps to explicitly account for the ascertainment scheme (e.g. Albrechtsen *et al.* 2010) when estimating genetic differentiation.

2.5.2 *Linkage map*

The consensus linkage map presented here shows a meaningful improvement over the single-family linkage map in chapter one (Waples *et al.* 2015b). A major addition is the inclusion of centromeric regions on 13 metacentric linkage groups. These 13 linkage groups are also the 13 longest in length (cM). In all cases the centromere was placed near the middle of the linkage group, suggesting a relatively even balance of recombination events on either side. There are specific sequence motifs that are associated with centromeric functions (Lamb & Birchler 2003) that we do not expect to find within our filtered RADseq data, instead we have likely observed nearby loci that are linked to the centromere by limited recombination. Within the consensus linkage map, paralogs were concentrated on the distal ends of 13 chromosomes (16 chromosome arms), similar to other salmonid species (Brieuc et al. 2014; Kodama et al. 2014), consistent with the notion that these regions are conserved across Pacific salmonids.

Also the adoption of *LEPmap* in place of *MSTmap* allowed the integration of segregation information from multiple families without exaggerating overall map length (McKinney et al. 2015). However, including loci segregating in three families provided only a modest increase in the number of loci on the consensus linkage map: 7,795 vs 6,162. This was likely due to the small number of offspring in each of the two additional families. Small families provide less accurate estimates of pairwise recombination fractions between loci, hindering linkage map construction.

2.5.3 *Genome scan*

In general, genomic regions selected for association with run timing did not overlap with regions with elevated F_{ST}, but they were adjacent in LG12 and LG13. This pattern of non-overlap suggests that run timing is not driving 'genomic islands of divergence' (Nosil *et al.* 2009), within the chum salmon in the Salish Sea.

We investigated the orthologous relationships for regions selected in the genome scan with elevated genetic differentiation or association with run timing. Chum LG21, with a region near one end of elevated divergence in both F_{ST} (W&C) and *BayeScan* q-value, is orthologous with *O. mykiss* chromosome Omy7 (Phillips et al. 2013). Loci on this chromosome have elevated divergence between migratory and non-migratory *O. mykiss* (Hale *et al.* 2013). The paralogous locus with the best statistical support for association with run timing (c*48610*) was placed on LG2, in a region of the linkage group dominated by paralogous loci. This region of LG2 is orthologous with Chinook chromosome Ots12. Studies in Chinook salmon from the Columbia River and Puget Sound have associated loci on Ots12 with run timing across multiple lineages (Hess & Narum 2011; Brieuc *et al.* 2014).

The genome scan including paralogous loci also highlights the difficulties inherent in working with paralogs. The peak of genetic association with run timing on LG2 ($p < .00001$, cM 125) is driven by multiple alleles from a single paralogous locus (c*48610*). The consensus sequence of this paralogous locus annotates to the *piggyBac* transposon (Li et al. 2013), a mobile genetic element useful in genome editing. Indeed, on our linkage map this locus is present in multiple copies, but they all co-locate near the same position on LG2. This is not consistent with residual tetrasomic inheritance, but together with the sequence annotation, strongly suggests a different type of duplication event.

Notably, the test statistics produced by both *BayeScan* and *LFMM* had a higher mean and more extreme values for loci not on the linkage map: *BayeScan* $F_{ST:}$ 0.0302 on map vs. 0.0310 overall; *LFMM* –log10pval on map: 0.0513, vs.0.0636 overall. There are at least two non-exclusive possible contributions to this pattern. First, loci are only included on the linkage map if they are observed to segregate within a pedigree; this serves as a strong validation step that ensures that we are observing real variation, not an alignment or sequencing error. As a result, loci on the linkage map have passed a more stringent validation process. Second, loci on the linkage map were ascertained in three individuals from the Hoodsport Hatchery collection

affecting the allele frequency spectrum (**Figure 2.6**). This has the potential to downwardly bias measures of population differentiation, which can influence the results of subsequent tests such as BayeScan, often in the form of an increased number of false positive results (Clark et al. 2005).

Local recombination rates can also affect genetic diversity (Hellmann et al. 2003); measures of genetic divergence such as F_{ST}, and outlier tests associated with them, are sensitive to local recombination rates (Cruickshank & Hahn 2014; Burri *et al.* 2015). Recombination rates along chromosomes are difficult to assess, but the local density of loci can provide a proxy for local recombination rate. If we assume that the restriction enzyme cuts randomly across the genome, then we expect the density of the linkage map to be negatively correlated with local recombination rate. However, loci must be variable to be included on the linkage map, and it is difficult to isolate the effect of recombination rate on either locus density or genetic diversity. It is known that distribution of recombination events in salmonids differ markedly between the sexes, with males having a greater fraction of their recombination events in the distal regions of chromosomes (Lien et al. 2011). We could not directly assess sex-specific recombination differences with only female mapping parents. We did not find a negative correlation between d_{xy} and F_{ST}, as would be predicted by region of low recombination driving increased F_{ST} (Cruickshank & Hahn 2014).

Genome scans, even those with significant results, will often fail to locate causative mutations. Genomes are large and techniques like RADseq only sample a fraction of the genome. Genome scans rely on genetic hitching, where limited recombination (genetic linkage) between observed variant and sites under selection, causes allele frequency changes at observed sites, even when these sites are not subject to selection. The expected strength of hitchhiking, and its impact on genome scans, is influenced by evolutionary processes such as recombination, selection, and effective population size (Tiffin & Ross-Ibarra 2014), but hitchhiking is generally stronger with low recombination, strong selection, and low population sizes.

The major variation in salmon run timing is between, rather than within, salmon populations (Quinn 2011). Consequently, there is covariation between neutral genetic structure and run timing, complicating investigations into the genetic basis of this life history trait. Neutral population structure confounded with life-history variation can cause false-positive signals of association among neutral loci and mask the true genetic basis (Altshuler *et al.* 2008), even after

statistical corrections for population structure (e.g. Hale *et al.* 2013). Case-control studies are designed to minimize this type of complication, but they are difficult to mimic in natural populations. In biological systems where population structure is dependent of the phenotype under investigation, sampling design is very important (Lotterhos & Whitlock 2014), as is replication of both phenotypes and genetic background.

Despite these potential difficulties, uncovering the genetic associations with run timing and other variation in salmon life history traits is important and has the potential to bolster conservation efforts by identifying genetic markers with high conservation value (Funk et al. 2012). Knowledge of the genetic basis for life-history variation will become crucial in guiding the management priorities of salmon and other species of conservation concern. It is evident that there is meaningful genetic variation within regions of the genome dominated by paralogs, and these regions should not be excluded from future studies.

2.5.4 *Population genetics with residual tetrasomy*

A number of outstanding issues remain in the genetic characterization of paralogous loci in Pacific salmon. It may be appealing to think that with enough sequence data, and long enough reads, we will be able to fully distinguish sequences derived from paralogous loci. But, as long as there is ongoing tetrasomic inheritance there is little hope of using short sequence data to resolve paralogs. Indeed, only a single homeologous recombination per generation is sufficient to homogenize the allele frequencies across loci undergoing residual tetrasomic inheritance (Allendorf & Danzmann 1997; Meirmans & Van Tienderen 2013). This implies that the assumption of full tetrasomy will not introduce significant bias in population genetic inference (Meirmans & Van Tienderen 2013) and that for the evolutionary process, the presence of residual tetrasomic inheritance is vitally important, while the rate at which it occurs is of secondary concern. Hence, the way forward may not be to strive to resolve and split paralogs still undergoing tetrasomic inheritance, but to deal with them as single locus, albeit one that exists in two locations.

Coupled with issues of how to treat paralogous genetic loci are questions on how to measure genetic diversity and divergence. Many measures of population divergence, including F_{ST}, are sensitive to ploidy, and F_{ST} values are not directly comparable between disomic loci and tetrasomic loci (see discussion by Meirmans & Van Tienderen (2013)). For salmonids, the ideal

measure of genetic divergence would allow a direct comparison of loci with varying inheritance patterns and be robust to variation and uncertainty in the rate of tetrasomic inheritance. This would allow genome-wide assessment of fundamental genetic parameters such as N_e, and it would aid integration into population genetic models built to test evolutionary hypotheses. To address these goals, Meirmans & Van Tienderen (2013) recommend the ρ statistic (Ronfort *et al.* 1998), a ploidy-independent F_{ST}-analogue. This statistic scales to different ploidies and shows little bias if the rate of tetrasomic inheritance is specified incorrectly.

Perhaps the biggest practical complication in genetic analysis of polyploids is the difficulty in assigning allele dosage. In the present study, we sidestepped this issue by only assigning the presence or absence of paralogous allele, but this approach came with many shortcomings. This approach hindered the calculations of genetic diversity and allele frequencies, and it did not make full use of all the information contained within the sequencing data. Many other methods assume the unique signal of each allele (e.g. sequence depth or hybridization florescence, depending on the study) is directly proportional to allele dosage. Models integrating this assumption have been implemented for large SNP chips in the software programs *fitTetra* (Voorrips *et al.* 2011), *SuperMASSA* (Serang *et al.* 2012), and *beadarrayMSV* (Gidskehaug *et al.* 2011). For high-throughput sequence data, polyploid-specific programs are just starting to emerge (e.g., *polyfreqs*, Blischak *et al.* 2015), and SNP callers *FreeBayes* (Garrison & Marth 2012) and *GATK* (McKenna *et al.* 2010) allow a specification of ploidy and will assign allele dosages to polyploids based on a positive correlation between sequencing depth and allele dosage.

Sequencing studies in polyploids require more sequence data in order to achieve the error rates that are currently tolerated in studies of diploids. Higher error rates occur because polyploid genotype calls must distinguish between more possible genotypes and thus require more information. Genotyping approaches that fully integrate genotype uncertainty into population genetic inferences have been successfully applied in studies where sequencing depth is insufficient to fully resolve genotypes, such as with ancient human DNA, e.g. (e.g. Raghavan *et al.* 2015). Based on the available sequence data, and any prior assumptions, all possible genotypes are assigned likelihoods. These genotype likelihoods are used to weight the possible genotypes when calculating population genetic statistics. This is in contrast to the common approach of 'calling' the mostly likely genotype and then proceeding as if it was known without

error. Most often this "-- new method --" is applied to very low-coverage genome resequencing studies, but it may also be very useful for polyploids. The software *ANGSD* (Korneliussen *et al.* 2014) can conduct these analyses under the assumption of diploidy, but is not yet able to accommodate higher ploidies.

2.5.5 *Conclusion*

We successfully demonstrated that paralogous loci can be characterized within wild salmon populations and that they can be used to understand the genetic relationships between populations and individuals. These paralogous loci reveal similar pattern of neutral population structure as the rest of the genome, providing confidence that they were accurately characterized and that paralogous loci are subject to similar evolutionary forces as the rest of the genome. The genome scan including paralogous loci showed significant associations with run timing across the genome, including within some paralog-dense regions, suggesting that studies that fully exclude paralogs may be blind to important signals of adaptive differentiation. Future work on the population genetics of paralogs should focus on resolving the outstanding genotyping difficulties and adapting population-genetics models to accommodate residual tetrasomic inheritance.

2.6 TABLES

Table 2.1. Sample size (n), aligned sequences, and genotyping rate for each collection.

Mean and standard deviation (sd) are given for the number of aligned sequences and genotyping rate for each collection.

				Aligned sequences		Genotyping rate	
Collection	Run timing	Year	n	mean	sd	mean	sd
Hamma Hamma River	Summer	2010	20	1,419,541	1,427,760	0.87	0.08
Lilliwaup Creek	Fall	2011	20	2,760,125	999,141	0.98	0.01
Hoodsport Hatchery*	Fall	2010	8	509,422	148,391	0.85	0.08
Sherwood River	Fall	1994	32	3,235,188	966,091	0.96	0.04
Sherwood River	Summer	1994	31	2,504,974	1,183,089	0.91	0.07
Skookum Creek	Fall	2010	11	1,644,932	637,844	0.95	0.09
Nisqually R. Kalama Creek	Winter	2003	17	2,270,022	1,432,866	0.96	0.03
Stillaguamish River	Fall	2010	13	710,538	269,873	0.91	0.06
Snohomish River	Fall	2010	14	1,135,085	495,888	0.94	0.07
Squakum Creek	Fall	2010	8	999,084	650,927	0.86	0.08

*paired-end sequencing

Table 2.2. Genetic Diversity.

Observed heterozygosity (Obs. Het) and estimated effective population size with the associated confidence intervals (CI) for collections of Salish Sea chum salmon.

Collection	Obs. Het.	$\widehat{N_e}$	Lower CI	Upper CI
Hamma Hamma River	0.30	339.4	316.5	365.8
Lilliwaup Creek	0.34	5959.0	3276.5	32651.6
Hoodsport Hatchery*	0.30	∞	NA	NA
Sherwood River (Fall)	0.33	319.9	310.8	329.5
Sherwood River (Summer)	0.31	145.2	142.9	147.6
Skookum Creek	0.33	1788.3	1121.0	4402.6
Nisqually River Kalama Creek	0.32	161.6	156.7	166.8
Stillaguamish River	0.30	1001.2	766.0	1443.7
Snohomish River	0.32	2122.8	1400.1	4379.3
Squakum Creek*	0.31	∞	NA	NA

*small sample size

2.7 FIGURES

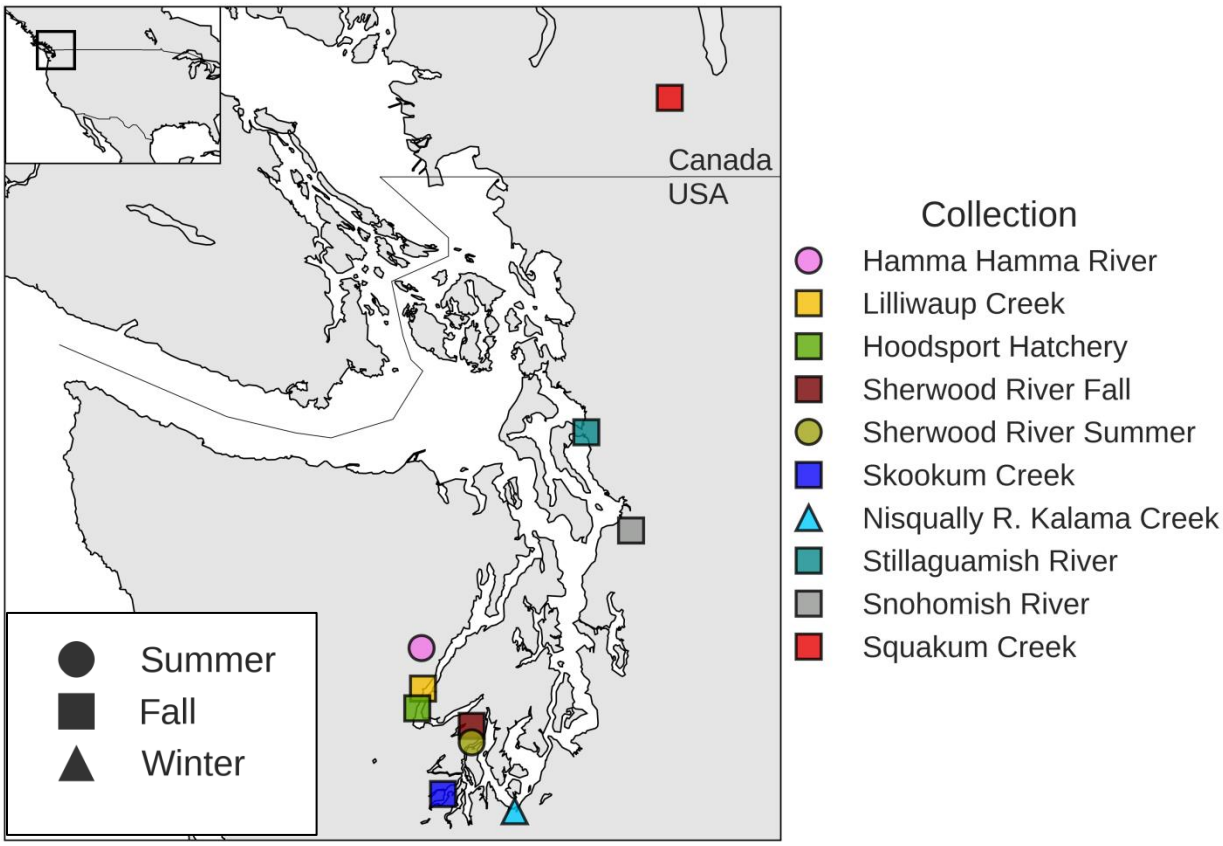

Figure 2.1. Collection locations within the Salish Sea

Locations and run timings of chum salmon collections in the Salish Sea.

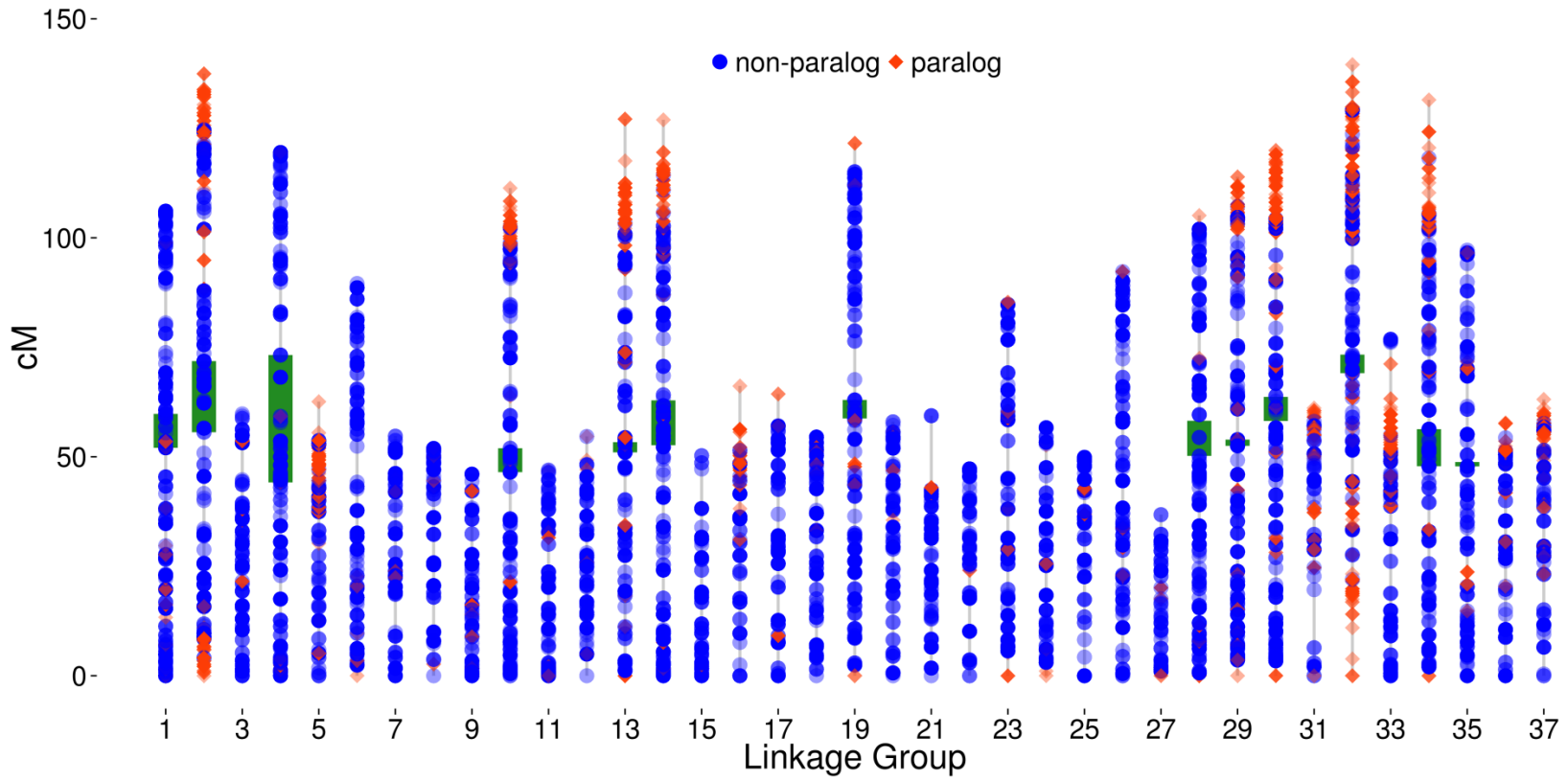

Figure 2.2. Consensus linkage map.

37 linkage groups, likely corresponding to the 37 chromosomes in the chum salmon karyotype. Paralogous loci are shown as red diamonds; non-paralogs are blue circles. Centromeric regions of 13 metacentric linkage groups are shown in green.

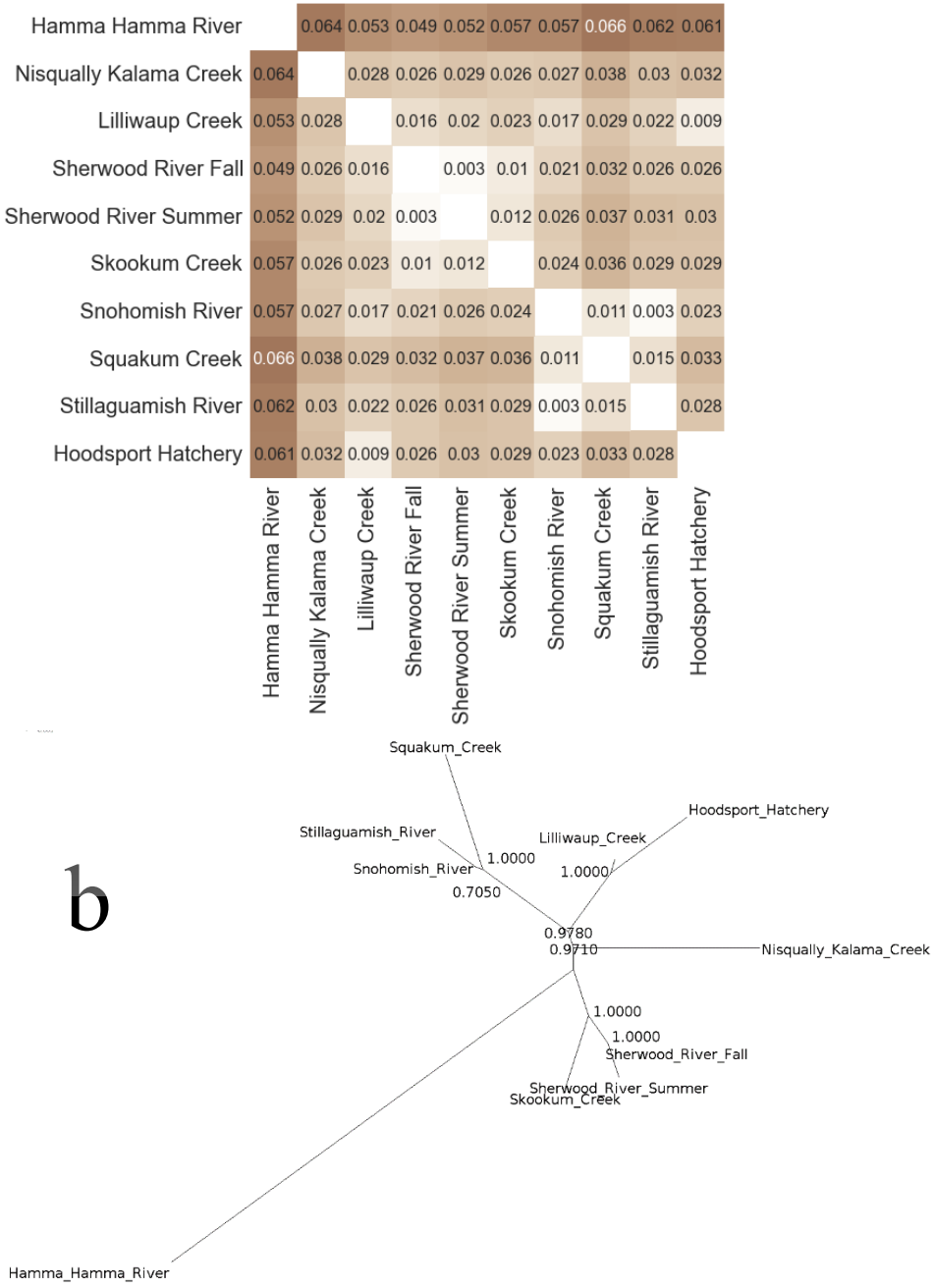

Figure 2.3. F_{ST} matrix and phylogenetic tree.

a) Pairwise F_{ST} matrix between collections. Intersecting cells are labeled with pairwise F_{ST} values. Darker shading shows pairs with more genetic divergence. The highest pairwise F_{ST} is highlighted with white text. b) Neighbor-joining tree constructed from pairwise F_{ST} matrix. Numbers on each node show the fraction of 1000 bootstrap replicates that support the shown topology.

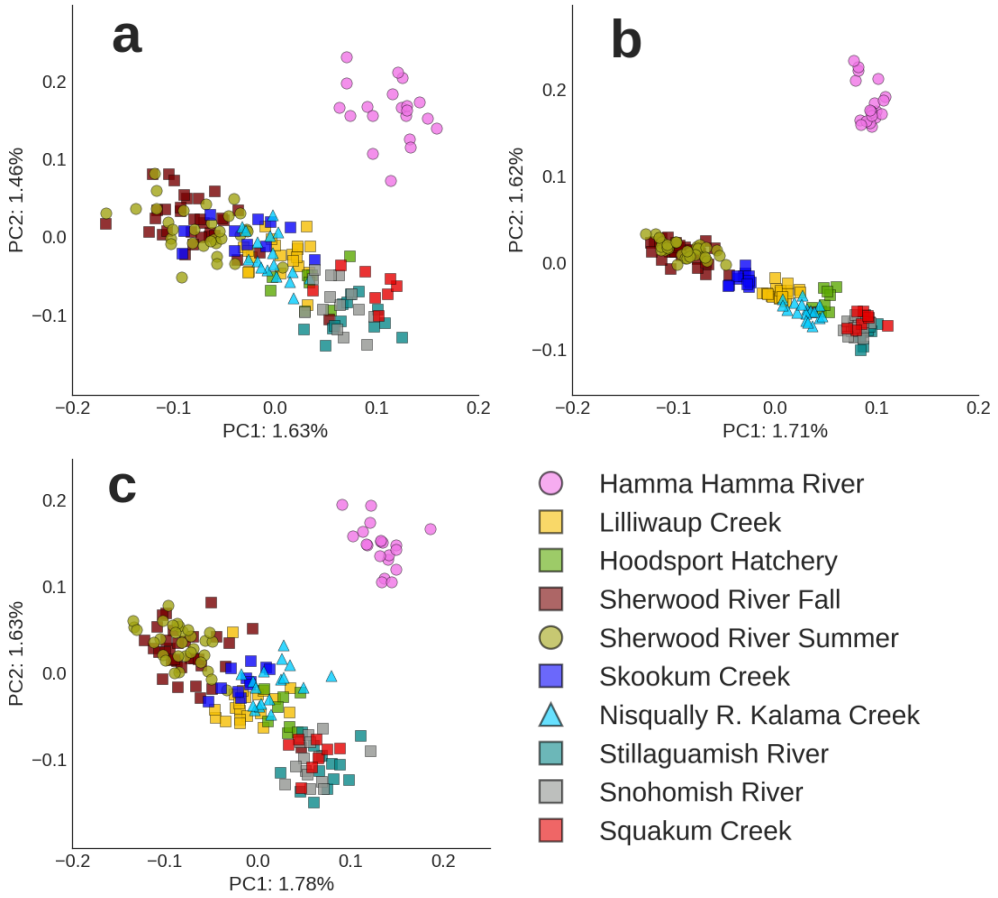

Figure 2.4. Population structure shown by individual-based PCAs

Individual-based PCA from ten populations (various colors) of chum salmon from the Salish Sea. Each dot is an individual; dots are colored according to collection location and run timing.

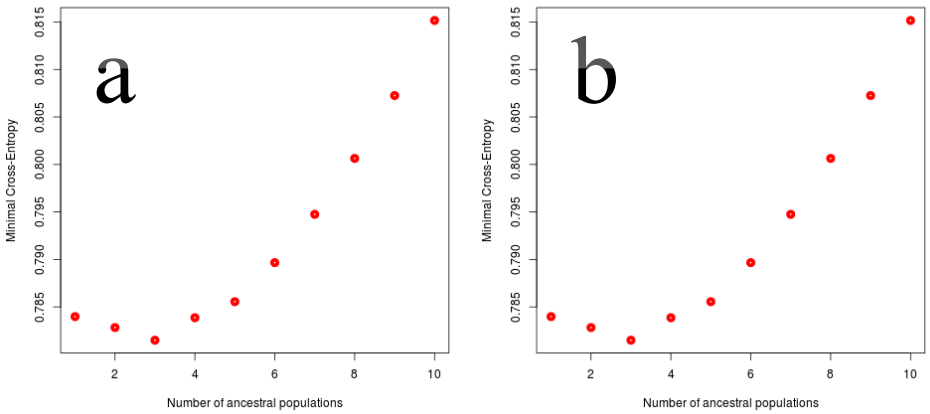

Figure 2.5. Cross-validation entropy at different value of K

Results of the cross-validation procedure of the sNMF to determine the value of K. The value of K with the minimum cross-validation entropy performs the best at predicting a set of holdout genotypes. Cross-validation entropy for a) non-paralogs, codominant genotypes b) all loci, including paralogs, alleles scored for presence/absence.

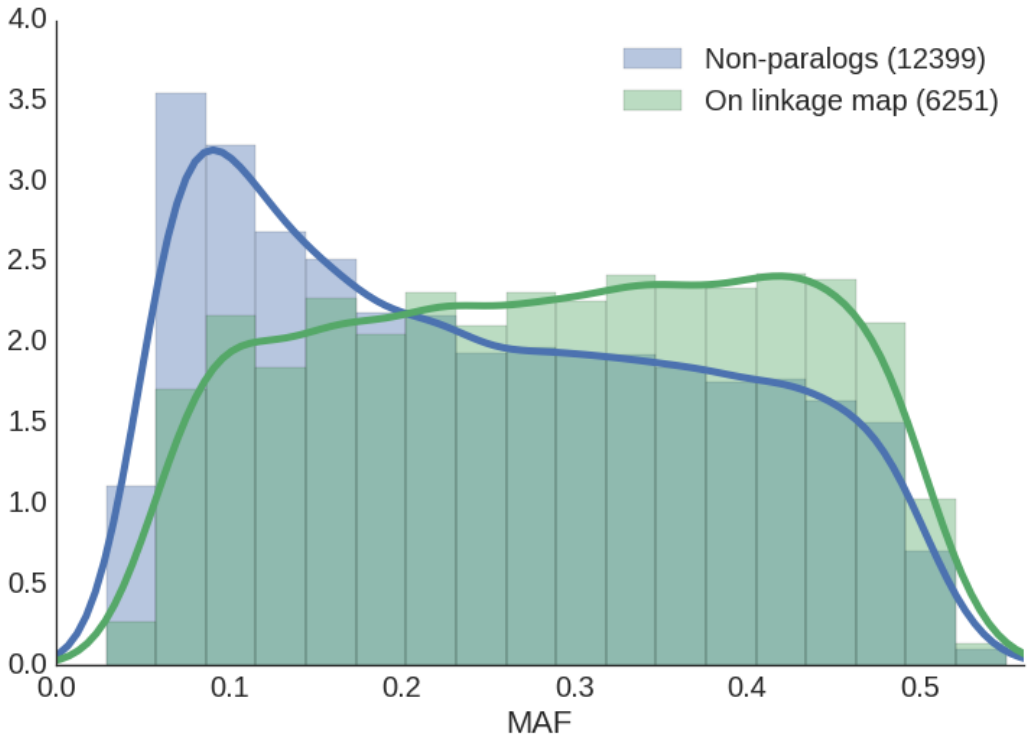

Figure 2.6. Ascertainment bias

Folded minor allele frequency (MAF) for all loci (blue shading) and the subset of loci placed on the linkage map (green shading). Solid lines are smoothed averages with a Gaussian kernel. The y-axis is density-scaled to accommodate the differing number of loci in each set.

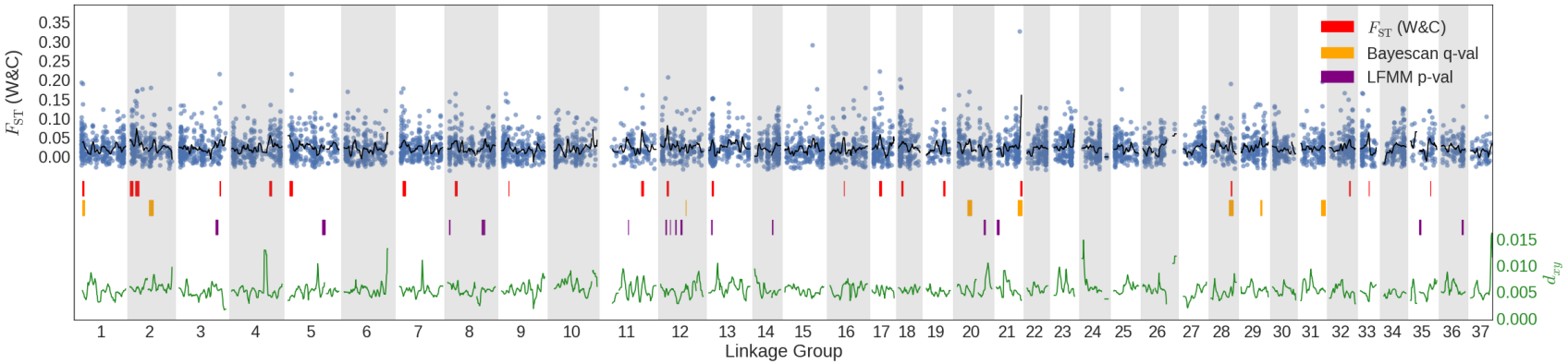

Figure 2.7. Manhattan plot of genetic differentiation.

Genetic differentiation across the chum salmon genome; only non-paralogous loci are shown. Alternating grey/white shading delineates linkage groups. Blue points show the global F_{ST} (W&C) value for each locus. The black line is a kernel-smoothed moving average of F_{ST} within a 5 cM region on either side of the focal point. The colored bars in the middle of the plot show cM positions above the 99th percentile in one of three statistics: red: F_{ST}, blue: *BayeScan* *q*-value, and purple: association with run timing (-log$_{10}$(p-value)). The green line at the bottom of the plot shows the kernel-smoothed d_{xy}, the mean absolute divergence of sequences sampled from different collections.

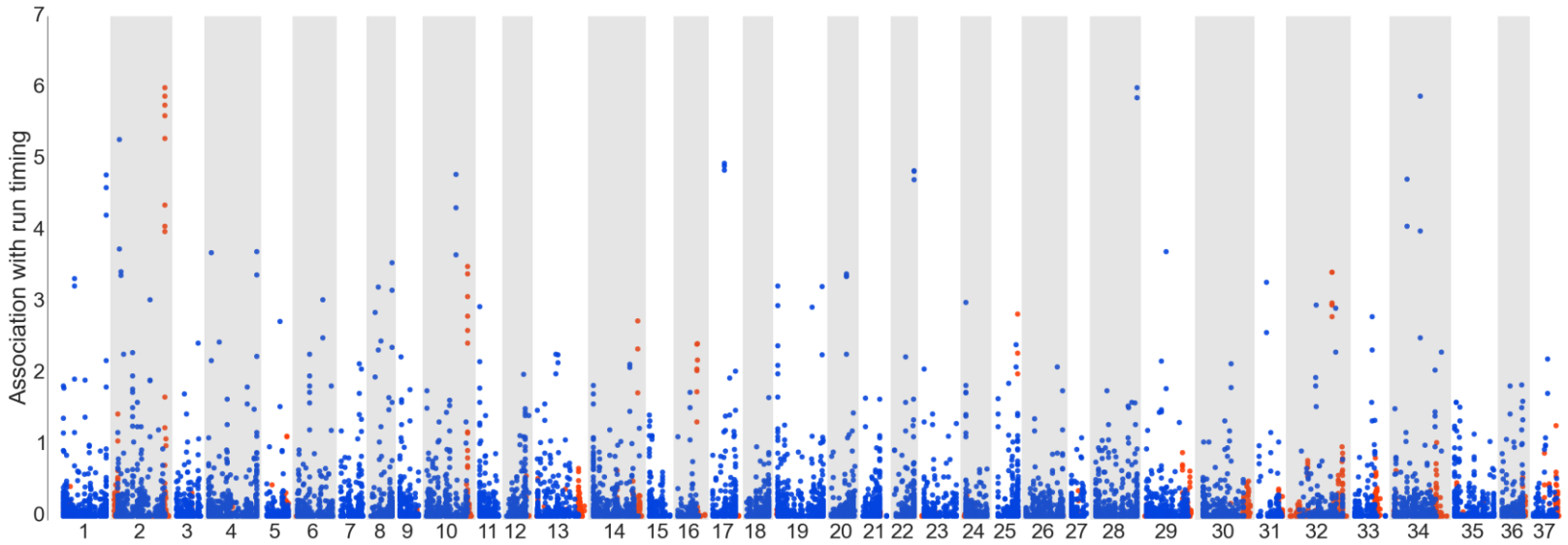

Figure 2.8. Genetic association with run timing at paralogous and non-paralogous loci

Genetic association with run timing as assessed by LFMM. Points are colored according to duplicate status; non-paralogous loci are in blue, paralogous loci are in red. Alternating grey/white shading delineates linkage groups. Loci are separated into distinct alleles and each point represents a separate allele. Y-axis shows the negative log10(p-value), a measure of statistical support for genetic association with run timing, measured per allele.

2.8 ACKNOWLEDGEMENTS

Many thanks to Carita Pascal and Meredith Everett for RADseq library prep. We would also like to thank, Morten Limborg, Paul Hohenlohe and Steven Roberts for constructive comments and discussion. We would also like to thank Ken Warheit, Sewall Young, and Maureen Small from the Washington Dept. of Fish and Wildlife for biological samples and stimulating conversation. Funding contributing to this research was from NOAA Saltonstall-Kennedy Award NA10NMF4270310, Pacific Salmon Commission Southern Boundary Restoration and Enhancement Fund, and the Gordon and Betty Moore Foundation.

2.9 SUPPLEMENTAL FIGURE LEGENDS

Figure S2.1 – Recombination fraction along each linkage group

The pairwise recombination fraction of each locus is shown relative to each end of the linkage group. Green dots show the recombination fraction (y) of that locus relative to rightmost locus. Blue dots show the recombination fraction (y) of that locus relative to the leftmost locus. Faint line shows the expected relationship between cM (x axis) and recombination fraction (y axis) along the 50cM from each end assuming complete recombination interference. Deviations from this line show recombination patterns that are not well predicted by the linkage map and complete interference within each chromosome arm. Only data from the largest family (n=175) is shown.

Figure S2.2 - Chum / Chinook oxford grid

Scatterplot of 756 loci that align between the linkage map of this study and the Chinook salmon linkage map of McKinney *et al.* (2015). Each dot shows a single locus. Dots are positioned according to their respective position on each linkage map. Some Chinook linkage groups are reversed in order to facilitate display.

Figure S2.3 - Subsequent PCA axes

Individual-based PCA from ten populations (colors match figure 2.4) of chum salmon from the Salish Sea. Subsequent axes for the three data sets: (a) paralogs, (b) non-paralogs, and (c) non-paralogs downsampled to match the number of loci in **a**.

Figure S2.4 – Number of significant PC axes

The number of significant PC axes for each of the three PCA analyses. Significant axes were determined by comparison of their eigenvalues to the Tracy-Widom distribution, p-values <0.05 were found significant.

Figure S2.5 - Ancestry coefficients for K=3 (non-paralogous loci)

Ancestry coefficients for each individual as inferred by the sNMF method applied to non-paralogous loci with K=3, the value of K with the lowest cross-validation entropy. Individuals

are shown along the x-axis, ordered as in supplemental file SF2.3. Each individual is modeled as a linear combination of the three ancestry coefficients, here shown in three colors. The sum of the ancestry coefficients for each individual must add up to one.

Figure S2.6 - Ancestry coefficients for K=3 (all loci)

Ancestry coefficients for each individual as inferred by the sNMF method applied to all loci with K=3, the value of K with the lowest cross-validation entropy. Individuals are shown along the x-axis, ordered as in supplemental file SF2.3. Each individual is modeled as a linear combination of the three ancestry coefficients, here shown in three colors. The sum of the ancestry coefficients for each individual must add up to one. Here the input data in scored as the presence absence of each allele.

Figure S2.7 Correlation between d_{xy} and F_{ST} (W&C)

Scatter plot of the FST (W&C) and d_{xy} (measures of genetic differentiation. The Spearman's rank correlation coefficient (ρ) is shown in the upper left.

Figure S2.8 Correlation between d_{xy} and F_{ST} (BayeScan)

Scatter plot of the FST (BayeScan) and d_{xy} (measures of genetic differentiation. The Spearman's rank correlation coefficient (ρ) is shown in the upper left.

Figure S2.9 Correlation between rolling-mean d_{xy} and F_{ST} (W&C)

Scatter plot of the sliding window rolling-mean values of the F_{ST} (W&C) and dxy measures of genetic differentiation at each cM position. The Spearman's rank correlation coefficient (ρ) is shown in the upper left.

Figure S2.10 Correlation between rolling-mean d_{xy} and F_{ST} (BayeScan)

Scatter plot of the sliding window rolling-mean values of the F_{ST} (BayeScan) and dxy measures of genetic differentiation at each cM position. The Spearman's rank correlation coefficient (ρ) is shown in the upper left.

Figure S2.11 Correlation between rolling-mean F_{ST} (W&C) and F_{ST} (BayeScan)

Scatter plot of the values of the F_{ST} (W&C) and F_{ST} (BayeScan) measures of genetic differentiation. The Spearman's rank correlation coefficient (ρ) is shown in the upper left.

Figure S2.12 Informativeness (I_n) across the genome

Informativeness for assignment (I_n) (Rosenberg et al. 2003) for each allele of each locus on the linkage map. Paralogs are shown in red, non-paralogous loci are shown in blue.

2.10 SUPPLEMENTAL FILES

File S2.1 "SF2_1 consensus linkage map.txt"

Tab-delimited text file, giving the position of all loci on the consensus linkage map

column	description
contig	Reference contig, matches chapter one "P1_consensus.fasta"
resolved_locus	Name of the contig, without the 'c'
stacks_CatID	Stacks catalog ID
stacks_SNP	Stacks SNP position
LEPname	LEPmap name
paper1_LG	Linkage group placement, LGs named as in chapter one
cM	Centimorgan position

File S2.2 "SF2_2 genome_stats.txt"

Tab-delimited text file, population genetic statistics for each locus on the linkage map

column	description
contig	Reference contig, matches chapter one "P1_consensus.fasta"
locus	Name of the contig, without the 'c'
stacks_CatID	Stacks catalog ID
stacks_SNP	Stacks SNP ID
LG	Linkage group placement
cM	Centimorgan position
pi_within	Mean genetic diversity within collection sites (haplotype)
dxy	Mean genetic diversity between collection sites (haplotype)
BSname	BayeScan locus ID
qval_bayescan	BayeScan q-value
Fst_bayescan	Mean Fst (Bayescan)
Fst_weir	Mean Fst (W+C)
cpvals	LFMM combined p-value
zscore	LFMM z-score
fdr_pval_LFMM	LFMM FDR-corrected pvalue

File S2.3 "SF2_3 genotypes.zip"

Zip file, Genotypes for all loci, all file are text.

- contig_to_stacks_id.txt: This file connects Stacks ID to contigs from chapter one "P1_consensus.fasta"
- complete.haplotypes.tsv: Stacks output file
- complete.dom.map: PLINK *.map format, one line per dominant-scored allele
- complete.dom.tsv: Presence/absence/missing = [1,0,9] for each dominant-scored allele in each individual. Individuals are ordered as in "non_paralogs.map"
- non_paralogs.map: PLINK *.map format – one line per non-paralogous loci
- non_paralogs.ped: PLINK *.ped format – genotypes for non-paralogous loci. Locus names given in non_paralogs.map. Use contig_to_stacks_id.txt to join to "P1_consensus.fasta"

BIBLIOGRAPHY

Aguiar D, Istrail S (2013) Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* **29**, i352-360. doi:10.1093/bioinformatics/btt213

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* **27**, 2534-2547. doi:10.1093/molbev/msq148

Allendorf FW (1978) Protein polymorphism and the rate of loss of duplicate gene expression. *Nature* **272**, 76-78.

Allendorf FW, Bassham S, Cresko WA, *et al.* (2015) Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J Hered* **106**, 217-227. doi:10.1093/jhered/esv015

Allendorf FW, Danzmann RG (1997) Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* **145**, 1083-1092.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403-410. doi:10.1016/S0022-2836(05)80360-2

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* **322**, 881-888. doi:10.1126/science.1156409

Arnold B, Bomblies K, Wakeley J (2012) Extending coalescent theory to autotetraploids. *Genetics* **192**, 195-204. doi:10.1534/genetics.112.140582

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one* **3**, e3376. doi:10.1371/journal.pone.0003376

Barson NJ, Aykanat T, Hindar K, *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*. doi:10.1038/nature16062

Berthelot C, Brunet F, Chalopin D, *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* **5**, 3657. doi:10.1038/ncomms4657

Blischak PD, Kubatko LS, Wolfe AD (2015) Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12493

Brenna-Hansen S, Li J, Kent MP, *et al.* (2012) Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis. *BMC genomics* **13**, 432. doi:10.1186/1471-2164-13-432

Brieuc MS, Waters CD, Seeb JE, Naish KA (2014) A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3 (Bethesda)* **4**, 447-460. doi:10.1534/g3.113.009316

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890. doi:10.1093/bioinformatics/btg112

Brown JD, O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annu Rev Genomics Hum Genet* **11**, 291-316. doi:10.1146/annurev-genom-082509-141554

Burri R, Nater A, Kawakami T, *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research* **25**, 1656-1665. doi:10.1101/gr.196485.115

Carbon S, Ireland A, Mungall CJ, *et al.* (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288-289. doi:10.1093/bioinformatics/btn615

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**, 3124-3140. doi:10.1111/mec.12354

Chang CC, Chow CC, Tellier LC, *et al.* (2014) Second-generation PLINK: rising to the challenge of larger and richer datasets. *arXiv preprint arXiv:1410.4803*.

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195-205. doi:10.1038/nrg2526

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* **15**, 1496-1502. doi:10.1101/gr.4107905

Comai L (2005) The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836-846. doi:10.1038/nrg1711

Crow KD, Wagner GP, Investigators ST-NY (2006) Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol* **23**, 887-892. doi:10.1093/molbev/msj083

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* **23**, 3133-3157.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**, 499-510. doi:10.1038/nrg3012

Davidson WS, Koop BF, Jones SJ, *et al.* (2010) Sequencing the genome of the Atlantic salmon *(Salmo salar*). *Genome Biol* **11**, 403. doi:10.1186/gb-2010-11-9-403

de Boer JG, Yazawa R, Davidson WS, Koop BF (2007) Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC genomics* **8**, 422. doi:10.1186/1471-2164-8-422

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**, e314. doi:10.1371/journal.pbio.0030314

del Carmen Calderón M, Rey M-D, Cabrera A, Prieto P (2014) The subtelomeric region is important for chromosome recognition and pairing during meiosis. *Scientific reports* **4**.

Dittman A, Quinn T (1996) Homing in Pacific salmon: mechanisms and ecological basis. *J Exp Biol* **199**, 83-91.

Dufresne F, Stift M, Vergilino R, Mable BK (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol Ecol* **23**, 40-69. doi:10.1111/mec.12581

Elfstrom CM, Smith CT, Seeb LW (2007) Thirty-eight single nucleotide polymorphism markers for high-throughput genotyping of chum salmon. *Molecular Ecology Notes* **7**, 1211-1215. doi:10.1111/j.1471-8286.2007.01835.x

Ellegren H, Smeds L, Burri R, *et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**, 756-760. doi:10.1038/nature11584

Engelhardt BE, Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**, e1001117. doi:10.1371/journal.pgen.1001117

Etter P, Bassham S, Hohenlohe P, Johnson E, Cresko W (2011a) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds. Orgogozo V, Rockman MV), pp. 157-178. Humana Press.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011b) Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PloS one* **6**, e18561. doi:10.1371/journal.pone.0018561

Everett MV, Seeb JE (2014) Detection and mapping of QTL for temperature tolerance and body size in Chinook salmon (*Oncorhynchus tshawytscha*) using genotyping by sequencing. *Evol Appl* **7**, 480-492. doi:10.1111/eva.12147

Fedoroff NV (2012) Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758-767. doi:10.1126/science.338.6108.758

Felcher KJ, Coombs JJ, Massa AN, *et al.* (2012) Integration of two diploid potato linkage maps with the potato genome sequence. *PloS one* **7**, e36347. doi:10.1371/journal.pone.0036347

Felsenstein J (1985) Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **39**, 783-791. doi:Doi 10.2307/2408678

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993. doi:10.1534/genetics.108.092221

Frichot E, Francois O (2015) LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6**, 925-929. doi:10.1111/2041-210x.12382

Frichot E, Schoville SD, Bouchard G, Francois O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol* **30**, 1687-1699. doi:10.1093/molbev/mst063

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L (2013) The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution* **67**, 2483-2497. doi:10.1111/evo.12075

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*, 1207.3907. doi:arXiv:1207.3907

Geller F, Ziegler A (2003) Detection rates for genotyping errors in SNPs using the trio design. *Human heredity* **54**, 111-117.

Gidskehaug L, Kent M, Hayes BJ, Lien S (2011) Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics* **27**, 303-310. doi:10.1093/bioinformatics/btq673

Graffelman J, Moreno V (2013) The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat Appl Genet Mol Biol* **12**, 433-448. doi:10.1515/sagmb-2012-0039

Gramfort A, Luessi M, Larson E, *et al.* (2013) MEG and EEG data analysis with MNE-Python. *Front Neurosci* **7**, 267. doi:10.3389/fnins.2013.00267

Hale MC, Thrower FP, Berntson EA, Miller MR, Nichols KM (2013) Evaluating adaptive divergence between migratory and nonmigratory ecotypes of a salmonid fish, *Oncorhynchus mykiss*. *G3 (Bethesda)* **3**, 1273-1285. doi:10.1534/g3.113.006817

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**, 1527-1535. doi:10.1086/375657

Hess JE, Narum SR (2011) Single-nucleotide polymorphism (SNP) loci correlated with run timing in adult Chinook salmon from the Columbia River basin. *Transactions of the American Fisheries Society* **140**, 855-864. doi:10.1080/00028487.2011.588138

Hill WG (1981) Estimation of Effective Population-Size from Data on Linkage Disequilibrium. *Genetical Research* **38**, 209-216.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**, e1000862. doi:10.1371/journal.pgen.1000862

Hohenlohe PA, Day MD, Amish SJ, *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol Ecol* **22**, 3002-3013. doi:10.1111/mec.12239

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877. doi:10.1101/gr.9.9.868

Jaillon O, Aury JM, Brunet F, *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957. doi:10.1038/nature03025

Johnson O, Grant W, Kope R, *et al.* (1997) Status review of chum salmon from Washington. *Oregon, and California NOAA Technical Memorandum NMFS-NWFSC-32, Seattle, WA.*

Keller I, Wagner CE, Greuter L, *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol Ecol* **22**, 2848-2863. doi:10.1111/mec.12083

Kodama M, Brieuc MS, Devlin RH, Hard JJ, Naish KA (2014) Comparative mapping between Coho Salmon (*Oncorhynchus kisutch*) and three other salmonids suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. *G3 (Bethesda)* **4**, 1717-1730. doi:10.1534/g3.114.012294

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics* **15**, 356. doi:10.1186/s12859-014-0356-4

Kosambi DD (1943) The estimation of map distances from recombination values. *Annals of Eugenics* **12**, 172-175.

Lamb JC, Birchler JA (2003) The role of DNA sequence in centromere formation. *Genome Biol* **4**, 214.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359. doi:10.1038/nmeth.1923

Larson WA, Seeb LW, Everett MV, *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon *(Oncorhynchus tshawytscha*). *Evol Appl* **7**, 355-369. doi:10.1111/eva.12128

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li X, Burnight ER, Cooney AL, *et al.* (2013) piggyBac transposase tools for genome engineering. *Proc Natl Acad Sci U S A* **110**, E2279-2287. doi:10.1073/pnas.1305987110

Lien S, Gidskehaug L, Moen T, *et al.* (2011) A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC genomics* **12**, 615. doi:10.1186/1471-2164-12-615

Limborg MT, Waples RK, Allendorf FW, Seeb JE (2015) Linkage Mapping Reveals Strong Chiasma Interference in Sockeye Salmon: Implications for Interpreting Genomic Data. *G3 (Bethesda)* **5**, 2463-2473. doi:10.1534/g3.115.020222

Limborg MT, Waples RK, Seeb JE, Seeb LW (2014) Temporally isolated lineages of pink salmon reveal unique signatures of selection on distinct pools of standing genetic variation. *J Hered* **105**, 741-751. doi:10.1093/jhered/esu063

Loh PR, Lipson M, Patterson N, *et al.* (2013) Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254. doi:10.1534/genetics.112.147330

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol Ecol* **23**, 2178-2192. doi:10.1111/mec.12725

Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B* **281**, 20132881. doi:10.1098/rspb.2013.2881

Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009. doi:10.1093/database/bar009

Makino T, McLysaght A (2012) Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. *Genome Research* **22**, 2427-2435. doi:10.1101/gr.131953.111

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351-372.

Mastretta-Yanes A, Arrigo N, Alvarez N, *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* **15**, 28-41. doi:10.1111/1755-0998.12291

May B, Wright JE, Stoneking M (1979) Joint Segregation of Biochemical Loci in Salmonidae - Results from Experiments with Salvelinus and Review of the Literature on Other Species. *Journal of the Fisheries Research Board of Canada* **36**, 1114-1128.

McKenna A, Hanna M, Banks E, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303. doi:10.1101/gr.107524.110

McKinney GJ, Seeb LW, Larson WA, *et al.* (2015) An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology Resources*. doi:10.1111/1755-0998.12479

Meirmans PG, Van Tienderen PH (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity (Edinb)* **110**, 131-137. doi:10.1038/hdy.2012.80

Miller MR, Brunelli JP, Wheeler PA, *et al.* (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Mol Ecol* **21**, 237-249. doi:10.1111/j.1365-294X.2011.05305.x

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**, 240-248. doi:10.1101/gr.5681207

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**, 5269-5273.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Mol Ecol* **18**, 375-402. doi:10.1111/j.1365-294X.2008.03946.x

Ohno S (1970) Enormous Diversity in Genome Sizes of Fish as a Reflection of Natures Extensive Experiments with Gene Duplication. *Transactions of the American Fisheries Society* **99**, 120-&. doi:Doi 10.1577/1548-8659(1970)99<120:Tedigs>2.0.Co;2

Oliphant TE (2007) Python for scientific computing. *Computing in Science & Engineering* **9**, 10-20. doi:Doi 10.1109/Mcse.2007.58

Ostberg CO (2015) *Genomic Consequences of Hybridization between Rainbow and Cutthroat Trout* Doctoral dissertation, University of Washington.

Otto SP (2007) The evolutionary consequences of polyploidy. *Cell* **131**, 452-462. doi:10.1016/j.cell.2007.10.022

Palti Y, Gao G, Miller MR, *et al.* (2014) A resource of single-nucleotide polymorphisms for rainbow trout generated by restriction-site associated DNA sequencing of doubled haploids. *Molecular Ecology Resources* **14**, 588-596. doi:10.1111/1755-0998.12204

Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. *New Phytol* **186**, 5-17. doi:10.1111/j.1469-8137.2009.03142.x

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* **2**, e190. doi:10.1371/journal.pgen.0020190

Pearse DE, Miller MR, Abadia-Cardoso A, Garza JC (2014) Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proc. R. Soc. B* **281**, 20140012. doi:10.1098/rspb.2014.0012

Peres-Neto PR, Jackson DA (2001) How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* **129**, 169-178.

Petrou EL, Hauser L, Waples RS, *et al.* (2013) Secondary contact and changes in coastal habitat availability influence the nonequilibrium population structure of a salmonid (*Oncorhynchus keta*). *Mol Ecol* **22**, 5848-5860. doi:10.1111/mec.12543

Petrou EL, Seeb JE, Hauser L, *et al.* (2014) Fine-scale sampling reveals distinct isolation by distance patterns in chum salmon (*Oncorhynchus keta*) populations occupying a glacially dynamic environment. *Conservation Genetics* **15**, 229-243. doi:10.1007/s10592-013-0534-3

Phillips R, Rab P (2001) Chromosome evolution in the Salmonidae (Pisces): an update. *Biol Rev Camb Philos Soc* **76**, 1-25.

Phillips RB, DeKoning J, Morasch MR, Park LK, Devlin RH (2007) Identification of the sex chromosome pair in chum salmon *(Oncorhynchus keta*) and pink salmon (*Oncorhynchus gorbuscha*). *Cytogenet Genome Res* **116**, 298-304. doi:10.1159/000100414

Phillips RB, Park LK, Naish KA (2013) Assignment of Chinook salmon (*Oncorhynchus tshawytscha*) linkage groups to specific chromosomes reveals a karyotype with multiple rearrangements of the chromosome arms of rainbow trout (*Oncorhynchus mykiss*). *G3 (Bethesda)* **3**, 2289-2295. doi:10.1534/g3.113.008078

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Quinn TP (2005) *The behavior and ecology of Pacific salmon and trout* UBC Press.

Raghavan M, Steinrucken M, Harris K, *et al.* (2015) Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884. doi:10.1126/science.aab3884

Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P (2013) Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* **29**, 3128-3134. doi:10.1093/bioinformatics/btt563

Rieseberg LH, Willis JH (2007) Plant speciation. *Science* **317**, 910-914. doi:10.1126/science.1137729

Riethman H (2009) Human subtelomeric copy number variations. *Cytogenet Genome Res* **123**, 244.

Robertson WRB (1916) Chromosome studies I. Taxonomic relationships shown in the chromosomes of Tettigidae and Acrididae : V-shaped chromosomes and their significance in Acrididae, Locustidae, and Gryllidae: Chromosomes and variation. *Journal of Morphology* **27**, 179-331. doi:DOI 10.1002/jmor.1050270202

Ronfort J, Jenczewski E, Bataillon T, Rousset F (1998) Analysis of population structure in autotetraploid species. *Genetics* **150**, 921-930.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**, 1402-1422. doi:10.1086/380416

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-1228.

Sasaki M, Hitotsumachi S, Makino S, Terao T (1968) A Comparative Study of the Chromosomes in the Chum Salmon, the Kokanee Salmon and their Hybrids. *Caryologia* **21**, 389-394. doi:10.1080/00087114.1968.10796319

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341-345. doi:10.1038/nature04562

Seeb JE, Pascal CE, Grau ED, *et al.* (2011a) Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources* **11**, 335-348. doi:10.1111/j.1755-0998.2010.02936.x

Seeb JE, Seeb LW (1986) Gene mapping of isozyme loci in chum salmon. *J Hered* **77**, 399-402.

Seeb LW, Templin WD, Sato S, *et al.* (2011b) Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Molecular Ecology Resources* **11 Suppl 1**, 195-217. doi:10.1111/j.1755-0998.2010.02966.x

Selmecki AM, Maruvka YE, Richmond PA, *et al.* (2015) Polyploidy can drive rapid adaptation in yeast. *Nature* **519**, 349-352. doi:10.1038/nature14187

Serang O, Mollinari M, Garcia AA (2012) Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PloS one* **7**, e30906. doi:10.1371/journal.pone.0030906

Shriner D (2011) Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity (Edinb)* **107**, 413-420. doi:10.1038/hdy.2011.26

Small MP, Johnson TH, Bowman C, Martinez E (2014) Genetic assessment of a summer chum salmon metapopulation in recovery. *Evol Appl* **7**, 266-285. doi:10.1111/eva.12118

Small MP, Rogers Olive SD, Seeb LW, *et al.* (2015) Chum Salmon Genetic Diversity in the Northeastern Pacific Ocean Assessed with Single Nucleotide Polymorphisms (SNPs): Applications to Fishery Management. *North American Journal of Fisheries Management* **35**, 974-987.

Smit A, Hubley R, Green P (2010) *RepeatMasker Open-3.0.* http://www.repeatmasker.org

Smith CT, Park L, Vandoornik D, Seeb W, Seeb E (2006) Characterization of 19 single nucleotide polymorphism markers for coho salmon. *Molecular Ecology Notes* **6**, 715-720. doi:10.1111/j.1471-8286.2006.01320.x

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends Ecol Evol* **14**, 348-352. doi:http://dx.doi.org/10.1016/S0169-5347(99)01638-9

Spruell P, Pilgrim KL, Greene BA, *et al.* (1999) Inheritance of nuclear DNA markers in gynogenetic haploid pink salmon. *J Hered* **90**, 289-296. doi:10.1093/jhered/90.2.289

Strimmer BKaK (2013) fdrtool: Estimation and control of (local) false discovery rates.

Takezaki N, Nei M, Tamura K (2014) POPTREEW: web version of POPTREE for constructing population trees from allele frequency data and computing some other quantities. *Mol Biol Evol* **31**, 1622-1624. doi:10.1093/molbev/msu093

Thorgaard GH, Allendorf FW, Knudsen KL (1983) Gene-Centromere Mapping in Rainbow Trout: High Interference over Long Map Distances. *Genetics* **103**, 771-783.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends Ecol Evol* **29**, 673-680. doi:10.1016/j.tree.2014.10.004

Tomekpe K, Lumaret R (1991) Association between Quantitative Traits and Allozyme Heterozygosity in a Tetrasomic Species - Dactylis-Glomerata. *Evolution* **45**, 359-370. doi:Doi 10.2307/2409670

Tracy CA, Widom H (1994) Level Spacing Distributions and the Bessel Kernel. *Communications in mathematical physics* **161**, 289-309. doi:Doi 10.1007/Bf02099779

Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC bioinformatics* **12**, 172. doi:10.1186/1471-2105-12-172

Wang X, Wang H, Wang J, *et al.* (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* **43**, 1035-1039. doi:10.1038/ng.919

Waples RK, Seeb LW, Seeb JE (2015a) Data from: Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). Dryad Data Repository. doi:doi:10.5061/dryad.5b64r.2

Waples RK, Seeb LW, Seeb JE (2015b) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*. doi:10.1111/1755-0998.12394

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics* **7**, 167-184. doi:10.1007/s10592-005-9100-y

Waples RS, Do C (2008) ldne: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* **8**, 753-756. doi:10.1111/j.1755-0998.2007.02061.x

Waples RS, Do C (2010) Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evol Appl* **3**, 244-262. doi:10.1111/j.1752-4571.2009.00104.x

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population-Structure. *Evolution* **38**, 1358-1370. doi:Doi 10.2307/2408641

Wright JE, Johnson K, Hollister A, 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**, 239-260.

Wu R, Ma CX, Casella G (2004) A mixed polyploid model for linkage analysis in outcrossing tetraploids using a pseudo-test backcross design. *J Comput Biol* **11**, 562-580. doi:10.1089/1066527041887393

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet* **4**, e1000212. doi:10.1371/journal.pgen.1000212

Young WP, Wheeler PA, Coryell VH, Keim P, Thorgaard GH (1998) A detailed linkage map of rainbow trout produced using doubled haploids. *Genetics* **148**, 839-850.

Zhao L, Yuanda L, Caiping C, *et al.* (2012) Toward allotetraploid cotton genome assembly: integration of a high-density molecular genetic linkage map with DNA sequence information. *BMC genomics* **13**, 539. doi:10.1186/1471-2164-13-539

APPENDIX

2.11 APPENDIX A

The likelihood of a parental genotype (G_p) was assessed as the probability of the observed offspring genotype counts, given the parental genotype, and accounting for genotyping error:

$$L(G_p) = P(h_i,\cdots,h_k)|G_p$$
$$= \binom{n}{h_i, h_{i+1}, h_{i+2}, \ldots, h_k} \prod_{i=1}^{k} P(h_i)|\, G_p \quad (1)$$

Where $[h_i,\cdots,h_k]$ is the vector counting the number of offspring with genotype h_i, summing to n, with k distinguishable offspring genotypes. The last term on the right is calculated as:

$$P(h_i)|\, G_p = \left(r + \frac{\varepsilon}{k} - (p * \varepsilon)\right)^{h_i} \quad (2)$$

Where r is the error-free probability of a parent of genotype G_p producing an offspring with genotype i (Supplemental File S1) and ε is the genotyping error rate in the offspring (the rate at which true genotypes are replaced by random genotypes). r is modified by two error terms: the $\frac{\varepsilon}{k}$ term represents genotyping errors that result in the genotype i, and the $r * \varepsilon$ term, represents the assignment of random genotype when the genotype specified by h_i is true.

2.12 APPENDIX B

A naïve estimate of the genotyping error rate $(\varepsilon_{\text{naive}})$ is the fraction of offspring genotypes that are impossible given the considered parental genotype:

$$\varepsilon_{\text{naive}}|\, G_p = \left(\sum_{i}^{n} h_i \in h_{\text{error}}|\, G_p\right)/n \quad (3)$$

where G_p is the parental genotype, there are n offspring, h_i is the genotype of offspring i, and h_{error} is the set of offspring genotypes impossible without error.

This naïve estimate is too low, however, because actual errors do not exclusively result in nonsense genotypes. We can correct bias if we assume that all possible combinations of alleles are equally likely to be the result of an error and then scaling our naïve estimate by the fraction of errors that we can observe. The total number of distinguishable combinations (h_{all}) of k alleles, given our inability to observe allele dosage, is:

$$h_{all} = \binom{k + k - 1}{k} \quad (4)$$

This allows up to k alleles to appear in each genotype and is not restricted to the bi-allelic case to allow for confounded loci. After correction, our estimate of the genotyping error rate ε is:

$$\varepsilon = \varepsilon_{\text{naive}} * \frac{h_{all}}{length(h_{\text{error}})} \quad (5)$$