MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12394

# Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*)

R. K. WAPLES, L. W. SEEB and J. E. SEEB
*School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Box 355020, Seattle, Washington 98195, USA*

## Abstract

**Gene sequence similarity due to shared ancestry after a duplication event, that is 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 rediploidized 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 subtelomeric 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 rediploidization 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 rediploidization process.**

*Keywords*: chum salmon, haploid, homeolog, linkage mapping, tetrasomic inheritance, whole-genome duplication

*Received 17 November 2014; revision received 13 February 2015; accepted 20 February 2015*

## 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.* 2011; Wang *et al.* 2011). Paralogs are especially difficult to identify and resolve in nonmodel 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), that is 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 nonmodel species is to use a distance-based metric to separate paralogs based on the underlying sequences (e.g. Seeb *et al.* 2011; 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.

Correspondence: J. E. Seeb, Fax: 206 685 7471;
E-mail: jseeb@uw.edu

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 rediploidization (Makino & McLysaght 2012). The rediploidization 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 rediploidized and have genes with both disomic and tetrasomic patterns, but tetrasomic inheritance has only been observed in males and is understood to not occur in females (Allendorf 1978; Wright *et al.* 1983; Allendorf & Danzmann 1997). It is not known if the rate of rediploidization is constant over time, but the 'extreme stability' of the retained salmonid chromosomes subsequent to the WGD (Berthelot *et al.* 2014) suggests 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 (i) develop a method to resolve confounded paralogous loci, (ii) build a chum salmon linkage map that includes the resolved loci, and (iii) 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 homologues 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.

## Materials and methods

### *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 (http://www.pac.dfo-mpo.gc.ca/science/aquaculture/sirp/incubwin-eng.html). 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, the 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.

### *Sequencing*

Haploid and diploid tissues were sequenced on 8 lanes of a HiSeq 2000 (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 analysed in a similar fashion (e.g. 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 (6 bp) 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.

### *Sequence analysis*

We quality-filtered the raw sequence reads and analysed 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 demultiplexed, stripped of the barcode, trimmed to 84 bp 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 (catalogue 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 catalogue using sstacks.

Within the Stacks analysis pipeline, *catalogue entries* are groups of aligned sequences. Individual catalogue entries nominally represent a unique locus, but they can also contain sequences originating from multiple loci. In these latter cases, we term the catalogue entry confounded as it no longer represents a distinct genetic locus. Genetic variation observed within confounded catalogue entries may be an artifact of aligning sequences from multiple loci, that is paralogous sequence variants. The Stacks method for constructing catalogue 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 catalogue entry as intra- or interlocus 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 'catalogue 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 catalogue 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 catalogue 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 catalogue 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. Catalogue entries with a no-call rate >0.25, with >4 alleles, or without variation, were also excluded.

### *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 catalogue entry based on the underlying parental genotype(s). Each parental genotype is expected to produce offspring with genotypes in a particular ratio (Fig. 1). We calculate the likelihood of the observed offspring genotype data given each of the parental genotypes using a multinomial sampling distribution (see Appendix I), a method similar to calculating genotype likelihoods from sequencing data using allele depths (Hohenlohe *et al.* 2010). Each catalogue 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 the 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 Fig. 1; the full list is available in Appendix S1 (Supporting information). 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.

### *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 (Appendix II). A separate error rate was estimated

| Maternal chromosomes | Maternal genotype | Haploid offspring genotypes | Offspring genotype ratio | # Loci with observable segregation |
|---|---|---|---|---|
| A◆B | AB | A, B | 1:1 | 1 |
| A◆B A– –A | AB/AA | AA, AB | 1:1 | 1 |
| A◆B A◆C | AB/AC | AA, AB, AC, BC | 1:1:1:1 | 2 |
| A◆B C◆D | AB/CD | AC, AD, BC, BD | 1:1:1:1 | 2 |
| A– –A B◆C | AA/BC | AB, AC | 1:1 | 1 |
| A– –B A– –B | AB/AB | AA, AB, BB | 1:2:1 | 0 |
| A– –A B– –B | AA/BB | AB | 1 | 0 |
| A– –A A– –A A– –A | AA/xx | Ax | 1 | 0 |

◆ = Observable segregation

**Fig. 1** Model relating parental and offspring genotypes. Confounded catalogue 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 (Appendix I). Figure 1 is incomplete, see supplemental Table S1 for a list of considered parental genotypes.

within each catalogue 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.

### *Resolving confounded loci*

Some confounded catalogue entries can be resolved into one or more constituent loci, while others remain unusable. We convert each catalogue entry into zero, one or two distinct loci by observing the segregation of locus-specific alleles (Fig. 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 ≤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 ml-psv.

### *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 colocated in alternate phase. We then identified pairs of LGs that contained many loci that would be colocated in alternate phase. Next, we switched the phase of loci in nonoverlapping 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 nonconfounded catalogue entries and *Map2* including an additional 1015 loci resulting from confounded catalogue entries (Appendix S3, Supporting information). LGs containing only a single locus were excluded from the final maps. Kendall's rank correlation ($\tau$) was calculated for each corresponding pair of LGs to compare the consistency of locus orders between *Map1* and *Map2*.

### *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 catalogue entries by alignment to the allelic sequences associated with all catalogue entries. Only alignments with full identity were accepted. Paired-end sequences assigned to catalogue 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 (Appendix S4, Supporting information). Many confounded catalogue 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 catalogue rather than each locus. The set of loci comprising a catalogue entry received at most a single annotation shared across them.

### *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 Catalogue 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. Catalogue 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 whether 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.

## Results

### *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 (doi: 10.5061/dryad.5b64r). 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 (Appendix S2, Supporting information).

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

### *Assignment of maternal genotype based on segregation patterns*

The segregation patterns of alleles within the 16 786 polymorphic catalogue entries were used to assign maternal

genotypes. We identified 6603 catalogue entries consisting of loci with observable segregation patterns (Table 1). Of these, 5321 consisted of a single segregating locus; the remaining 1282 consisted of confounded loci. In 154 of the confounded catalogue entries, both loci had observable segregation patterns, allowing the recovery of two loci. In all other cases (1128), only one of the confounded loci had observable segregation. Of the 6757 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; 6236 loci were found eligible for linkage map construction.

Many polymorphic catalogue entries could not be resolved into distinct, segregating loci. The female parent was determined to be homozygous for 8252 (49%) catalogue entries; the observed variation was probably due to genotyping errors in one of the offspring. A total of 669 catalogue 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 catalogue 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 catalogue entries (Appendix S3).

### *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 6159 loci. A total of 3075 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 nonconfounded catalogue entries; *Map2* also included paralogous loci resolved from confounded catalogue entries. Within *Map1,* 5221 loci were placed at 3401 unique locations on 37 LGs. In *Map2,* 37 LGs contained a total of 6162 loci in 4006 unique locations (Table 2), with an additional 74 unplaced loci not considered further. The 37 LGs probably 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.

**Table 1** Catalogue entries are classified into four categories. Nonconfounded catalogue corresponds to individual loci; confounded catalogue entries can be resolved into one or two constituent loci. Some catalogue 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 | | 6603 |
| Nonconfounded | AB | 5321 |
| Confounded, recover one | AB/AA, AA/BC | 1128 |
| Confounded, recover two | AB/AC, AB/CD | 154 |
| Unresolved | All others | 10 183 |

### *Comparison of Map1 and Map2*

*Map2* is longer than *Map1* (3728 vs 3246 cM) and has a slightly increased marker density (1.65 vs. 1.60 loci per cM) (Table 2). LG assignment was consistent for all loci between the two maps (Fig. 2). There were some changes in the order and spacing of loci within LGs, shown as crossed blue lines in Fig. 2. The few discrepancies in order that do occur are small in scale; mean Kendall's $\tau$ rank correlation coefficient of the locus orders was 0.971 (range 0.889–0.999) (Appendix S3). MSTmap and other linkage mapping algorithms do not produce confidence/credible intervals for mapping results, complicating the interpretation of between-map comparisons.

### *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%) (Fig. 2; cf., Brieuc *et al.* 2014). Within these 14 LGs, paralogs are concentrated near chromosome ends (Fig. 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] (Fig. 3). These paired regions probably 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 probably 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 (Figs 2 and 3) which is consistent with the

**Table 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* | 5221 | 3401 | 37 | 141.1 | 3246 | 51–145 | 1.60 |
| *Map2* | 6162 | 4006 | 37 | 166.5 | 3728 | 54–174 | 1.65 |

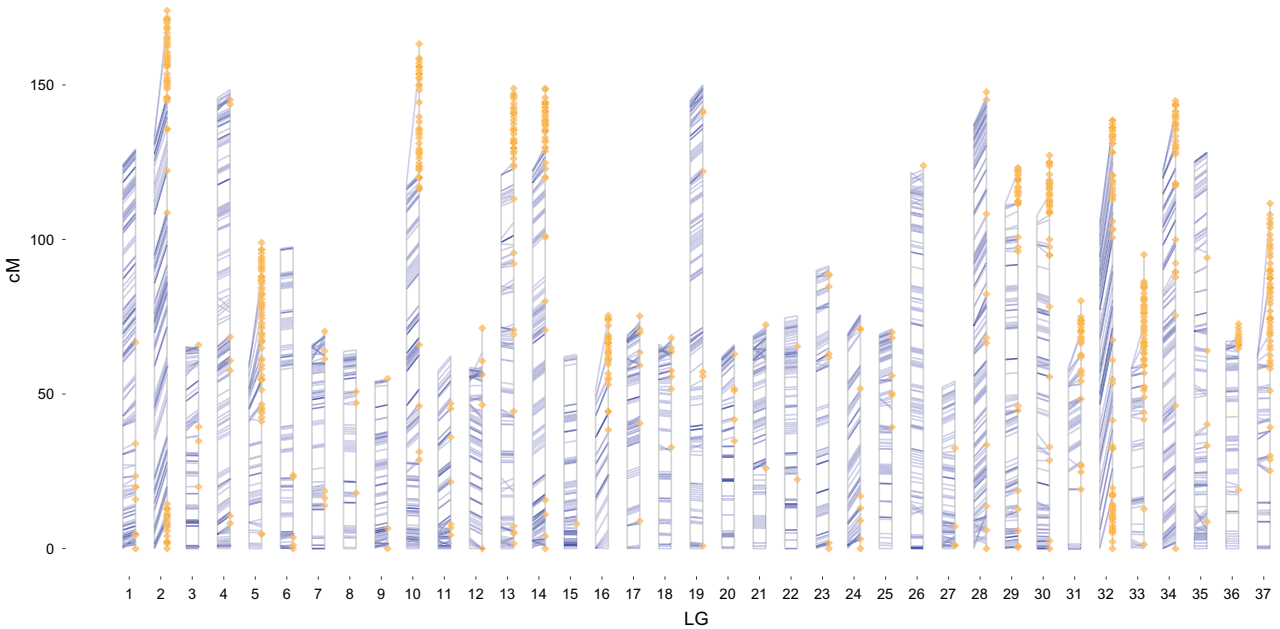


**Fig. 2** Linkage maps constructed with and without the inclusion of paralogs. Each of 37 linkage groups is represented by two adjacent parallel vertical lines: *Map1* (no paralogs) on the left and *Map2* (with paralogs) on the right. Blue lines connect the positions of loci appearing on both maps. Yellow diamonds are loci resolved from confounded catalogue entries (paralogs) and appear only on *Map2*. LGs are numbered 1–37 according to the number of loci present on each LG in *Map2*.

known concentration of segmental duplications (Riethman 2008; del Carmen Calderón *et al.* 2014).

### *Paired-end assembly*

A total of 710 775 (~1% of 69.5 M) paired-end sequences were assigned to a catalogue entry included on *Map2*, leading to an average of 117.8 (SD 44.3) sequence pairs available for assembly at each catalogue entry. Local assembly was successful, with >99% of catalogue entries producing contigs. These assemblies resulted in 14 157 total contigs representing 6034 loci on *Map2* and ranged in length from 93 to 595 bp.

### *BLAST annotation and GO enrichment*

We successfully annotated 1049 catalogue entries with proteins found in the SwissProt database (Appendix S4, Supporting information). Tc1/mariner transposable element-associated (TE) sequences were the most common annotation, constituting 98 (9.3%) of the annotations. A total of 104 catalogue 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 nonparalogous loci (<0.05), with 10 being enriched in paralogous loci, whereas seven 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), all terms related to TE-associated sequences. We are more interested in general patterns of functional differences between paralogous and nonparalogous 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.

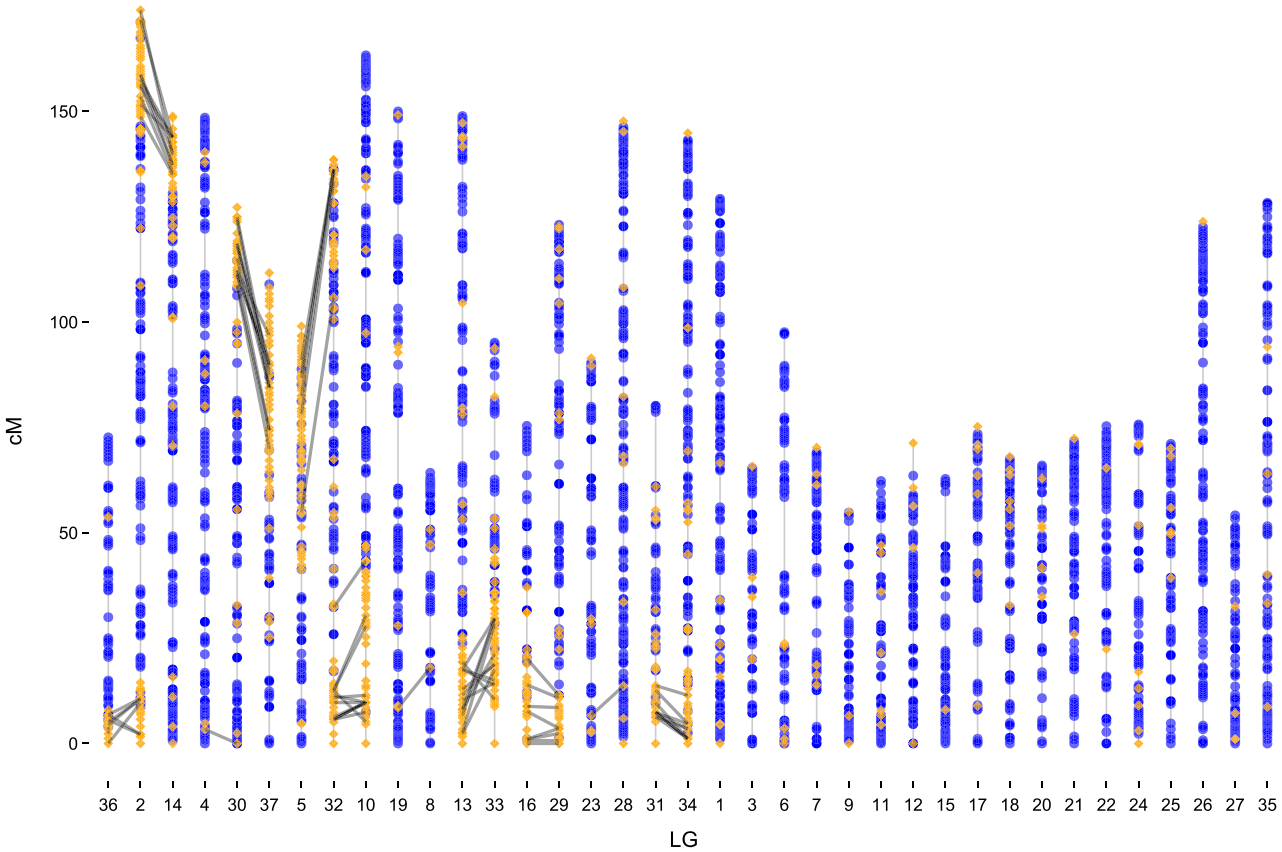


**Fig. 3** Identification of homeologous chromosomes and regions of residual tetrasomic inheritance in *Map2,* which includes resolved paralogous loci. Nonduplicated loci are shown as blue circles and duplicated loci presented as yellow diamonds. Black lines connect confounded paralogs that have been resolved into two loci. The 16 subtelomeric concentrations of paralogs form 8 pairs; notice LGs 2 and 32 form distinct pairings on each end. LGs have been reordered from Fig. 2 to facilitate illustration.

## 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 subtelomeric 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, Fig. 3), and the presence of identical alleles at paralogous loci (Table 1), suggest regions of 14 chromosomes are still undergoing residual tetrasomic inheritance due to incomplete rediploidization (May *et al.* 1979; Allendorf & Danzmann 1997).

Subtelomeres are known to harbour segmental duplicates in the most distal 500 kb (cf., Riethman 2008); 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 (cf., Lien *et al.* 2011; Brieuc *et al.* 2014). These pairs probably represent 16 (2*8) ancestral chromatids that joined into 14 chromosomes through 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 probably metacentric chromosomes formed by the fusion of at least two ancestral chromosome arms.

### *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).

### *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, rediploidization (Parisod *et al.* 2010). It is not clear whether the residual tetrasomic inheritance in salmonids is stable, or whether rediploidization 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 rediploidized taxa.

### *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 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 eight catalogue 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 colocated 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 rediploidization. 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).

## Acknowledgements

The authors thank 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.

## References

Allendorf FW (1978) Protein polymorphism and the rate of loss of duplicate gene expression.

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. *Journal of Molecular Biology*, **215**, 403–410.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Berthelot C, Brunet F, Chalopin D *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, **5**, 3657.

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.

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.

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: Genes|Genomes|Genetics*, **4**, 447–460.

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

Brown JD, O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annual Review of Genomics and Human Genetics*, **11**, 291–316.

Carbon S, Ireland A, Mungall CJ *et al.* (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

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**, 6488.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davidson WS, Koop BF, Jones SJ *et al.* (2010) Sequencing the genome of the Atlantic salmon *(Salmo salar)*. *Genome Biology*, **11**, 403.

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, **3**, e314.

Dittman A, Quinn T (1996) Homing in Pacific salmon: mechanisms and ecological basis. *Journal of Experimental Biology*, **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. *Molecular Ecology*, **23**, 40–69.

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.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, **491**, 756–760.

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, New York.

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.

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. *Evolutionary Applications*, **7**, 480–492.

Fedoroff NV (2012) Presidential address. Transposable elements, epigenetics, and genome evolution. *Science*, **338**, 758–767.

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.

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.

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

Gramfort A, Luessi M, Larson E *et al.* (2013) MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, **7**, 267.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

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. *Molecular Ecology*, **22**, 3002–3013.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Jaillon O, Aury J-M, Brunet F *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.

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. *Molecular Ecology*, **22**, 2848–2863.

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: Genes|Genomes|Genetics*, **4**, 1717–1730.

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

Larson WA, Seeb LW, Everett MV *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon *(Oncorhynchus tshawytscha)*. *Evolutionary Applications*, **7**, 355–369.

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.

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. *Journal of Heredity*, **105**, 741–751.

Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings. Biological Sciences*, **281**, 20132881.

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

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.

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 Board of Canada*, **36**, 1114–1128.

Meirmans PG, Van Tienderen PH (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity*, **110**, 131–137.

Miller MR, Brunelli JP, Wheeler PA *et al.* (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, **21**, 237–249.

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.

Oliphant TE (2007) Python for scientific computing. *Computing in Science & Engineering*, **9**, 10–20.

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.

Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. *New Phytologist*, **186**, 5–17.

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*). *Molecular Ecology*, **22**, 5848–5860.

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*). *Cytogenetic and Genome Research*, **116**, 298–304.

Riethman H (2008) Human subtelomeric copy number variations. *Cytogenetic and Genome Research*, **123**, 244.

Robertson WR (1916) Chromosome studies. *Journal of Morphology*, **27**, 179–331.

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.

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.

Seeb JE, Seeb LW (1986) Gene mapping of isozyme loci in chum salmon. *Journal of Heredity*, **77**, 399–402.

Seeb JE, Pascal CE, Grau ED *et al.* (2011) Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources*, **11**, 335–348.

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

Smith CT, Park L, Vandoornik D, Seeb LW, Seeb JE (2006) Characterization of 19 single nucleotide polymorphism markers for coho salmon. *Molecular Ecology Notes*, **6**, 715–720.

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution*, **14**, 348–352.

Spruell P, Pilgrim KL, Greene BA *et al.* (1999) Inheritance of nuclear DNA markers in gynogenetic haploid pink salmon. *Journal of Heredity*, **90**, 289–296.

Strimmer B (2013) FDRTOOL: Estimation and control of (local) false discovery rates.

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

Tomekpe K, Lumaret R (1991) Association between quantitative traits and allozyme heterozygosity in a tetrasomic species: Dactylis glomerata. *Evolution*, **45**, 359–370.

Wang X, Wang H, Wang J *et al.* (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, **43**, 1035–1039.

Wright JE, Johnson K, Hollister A, May B (1983) Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. In: Isozymes: *Current Topics in Bological and Medical Research* (eds Rattazzi MC, Scandalios JG, Whitt GS), pp. 239–260. Alan R. Liss, New York.

Wu R, Ma CX, Casella G (2004) A mixed polyploid model for linkage analysis in outcrossing tetraploids using a pseudo-test backcross design. *Journal of Computational Biology*, **11**, 562–580.

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 Genetics*, **4**, e1000212.

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.

J.S. and L.S. conceived the study, R.W. developed methods and analysed the data, and R.W., J.S. and L.S. wrote the study.

## 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 'ml-psv'.

Illumina reads are available on the NCBI Short Read Archive (BioProject Submission ID: SUB745855).

## Appendix I

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:

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

where [$h_i$, …, $h_k$] 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 follows:

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

where $r$ is the error-free probability of a parent of genotype $G_p$ producing an offspring with genotype $i$ (Supplemental File s1), and $\varepsilon$ 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 $p * \varepsilon$ term represents the assignment of a random genotype when the genotype specified by $h_i$ is true.

## Appendix II

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:

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

where $G_p$ is the parental genotype, there are $n$ offspring, $h_i$ is the genotype of offspring $i$, and $h_{\text{error}}$ 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 this bias if we assume that all possible combinations of alleles are equally likely to be the result of an error and then scale our naïve estimate by the fraction of errors that we can observe. The total number of distinguishable combinations ($h_{\text{all}}$) of $k$ alleles, given our inability to observe allele dosage, is as follows:

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

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 $\varepsilon$ is as follows:

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

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Appendix S1** Parental allele distribution models.

**Appendix S2** Sequencing read counts and quality filtering.

**Appendix S3** Linkage mapping.

**Appendix S4** Paired-end assembly and annotation.