# Supplement: Detecting Differential Copy Number Variation Between Groups of Samples

Craig B. Lowe*[1,2], Nicelio Sanchez-Luege*[1], Timothy R. Howes[1], Shannon D. Brady[1], Rhea R. Daugherty[1,3], Felicity C. Jones[1,4], Michael A. Bell[5], and David M. Kingsley†[1,2]

[1]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA

[2]Howard Hughes Medical Institute, Stanford University, Stanford, CA

[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA

[4]Current address: Friedrich Miescher Laboratory of the Max Planck Society, Tuebingen, Germany

[5]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY

*These authors contributed equally

†Corresponding author: kingsley@stanford.edu

# Supplement

Supplemental text, figures, and tables.

## S1 Supplemental Text

### Calculating Read Depth

We used sequencing data from a recent study of single nucleotide polymorphisms in geographically widespread freshwater and marine stickleback populations (Jones et al. 2012b). The reads are single-end and 36bp in length. We mapped these reads to the stickleback reference genome (Jones et al. 2012b) with BWA (Li and Durbin 2010). Reads that could not be uniquely mapped to the genome were randomly placed at one of the top scoring locations. We counted matches and mismatches towards read depth at a given base position, but alignment gaps in the read or reference were not counted towards read depth.

### Model notation

- $i$ = index over all base positions in the genome
- $s$ = the number of samples
- $n$ = index over samples such that $n \in \{1, 2, \cdots, s-1, s\}$
- $k$ = length of read
- $j$ = index over all reads that could cover a base position, $j \in \{1, 2, \cdots, k-1, k\}$
- $\mathbf{r} = (r_1, r_2, \cdots, r_s)$ = the total number of reads used to sequence each sample
- $\mathbf{d} = (d_1, d_2, \cdots, d_s)$ = the read depth for each sample at a base position. Each base position will have a separate instance of $\mathbf{d}$.

- $t$ = total number of positions in the genome where a k-mer could begin (approximately the size of the genome assembly)

- $\mathbf{b} = (b_1, b_2, \cdots, b_k)$ = the number of times each k-mer that overlaps a base position appears elsewhere in the genome. Each base position in the genome will have a separate instance of $\mathbf{b}$.

- $\mathbf{G}$ = a matrix of correction factors for how likely a k-mer is to appear as a sequencing read based on GC content. $G_{nj}$ is the GC correction for the $j^{th}$ read that overlaps a base position for sample number $n$. Each base position has its own $\mathbf{G}$ matrix.

**Definition of model states**

The transducer has 25 states representing all combinations of canonical copy numbers for the two sample groups (named marine and freshwater in this study): homozygous deletion, heterozygous deletion, consistent with reference, heterozygous duplication, homozygous duplication (Figure 2). We denote these 25 states as $\psi_{m,f}$, where $m$ represents the canonical copy number of marine fish $m \in \{0, 1, 2, 3, 4\}$ and $f$ represents that of the canonical freshwater fish $f \in \{0, 1, 2, 3, 4\}$. Each state can be parameterized by: a vector $\mathbf{c}$ that is comprised of the copy number for each sample such that $c_n = \{m, f\}$, a vector $\mathbf{r}$, which contains the total number of reads for each sample, and a value $t$ that holds the number of positions in the genome where a read of length $k$ could begin: $\psi_{m,f} = (\mathbf{c}, \mathbf{r}, t)$.

The transducer has both input tapes, $\mathbf{x}$, and output tapes, $\mathbf{y}$. $x_i$ is a tuple of length $s \cdot k + k$ representing the values in all the input tapes at base $i$. Of these input tapes, $k$ are used to encode $\mathbf{b}$ for each base and $s \cdot k$ are used to encode $\mathbf{G}$ for each base. A single position in the output tapes $y_i$ uses $s$ tapes to encode a tuple representing $\mathbf{d}$ for each base.

Each state $\psi_{m,f}$ has a set of emission probabilities. The probability of emitting the entire column of depths $\mathbf{d}$ encoded by $y_i$ can be expressed as the joint probability of emitting each

depth individually.

$$\Pr(y_i|x_i, \psi_{m,f}) = \prod_{n=1}^{s} \Pr(d_n|x_i, \psi_{m,f}) \tag{S1}$$

The read depth for each sample is modeled using the binomial distribution, which has parameters for the number of trials and the probability of success in each trial. The number of trials is different for each sample and is equal to the number of mapped reads for that sample, $r_n$. The probability of success, $p_n$, is the chance that a randomly chosen read will cover the base of interest. The equation for $p_n$ is described in detail leading up to equation Equation (6), but with the current annotation can be written as follows:

$$p_n = \sum_{j=1}^{k} \frac{\frac{1}{2} \cdot c_n + (b_j + 0.01)}{t} \cdot \frac{1}{(b_j + 0.01) + 1} \cdot g_{nj} \tag{S2}$$

The number of trials for the binomial is equal to the number of mapped reads for the sample, and the probability of the entire tuple of depths being emitted is the joint probability of the depth for each sample.

$$\Pr(y_i|x_i, \psi_{m,f}) = \prod_{n=1}^{s} Binom(trials = r_n, p = p_n) \tag{S3}$$

**Simulating main data set**

For understanding mismapping probabilities, optimizing model parameters, and quantifying model performance, we repeatedly simulated data analogous to our main data set. We used the stickleback genome assembly as the ancestral state and evolved this sequence, using only substitutions, to recapitulate the divergences seen in our data set. To estimate the probability of mismapping or false positives with the model, we mutated the genomes no further. However, to assess the model's ability to detect consistent copy number differences between marine and freshwater genomes, we created 1000 randomly placed deletions or

duplications of lengths between 30bp and 2500bp in all freshwater or marine genomes. We then simulated reads from each of the 21 evolved genomes with ART (Huang et al. 2012), using the default parameters specific to Illumina sequencing technology. The reads were 36 bases long and were mapped to the genome with BWA (Li and Durbin 2010). To have the same coverage, we then downsampled the number of mapped reads for each of the 21 libraries to match the number of mapped reads in our sample for that individual.

To estimate the mapping and mismapping rates in our data set we calculated the fraction of mapped reads that mapped back to either the same locus from which they were created (0.91), or a location different from which they were created (0.09).

**Optimizing transition probabilities**

Even though we allow all states to transition to all other states, we have reduced the number of free parameters in the model to 14. We will first describe the transition parameters out of the state consistent with the reference genome, $\psi_{2,2}$ (Figure 2). We have a similar number of samples for marine and freshwater populations so we use the same transition probabilities to state $\psi_{x,y}$ as we do to $\psi_{y,x}$. This reduces the number of free parameters describing transitions to off diagonal states from 20 to 10. We then have four other states on the diagonal ($\psi_{0,0}$, $\psi_{1,1}$, $\psi_{3,3}$, and $\psi_{4,4}$), which gives us 14 free parameters. The self-transition to state $\psi_{2,2}$ is not free because the probabilities must sum to one. We use the same parameters for transitioning out of the models start state.

We then use the same 14 parameters to define the transitions from an off-diagonal state without introducing any new variables. From an off-diagonal state we expect the model to stay in a state for 50bp, which defines the self-transition probability ($\frac{1}{50}$). We then use the 14 already defined parameters as weights to define the remaining transition probabilities from an off-diagonal state to any of the other states.

We simulated data sets in order to calculate the transducer's ability to detect copy

number variation correlated with the ecological variable of marine or freshwater habitat while constraining the expected number of false positives to be less than or equal to 0.2 for a genome-wide analysis. To understand the expected number of false positives, we simulated a data set analogous to the actual data set, but with no copy number variation (see Simulating main data set). We did this for five different simulated data sets lacking copy number variation and only considered parameter sets that had zero or one false positives cumulatively across these five data sets.

Staying within the set of parameters that satisfies the constraint on false positives, we maximized the number of true positives detected in the simulated data sets. Similar to how we generated the data sets with no copy number variation, we generated data sets with known copy number variation by making deletions or duplications across all simulated marine or freshwater genomes (see Simulating main data set). We made 1000 of these mutations across the genome of various lengths (30bp, 50bp, 100bp, 150bp, 200bp, 300bp, 400bp, 500bp, 750bp, and 1000bp). We then maximized the area under the curve (AUC) for the length of the mutation versus sensitivity functions while maintaining an expectation of 0.2 false positives or less. To maximize the AUC we randomly selected a transition penalty and adjusted it up or down by 10 (in log space). If the new parameterization had a greater AUC and did not return more than one false positive on the five data sets with no copy number variation, this adjustment was accepted and otherwise rejected. This search repeated until convergence. Evaluating the sensitivity of the model and verifying that the false positive rate in the absense of copy number variation is less than or equal to 0.2 was performed on similated data sets that were generated with the same methods as the training data, but had not previously been seen by the transducer (Figure 3). These parameters that were learned on the training data sets and evaluated on the test data sets are the same parameters used in the true data set of sequenced stickleback fish.

New parameterizations should be used for entirely new data sets, but small changes to

the data set such as using different read mapping software or adding an individual to a group should not require a new parameter search (Figure S4).

### Binomial Mixture Model

A strength of the described method is that we do not expect the same distribution of read depths for every base; we tailor the distribution to the uniqueness and GC-content of reads that would cover the base. However, to compare to other methods that look at coverage over large regions, we expect the read depth to follow a mixture model where each base position contributes an equally weighted component to the model that is a binomial with a probability of success described by Equation (6). Where $p_i$ is the evaluation of equation Equation (6) at base position $i$ and $reads$ is the number of mapped reads for the sample, the mixture model over a set of $n$ bases would be:

$$\Pr(depth) = \frac{1}{n}\sum_{i=1}^{n} Binom(p = p_i,\ trials = reads) \tag{S4}$$

### Comparison to existing descriptions of read depth

When comparing to the binomial distribution, we calculate the probability of success in the binomial using the genome-wide mapping (0.91) and mismapping (0.09) rates, the number of ungapped bases in the assembly, $sizeOfGenome$, and the copy number of the locus in a diploid genome.

$$binomialProb = 0.91 \cdot \frac{\frac{1}{2} \cdot copies}{sizeOfGenome} \cdot +0.09 \cdot \frac{1}{sizeOfGenome} \tag{S5}$$

This equation compensates for mismapping, but as a genome-wide average and not at per-base resolution. The trials of the binomial is the number of mapped reads times the read length.

We parameterize the negative binomial based on the mean and size parameters. The mean is calculated as the above binomial probability, times the number of mapped reads, times the read length. To be conservative, we directly calculate the size parameter from moment matching with the read depth variance seen in simulated data.

**Comparison to other methods**

We compared the transducer's performance on simulated data sets to analyses using CNVnator (Abyzov et al. 2011), rSW-seq (Kim et al. 2010), cnMOPS (Klambauer et al. 2012), and Genome STRiP (Handsaker et al. 2015). We tested these programs on both individuals and pseudo-individuals created by pooling all marine samples into a single file and all freshwater samples into a different file. We tuned parameters on one simulated data set and evaluated the performance of all models on a second never-before-seen data set. We tuned parameters with the goal of keeping the number of false positives to 0.2 per genome-wide analysis (1 per 5 simulated data sets with no copy number variation present). When we evaluated method sensitivity on new data sets, we also generated new data sets with no copy number variants to ensure that not more than one false positive was annotated per five genome-wide runs.

We note the methods we are comparing against were primarily designed and optimized to detect copy number variants larger than 1kb and with greater sequencing coverage, so are being tested on simulated data that is quite different from their intended use case. However, these comparisons are likely to illustrate how existing methods may perform in detecting repeated copy number variation less than 1kb in length when there is low sequencing coverage.

When running CNVnator, we tuned the window size so that the mean was greater than the standard deviation by a factor of four to five, as advised by the supplement of that publication (Abyzov et al. 2011). When analyzing individual fish we adjusted the window size for each library to be as close as possible to four-and-a-half. The window sizes for freshwater fish were as follows: BEPA 575bp, BIGL 300bp, FTC 300bp, HUTU 250bp, MATA 325bp,

MUDL 500bp, NOST 1350bp, PAXB 375bp, SCX 600bp, SHEL 450bp, TYNE8 325bp. For marine fish: ANTL 950bp, BDGB 500bp, BIGR 450bp, GJOG 375bp, GORT 425bp, JAMA 325bp, NEU 2000bp, RABS 350bp, SALR 600bp, TYNE1 725bp. The lack of reads mapping to gaps in the assembly caused those regions to often be marked as deletions by CNVnator, so we filtered out all calls overlapping an assembly gap. We then identified continuous genomic regions where, using Fisher's exact test, there was a difference in the number of individuals with copy number variants between the two groups. We searched for a p-value threshold that would allow us to reliably detect mutations less than 1kb, while yielding no more than 10 false positives, but we were unable to find such a threshold. We believe that the sequencing coverage in our data set is too sparse to allow for current methods to call deletions and duplications of less than 1kb, when considering individuals in isolation.

Both cnMOPS (Klambauer et al. 2012) and Genome STRiP (Handsaker et al. 2015), while providing individual copy number variation calls, process all samples together and leverage the knowledge of other samples when annotating an individual. We maximized the area under the curve for the functions plotting size of deletions and duplications (30bp to 1kb) compared to sensitivity of the method. The final parameterization we used for cnMOPs was a window length of 125, a prior impact of 0.1, and a minimum width of 1. Genome STRiP needed the most modification since a 1kb lower limit for deletions and a 2kb lower limit for duplications is hard-coded into the software pipeline. We reduced both of these limits to 700 so that we could test the method against 750bp and 1kb mutations without greatly changing the program's intended use. We kept tiling window size equal to the maximum reference gap length, both the tiling window overlap and minimum refined length at half of this value, and the boundary precision at 100. These constraints were based on reading the online documentation which maintained these ratios in the most sensitive suggested parameterization and stated that boundary precisions below 100 would have a limited effect. Our final parameterization was a tiling window size of 700, maximum reference gap length

of 700, tiling window overlap of 350, and minimum refined length of 350. Genome STRiP has also been optimized for the human genome, so we added an artificial Y chromosome to the stickleback genome assembly when mapping simulated reads so that the pipeline could compare coverages on chrX and chrY to infer a male or female sample and continue the analysis.

Pooling many individuals into one pseudo-individual has drawbacks for real data sets, such as a single individual duplicating a region many times appearing the same as all individuals duplicating the same region to a lesser extent. However, pooling all individuals from one group, pooling all individuals from the other group, and identifying copy numbers that are different between the two pools is a way to detect copy variants correlated with group membership using current methods. After pooling the reads we ran CNVnator (Abyzov et al. 2011) on each pool. We kept the window size equal for both the marine and freshwater pools since they have similar sequencing coverage and changing the window size may create region of differing copy number as an artifact of the window starts and stops not occurring at the same genomic locations. We used a window size of 280 to maximize sensitivity while limiting the number of false positives to 0.2 per genome-wide analysis. We also used rSW-seq (Kim et al. 2010) to analyze both pools at once since it is limited to analyzing two samples at once and detecting differences. We found that the performance was quite different when the sample order was swapped. Because of this we found that we achieved the best performance when we ran the software once to detect deletions and then again with the sample order switched to detect duplications. We used parameters of 45 and 325 after optimizing both input orders. See Table S1 for a summary of parameters used for all methods.

**Multi-species alignment and cross-species comparison**

We created a multi-species alignment referenced on the threespine stickleback (gasAcu1) including the following other species: medaka (oryLat2), tetraodon (tetNig1), fugu (fr2),

zebrafish (danRer5), human (hg18), mouse (mm9), cow (bosTau4), chicken (galGal3), and opossum (monDom4). We created pairwise alignments with LASTZ (Harris 2007), filtered the initial alignments for regions showing conserved synteny (Kent et al. 2003), and used Multiz (Blanchette et al. 2004) to construct a multi-species alignment. We identified regions of the alignment that were evolving under constraint, based on frequency of substitutions using a phylo-HMM (Siepel et al. 2005). The evolutionary history of these regions in stickleback is similar to other teleosts and shares many characteristics with mammals (Lowe et al. 2011).

### Sequencing individuals from the Little Campbell River and Fish Trap Creek

To investigate signatures of selection near consistent, derived deletions in marine stickleback, we sequenced additional fish from the mouth of the Little Campbell River in British Columbia (marine) and Fish Trap Creek in Washington State (freshwater). The reads are paired-end and 76bp in length. We mapped the reads to the stickleback assembly (Jones et al. 2012b) using BWA (Li and Durbin 2010) and removed duplicates with Picard. This resulted in 9x coverage for both fish. To identify heterozygous sites, we used GATK (McKenna et al. 2010; DePristo et al. 2011) according to the published best practices (Van der Auwera et al. 2013).

### Simulating selective sweeps

We estimated how long ago the selective sweeps in marine populations occurred by simulating selective sweeps and the return to normal levels of heterozygosity. To accomplish this, we used Cosi2, which allows for coalescent simulation and positive selection (Shlyakhter et al. 2014).

We used a mutation rate estimate of $2.5 \cdot 10^{-8}$ from humans (Nachman and Crowell 2000), a genome-wide recombination rate from stickleback (Roesti et al. 2013), and assumed a single population with constant size. We used the population size as a free parameter

to fit the observed genome-wide level of heterozygosity in our marine fish from the Little Campbell River. A population size of 40,000 closely fits the observed data, with an intersection distance (Swain and Ballard 1991) of 0.12 (Figure S10). We then repeatedly simulated 20kb windows with the advantageous mutation in the center. For each repetition of the simulation, the 20kb fragment was randomly assigned to have the recombination rate of a segment containing a derived marine deletion. After each repetition, we randomly selected one individual from the population of 40,000 and calculated the number of heterozygous sites in that individual.

We have used a mutation rate estimate that is two times higher than other reported rates in humans (Kong et al. 2012) to be conservative when dating the likely age of the deletions. There is a report of stickleback having an even lower mutation rate by a factor of 10 (Colosimo et al. 2005); however, this was estimated before whole-genome sequencing was widely available. If the true mutation rate in stickleback is lower, this will only increase the number of generations needed to restore a baseline level of heterozygosity after it has been lost due to a sweep. This makes it even more unlikely that these selective sweeps in marine populations would have occurred more recently than the last glacial maximum.

The population size that fits the observed heterozygosity is two orders of magnitude higher than what has previously been reported for marine stickleback in northern Europe, although there was noise in the data that caused many of the confidence intervals for effective population size to include infinity (DeFaveri and Merila 2015). The effective population size of 40,000 is in line with what has been seen for other ocean fish, even those experiencing high levels of fishing (Therkildsen et al. 2010). An effective population size of 40,000 is also more in line with $\frac{N_e}{N_c}$ ratios reported for other wild animal populations (Frankham 1995).

**Analyzing the 1000 Genomes data set**

To investigate the performance of the transducer method on a well-studied human data set, we selected 10 individuals from the 1000 Genomes Project for which the consortium has recently annotated copy number variants (Sudmant et al. 2015). We defined one group as consisting of five east Asian individuals (NA18623, NA18624, NA18632, NA18984, and NA19058) and a second group consisting of five African individuals (NA18498, NA18867, NA18912, NA19474, NA19901).

A current method to identify CNVs differing between these two groups is to observe where the current annotation (Sudmant et al. 2015) shows a CNV that is significantly more prevalent in one of the groups compared to the other ($p < 0.01$, uncorrected Fisher's exact test). There are 72 such regions with annotated beginning and end coordinates.

To run the transducer on this data set we treated all reads as single-end and reduced each read to the 75 continuous basepairs with the highest sequencing quality. To identify the most dramatic differences between these two groups we only consider copy numbers of homozygous deletion, reference, and homozygous duplication, which leaves the model with 9 states, instead of the full 25. We reuse all the parameters from the stickleback data set with the exception of recalculating a single penalty for transitioning into a deletion or duplication state. We fit this parameter by simulating a data set with the same sequencing depth and no copy number variation between the samples. We then make the penalty progressively larger until no false positives were detected on the simulated data set. We use $-5000$ (in log space) as the penalty.

We then ran the transducer method on the 10 individuals from the 1000 Genomes data set. The transducer identified 368 regions where the canonical copy number of the two groups is annotated as being different. These 368 regions overlap 29 of the 72 (40%) that we would expect from the current annotation. A 40% overlap is slightly higher than what was observed between the methods used by the consortium to annotate CNVs, where on

average the larger set would cover 35% of the smaller set (median of 31% coverage). There is a significant amount of disagreement between the existing methods for CNV calling, so it is difficult to estimate the transducer's false negative rate. However, the slightly higher than expected overlap between the transducer and the existing set of annotations suggests that the method is likely not missing more true positives than the other existing methods, and may in fact have a lower false negative rate.

Due to the transducer calls at times being fragmented compared to the evolutionary events (see Main Text) the 368 regions overlap 29 of the expected calls, but the expected calls overlap 31 of the transducer regions. This leaves 337 CNVs found by the transducer that were not found by applying Fisher's exact test to the allele counts from the consortium. Most of these 337 CNVs, while not passing the Fisher's exact test threshold, do overlap CNVs found by the consortium, either in the 10 individuals or the other 2494 genomes. However, there are 157 regions annotated by the transducer that are not identified as existing in any individuals by the consortium. These regions may be false positives of the transducer, or CNVs that the transducer detects, but were missed by existing methods. To further test this experimentally, we designed primers and verified two of these regions by PCR amplification and Sanger sequencing (Table S16 and Figure S11). While these data do not address the exact false positive rate, our results do demonstrate that some common copy number variants are missed by existing methods, yet can be discovered by the transducer method.

After sequencing, both of these newly detected regions appear to be not only deletions, but to also have sequence inserted at the deletion breakpoints. It is possible that this pairing of a deletion and an insertion makes these events more difficult for some other methods to detect. There will be no sequencing reads that span the expected junction of the deletion ends because of the inserted sequence. The insertion will also reduce, or even eliminate, the apparent change in insert size of paired-end reads flanking a deletion. The verified deletion on chr14 was also likely not in the existing data set because it is located in a region of the

genome associated with CNVs found in DNA isolated from lymphoblast cell lines (Sudmant et al. 2015). However, we also see presence or absence of this deletion in DNA samples isolated from whole blood (Figure S11), so it is unlikely to be an artifact of cell culture.

It is possible that these two newly verified deletions are functionally important in humans. The deletion on chr15 overlaps a DNase hypersensitivity site in villous mesenchymal fibroblast cells from human placentas at P0 (Thurman et al. 2012). DNase hypersensitivity sites are used to identify tissue-specific regulatory elements. The loss of a regulatory element may increase or decrease the expression of a nearby gene, likely SMAD3. SMAD3 expression is associated with increased proliferation in vascular-related tissues (Tsai et al. 2009), raising the possibility that the polymorphic deletion of a villous regulatory region may affect placental development.

## S2 Supplemental Figures

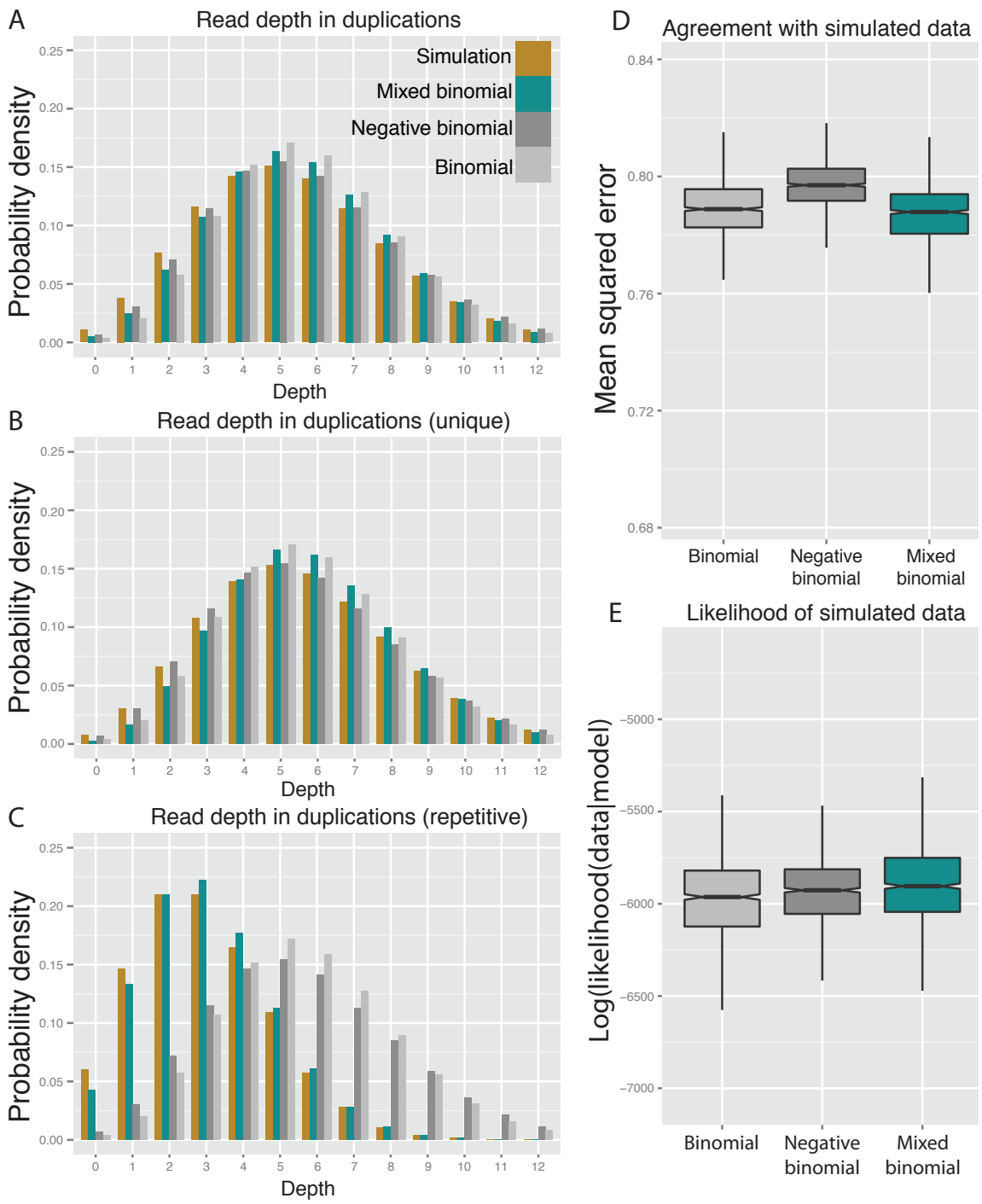


Figure S1: **The binomial mixture model shows modest improvements for modeling depth in duplications.** The binomial mixture model accounts for the fact that bases in repetitive regions, when duplicated, will not show as dramatic of a read depth increase since some of the reads will mismap to other identical regions of the genome. However, this improvement is modest compared to the improvement seen in deleted regions (Figure S2).

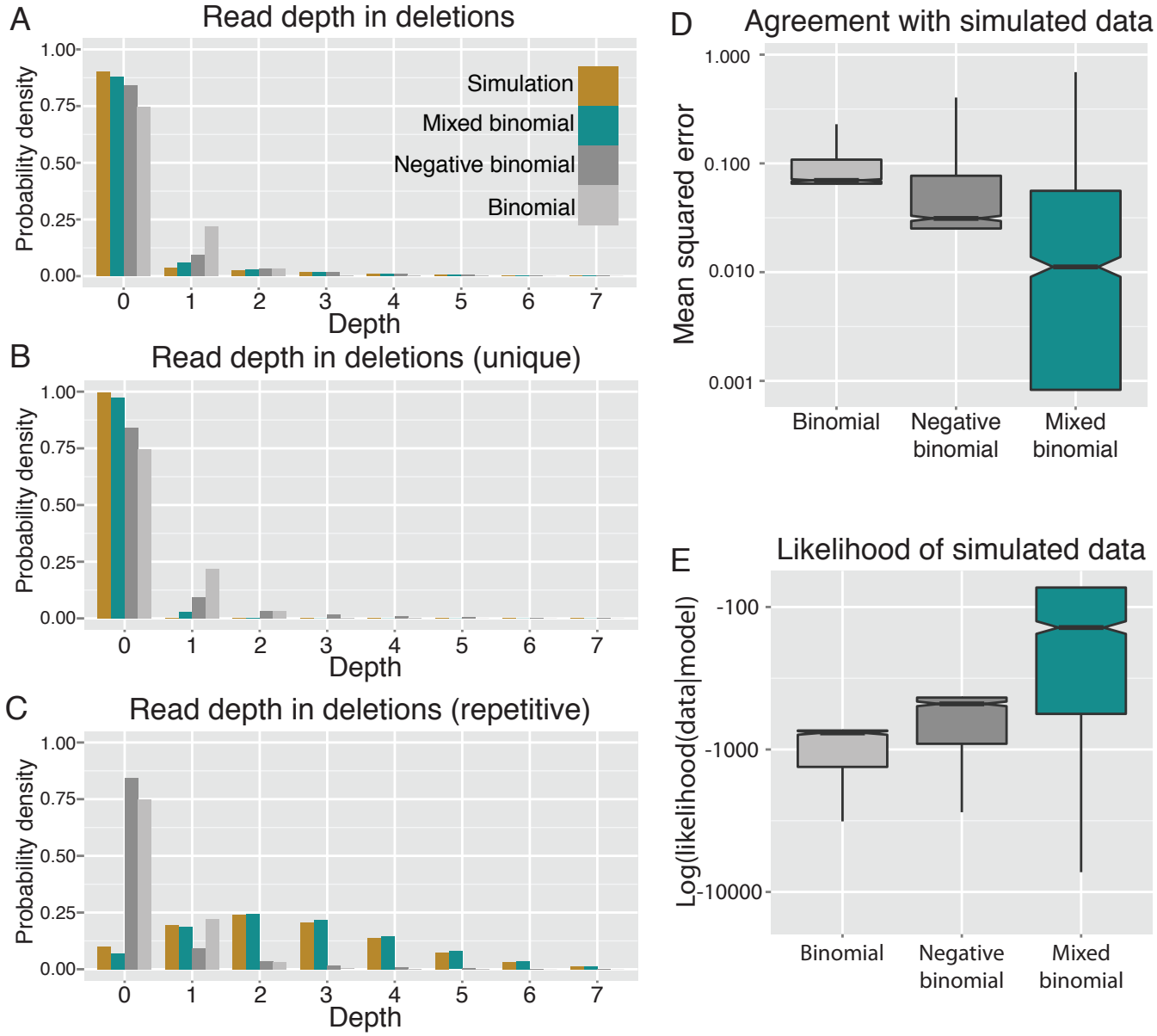


Figure S2: **Performance of the binomial mixture model.** We compare the performance of our binomial mixture model to previously used methods of modeling read depth: the binomial distribution and the negative binomial distribution (see Supplemental Methods). (A) Our method provides a closer fit to the distribution of simulated read depths occurring across 1000 randomly placed 2.5kb deletions (see Supplemental Methods). While other methods often apply the same distribution to all bases, the model we present adapts with single-base resolution, which is advantageous when a number of the reads potentially covering a base are either (B) unique in the genome or (C) repetitive. Both the (D) mean squared error and the (E) likelihood are calculated for each model over the 1000 randomly placed deletions, with the mixture model having a much closer fit to the read depth seen in simulated data.

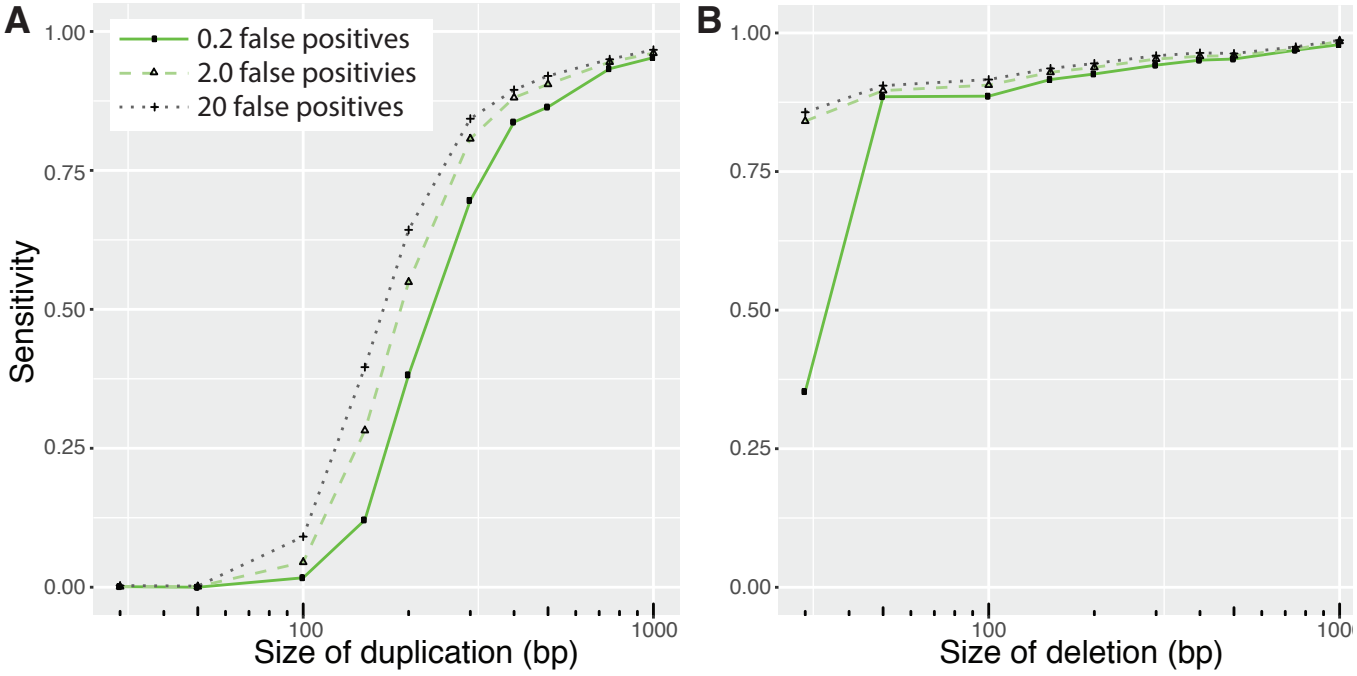


Figure S3: **Effect of changing specificity.** We performed the main analysis with a false positive rate of 0.2 per genome-wide run. By increasing the expected rate of false positives, we are able to recover more true positives for both (A) duplications and (B) deletions.

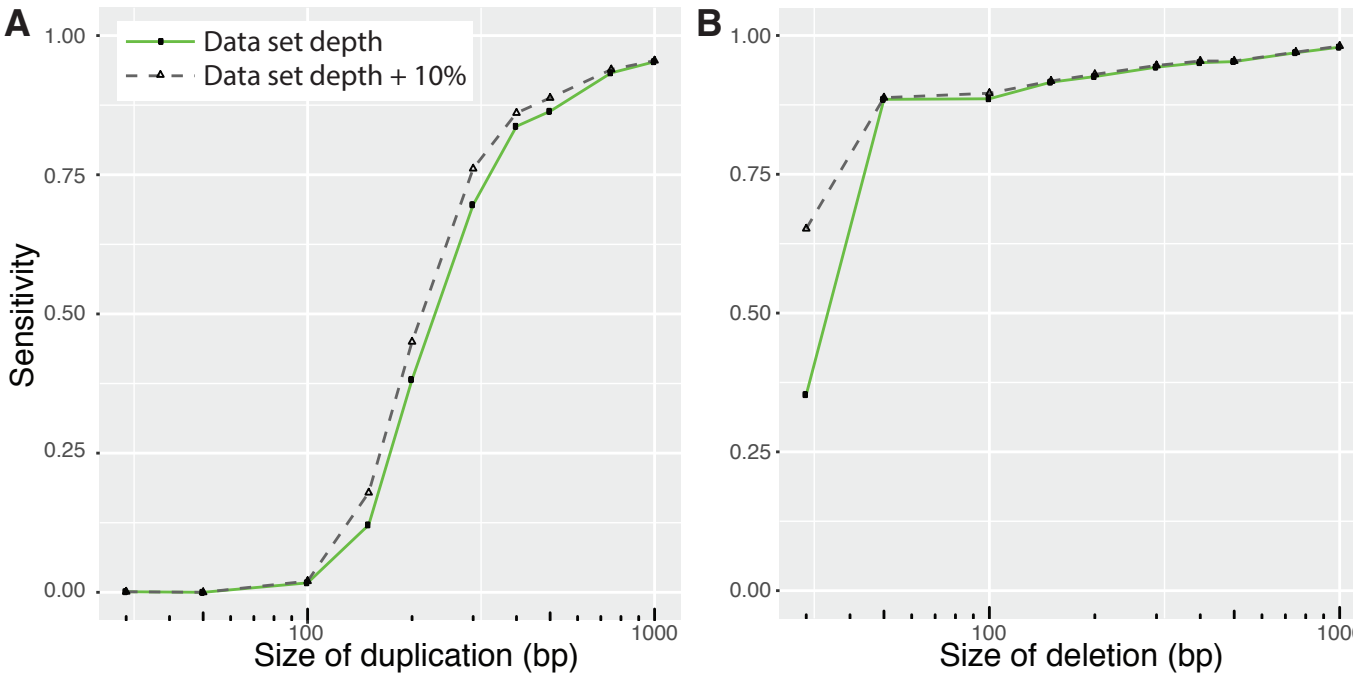


Figure S4: **Performance with ten percent more reads.** We added ten percent more reads to the simulated data sets without modifying transition parameters that had been learned on the lower depth data set. The parameterization is not highly-sensitive to the data set, and the performance of the model improves with the additional reads. There is an increased ability to detect both (A) duplications and (B) deletions. There is no noticeable decrease in specificity with the model detecting no false positives on a simulated data set with no copy number variants, but ten percent more reads.

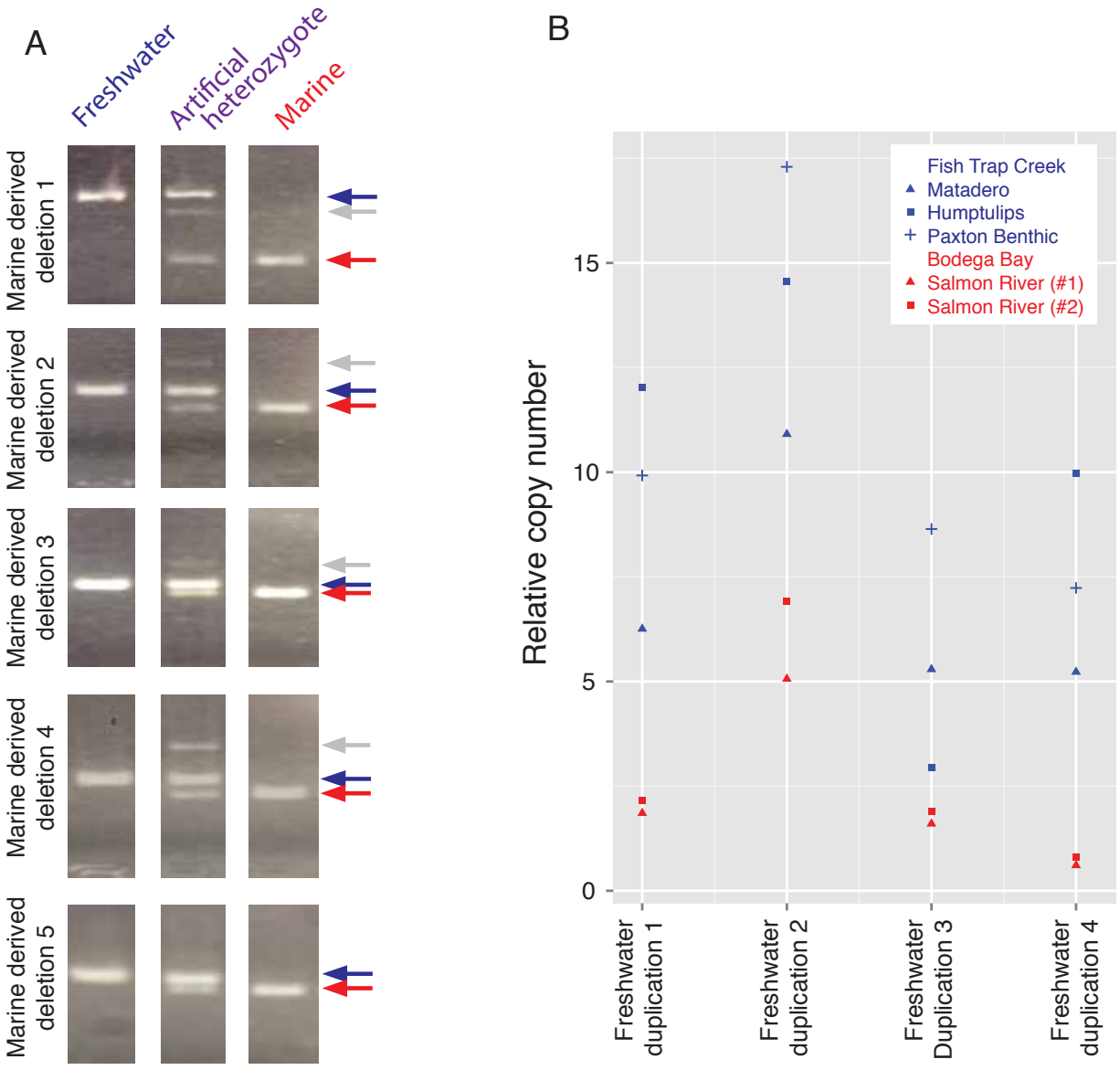


Figure S5: **Validation of proposed deletions and duplications.** (A) We designed PCR primers flanking five teleost conserved regions that were usually present in freshwater stickleback, but appeared consistently deleted in marine fish, as we initially believed this pattern of evolution to be unlikely and therefore anticipated possible false positives. However, the experimental amplifications validated the model's predictions by giving large bands in freshwater fish (blue arrows) and shorter bands in marine fish (red arrows). To ensure that we could detect heterozygotes, we mixed marine and freshwater DNA samples in a 1:1 ratio to create an artificial heterozygote, which gave both the large and small bands, as well as non-specific background bands in four reactions (gray arrow), which did not limit our ability to detect heterozygous individuals. (B) We used qPCR to validate four genomic regions that the transducer annotated as having greater copy number in freshwater fish. We performed the assay on three marine (red symbols) and four freshwater (blue symbol) fish. Each point on the graph represents three reactions that were averaged and then normalized against a control region (Table S3). The results are consistent with the model's prediction that freshwater fish have increased copy number relative to marine fish.

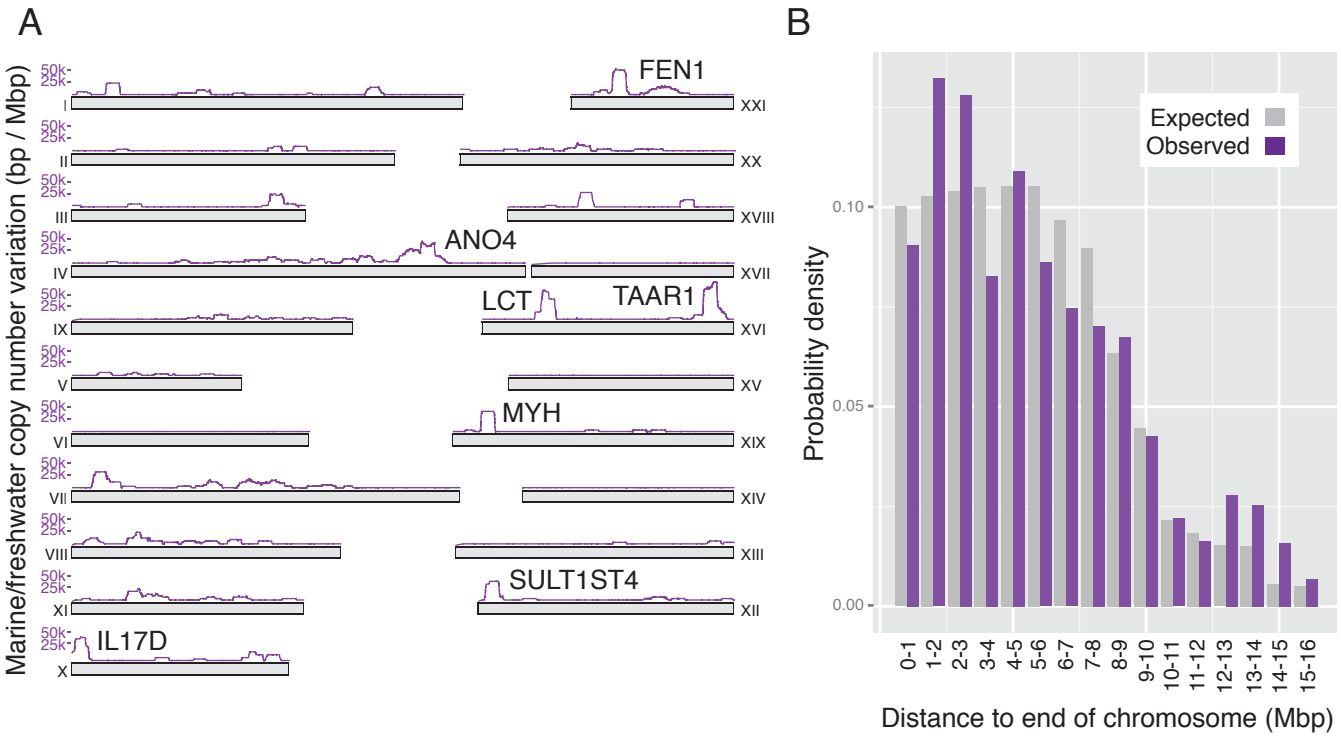


Figure S6: **Genomic locations of repeated marine/freshwater copy number variation.** (A) The number of bases affected by repeated copy number variation that correlates with the marine versus freshwater ecotypes is plotted across the genome in 1Mbp windows. The events occur throughout all chromosomes. Windows with the highest density of affected bases ($\geq 40000$) are annotated with the gene closest to the peak. While there are some large events towards the ends of chromosomes, (B) the distribution of affected bases does not show any large deviations from the expected uniform distribution across ungapped regions in the assembly.

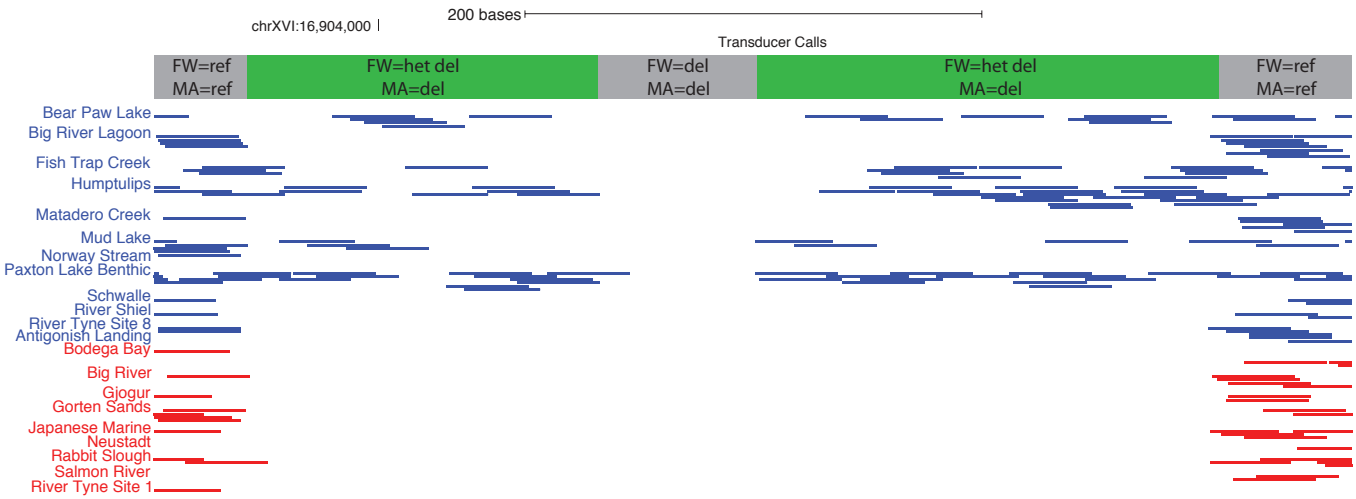


Figure S7: **Complex transducer states and transitions.** This region of the genome illustrates what could be an insertion in the reference assembly fish, or a deletion in all other populations, that breaks up an otherwise larger region of copy number difference between marine and freshwater populations. The outer edges are regions where both freshwater and marine individuals are consistent with the reference genome. Within those boundaries, there is a region that is present in many freshwater individuals, but none of the marine individuals (green). This green region of differing copy number is broken in two by a region where both freshwater and marine individuals appear to have a deletion relative to the reference assembly, possibly due to an insertion in the reference individual.

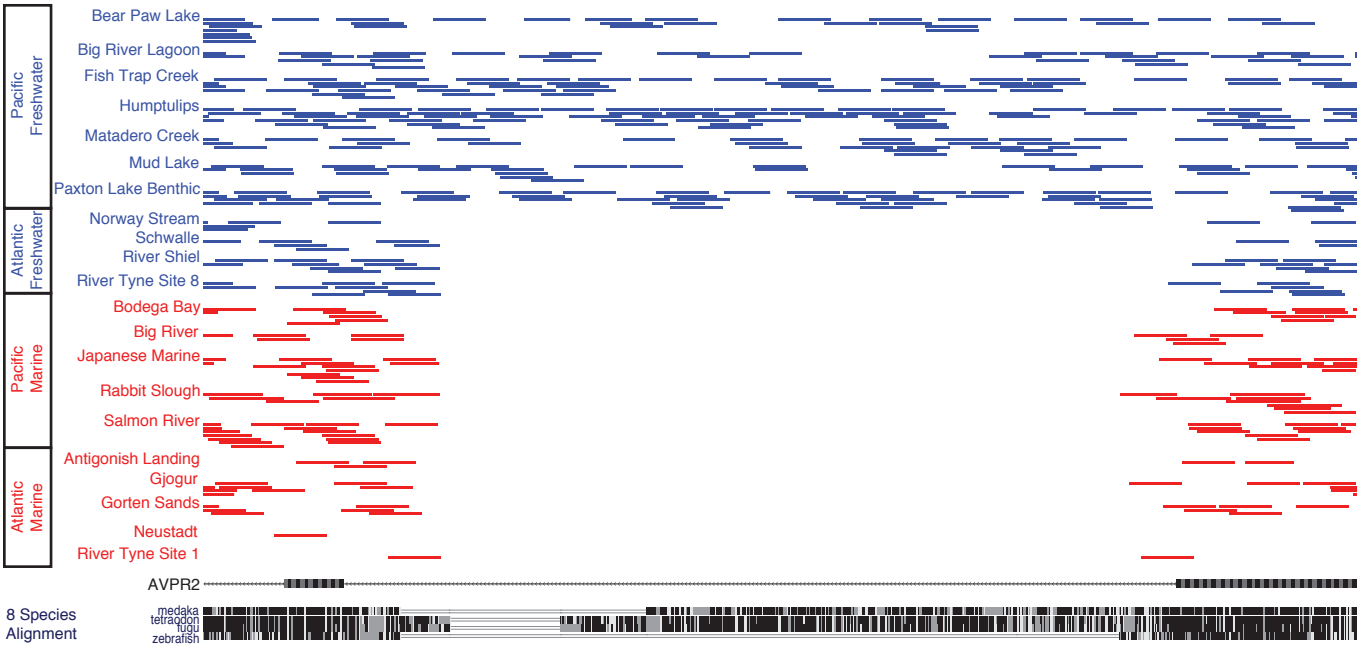


Figure S8: **Copy number correlating with both ecotype and geography.** In this region of the genome, marine individuals have a deletion, while freshwater individuals near the Pacific Ocean have an intact allele. The intact allele may not be found in Europe either due to limited gene flow between Pacific and European freshwater populations, or because the intact allele is not advantageous in Europe. This region is identified in both the Pacific-Atlantic and marine-freshwater analyses.

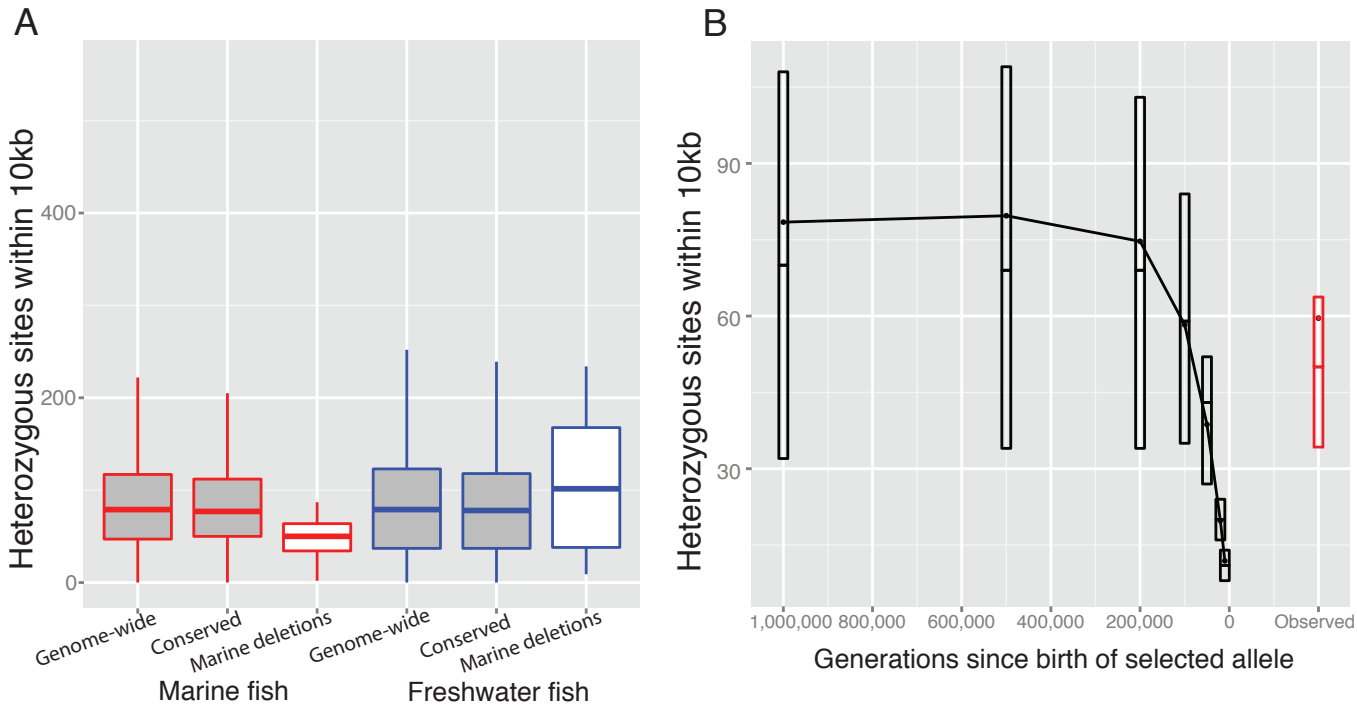


Figure S9: **Regions flanking marine deletions show signatures of selection.** There are 86 regions where marine populations have repeatedly deleted an element of the genome showing cross-species conservation, implying that the marine fish are derived with respect to the freshwater stickleback and other teleosts. We sequenced a marine fish to high coverage, identified breakpoints of the deletions, and summed the number of heterozygous sites within 10kb for each deletion (see Supplemental Methods). (A) These flanking regions had significantly less heterozygosity than windows either randomly placed in the genome or flanking other conserved elements ($p \leq 0.003$ and $p \leq 0.002$). This was not the case for the freshwater fish, whose regions flanking the marine deletions showed a possible increase in heterozygosity compared to the genome-wide distribution or other regions flanking other conserved elements ($p \leq 0.11$ and $p \leq 0.09$). This is consistent with selective sweeps occurring in marine populations while freshwater fish maintain ancient, ancestral alleles at these positions. (B) To estimate how long ago the selective sweeps may have occurred, we compared the decreased level of heterozygosity that we observed to that which we would expect from sweeps of varying ages. Boxes in the graph show first quartile, median, and third quartile, while the mean is depicted by dots and connecting lines. It is likely that many of the sweeps happened around 100,000 generations ago (200,000 years).

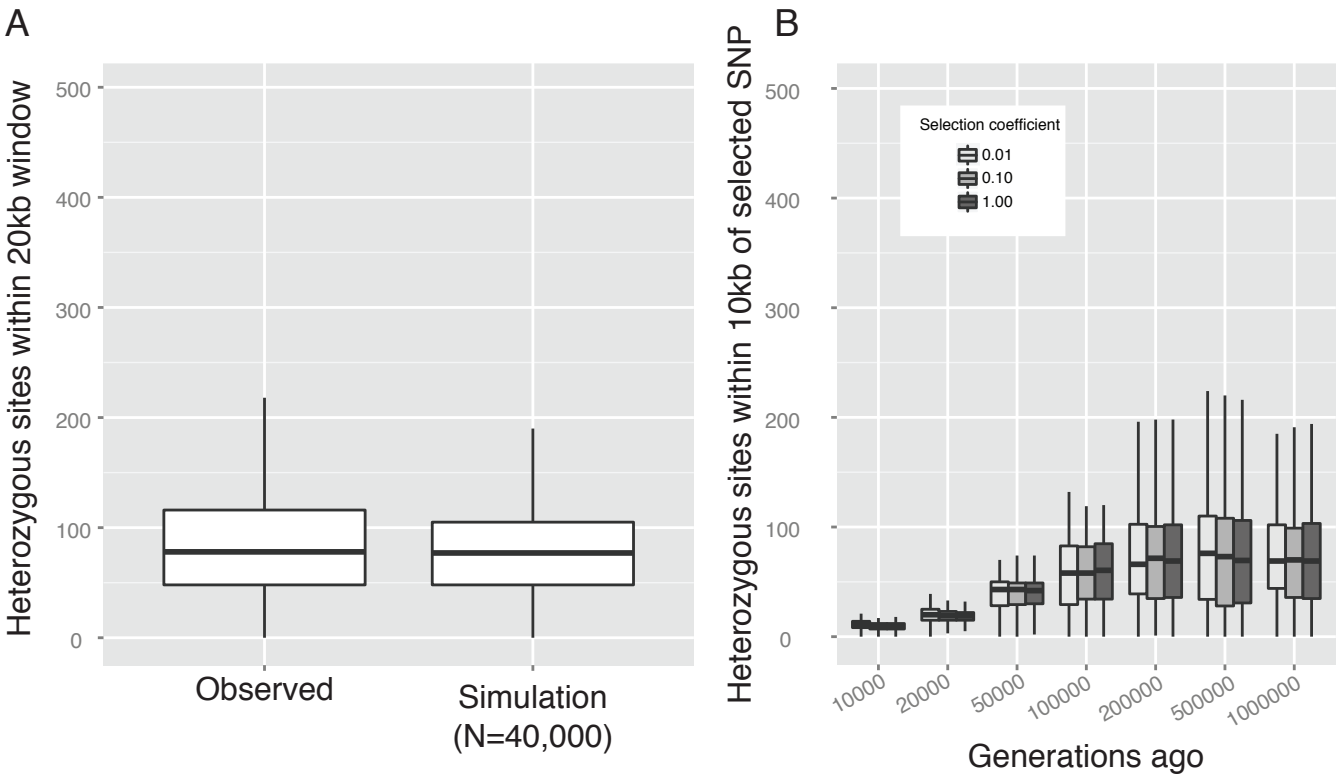


Figure S10: **Fitting a population genetic model to heterozygosity in marine stickleback.** (A) We compare the number of heterozygous sites in 20kb windows (analogous to looking at the 10kb upstream and 10kb downstream of an allele of interest) observed in a marine fish from the mouth of the Little Campbell River to that from our population genetic model with a population size of 40,000. The intersection distance (Swain and Ballard 1991) between these distributions is 0.12. The majority of the disagreement is in the tail, representing windows of high heterozygosity; the observed data have more windows of high heterozygosity than would be expected from the simulation. This tail is likely to be caused, at least in part, by the fact that the model does not incorporate migration. The marine fish near the mouth of the Little Campbell River have constant gene flow with the freshwater population located upstream, where there are likely to be a number of genomic regions that are highly differentiated (Jones et al. 2012b). (B) Since we are not sure of the strength of selection that acted on the deletion alleles, we tested how sensitive the model is to changes in that parameter. Changing the selection coefficient by two orders of magnitude did not create a difference in how quickly the flanking heterozygosity returns at the time scales we investigated.

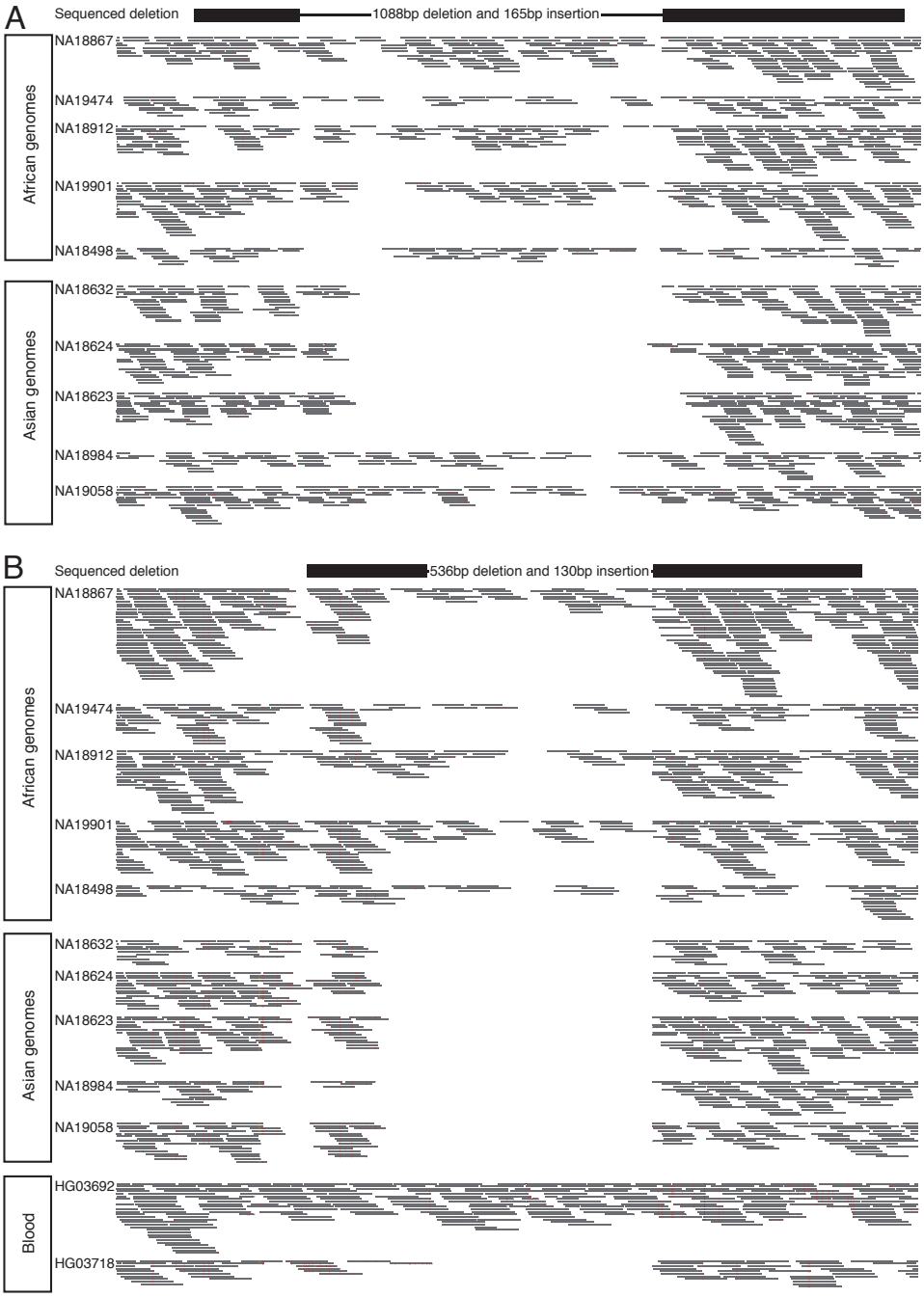


Figure S11: **Human deletions verified by sequencing.** We verified two deletions that were predicted by the transducer method to be largely present in five African samples, but largely absent from five east Asian samples. Both the deletion on chr15 (A) and chr14 (B) also had small insertions at the site of the deletion (Table S16). We show the mapped reads for the five African and five east Asian samples used in the analysis. For the samples on chr14, which is located in a region often associated with CNVs in cultured lymphobast cell lines, we also show sequencing reads from whole blood samples, instead of cell lines, where one individual appears to have the intact allele and the other a deletion. Neither of these deletions appear in the recent analysis of CNVs in the 1000 Genomes individuals (Sudmant et al. 2015).

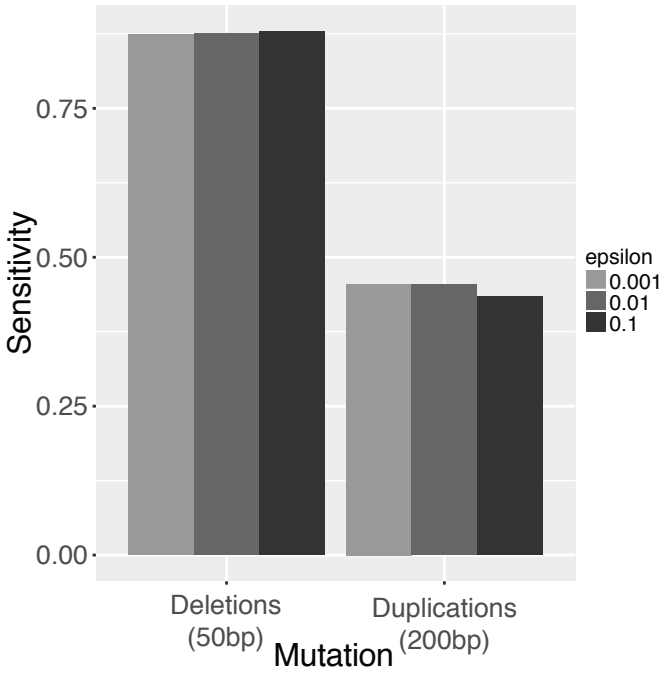


Figure S12: **Limited sensitivity to changes in epsilon.** The transducer method has a parameter, epsilon, which is used as a pseudocount when estimating the probability of mismapping a read. In the Methods section of the main text we set this parameter to 0.01 so that the genome-wide mismapping rate will be equal to that which is seen in simulations. However, the performance of the method is not heavily influenced by perturbing this value. An increase or decrease by an order of magnitude, while holding all other parameters constant, did not have a large effect on the ability of the method to detect 50bp deletions or 200bp duplications. The largest effect was a ten fold increase in epsilon leading to a 0.044 fold decrease in the ability of the method to detect 200bp deletions, which was not a significant change when tested with Fisher's exact test (uncorrected $p \approx 0.4$). We chose these mutation sizes since they represent places where the method is able to detect some, but not all, mutations, which should enable us to detect a decrease in performance. We do not believe these order of magnitude changes of epsilon had a large effect on the false positive rate either since running on a simulated data set with no copy number variation yielded no false positives.

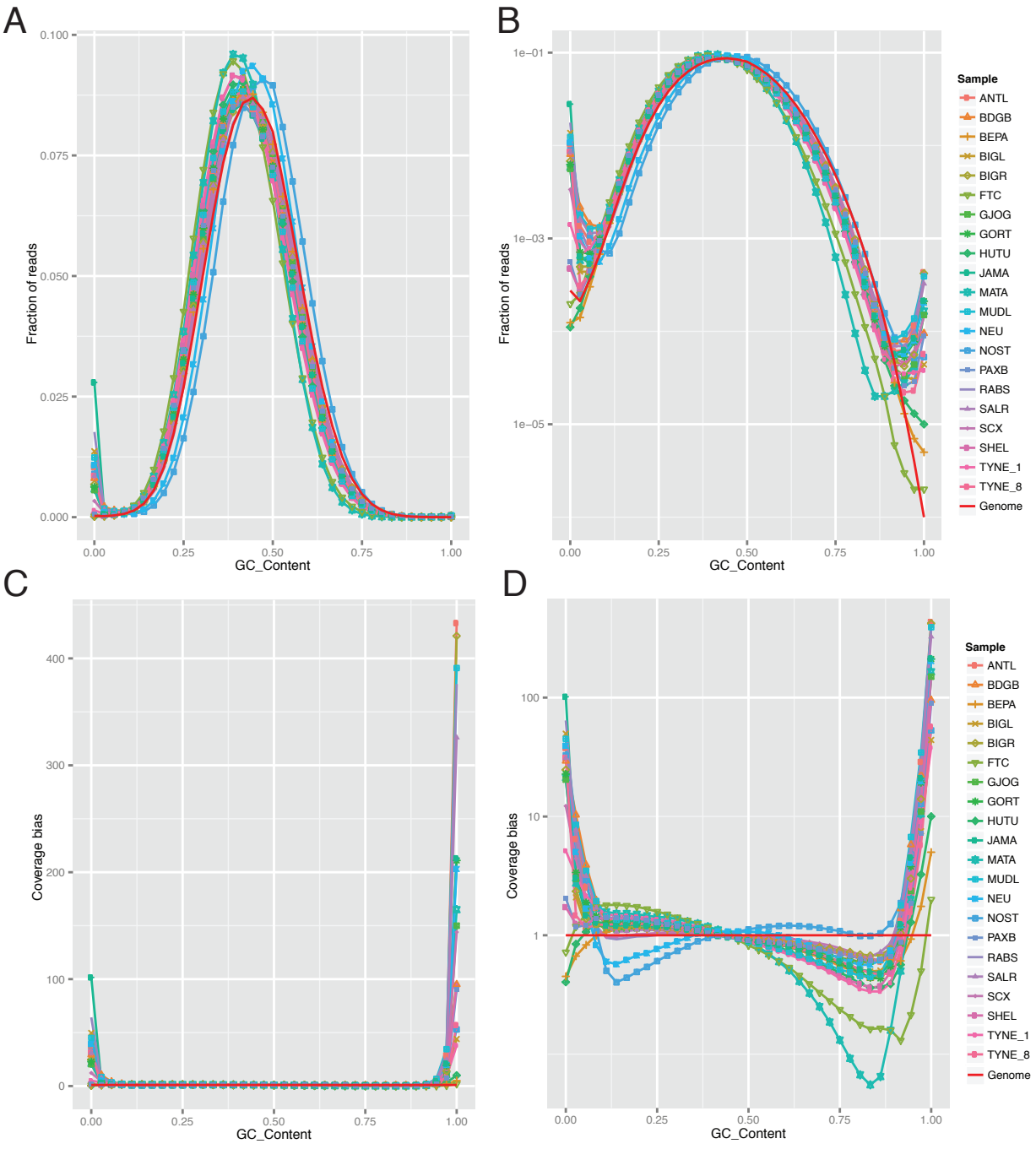


Figure S13: **Fraction of reads by GC content.** The fraction of reads for each bin of GC content is shown for all sequencing libraries as well as the stickleback assembly, which has an overall GC-content of 43% (red line). The data is visualized on both a (A) linear axis and (B) log axis to show differences at the GC extremes. To correct for GC content in our model, we calculate the coverage bias, which is the fraction of reads in a bin for a given sample divided by the fraction of k-mers in that bin for the assembly. This too is visualized on both a (C) linear and (D) log scale.

## S3 Supplemental Tables

Table S1: Parameters used when optimizing software

| Software | Parameter | Value |
|---|---|---|
| cnMOPS | window length | 125 |
| | prior impact | 0.1 |
| | minimum width | 1 |
| | pvalue threshold (Fisher's exact) | 0.01 |
| Genome STRiP | minimum deletion | 700 |
| | minimum duplication | 700 |
| | boundary precision | 100 |
| | tiling window size | 700 |
| | max reference gap length | 700 |
| | tiling window overlap | 350 |
| | minimum refined length | 350 |
| | pvalue threshold (Fisher's exact) | 0.0001 |
| CNVnator | window size | 280 |
| rSW-seq | threshold | {47, 375} |

We changed and optimized parameters of other software packages to compare performance at a similar false positive rate to our method.

Table S2: Primers used to test for the presence or absence of a deletion

| Region Name | Location | Nearby Gene | Primers |
|---|---|---|---|
| Derived marine deletion 1 | chrXII:10795677-10796335 | AVPR2 | ACGTCACCACCCTTTCTGAC<br>TTCCTGCCCTTATCATCACC |
| Derived marine deletion 2 | chrVII:16835340-16835668 | PCDHGC5 | TTCGCTACCTACTTCATATCAAAGG<br>TGATACTCTCCATGCCGTAGAA |
| Derived marine deletion 3 | chrXI:5490849-5491226 | CNTNAP1 | CCGTCTTTACCTGCACATCA<br>TGCCAGTGCAGATTATCCAG |
| Derived marine deletion 4 | chrXI:5847930-5848275 | KCNH4 | GGGAGGACAATTCTGAACCA<br>AAGGCCTTGGAGATGCTGTA |
| Derived marine deletion 5 | chrIV:24697371-24697776 | PSMC2 | CTCTCATGCCCTCCTCGAT<br>AGTGGCATCAGAGATGTGTCA |

We used these primers to interrogate the genomes of marine fish for the presence of the freshwater allele. The primers flank the deletion breakpoints, so alleles with the deletion produce a small band, and intact alleles produce a large band (Figure S5).

Table S3: Primers used to test for additional copies of a genomic segment

| Region Name | Location | Nearby Gene | Primers |
|---|---|---|---|
| Freshwater duplication 1 | chrXXI:7992834-7992991 | Similar to stonustoxin subunit beta | GGGCCTAATTGCCTTTCATT<br>CAATTGTCTGTATTTTGTTCAAGC |
| Freshwater duplication 2 | chrXI:15609248-15609432 | APOL4 | GCAGACCAGTAAAACGTCTACAAA<br>GGGATTGATTTTAGGGATCCTG |
| Freshwater duplication 3 | chrXI:15632100-15632296 | CAGNG2A | CCTTTTCTCCGACTCGACAG<br>AGGGAAAGGAAAGGAAACGA |
| Freshwater duplication 4 | chrI:2971330-2971513 | VWA5A | TTTGGCACAATCTAATGTGGT<br>ACTGGGGGATCAATACAAACA |
| Normalizer | chrVI:12802874-12803047 | EDA | GCCGTACTGCAAACCAAAA<br>ATCGTCAGCACCACTCAGC |

We used these primers in qPCR assays to assess the model's prediction that regions were of consistently greater copy number in the freshwater populations (Figure S5).

Table S4: Genome coordinates and additional information for CNVs that correlate with ecotype
Table available as attached electronic spreadsheet.

Table S5: Genes showing deletions overlapping their protein-coding exons, relative to reference genome
Table available as attached electronic spreadsheet.

Table S6: Genes showing duplications overlapping their protein-coding exons, relative to the reference genome
Table available as attached electronic spreadsheet.

Table S7: Functional enrichments for genes showing deletions overlapping their protein-coding exons, relative to the reference genome
Table available as attached electronic spreadsheet.

Table S8: Functional enrichments for genes showing duplications overlapping their protein-coding exons, relative to the reference genome
Table available as attached electronic spreadsheet.

Table S9: Functional enrichments for the genes closest to noncoding copy number variation
Table available as attached electronic spreadsheet.

Table S10: Primers used to amplify deleted region

| Name | Primer |
|---|---|
| DCHS1 fwd | TTACCTTTCCAGAATCCATGC |
| DCHS1 rev | GACGCTGCCATCTCCATTA |

We used these primers to clone a genomic region from an intron of DCHS1 into an expression vector (Figure 4).

Table S11: Genome coordinates of marine-freshwater copy number differences that overlap a region showing cross-species conservation.
Table available as attached electronic spreadsheet.

Table S12: Genome coordinates where freshwater populations completely remove a conserved sequence
Table available as attached electronic spreadsheet.

Table S13: Genome coordinates where marine populations completely remove a conserved sequence
Table available as attached electronic spreadsheet.

Table S14: Functional enrichments for genes closest to a marine deletion of conserved sequence
Table available as attached electronic spreadsheet.

Table S15: Frequency of the freshwater allele in a marine population

| Region Name | Nearby Gene | Samples Called | Marine Homozygotes | Heterozygotes | Freshwater Homozygotes | FW Allele Frequency |
|---|---|---|---|---|---|---|
| Derived marine deletion 1 | AVPR2 | 264 | 262 | 2 | 0 | 0.0038 |
| Derived marine deletion 2 | PCDHGC5 | 266 | 262 | 2 | 2 | 0.0113 |
| Derived marine deletion 3 | CNTNAP1 | 264 | 261 | 3 | 0 | 0.0057 |
| Derived marine deletion 4 | KCNH4 | 265 | 262 | 3 | 0 | 0.0057 |
| Derived marine deletion 5 | PSMC2 | 262 | 260 | 1 | 1 | 0.0057 |
| EDA | EDA | 644 | 635 | 8 | 1 | 0.0078 |
| ATP1A1 | ATP1A1 | 651 | 582 | 64 | 5 | 0.0568 |

We used primers flanking the deletion breakpoints to assay for the presence of the intact allele in hundreds of fish from Resurrection Bay, Alaska (Table S2 and Figure S5). We detected the ancestral freshwater allele at low frequency in this marine population, but not at more than 1%, which is significantly below the frequency seen for the ATP1A1 allele (Jones et al. 2012a); many of the frequencies were also below the frequency seen for the freshwater allele at the EDA locus (O'Brown et al. 2014).

Table S16: Primers used to amplify deletion alleles in humans

| Deletion Location (hg19) | Nearby Gene | Primers |
|---|---|---|
| chr14:107174392-107174928 | IGHV2-70 | ACGTCACCACCCTTTCTGAC<br>TTCCTGCCCTTATCATCACC |
| chr15:67404504-67405592 | SMAD3 | TTCGCTACCTACTTCATATCAAAGG<br>TGATACTCTCCATGCCGTAGAA |
| | | CCGTCTTTACCTGCACATCA<br>TGCCAGTGCAGATTATCCAG |
| | | CTCTCATGCCCTCCTCGAT<br>AGTGGCATCAGAGATGTGTCA |

We used these primers to amplify the deletion alleles from human DNA samples. We then Sanger sequenced the PCR product. The primer pair for the deletion on chr14 will give a larger band for the intact allele while the primer pairs for the deletion on chr15 will not amplify the intact allele.

Table S17: Regions not overlapping previous CNV annotations
Table available as attached electronic spreadsheet.

# References


Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M., 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**(6):974–984.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**(4):708–715.

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Villarreal, G., Dickson, M., Grimwood, J., Schmutz, J., Myers, R. M., Schluter, D., and Kingsley, D. M., *et al.*, 2005. Widespread parallel evolution in sticklebacks by repeated fixation of *Ectodysplasin* alleles. *Science*, **307**(5717):1928–1933.

DeFaveri, J. and Merila, J., 2015. Temporal stability of genetic variability and differentiation in the three-spined stickleback (*Gasterosteus aculeatus*). *PLoS ONE*, **10**(4):e0123891.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., *et al.*, 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**(5):491–498.

Frankham, R., 1995. Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.*, **66**:95–107.

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., and McCarroll, S. A., 2015. Large multiallelic copy number variations in humans. *Nat. Genet.*, **47**(3):296–303.

Harris, R. S., 2007. *Improved pairwise alignment of genomic DNA*. PhD thesis, The Pennsylvania State University.

Huang, W., Li, L., Myers, J. R., and Marth, G. T., 2012. ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**(4):593–594.

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., Absher, D. M., Myers, R. M., Reimchen, T. E., Deagle, B. E., *et al.*, 2012a. A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Curr. Biol.*, **22**(1):83–90.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., *et al.*, 2012b. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**(7392):55–61.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D., 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(20):11484–11489.

Kim, T. M., Luquette, L. J., Xi, R., and Park, P. J., 2010. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics*, **11**:432.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S., 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**(9):e69.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., *et al.*, 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, **488**(7412):471–475.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5):589–595.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K., and Haussler, D., 2011. Three periods of regulatory innovation during vertebrate evolution. *Science*, **333**(6045):1019–1024.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.*, 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**(9):1297–1303.

Nachman, M. W. and Crowell, S. L., 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**(1):297–304.

O'Brown, N. M., Summers, B. R., Jones, F. C., Brady, S. D., and Kingsley, D. M., 2014. A recurrent regulatory change underlying altered expression and Wnt response of the stickleback armor plates gene EDA. *Elife*, **4**:e05290.

Roesti, M., Moser, D., and Berner, D., 2013. Recombination in the threespine stickleback genome–patterns and consequences. *Mol. Ecol.*, **22**(11):3014–3027.

Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F., 2014. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, **30**(23):3427–3429.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.*, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**(8):1034–1050.

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe,

B. P., Baker, C., Nordenfelt, S., Bamshad, M., *et al.*, 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**(6253):aab3761.

Swain, M. J. and Ballard, D. H., 1991. Color indexing. *Int. J. Comput. Vision*, **7**(1):11–32.

Therkildsen, N. O., Nielsen, E. E., Swain, D. P., and Pedersen, J. S., 2010. Large effective population size and temporal genetic stability in Atlantic cod (*Gadus morhua*) in the southern Gulf of St. Lawrence. *Can. J. Fish. Aquat. Sci.*, **67**:1585–1595.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., *et al.*, 2012. The accessible chromatin landscape of the human genome. *Nature*, **489**(7414):75–82.

Tsai, S., Hollenbeck, S. T., Ryer, E. J., Edlin, R., Yamanouchi, D., Kundi, R., Wang, C., Liu, B., and Kent, K. C., 2009. TGF-beta through Smad3 signaling stimulates vascular smooth muscle cell proliferation and neointimal formation. *Am. J. Physiol. Heart Circ. Physiol.*, **297**(2):H540–549.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., *et al.*, 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, **11**(1110):1–11.