Hi-reComb: constructing recombination maps from bulk gamete Hi-C sequencing

Authors:
Milan Malinsky[1,2], Marion Talbi[1,2], Chenxi Zhou[3], Nicholas Maurer[4,5], Samuel Sacco[4,5], Beth Shapiro[4,5], Catherine L. Peichel[1], Ole Seehausen[1,2], Walter Salzburger[6], Jesse N. Weber[7], Daniel I. Bolnick[8], Richard E. Green[4,5] and Richard Durbin[3]

Affiliations:
[1] Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland
[2] Department of Fish Ecology and Evolution, EAWAG, 6047 Kastanienbaum, Switzerland
[3] Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK
[4] Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA
[5] UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA
[6] Department of Environmental Sciences, Zoological Institute, University of Basel, 4051 Basel, Switzerland
[7] Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI 53706, USA
[8] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA

Corresponding author:
Milan Malinsky: millanek@gmail.com

Summary:

Recombination is central to genetics and to evolution of sexually reproducing organisms. However, obtaining accurate estimates of recombination rates, and of how they vary along chromosomes, continues to be challenging. To advance our ability to estimate recombination rates, we present `Hi-reComb`, a new method and software for estimation of recombination maps from bulk gamete chromosome conformation capture sequencing (Hi-C). Simulations show that `Hi-reComb` produces robust, accurate recombination landscapes. With empirical data from sperm of five fish species we show the advantages of this approach, including joint assessment of recombination maps and large structural variants, map comparisons using bootstrap, and workflows with trio phasing vs. Hi-C phasing. With off-the-shelf library construction and a straightforward rapid workflow, our approach will facilitate routine recombination landscape estimation for a broad range of studies and model organisms in genetics and evolutionary biology. `Hi-reComb` is open-source and freely available at https://github.com/millanek/Hi-reComb.

Keywords (three to ten keywords that represent the content of the article):
Recombination map, genetics, Hi-C, gametes, sperm, software, cichlids, stickleback

Introduction

Meiotic recombination is a hallmark of sexual reproduction. While mutations give rise to new genetic variants, recombination shuffles them to generate new haplotypes – that is, chromosomes with novel combinations of existing alleles which selection can act on. Therefore, recombination impacts key evolutionary processes such as adaptation and speciation, and it shapes the distribution of genetic variation along the genomes. Because organismal traits are usually influenced by many genetic variants (Shi et al. 2016; Boyle et al. 2017; Barton 2022), often with non-linear epistatic interactions among them (Phillips 2008; Domingo et al. 2019; Johnson et al. 2023), and because adaptation and speciation commonly require co-evolution of a whole suite of traits (Phillips and Arnold 1989; White and Butlin 2021), evolution is increasingly seen as multidimensional and combinatorial (Marques et al. 2019; Barton 2022), with recombination in a central role. Yet, routine and accurate genome-wide reconstruction of recombination maps remains a major challenge, and there is a need for novel approaches and methods for recombination inference.

Recombination rates vary by several orders of magnitude along the genome (Coop and Przeworski 2007; Stapley et al. 2017; Halldorsson et al. 2019). Accurate inference and representation of this variation requires ascertaining the chromosomal locations of many crossover events. Depending on how this is done, recombination rate inference methods can be divided into three catogories (Peñalba and Wolf 2020): (i) genetic linkage maps based on the transmission of polymorphic markers in crosses and pedigrees; (ii) population genetic approaches based on patterns of linkage disequilibrium (LD); and (iii) gamete sequencing, based on finding breakpoints in gamete haplotypes compared to the donor genome. The recombination maps inferred by the three approaches differ in several important respects, perhaps most notably in resolution, the measured time interval, and in how they are impacted by selection (Peñalba and Wolf 2020). For example, LD-based maps typically integrate over

thousands of generations of recombination, deliver high resolution, but are influenced by selection occuring over that timeframe. In contrast, studies using crosses and pedigrees reflect recombination over a small number of generations and rarely include the thousands of samples needed to achieve resolution comparable to LD-based maps [but see (Morgan et al. 2017; Halldorsson et al. 2019) for high resolution maps in humans and mice]. Approaches based on sequencing of sperm, egg, or pollen reveal all recombination during successful gametogenesis, and thus deliver a direct snapshot, not affected by selection acting at other life stages. The resolution of these gamete-based maps is largely determined by the number of gametes that can be typed.

Single sperm whole-genome sequencing has been used for well over a decade (Lu et al. 2012; Wang et al. 2012) and continues to deliver insights into the factors that influence meiotic recombination (Hinch et al. 2019; Yang et al. 2022). However, practical limitations have restricted studies to typing at most a few hundred sperm cells. One exception is the Sperm-seq (Bell et al. 2020) protocol – a sperm specific variation of high-thoughput single cell sequencing (Macosko et al. 2015) – which scaled to over 30,000 sperm cells. Nevertheless, the adoption of this approach beyond the original study appears to be limited, likely due to the custom and relatively complex laboratory procedure that is required. Bulk sequencing approaches avoid the need for the isolation of individual cells. Three studies introduced linked-read sequencing of bulk sperm and pollen (Dréau et al. 2019; Sun et al. 2019; Xu et al. 2019), demostrating that single cell sequencing of individual gametes is not necessary to infer crossover events. These studies use read-linkage information to find recombination breakpoints by comparing individually barcoded gamete DNA fragments against the haplotypes of the donor individual.

Here we present `Hi-reComb`, a new method and software for estimating individual recombination maps from bulk gamete chromosome conformation capture (Hi-C) sequencing

data (Lieberman-Aiden et al. 2009; Rao et al. 2014; Oksuz et al. 2021), using a standard Hi-C library preparation protocol. By taking advantage of the long insert sizes delivered by Hi-C, this approach can achieve substantially greater effective coverage by crossover-informative read pairs than has been possible with comparable amount of linked-read sequencing. Moreover, Hi-C data provide information regarding large scale structural variation in the donor individual, allowing for simultaneous improvements of the reference genome and better interpretation of the genetic map constructed from this data. First, we show the accuracy of `Hi-reComb` by reconstructing genetic maps from simulated data. Next, to demonstrate the real-world utility of this approach, we constructed and sequenced several sperm Hi-C libraries, used the data to scaffold two reference genomes, and then inferred and evaluated several recombination maps for cichlid and stickleback fish. We show that the maps correspond well to LD-based maps and that the effects of donor haplotype phasing errors are limited. This approach is applicable to any species / individuals that produce at least hundreds of thousands of gametes. Many such species exist; therefore, `Hi-reComb` will be of utility to a broad range of researchers interested in genetics and recombination.

Materials and Methods

The Hi-reComb approach

Recombination map reconstruction using `Hi-reComb` starts with preparing a standard commercially available Hi-C library. We choose a library based on endonuclease digestion to provide a relatively uniform sequencing coverage (**Figure S1**). After sequencing and alignment to a reference genome, `Hi-reComb` detects meiotic crossovers by comparison of Hi-C fragments against donor haplotypes (**Figure 1**). Because the two reads from each pair of Hi-C fragments originate from the same haploid gamete, cases where the two fragments match different donor haplotypes indicate that a crossover took place between them. We denote this event as $X = C$. Conversely, if the two fragments match the same donor haplotype,

this suggests the absence of any crossover between them, denoted $X = N$. The match of Hi-C fragments to either of the two donor haplotypes is determined by sites at which the donor is heterozygous. Therefore, if the Hi-C fragments do not cover at least one heterozygous site each, such a read pair is not informative with regards to crossovers.

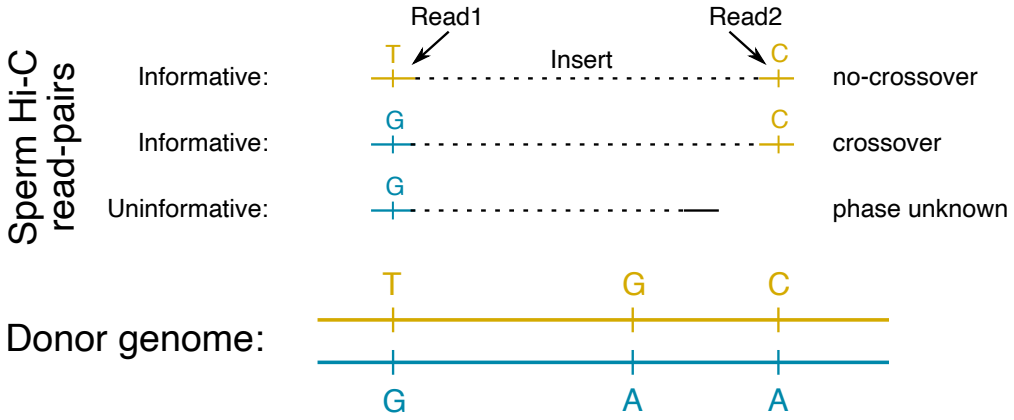

Figure 1: The principle of detecting crossovers with `Hi-reComb`. The donor genome is shown with heterozygous sites separating the two haplotypes indicated. For each sperm Hi-C read pair, if both fragments match the same donor haplotype, this indicates an absence of crossover (no-crossover). If the fragments match different donor haplotypes, this indicates a crossover between them. Finally, if both fragments do not cover haterozygous sites in the donor genome, the read pair is not informative with regards to recombination.

`Hi-reComb` estimates crossover likelihoods for each informative read pair i considering potential base-calling errors (base quality scores) and phasing errors (phase quality scores). Let c_i be a parameter indicating if a crossover did ($c_i = 1$) or did not ($c_i = 0$) take place between a pair of reads; thus $\theta = \{c_i | c_i \in \{0,1\}\}$. We then define a crossover likelihood $l_{ci} = P(X = C \mid c_i = 1)$ and non-crossover likelihood $l_{ni} = P(X = C \mid c_i = 0)$. While these likelihoods account for base calling and phasing, other sources of error, e.g. arising in variant calling/filtering or from non-crossover gene conversion events, are unaccounted for. We found that false positive crossover read pairs – i.e. overestimating l_{ci} and underestimating l_{ni} – can have substantial impact on the inferred recombination landscapes, creating false 'spikes' of recombination. This is true especially for read pairs with short inserts; that is, short genomic distances between the Hi-C fragments.

While false positives are equally likely across the insert length distribution, the proportion of true positives increases linearly with insert length: on average, there are 1,000x more true positives for read pairs with insert size of 1Mb compared with read pairs with insert size 1kb. This is because the further away two loci are from each other along the chromosome (the greater the insert length), the higher the probability that a crossover really occurred between these two loci (i.e., the greater the true positive rate). To estimate the overall false positive rate and to apply a correction to the likelihoods, we first use likelihoods of long-insert read pairs (default: > 1Mb), where the ratio of true to false positives is highly favorable, to estimate r_l^*, the average per-bp crossover rate for each chromosome. Then we calculate a correction factor by evaluating the crossover likelihoods of short-insert read pairs (default: < 1kb) against this baseline. Formally, the correction factor f is defined as $f = \frac{\sum_i l_{ci} - \sum_i d_i * r_l^*}{n_s}$ where i sums over all n_s short-insert informative read pairs with $X = C$ and where d_i is the insert size of the read pair i.

To calculate crossover probabilities, we incorporate the corrected likelihoods together with insert-length dependent priors in a Bayesian framework. We assign priors to read pairs under the assumption that the recombination rate is uniform along each chromosome. Therefore, the longer the insert length d_i the higher is the prior for each informative read pair i. The prior probability of crossover p_{ci} for each informative read pair i is then simply the product of r_l^* and of the insert length:

$$p_{ci} = r_l^* * d_i$$

Finally, the posterior crossover probability for each informative read pair i where $X = C$ is:

$$\mathbb{P}(c_i) = \frac{(l_{ci} - f) * p_{ci}}{(l_{ci} - f) * p_{ci} + (l_{ni} + f) * (1 - p_{ci})}$$

For read pairs that do not indicate crossovers, we set $l_{ci} = P(X = N|c_i = 1) = 0$ and $l_{ni} = P(X = N|c_i = 0) = 1$. This is highly 'conservative' in terms of avoiding any risk of introducing

additional false positives. While it results in false negatives – i.e. underestimating l_{ci} and overestimating l_{ni} – this approach does not affect the shape of the estimated recombination landscapes because these false negatives are randomly distributed along the chromosomes.

To reconstruct the recombination map, we initially divide the crossover probability of each Hi-C read pair uniformly along its insert length. That is, we assign the same probability to all basepairs between the informative sites that determine the crossover. Then, we construct an initial recombination map by dividing the sum of crossover and non-crossover probabilities at each bp of the chromosome. While this initial map reflects crossover probabilities of individual read pairs, our goal is to produce a recombination map that integrates probabilities over all read pairs. To achieve this, `Hi-reComb` employs an Expectation-Maximization (EM) procedure (Dempster et al. 1977), which is very similar to the EM procedure employed by (Halldorsson et al. 2019). For a detailed description of the initial map construction and of our EM procedure see **Supplementary Note 1**.

The core algorithm is supported by several practical heuristics that we found improve its overall performance. The first heuristic concerns an 'edge effect' whereby the *effective coverage* – i.e., the number of informative read pairs that span a given genomic interval – drops towards chromosome ends (**Figure S1**). Reduced effective coverage leads to increased sampling noise, which propagates through the EM algorithm and can affect genomic regions far beyond the chromosome edges. To alleviate this problem, we have introduced a minimum effective coverage cutoff around chromosome edges, whereby the genomic intervals beyond the cutoff limits are excluded from the EM procedure and from the resulting recombination maps. Second, if unexpectedly many read pairs are all bounded by the same SNP and/or if unexpectedly many of these read pairs indicate crossovers (i.e., $X = C$), we eliminate these read pairs and SNPs from consideration before starting the recombination map reconstruction. This reduces errors arising from incorrect variant calling and/or phasing. Finally, the third

heuristic involves adjusting crossover likelihoods for long-insert read pairs (default >1Mb) to consider the probability of double crossovers. In this calculation, the probability p_d of two crossovers is given by a Poisson mass function: $p_d = \lambda^2 * e^{-\lambda}/2$, where $\lambda = r_l^*$.

Implementation

The `Hi-reComb` package is efficiently coded in C++, does not have any external dependencies, and is straightforward to install, compile and use. It is open source and freely available from https://github.com/millanek/Hi-reComb. `Hi-reComb` currently contains two core modules and two additional utilities. The first core module, `FindInfoPairs` processes aligned Hi-C read pairs to find the pairs that are informative with regards to crossovers (**Figure 1**). It takes as an input a set of phased heterozygous sites in the HapCUT2 format (Edge et al. 2017) and Hi-C read pairs in SAM / BAM format (Danecek et al. 2021) and outputs only pairs of Hi-C reads that cover at least one phased heterozygous site each. The second core module, `RecombMap` performs recombination map inference as described above, with run times of the order of minutes for each chromosome. In addition, `RecombMap` provides a bootstrap option to evaluate uncertainty, whereby informative Hi-C read-pairs are resampled with replacement to estimate additional genetic maps. This option also allows taking an average across the bootstraps, which tends to result in smoother maps than a single run estimate.

The `Hi-reComb` package also includes a tool to simulate informative Hi-C read pairs reflecting a known recombination map. This `Simulate` utility enables users to evaluate the accuracy of recombination map inference for a given effective coverage, error rate, insert-size distribution, and map profile. The utility matches the insert size distribution of the simulated pairs to real Hi-C read pair dataset provided as an input. Each simulated read pair is placed randomly onto a chromosome, and its crossover status (i.e., $X = C$ or $X = N$) is determined by the centimorgan (cM) distance from the input recombination map with an error rate (both false

positive and false negative) determined by the `--errorRate` parameter. Read pairs are simulated until reaching a target effective coverage, which is specified by the `--targetCoverage` parameter. The simulated reads are used to reconstruct a recombination map as in the `RecombMap` module. The accuracy of map reconstruction for the given parameters can then be evaluated by comparison with the input map. This procedure can be repeated multiple times using the `--replicates` parameter.

Because the vast majority of read pairs do not have a crossover between them, it is possible to use the sperm Hi-C reads to determine donor haplotype phasing, for example using HapCUT2 (Edge et al. 2017). In this case, no other data is required. This approach works well in practice, in part because of the robustness of `Hi-reComb` to errors, including phasing errors, as we demonstrate in the Results section below. However, more accurate phasing can be obtained, for example with Hi-C data from a somatic tissue or by using mother-father-offspring trio data. For the latter purpose `Hi-reComb` provides the `TrioPhase` utility, which takes as input a VCF file specifying heterozygous sites in the donor and another VCF file with his parents who have been genotyped at the same loci. The `TrioPhase` utility outputs phased haplotypes in the same format used by the core modules of `Hi-reComb`.

Sperm Hi-C data

We obtained sperm Hi-C data from one individual each of cichlid fish species *Aulonocara stuartgranti* and *Astatotilapia calliptera* (from Lake Malawi), *Neolamprologus multifasciatus* (from Lake Tanganyika), *Astatotilapia nubila* (from the swamps of Lake Victoria), and threespine stickleback *Gasterosteus aculeatus* (from Walby Lake in Alaska), all obtained from laboratory aquarium stocks. A single sexually mature male was sacrificed from each species and his freshly harvested testes were flash-frozen in liquid nitrogen. Upon thawing at room temperature, we cut open the testes and suspended the sperm in $\sim 100 \mu l$ of TE buffer. Sperm cells were counted with hemocytometer, aiming for between 100k and 1 million cells. This was

then used as input into the standard Dovetail Omni-C library preparation, following the standard protocol. All libraries were sequenced on the Illumina Novaseq 6000 instrument, obtaining paired end 2x150bp reads.

Reference genomes and Hi-C scaffolding

We first used to Hi-C reads from *A. calliptera* and from *N. multifasciatus* to produce accurate reference genome assemblies with chromosome-scale scaffolds. For *A. calliptera*, we used as a starting point the fAstCal1.2 genome (GenBank: GCA_900246225.3). This genome assembly was already chromosome-scale, with scaffolding using 10x Genomics Chromium linked reads, BioNano Irys optical maps and two low-resolution genetic maps (Quin et al. 2013; Albertson et al. 2014). However, we found about a hundred disagreements between that assembly and the Hi-C contact map. These discrepancies were corrected manually using the PretextView software (https://github.com/sanger-tol/PretextView). The new manually curated genome was used as a reference for recombination analyses and was deposited as fAstCal1.5 under GenBank accession GCA_900246225.6. For *N. multifasciatus*, we used as a starting point the fNeoMul1.1 genome (GCA_963576455.1), which was not chromosome-scale but was fragmented in 378 large contigs. To produce chromosome scale scaffolds for this species, we used the YaHS Hi-C scaffolding tool with default parameters (Zhou et al. 2022). The new scaffolded genome fNeoMul1.2 (GCA_963576455.2) was then used for recombination analyses. Finally, for *G. aculeatus* we used the stickleback v5 reference under GCA_016920845.1.

Alignment and Hi-C contact maps

Hi-C reads were mapped to the reference genomes using `bwa mem` v 0.7.17 (Li 2013) with the `-5SP` and `-T0` options. To generate Hi-C contact maps, we used the `pairtools` (Open2C et al. 2024) software (v. 1.1.0), using the `--min-mapq 30`, `--walks-policy 5unique`, `--max-inter-align-gap 30`, and `--chroms-path` options for the `parse`

command, and using the `dedup` command with default parameters. The 'pairs' files were then used as input into the `juicer_tools` (v 1.22.01) `pre` command from the Juicebox package (Durand et al. 2016) which was also used for contact map visualization.

Variant calling, filtering, phasing

To remove duplicates for all purposes other than the Hi-C contact maps above, we used the `MarkDuplicates` command from the `picard` package (v 2.26.6) with the option `REMOVE_DUPLICATES=true`. Variant calling was done separately for each individual with bcftools v.1.16 (Danecek et al. 2021) using `mpileup --count-orphans -Ou` output piped into the `call` program with `-mv -Oz` options. For variant filtering we used a mappability mask, whereby we broke down the genome into overlapping k-mers of 150bp (matching the read length), mapped these k-mers back to the genome, and masked all sites where fewer than 90% of k-mers mapped back to their original location perfectly and uniquely. Next, we filtered variants based on sequencing depth, with limits based on examining the coverage histogram for each sample: removing variants with depth ≤ 12 and ≥ 75 for *A. calliptera*, ≤ 35 and ≥ 140 for *A. stuartgranti*, ≤ 20 and ≥ 120 for *N. multifasciatus*, ≤ 35 and ≥ 120 for *A. nubila*, and ≤ 8 and ≥ 100 for *G. aculeatus*. Finally, we applied the following hard filters, removing variants where any of these applied: `%QUAL<20`, `MQ<40`, `MQ0F>0.4`, `RPBZ<-5.0`, or `RPBZ > 5.0`. After variant filtering, we kept only biallelic SNPs with heterozygous genotypes. To estimate the allelic phase of these SNPs, we used `hapcut2` (v 1.3.4). After dividing the VCF files and the alignment bam files per chromosome, for each chromosome we ran first the `extractHAIRS` and then the `hapcut2` commands, both with the `--hic 1` option.

Stickleback trio-based phasing – data and processing

The *G. aculeatus* sperm donor individual originated from an aquarium F1 cross from wild-caught parents. We sequenced the DNA of his parents to trio-phase the heterozygous SNPs in this individual using Mendelian inheritance logic (trio phasing). The sequencing was done

on the Illumina Novaseq 6000 instrument with 2x150bp reads as a part of a larger DNA sequencing of 24 individuals, each at approximately 25x coverage. For processing this dataset we used alignment with `bwa mem` v 0.7.17 (Li 2013) with default options, followed by the `MarkDuplicates` command from the `picard` package (v 2.26.6) with default options. Then we used GATK v 4.2.3 (DePristo et al. 2011) to call variants, using `HaplotypeCaller` in GVCF mode for each individual separately followed by joint genotyping using `GenotypeGVCFs` with the `--include-non-variant-sites` option. For variant filtering, we generated a callability mask to identify and filter out regions where we were unable to confidently call variants. This included: (i) sites determined by overall read depth cutoffs based on examining a depth histogram (≤ 300 and ≥ 700 on autosomes; ≤ 200 and ≥ 600 on the X chromosome), (ii) sites where > 6 individuals had missing genotypes, (iii) sites identified by GATK as low quality (with the `LowQual` tag) and (iv) sites with poor mappability. The mappability mask was determined in the same way as for the Hi-C analyses above: we broke down the genome into overlapping k-mers of 150bp (matching the read length), mapped these k-mers back to the genome, and masked all sites where fewer than 90% of k-mers mapped back to their original location perfectly and uniquely. Finally, we used several hard filters based on GATK best practices, specifically focusing on overall genotype quality (`QUAL<20`), mapping quality (`MQ<40`), mapping strand bias (`FS>40`), variant quality normalized by depth (`QD<2`) and excess heterozygosity when compared with Hardy-Weinberg equilibrium (`ExcessHet>40`).

Stickleback trio-based phasing – Hi-reComb trio phasing module

We used the utility `TrioPhase` from `Hi-reComb` to estimate the haplotype phase of the *G. aculeatus* sperm donor individual using mother-father-offspring trio genotype calls and simple Mendelian logic. For each heterozygous SNP in the offspring, the utility first checks that both alleles are present in the parents at the same site. If both parents are heterozygous then the SNP cannot be phased because the parental genotypes are not informative. In all other cases,

the allele inherited from the first parent is added to haplotype 1 and the allele inherited from the second parent is added to haplotype 2. The phased haplotypes are output as a single block in the `hapcut2` format.

Stickleback LD-based map

We used a dataset of 334 Alaskan stickleback individuals sequenced to mean 19.6x coverage (min: 13.4x; max: 25.8x) on the Illumina Novaseq 6000 instrument with 2x150bp reads. Alignment, variant calling, and variant filtering was done in the same way as for the trio-based phasing dataset (see above), except that the overall depth filter was set at ≤ 4000 and ≥ 8000 on autosomes; ≤ 3000 and ≥ 7000 on the X chromosome and the maximum number of missing genotypes to 66 (i.e. $\leq 20\%$ missingness).

This dataset included 23 individuals from Walby Lake, the same population as the sperm donor. To estimate changes in effective population size (N_e) through time, we used smc++ v.1.15.4 (Terhorst et al. 2017), with the commands: vcf2smc -> estimate. Then we used the pyrho (Spence and Song 2019) software to infer recombination rates along the genome based on patterns of LD. To build likelihood tables for pairs of biallelic sites, we used the `make_table` command with demographic history as inferred by smc++, and the Moran approximation specified by the `--approx` and `--moran_pop_size N` flags where N equals 1.5× the number of haplotypes. This was followed by the pyrho `optimize` command to infer the recombination maps with a window size of 50 SNPs and block penalty of 15.

Results

Sperm Hi-C datasets, contact maps, and genome scaffolding

Key characteristics of the sperm Hi-C datasets presented in this manuscript are summarized in **Table 1**. A statistic that has a crucial effect on recombination inference is the *effective coverage*. Effective coverage reflects the total length of DNA segments that can be assessed

for presence of crossovers and is determined not only by the depth and quality of the Hi-C library but is also substantially influenced by the heterozygosity of the donor individual. The greater the heterozygosity, the greater is the chance that each of the two fragments of a Hi-C read pair covers a heterozygous site and thus is informative, as illustrated in **Figure 1**. This effect is clearly seen in the *G. aculeatus* dataset, where the Hi-C library was of relatively low quality and contained only ~2.6 million read pairs mapping to the same chromosome with >1kb insert, an order of magnitude lower than all the other samples, likely due to the much lower amount of sperm cells used as a starting material. Despite this, the effective coverage for *G. aculeatus* is comparable to the other datasets.

Table 1: Sperm Hi-C data overview.

Species	Raw coverage	Cis >1kb reads pairs (millions)	Heterozygous sites per kb	Effective coverage (chr 1)
A. calliptera	92x	18.7	1.16	292x
A. nubila	140x	48.2	1.55	3,587x
A. stuartgranti	185x	60.1	0.94	1,253x
G. aculeatus	270x	2.6	2.61	655x
N. multifasciatus	165x	40.4	1.55	1,369x

The use of the Hi-C data for scaffolding of the *N. multifasciatus* genome resulted in 863.6Mb (98.2%) of sequence being assigned to 22 chromosomes, while the remaining 67 unplaced scaffolds comprise 16Mb of sequence. The *N. multifasciatus* Hi-C contact map mapped to this new fNeoMul1.2 assembly (GCA_963576455.2) is shown in **Figure S2**. The use of the Hi-C data for scaffolding of the *A. calliptera* genome resulted in 863.0Mb (98.3%) of sequence being assigned to 22 chromosomes, while the remaining 122 unplaced scaffolds comprise 14.7Mb of sequence. The *A. calliptera* Hi-C contact map mapped to this new fAstCal1.5 assembly (GCA_900246225.6) is shown in **Figure S3A** and examples of how disagreements between the Hi-C contact map and the previous version of the assembly were resolved are shown in **Figure S3B**. For *A. calliptera*, the chromosome count is as expected, based on known karyotypes of Lake Malawi cichlids (Poletto et al. 2010; Conte et al. 2019), and matches the previous assembly. On the other hand, karyotypes of other species of the cichlid tribe

Lamprologini, to which *N. multifasciatus* belongs, showed only 21 chromosomes (Ozouf-Costaz et al. 2017); therefore, our results reveal previously unknown chromosome number polymorphism in this tribe.

Hi-C contact maps provide information regarding large-scale structural variation present in the donor individual. Perhaps the most prominent of these is an inversion in *A. stuartgranti* with respect to the *A. calliptera* reference, located in the middle of chromosome 2 (~ 12.7Mb – 16.7Mb), corresponding to the 'small' inversion previously reported by (Blumer et al. 2024). Notably, we found that this inversion is surrounded by an extended region of very low recombination in *A. stuartgranti* (**Figure S4**), illustrating how sperm-based Hi-C contact maps and recombination maps can be used together to better understand the interaction between structural variation and recombination.

Hi-reComb recombination inference from simulations

To evaluate the accuracy of recombination map inference with `Hi-reComb`, we used the `Simulate` utility and explored how the performance is influenced by key parameters: the error rate and the effective coverage. For each run, we supplied the `Hi-reComb Simulate` utility with a reference map, simulated ten replicate Hi-C datasets from this map, and then ran recombination map inference for each replicate.

Figure 2A shows an example of ten replicate maps reconstructed from simulations with 1% error rate and 3,000x effective coverage, with chr 2 of *A. stuartgranti* as a reference map. With these parameters, the reconstructed maps showed correlation with the truth of between 0.75 and 0.97 (orange line in **Figure 2B**), depending on the resolution at which we measured the correlations. In subsequent runs, we varied the effective coverage between 500x and 3,000x, matching approximately the range of coverage found in our empirical datasets (**Table 1**), while keeping the error rate at 1%. The results, shown in **Figure 2B**, revealed that decreasing the

effective coverage reduces the accuracy of inferred maps, as expected. However, even at the relatively low effective coverage of 500x, the reconstructed maps show substantial positive correlation with the truth – on average 0.53 at 2kb scale and 0.90 at 5Mb scale.

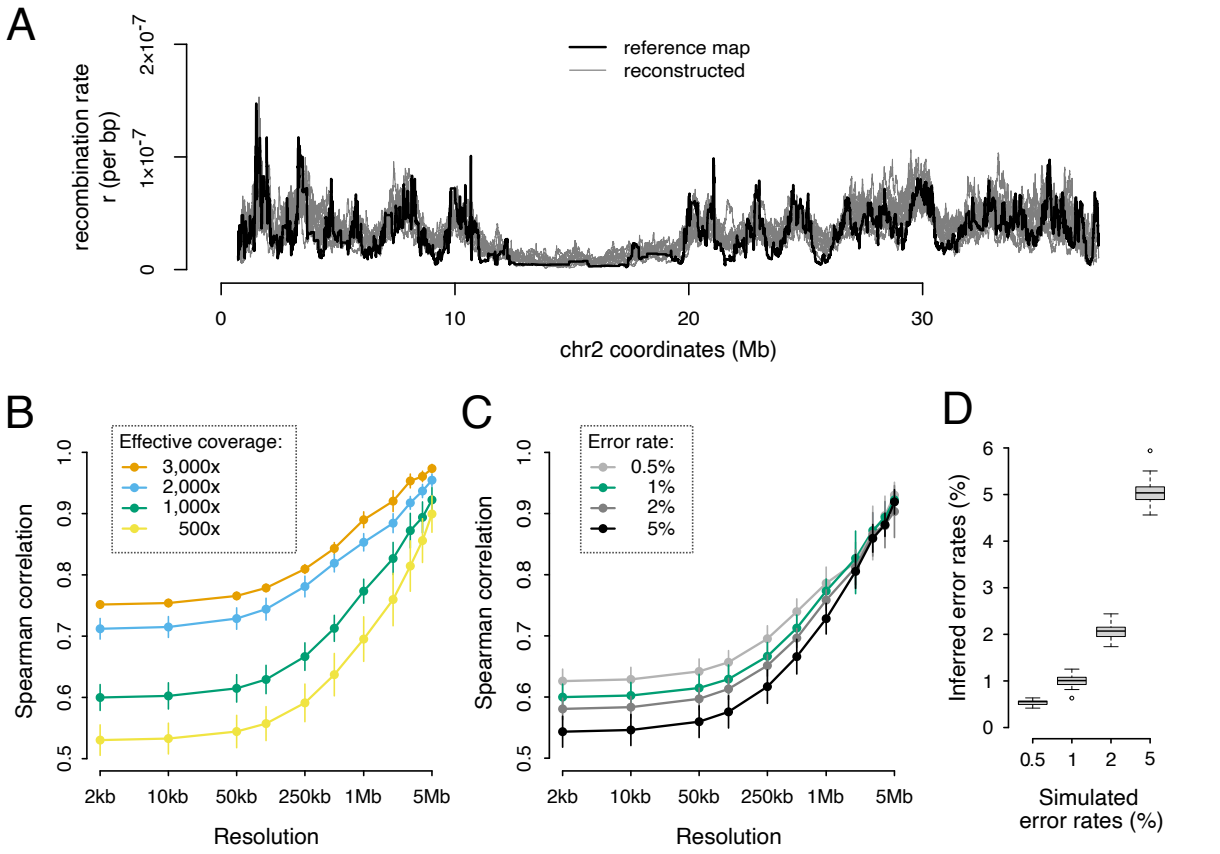

Figure 2: Hi-reComb recombination map inference from simulated data. (**A**) A comparison of a reference map against ten maps reconstructed by `Hi-reComb` from simulated data at 3,000x effective coverage and 1% error rate. (**B**) The dependence of accuracy of recombination map reconstruction on the effective coverage. (**C**) The dependence of accuracy of recombination map reconstruction on the error rate. (**D**) The accuracy of error rate estimation by `Hi-reComb`.

Next, we explored the impact of the error rate on recombination map inference. We fixed the coverage at 1,000x and varied the error rate from 0.5% to 5%. We found that at the high error rate of 5%, `Hi-reComb` still infers maps with highly positive correlation with the truth (0.54 at 2kb; 0.92 at 5Mb), which is only a fraction lower than at the 0.5% error rate where the correlations are 0.63 at 2kb and 0.93 at 5Mb scale (**Figure 2C**). The resilience to a relatively high degree of error is partly due to the ability of `Hi-reComb` to accurately infer the error rate

from the data: at 1,000x effective coverage, all inference runs reported error rate estimates within a very narrow range of the truth (**Figure 2D**).

Empirical recombination maps

We used `Hi-reComb` to infer genetic maps for the five datasets described in **Table 1**. Error rates estimated from these empirical datasets were within the range where simulations demonstrated reliable performance, with means between 0.8 and 1.9% in cichlids and 0.6% in stickleback (**Figure 3A**). The highest error rates of over 4% were found in cichlids on chromosome 3, which contains by far the most highly repetitive sequence where variant calling is difficult (**Figure S5**). This, along with the lower error rate found in sticklebacks, which have a genome with a much lower proportion of repetitive elements than cichlids (**Figure S5**), shows the extent to which error rates are affected by miscalled SNPs in repetitive regions of the genome. Among the cichlids, we see a link between error rates and effective coverage, with the low-coverage *A. calliptera* having the highest error rate, followed by the medium coverage *A. stuartgranti* and *N. multifasciatus*, and the lowest error rate among cichlids is in the high-coverage *A. nubila*.

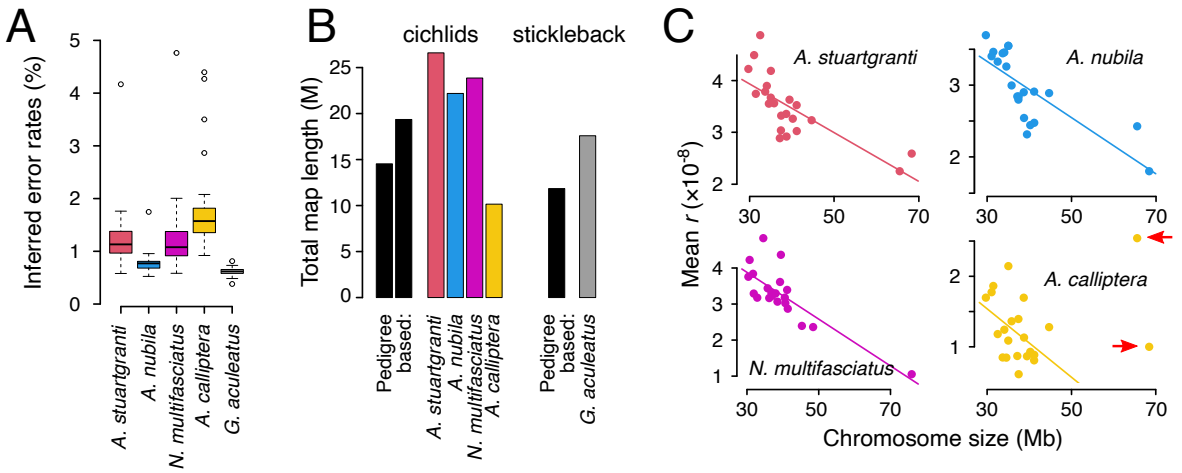

Figure 3: An overview of Hi-C inferred maps. (**A**) The error rates estimated by `Hi-reComb` from empirical data. (**B**) Recombination map lengths in Morgans (M) compared against three previously published pedigree-based maps (for cichlids, left: Albertson *et al. (Albertson et al. 2014)* ; right: O'Quin *et al. (Quin et al. 2013)*; for stickleback: Roesti *et al. (Roesti et al. 2013)*. (**C**) The negative relationship between the chromosome size and the mean recombination rate in cichlids. The two large chromosomes that were excluded in *A. calliptera* as outliers are highlighted with arrows.

The inferred cichlid recombination maps varied in length from 1,015 centimorgan (cM) for *A. calliptera* to 2,660 cM for *A. stuartgranti*, which falls both below and above two previously published pedigree-based maps for Lake Malawi cichlids that had lengths of 1,453 cM (Albertson et al. 2014) and 1,935 cM (Quin et al. 2013). For stickleback, the inferred map had a length of 1,758 cM which is somewhat above the 1,184 cM previously reported from a pedigree-based map (Roesti et al. 2013) (**Figure 3B**). However, we note that the absolute map lengths should be interpreted with caution for at least two reasons. First, differing levels of somatic cell 'contamination' across samples would affect the overall rate comparisons. While we tried to separate sperm from the surrounding tissues, examination under the microscope showed that our biological samples contained a small number of somatic cells, and we did not conduct further cell purification or sorting. Second, the edge effect where effective coverage decreases towards the ends of each chromosome (see **Methods**) can have different impact across individuals with different mean effective coverage.

It is known that in many species, there is a negative association between the chromosome length (in Mb) and the mean per-bp recombination rate (mean r)(Haenel et al. 2018; Brazier and Glémin 2022) . Consistent with these previous studies, our results also indicate a strong negative correlation in both cichlids and in stickleback. In cichlids (**Figure 3C**), the negative link was very clear across all chromosomes for *A. stuartgranti* ($r^2 = 0.55$; $p = 7.9 \times 10^{-5}$), *A. nubila* ($r^2 = 0.59$; $p = 3.1 \times 10^{-5}$), and *N. multifasciatus* ($r^2 = 0.64$; $p = 7.1 \times 10^{-6}$). In *A. calliptera*, the two large chromosomes (chr 3 and chr 7) were outliers and the negative association between chromosome length and the mean recombination rate was present only if these two chromosomes were excluded (complete dataset: $r^2 = 0.02$; $p = 0.53$; outliers excluded: $r^2 = 0.20$; $p = 0.049$). In stickleback, the negative association between the chromosome length and mean r was very strong in maps based on the trio phasing, which will be described below ($r^2 = 0.71$; $p = 3.0 \times 10^{-6}$; **Figure S6**).

To estimate uncertainty in reconstructed recombination landscapes, `Hi-reComb RecombMap` includes a bootstrap procedure whereby informative read pairs are resampled with replacement. As an example, **Figure 4A** shows a recombination landscape for chr 4 of *A. stuartgranti* with 95% confidence intervals (95% CIs) estimated based on 50 bootstrap replicate runs. Importantly, the bootstrap estimates facilitate comparisons among recombination landscapes. We define the areas where 95% CIs of two maps do not intersect as areas of significant recombination rate differences, or $\Delta(r)$ regions. To illustrate this functionality, **Figure 4B** shows comparisons between the maps of *A. stuartgranti* and *A. nubila*. Overall, we found that $\Delta(r)$ regions between these two species comprised 62.3 Mb of sequence, or 7.23% of the genome, with variation across chromosomes between 3.1% and 14.7%. To account for the fact that mean rates differ between the two maps, we also normalized the means before calculating the $\Delta(r)$ regions. On these mean-normalized maps, $\Delta(r)$ regions comprised 45.7 Mb of sequence, or 5.30% of the genome, with variation across chromosomes between 2.8% and 11.0%.

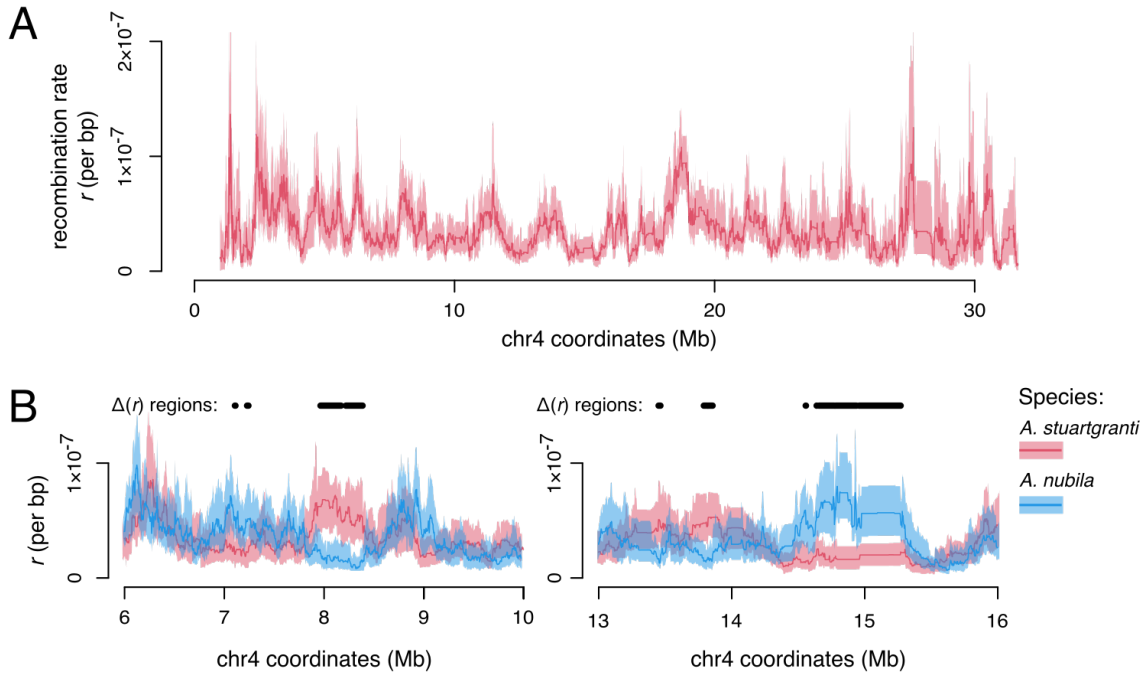

Figure 4: Bootstrap facilitates comparisons between recombination landscapes. (**A**) Chromosome 4 of *A. stuartgranti* with recombination landscapes based on the mean of 50 bootstrap replicates (thick line) and 95% confidence intervals (shaded areas). (**B**) Examples of comparisons between the maps of *A. stuartgranti* and *A. nubila*, highlighting areas of significant recombination rate differences between the maps (the $\Delta(r)$ regions).

Comparisons with trio phasing and with LD-based maps

Our standard workflow uses the same Hi-C dataset for both haplotype phasing and for recombination map inference (see **Methods**). To compare this approach with independent trio-based phasing, we took advantage of the fact that the *G. aculeatus* donor individual was bred in an aquarium from known parents. We obtained short read whole genome data from the parents and, after variant calling and filtering, we used `Hi-reComb TrioPhase` to obtain haplotype phase information for ~840 thousand SNPs across the 20 stickleback autosomes. We found that this approach reduced the crossover (false positive/negative) error rate estimated by `Hi-reComb` by more than a third, down to below 0.4% for the *G. aculeatus* data (**Figure 5A**). This result suggests that about a third of the errors in the hapcut2-phased datasets arose due to incorrect phasing.

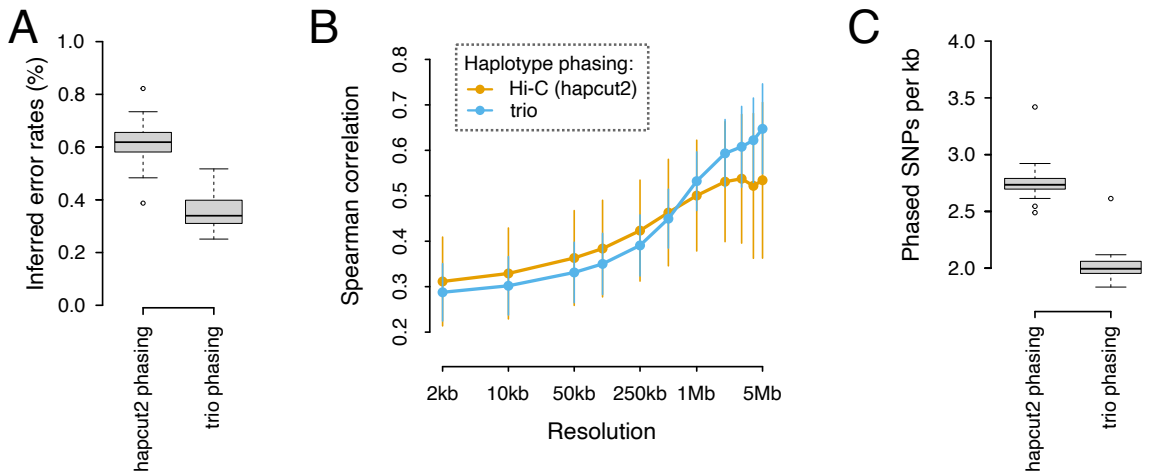

Figure 5: Hapcut2 vs. trio phasing of stickleback data. (**A**) The error rates estimated by `Hi-reComb` for the hapcut2 and trio-phased *G. aculeatus* datasets. (**B**) Correlations between gamete based `Hi-reComb` maps and an LD-based map from the same stickleback population. (**C**) The average density of SNPs phased with the hapcut2 approach and the trio approach. Each datapoint corresponds to one chromosome.

To evaluate the accuracy of the recombination maps inferred by `Hi-reComb` from the hapcut2 and trio-phased datasets, we compared both against a recombination map obtained for the same stickleback population (Walby Lake, Alaska) using linkage disequilibrium (LD) patterns in an independent population genetic dataset (see **Methods**). While the LD-based map does not represent the truth, we can nevertheless draw some conclusions from the map

414 correlations. As expected, given its lower error rate, the trio-phased dataset delivers more
415 accurate maps reflected in better correlations with the LD-based dataset at resolutions ≥1Mb
416 (**Figure 5B**). However, surprisingly, the pattern changes at finer resolutions (2kb to 500kb),
417 where the hapcut2-phased dataset seems to deliver a more accurate recombination map. The
418 apparent greater accuracy at high-resolution is likely explained by the fact that hapcut2-based
419 phasing, although less accurate, delivers a greater density of phased heterozygous sites along
420 the genome and thus more fine-scale information (**Figure 5C**).

421 ## Discussion

422 Mapping the distribution of meiotic recombination along chromosomes is a crucial step in
423 many genomic analyses. The `Hi-reComb` software provides a new, straightforward, and cost-
424 effective approach for inferring recombination maps, based on sequencing of a Hi-C library
425 from gametes from a single individual. In this manuscript we show results based on fish sperm,
426 but the method will be applicable to a broad range sexually reproducing species. Nevertheless,
427 at least two factors can limit the applicability of `Hi-reComb`. First, it is necessary to obtain a
428 relatively large number of gametes from a single individual, in the range between 100
429 thousand and 1 million for the Hi-C protocol we used. While this is easily achievable for males
430 of many larger species (e.g., most vertebrates), the cell count requirement will be challenging
431 to fulfil for males of smaller species and almost always for females who rarely produce such
432 large numbers of gametes. Second, it is necessary that the donor individual has sufficient
433 heterozygosity. We have demonstrated that heterozygosity of ~1 SNP per thousand basepairs
434 is sufficient (**Table 1**), a value that is towards the lower end of nucleotide diversity range across
435 sexually reproducing species(Leffler et al. 2012; Romiguier et al. 2014). Therefore, individuals
436 from most natural populations will be sufficiently heterozygous. However, the heterozygosity
437 requirement will pose a limitation for obtaining recombination maps from individuals who are
438 inbred in nature or due to human manipulation.

Hi-reComb is well suited for comparisons of recombination among individuals of the same or closely related species, facilitated by the bootstrapping option as illustrated in **Figure 4**. It will be interesting to learn more about how much variation there is across individuals, where current knowledge of recombination landscape variation is limited(Johnston et al. 2016; Peñalba and Wolf 2020), with studies mostly focusing on genome-wide crossover counts(Payseur 2024). An important aspect of inter-individual variation in recombination is the difference between sexes (known as heterochiasmy), which is known to be considerable, at least in some species(Sardell and Kirkpatrick 2019). In this context, it should be noted that the Hi-reComb approach is limited to diploid sequences and, therefore, we were not able to obtain the X or Y chromosome maps for *G. aculeatus* in which there are large regions of hemizygosity due to degeneration on the Y chromosome(Peichel et al. 2020) .

We used a straightforward protocol to illustrate the potential of Hi-reComb for routine recombination map inference. At the same time, it is possible to envisage several improvements to the protocol. For example, cell sorting or purification prior to library preparation that ensures that only gametes are used would deliver recombination rate quantification that is more accurate and comparable across individuals. It could also be beneficial to take into account that chromatin in gametes can be distinct from somatic cells, especially in sperm where chromatin is highly condensed by protamines(Okada 2022) or by specific histone proteins in flowering plants(Buttress et al. 2022). Chromatin decompaction treatment of sperm cells could deliver an even more uniform coverage and a higher quality of the Hi-C library (e.g., fewer duplicates, larger insert sizes).

We envisage that Hi-reComb will contribute to our understanding of patterns and of the ultimate causes of recombination rate variation by substantially easing the production of gamete-based recombination maps. Given the strengths and weaknesses of this approach, we also see great potential in combining Hi-reComb maps with pedigree- and LD-based

maps. Measuring recombination in gametes while assessing which haplotypes are transmitted across generations will shed light on the multifaceted interaction between recombination and selection.

Funding

This work was supported by a Swiss National Science Foundation (SNSF) award to M.M. (grant: 193464), an SNSF award to C.L.P. (grant: TMAG-3_209309 / 1), and NIH grants (NIAID 1R01AI123659-01A1) to D.B. and (NIGMS 5R35GM142891-04) to J.N.W..

Data Availability Statement

All raw sequence data are available on NCBI under the following accessions: BioProject PRJNA1133007 (sperm Hi-C of cichlids), BioProject PRJNA1192732 (stickleback sperm Hi-C and whole genome sequences of parents for trio based phasing), BioProject PRJEB49185 (stickleback population genetic data). The Hi-C scaffolded cichlid genomes are available under GCA_900246225.6 for *A. calliptera* and GCA_963576455.2 for *N. multifasciatus*.

Author contributions

M.M., R.D., B.S., and R.E.G. conceived and designed the study; M.M. developed the Hi-reComb software and conducted the analyses; M.T., M.M., N.M., and S.S. prepared the Hi-C libraries; C.Z. performed scaffolding of the fNeoMul1.2 genome; M.T. produced the stickleback LD-based maps; J.N.W. bred stickleback for Hi-C sequencing and trio phasing; J.N.W., C.L.P. and D.B. led the fieldwork obtaining stickleback population genetic data; W.S. and O.S. bred cichlid fish for Hi-C sequencing; M.M. wrote the manuscript with comments and input from other co-authors.

Acknowlegements

We would like to thank Daniel Jeffries, Simon H. Martin, and Aurora Ruiz-Herrera for helpful discussions, Matthew Chotlos and Adrian Indenmaur for fish husbandry, Carolin Sommer-

488 Trembo and Pamela Nicholson for assistance with lab work, and the Welcome Sanger Institute
489 sequencing core for DNA sequencing.

490 **Competing interests**

491 R.E.G. is a co-founder of Dovetail Genomics.

492

Supplementary Figures

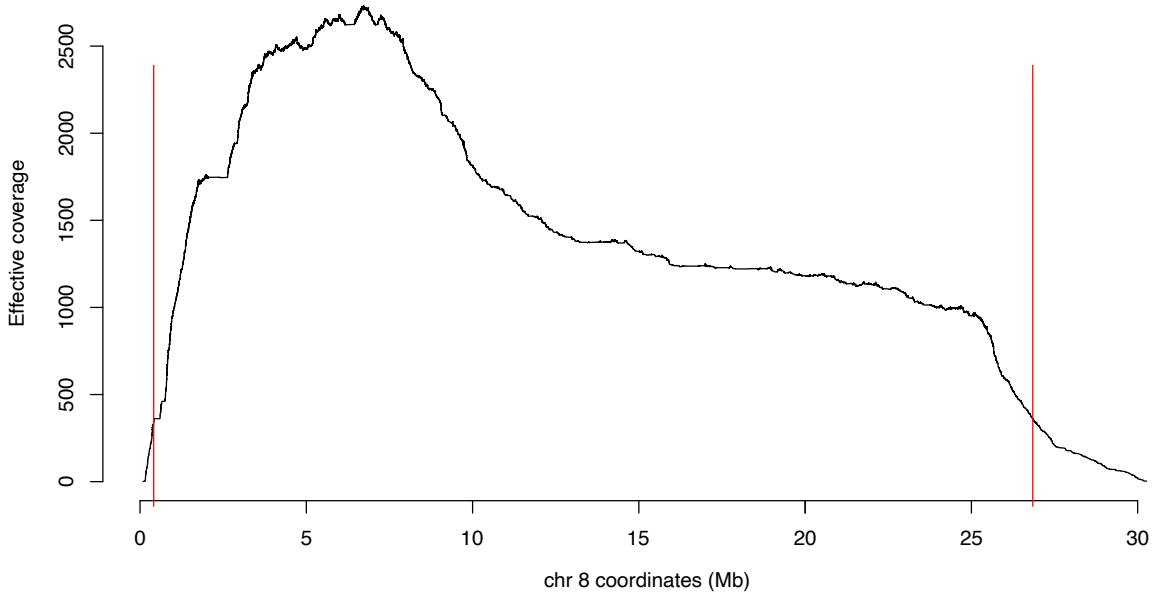

Figure S1: Effective coverage along the chromosome and the 'edge effect'.

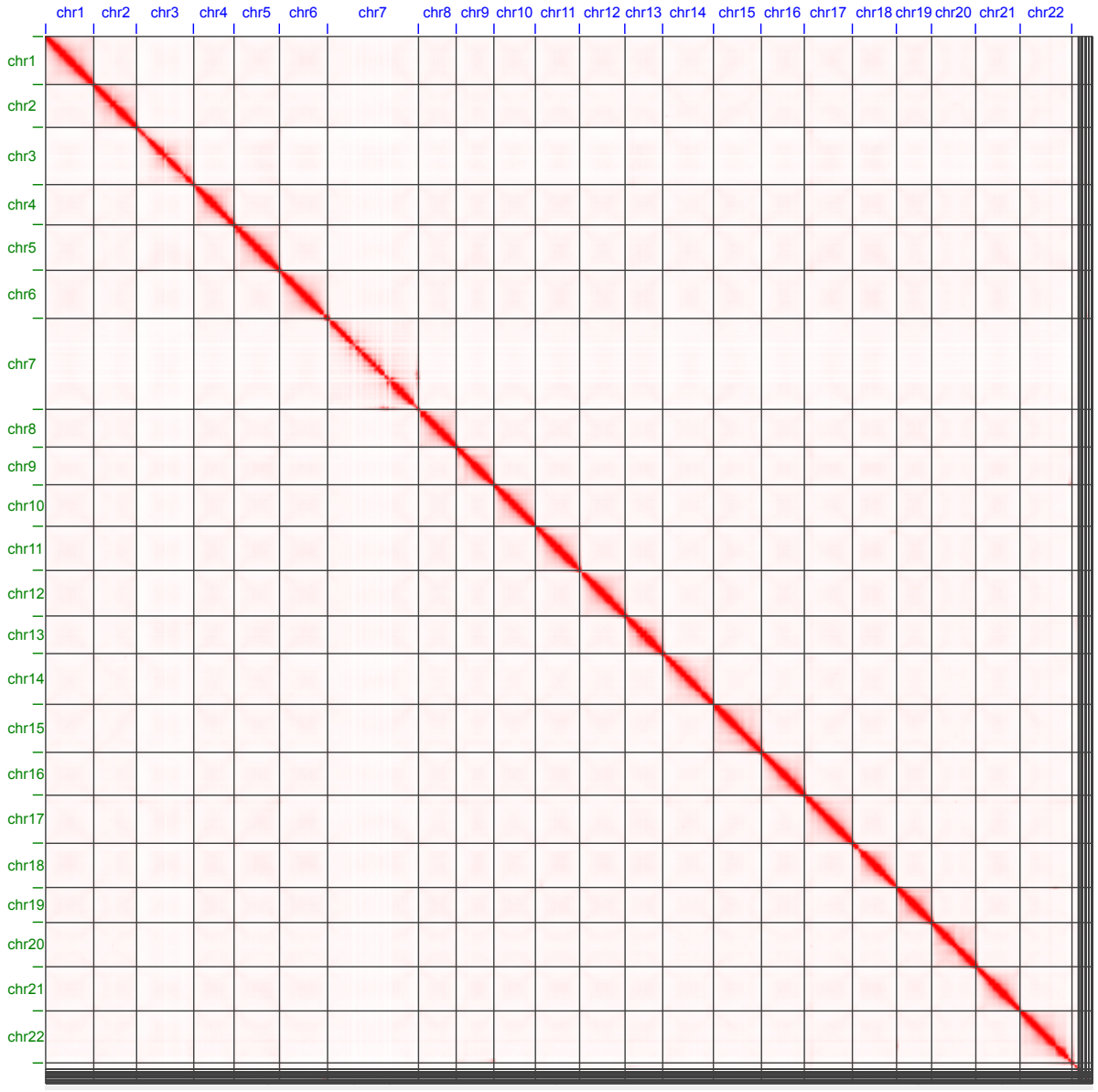

Figure S2: Hi-C contact map for read pairs mapped to the newly scaffolded *Neolamprologus multifasciatus* fNeoMul1.2 reference genome.

495

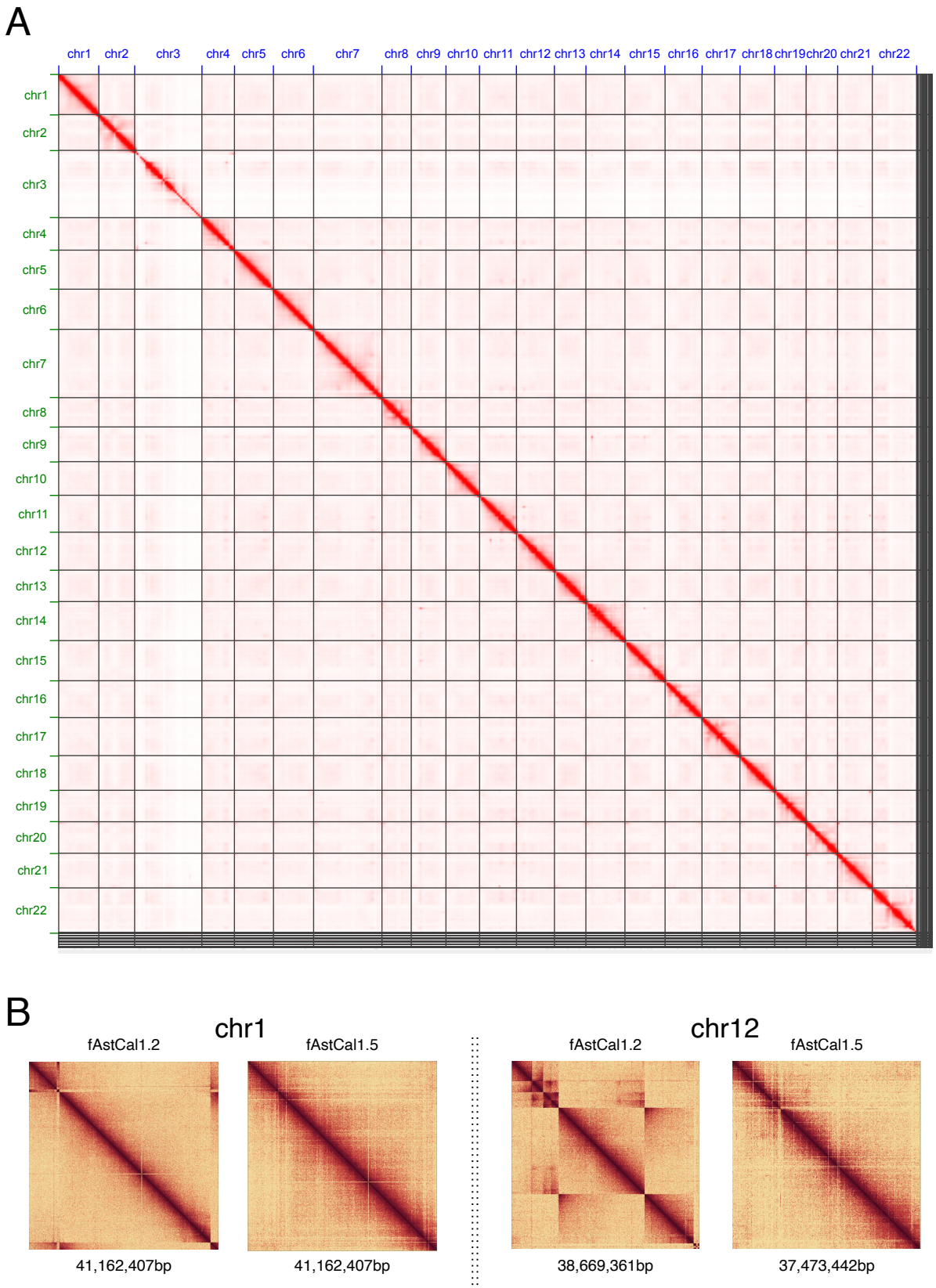

Figure S3: (A) Hi-C contact map for read pairs mapped to the newly scaffolded *Astatotilapia calliptera* fAstCal1.5 reference genome. (B) A comparison of the original fAstCal1.2 vs. the new fAstCal1.5 assemblies for two example chromosomes, showing the resolution of disagreements between the Hi-C contact map and the original assembly.

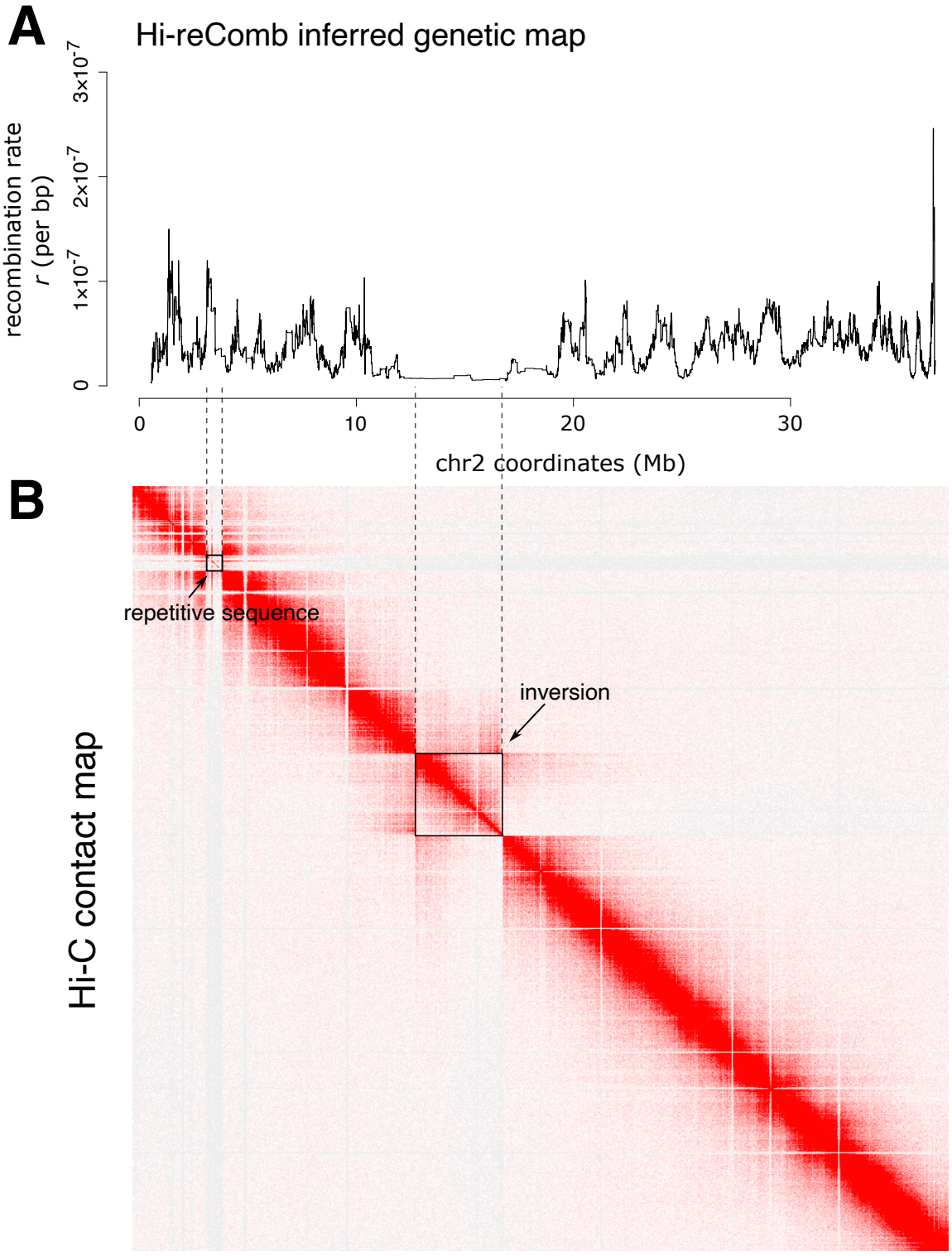

Figure S4: An inversion and suppressed recombination on *A. stuartgranti* chromosome 2. (**A**) The `Hi-reComb` inferred genetic map. (**B**) Hi-C contact map. Both panels are aligned along chr2 coordinates of the *A. calliptera* fAstCal1.5 reference. The figure illustrates a useful advantage of the Hi-C approach: at the same time as being used for recombination inference, the Hi-C contact map provides information about structural variation in the donor individual. Thus, we can see that the region of low recombination in the centre of the chromosome corresponds to and extends beyond the inversion.

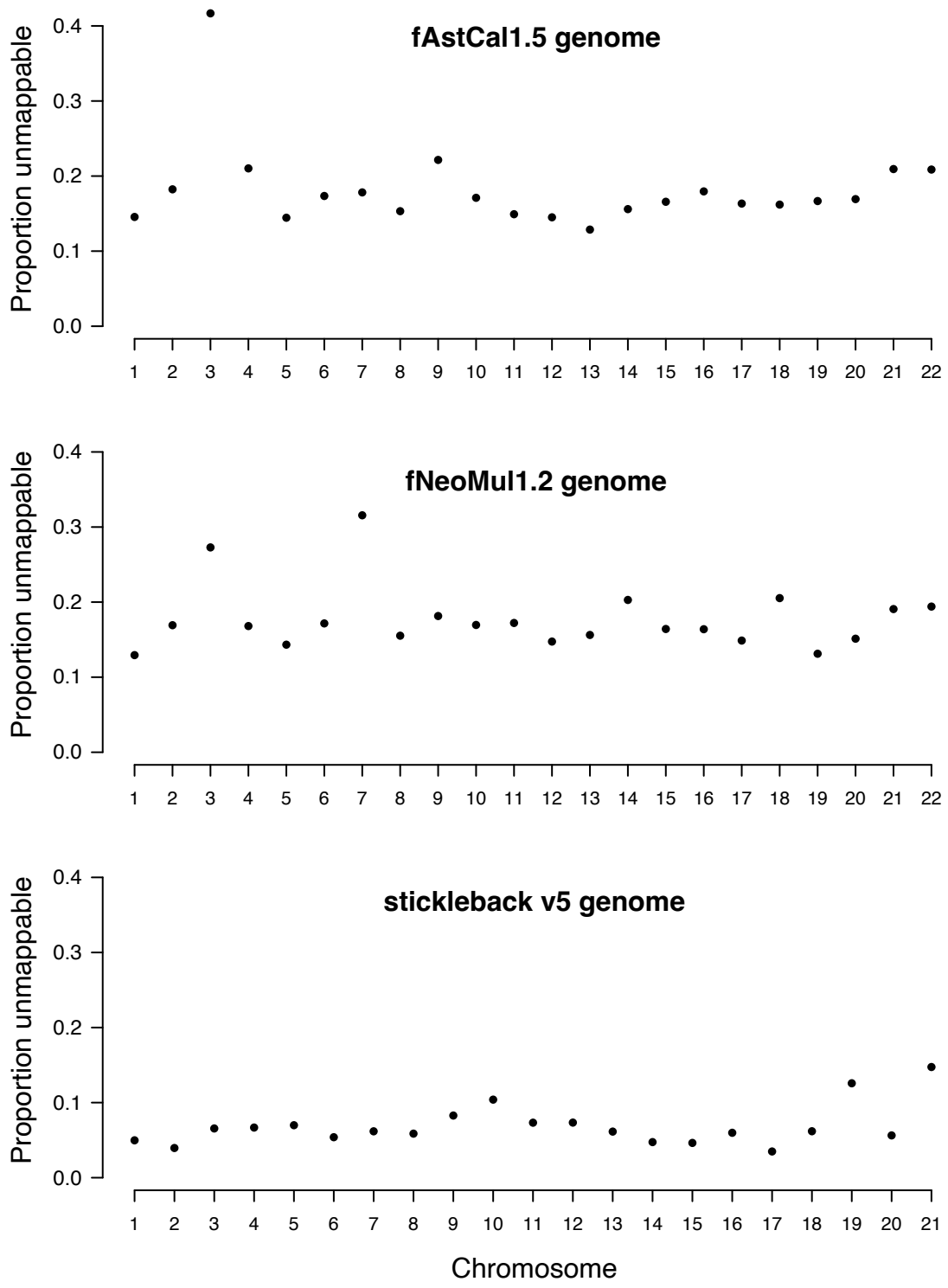

Figure S5: The proportion of unmappable sequence on different chromosomes in the three genome asseblies used in this study. Unmappable sequences are defined as regions covered by the mappability mask (see **Methods**). In the fAstCal1.5, chromosome 3 has by far the most unmmapable sequence (42%), while the next highest number is 22% on chromosome 9. It is also notable that the stickleback assembly contains much less unmappable sequence than the cichlid genomes (stickleback mean: 6.9%; cichlid 17.7% in fNeoMul1.2 and 18.2% in fAstCal1.5).

498

499

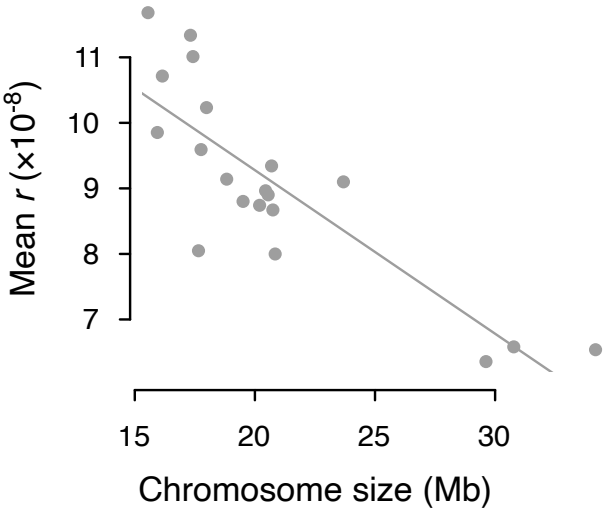

Figure S6: The negative relationship between the chromosome size and the mean recombination rate in stickleback. These results reflect recombination maps based on trio-phased data.

500

Supplementary Note 1: Map inference and the EM procedure

For each chromosome, `Hi-reComb` produces a recombination map that integrates crossover probabilities from all informative read pairs mapping onto that chromosome. The procedure is similar to the algorithm used by Halldorsson et al. (2019)(Halldorsson et al. 2019). Let $\mathbb{P}(c_i)$ be the crossover probability for a read pair i between a pair of defining phase-informative SNPs at physical positions along the chromosome g_{li} and g_{ui}. We use d_i to denote the physical distance between g_{li} and g_{ui}. The per-bp crossover probability p_i^{bp} for the read pair i is then $p_i^{bp} = \mathbb{P}(c_i)/d_i$. Next, let $(g_j)_0^N$ be the ordered sequence of defining SNPs for all informative read pairs and, within this sequence, let l_i and u_i be the indices of SNPs that define the informative read pair i.

We estimate the per-bp recombination rate r_j^{bp} for each genomic interval $I_j = [g_j, g_{j+1})$. We initialize the genetic map by finding all the K read-pairs that span each I_j and setting $r_j^{bp} = \sum_{k=1}^{K} p_k^{bp} / \sum_{k=1}^{K}(1 - \mathbb{P}(c_k))$. The proportion of recombination expected to occur within the interval, the recombination fraction r_j, is estimated as $r_j = r_j^{bp} * (g_{j+1} - g_j + 1)$. The sequence of recombination rate values $(r_j^{bp})_0^{N-1}$ specifies the entire genetic map m on the interval $[g_0, g_N)$.

We use the initial genetic map as an input into the EM procedure. First, we calculate for each read pair with positive crossover probability (i.e., $p_i^{bp} > 0$), the probability that the crossover occurred within a genomic interval I_j. This probablity p_i^j depends on the recombination rates in all the intervals that are covered by the read pair; specifically:

$$p_i^j = \frac{r_j}{\sum_{k=l_i}^{u_i} r_k}$$

An updated genetic map is calculated from these probabilities, with recombination fraction values given by $r_j = \sum_{k=1}^{K} p_k^j / \sum_{k=1}^{K}(1 - \mathbb{P}(c_k))$, where the sums are again over all the K read-

pairs that span the interval I_j. With the updated map, we then calculate new values for p_i^j then update genetic map again, and continue for `--maxEM` iterations or until convergence is reached, as defined by $\sum_{j=0}^{N-1}\left|r_j^m - r_j^{m-1}\right| < \varepsilon$, where the superscript m denotes iteration number m and ε is provided by the user via the `--epsilon` parameter.

528 **References**

529

530 Albertson RC, Powder KE, Hu Y, Coyle KP, Roberts RB, Parsons KJ. 2014. Genetic basis of
531 continuous variation in the levels and modular inheritance of pigmentation in cichlid fishes.
532 Molecular Ecology. 23(21):5135–5150. doi:10.1111/mec.12900.

533 Barton NH. 2022. The “New Synthesis.” Proc National Acad Sci. 119(30):e2122147119.
534 doi:10.1073/pnas.2122147119.

535 Bell AD, Mello CJ, Nemesh J, Brumbaugh SA, Wysoker A, McCarroll SA. 2020. Insights into
536 variation in meiosis from 31,228 human sperm genomes. Nature. 583(7815):259–264.
537 doi:10.1038/s41586-020-2347-0.

538 Blumer LM, Burskaia V, Artiushin I, Saha J, Garcia JC, Elkin J, Fischer B, Zhou C, Gresham
539 S, Malinsky M, et al. 2024. Introgression dynamics of sex-linked chromosomal inversions
540 shape the Malawi cichlid adaptive radiation. bioRxiv.:2024.07.28.605452.
541 doi:10.1101/2024.07.28.605452.

542 Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic
543 to Omnigenic. Cell. 169(7):1177–1186. doi:10.1016/j.cell.2017.05.038.

544 Brazier T, Glémin S. 2022. Diversity and determinants of recombination landscapes in
545 flowering plants. PLoS Genet. 18(8):e1010141. doi:10.1371/journal.pgen.1010141.

546 Buttress T, He S, Wang L, Zhou S, Saalbach G, Vickers M, Li G, Li P, Feng X. 2022. Histone
547 H2B.8 compacts flowering plant sperm through chromatin phase separation. Nature.
548 611(7936):614–622. doi:10.1038/s41586-022-05386-6.

549 Conte MA, Joshi R, Moore EC, Nandamuri SP, Gammerdinger WJ, Roberts RB, Carleton
550 KL, Lien S, Kocher TD. 2019. Chromosome-scale assemblies reveal the structural evolution
551 of African cichlid genomes. GigaScience. 8(4):288. doi:10.1093/gigascience/giz030.

552 Coop G, Przeworski M. 2007. An evolutionary view of human recombination. Nature reviews
553 Genetics. 8(1):23–34. doi:10.1038/nrg1947.

554 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
555 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools.
556 Gigascience. 10(2):giab008. doi:10.1093/gigascience/giab008.

557 Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data Via
558 the EM Algorithm. J Royal Statistical Soc Ser B Methodol. 39(1):1–22. doi:10.1111/j.2517-
559 6161.1977.tb01600.x.

560 DePristo MAM, Banks EE, Poplin RR, Garimella KVK, Maguire JRJ, Hartl CC, Philippakis
561 AAA, Angel GG del, Rivas MAM, Hanna MM, et al. 2011. A framework for variation
562 discovery and genotyping using next-generation DNA sequencing data. Nature Genetics.
563 43(5):491–498. doi:10.1038/ng.806.

564 Domingo J, Baeza-Centurion P, Lehner B. 2019. The Causes and Consequences of Genetic
565 Interactions (Epistasis). Annu Rev Genom Hum Genet. 20(1):1–28. doi:10.1146/annurev-
566 genom-083118-014857.

567 Dréau A, Venu V, Avdievich E, Gaspar L, Jones FC. 2019. Genome-wide recombination
568 map construction from single individuals using linked-read sequencing. Nature
569 communications. 10(1):1–11. doi:10.1038/s41467-019-12210-9.

570 Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016.
571 Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell
572 Syst. 3(1):99–101. doi:10.1016/j.cels.2015.07.012.

573 Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for
574 diverse sequencing technologies. Genome Res. 27(5):801–812. doi:10.1101/gr.213462.116.

575 Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale
576 crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol
577 Ecol. 27(11):2477–2497. doi:10.1111/mec.14699.

578 Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP,
579 Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, et al. 2019. Characterizing mutagenic
580 effects of recombination through a sequence-level genetic map. Science (New York, NY).
581 363(6425):eaau1043. doi:10.1126/science.aau1043.

582 Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, Bowden R, Donnelly P.
583 2019. Factors influencing meiotic recombination revealed by whole-genome sequencing of
584 single sperm. Science (New York, NY). 363(6433):eaau8861. doi:10.1126/science.aau8861.

585 Johnson MS, Reddy G, Desai MM. 2023. Epistasis and evolution: recent advances and an
586 outlook for prediction. BMC Biol. 21(1):120. doi:10.1186/s12915-023-01585-3.

587 Johnston SE, Bérénos C, Slate J, Pemberton JM. 2016. Conserved Genetic Architecture
588 Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep
589 (Ovis aries). Genetics. 203(1):583–598. doi:10.1534/genetics.115.185553.

590 Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P,
591 Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within
592 species? PLoS biology. 10(9):e1001388. doi:10.1371/journal.pbio.1001388.

593 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-
594 MEM. arXiv.org. q-bio.GN. arXiv.org.

595 Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I,
596 Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive Mapping of Long-Range
597 Interactions Reveals Folding Principles of the Human Genome. Science. 326(5950):289–
598 293. doi:10.1126/science.1181369.

599 Lu S, Zong C, Fan W, Yang M, Li J, Chapman AR, Zhu P, Hu X, Xu L, Yan L, et al. 2012.
600 Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome
601 Sequencing. Science. 338(6114):1627–1630. doi:10.1126/science.1229112.

602 Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR,
603 Kamitaki N, Martersteck EM, et al. 2015. Highly Parallel Genome-wide Expression Profiling
604 of Individual Cells Using Nanoliter Droplets. Cell. 161(5):1202–1214.
605 doi:10.1016/j.cell.2015.05.002.

606 Marques DA, Meier JI, Seehausen O. 2019. A Combinatorial View on Speciation and
607 Adaptive Radiation. Trends in ecology & evolution. 34(6):531–544.
608 doi:10.1016/j.tree.2019.02.008.

609 Morgan AP, Gatti DM, Najarian ML, Keane TM, Galante RJ, Pack AI, Mott R, Churchill GA,
610 Villena FP-M de. 2017. Structural Variation Shapes the Landscape of Recombination in
611 Mouse. Genetics. 206(2):603–619. doi:10.1534/genetics.116.197988.

612 Okada Y. 2022. Sperm chromatin structure: Insights from in vitro to in situ experiments. Curr
613 Opin Cell Biol. 75:102075. doi:10.1016/j.ceb.2022.102075.

614 Oksuz BA, Yang L, Abraham S, Venev SV, Krietenstein N, Parsi KM, Ozadam H, Oomen
615 ME, Nand A, Mao H, et al. 2021. Systematic evaluation of chromosome conformation
616 capture assays. Nat Methods. 18(9):1046–1055. doi:10.1038/s41592-021-01248-7.

617 Open2C, Abdennur N, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko A, Imakaev M,
618 Venev SV. 2024. Pairtools: From sequencing data to chromosome contacts. PLOS Comput
619 Biol. 20(5):e1012164. doi:10.1371/journal.pcbi.1012164.

620 Ozouf-Costaz C, Coutanceau JP, Bonillo C, Mercot H, Fermon Y, Guidi-Rotani C. 2017.
621 New insights into the chromosomal differentiation patterns among cichlids from Africa and
622 Madagascar. Cybium.

623 Payseur BA. 2024. Genetics of recombination rate variation within and between species. J
624 Evol Biol.:voae158. doi:10.1093/jeb/voae158.

625 Peichel CL, McCann SR, Ross JA, Naftaly AFS, Urton JR, Cech JN, Grimwood J, Schmutz
626 J, Myers RM, Kingsley DM, et al. 2020. Assembly of the threespine stickleback Y
627 chromosome reveals convergent signatures of sex chromosome evolution. Genome Biol.
628 21(1):177. doi:10.1186/s13059-020-02097-x.

629 Peñalba JV, Wolf JBW. 2020. From molecules to populations: appreciating and estimating
630 recombination rate variation. Nat Rev Genet. 21(8):476–492. doi:10.1038/s41576-020-0240-
631 1.

632 Phillips PC. 2008. Epistasis — the essential role of gene interactions in the structure and
633 evolution of genetic systems. Nat Rev Genet. 9(11):855–867. doi:10.1038/nrg2452.

634 Phillips PC, Arnold SJ. 1989. Visualizing Multivariate Selection. Evolution. 43(6):1209–1222.
635 doi:10.1111/j.1558-5646.1989.tb02569.x.

636 Poletto AB, Ferreira IA, Cabral-de-Mello DC, Nakajima RT, Mazzuchelli J, Ribeiro HB,
637 Venere PC, Nirchio M, Kocher TD, Martins C. 2010. Chromosome differentiation patterns
638 during cichlid fish evolution. BMC Genet. 11(1):50–50. doi:10.1186/1471-2156-11-50.

639 Quin CTO, Drilea AC, Conte MA, Kocher TD. 2013. Mapping of pigmentation QTL on an
640 anchored genome assembly of the cichlid fish, Metriaclima zebra. BMC genomics.
641 14(1):287. doi:10.1186/1471-2164-14-287.

642 Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL,
643 Machol I, Omer AD, Lander ES, et al. 2014. A 3D Map of the Human Genome at Kilobase
644 Resolution Reveals Principles of Chromatin Looping. Cell. 159(7):1665–1680.
645 doi:10.1016/j.cell.2014.11.021.

646 Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome—
647 patterns and consequences. Molecular Ecology. 22(11):3014–3027.
648 doi:10.1111/mec.12322.

649 Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R,
650 Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the
651 determinants of genetic diversity. Nature. 515(7526):261–263. doi:10.1038/nature13685.

652 Sardell JM, Kirkpatrick M. 2019. Sex Differences in the Recombination Landscape. Am Nat.
653 195(2):361–379. doi:10.1086/704943.

654 Shi H, Kichaev G, Pasaniuc B. 2016. Contrasting the Genetic Architecture of 30 Complex
655 Traits from Summary Association Data. Am J Hum Genetics. 99(1):139–153.
656 doi:10.1016/j.ajhg.2016.05.013.

657 Spence JP, Song YS. 2019. Inference and analysis of population-specific fine-scale
658 recombination maps across 26 diverse human populations. Sci Adv. 5(10):eaaw9206.
659 doi:10.1126/sciadv.aaw9206.

660 Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in
661 recombination frequency and distribution across eukaryotes: patterns and processes.
662 Philosophical Transactions of the Royal Society B: Biological Sciences.
663 372(1736):20160455. doi:10.1098/rstb.2016.0455.

664 Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, Michelmore RW, Huettel B,
665 Schneeberger K. 2019. Linked-read sequencing of gametes allows efficient genome-wide
666 analysis of meiotic recombination. Nat Commun. 10(1):4310. doi:10.1038/s41467-019-
667 12209-2.

668 Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history
669 from hundreds of unphased whole genomes. Nat Genet. 49(2):303–309.
670 doi:10.1038/ng.3748.

671 Wang J, Fan HC, Behr B, Quake SR. 2012. Genome-wide Single-Cell Analysis of
672 Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell. 150(2):402–
673 412. doi:10.1016/j.cell.2012.06.030.

674 White NJ, Butlin RK. 2021. Multidimensional divergent selection, local adaptation, and
675 speciation. Evolution. 75(9):2167–2178. doi:10.1111/evo.14312.

676 Xu P, Kennell T, Gao M, Consortium HGSV, Kimberly RP, Chong Z. 2019. MRLR:
677 unraveling high-resolution meiotic recombination by linked reads. Bioinformatics. 36(1):10–
678 16. doi:10.1093/bioinformatics/btz503.

679 Yang Liu, Gao Y, Li M, Park K-E, Liu S, Kang X, Liu M, Oswalt A, Fang L, Telugu BP, et al.
680 2022. Genome-wide recombination map construction from single sperm sequencing in
681 cattle. BMC Genom. 23(1):181. doi:10.1186/s12864-022-08415-w.

682 Zhou C, McCarthy SA, Durbin R. 2022. YaHS: yet another Hi-C scaffolding tool.
683 Bioinformatics. 39(1):btac808. doi:10.1093/bioinformatics/btac808.

684