Title: Versatile simulations of admixture and accurate local ancestry inference with *mixnmatch* and *ancestryinfer*

Authors: Molly Schumer[1,2,3*], Daniel L. Powell[1,2,4], Russ Corbett-Detig[5,6*]

*correspondence to: schumer@stanford.edu and rucorbet@ucsc.edu

[1]Department of Biology, Stanford University
[2]Centro de Investigaciones Científicas de las Huastecas "Aguazarca"
[3]Hanna H. Gray Fellow, Howard Hughes Medical Institute
[4] Department of Biology, Texas A&M University
[5]Genomics Institute, University of California, Santa Cruz
[6]Department of Biomolecular Engineering, University of California, Santa Cruz

Keywords: hybridization, admixture, local ancestry inference, hidden Markov model

Abstract

It is now clear that hybridization between species is much more common than previously recognized. As a result, we now know that the genomes of many modern species, including our own, are a patchwork of regions derived from past hybridization events. Increasingly researchers are interested in disentangling which regions of the genome originated from each parental species using local ancestry inference methods. Due to the diverse effects of admixture, this interest is shared across disparate fields, from human genetics to research in ecology and evolutionary biology. However, local ancestry inference methods are sensitive to a range of biological and technical parameters which can impact accuracy. Here we present paired simulation and ancestry inference pipelines, *mixnmatch* and *ancestryinfer*, to help researchers plan and execute local ancestry inference studies. *mixnmatch* can simulate arbitrarily complex demographic histories in the parental and hybrid populations, selection on hybrids, and technical variables such as coverage and contamination. *ancestryinfer* takes as input sequencing reads from simulated or real individuals, and implements an efficient local ancestry inference pipeline. We perform a series of simulations with *mixnmatch* to pinpoint factors that influence accuracy in local ancestry inference and highlight useful features of the two pipelines. Together, *mixnmatch* and *ancestryinfer* are powerful tools for predicting the performance of local ancestry inference methods on real data.

Introduction

Since the advent of inexpensive whole genome sequencing, it has become increasingly clear that hybridization is an important part of the evolutionary history of many species. This has made methods to study hybridization fundamental tools in the fields of genetics and evolutionary biology. In addition to methods for inferring the genome-wide history of admixture (Alexander, Novembre, & Lange, 2009; Patterson et al., 2012; Pritchard, Stephens, & Donnelly, 2000), researchers have recently taken advantage of methods that make it possible to infer ancestry at a small spatial scale along the genome (i.e. "local ancestry inference"). Originally applied to study genetic diseases in humans (Hoggart, Shriver, Kittles, Clayton, & McKeigue, 2004; Montana & Pritchard, 2004; Patterson et al., 2004; Winkler, Nelson, & Smith, 2010), applications of local ancestry inference methods have become a cornerstone of studies of genome evolution (Sankararaman et al., 2014; Schumer et al., 2018), population history (Baharian et al., 2016; Corbett-Detig & Nielsen, 2017), and trait evolution (Heliconius Genome, 2012; Jones et al., 2018; Oziolor et al., 2019).

Particularly popular among local ancestry inference methods are methods that use a hidden Markov model (HMM) to infer ancestry as a hidden state based on observations of genotypes or sequencing read counts. For autosomal loci in diploid individuals, there are three possible ancestry states (homozygous parent species 1, homozygous parent species 2, and heterozygous for the two ancestries). Ancestry HMMs allow the probability of a given local ancestry state to be modeled as a function of the observed data in the region, the ancestry state probability at the previous site, and the recombination distance between adjacent ancestry informative sites, among other possible parameters. The output of these methods is typically posterior probabilities for each possible ancestry state at each ancestry informative site along the chromosome (Alexander et al., 2009; Andolfatto et al., 2011; Corbett-Detig & Nielsen, 2017).

As local ancestry methods have been applied in more and more species (e.g. Cande, Andolfatto, Prud'homme, Stern, & Gompel, 2012; R. Li et al., 2018; Sankararaman et al., 2014; Schumer et al., 2014; Slotte et al., 2013), each with their own population genetic properties and demographic histories, simulation tools to evaluate performance have not kept apace. While some studies have carefully modeled the demographic history of hybridizing populations and the impact of this history on the accuracy of ancestry inference (e.g. Sankararaman et al., 2014;

Medina, Thornlow, Nielsen, & Corbett-Detig, 2018), this typically requires the development of custom computational pipelines. As a result, many studies do not evaluate the expected performance of local ancestry inference methods and use default parameter sets.

While the majority of tools for local ancestry inference report performance under parameters relevant to human populations (e.g. Maples, Gravel, Kenny, & Bustamante, 2013), these tools are being applied much more broadly. This is concerning because local ancestry inference approaches that perform well in one context may perform poorly in others, as their performance is sensitive to a number of biological and technical variables (Medina et al., 2018; Sankararaman et al., 2014; Schumer, Cui, Rosenthal, & Andolfatto, 2016). Because of the importance of accurate local ancestry inference in evolutionary biology and genetics, simulation tools that allow researchers to systematically evaluate their accuracy and common biases are needed.

Here, we present a hybrid genome simulation pipeline called *mixnmatch* that can be used to evaluate the accuracy of local ancestry inference under a range of biological and technical parameters. Our pipeline builds on previous simulation tools developed by us and others to implement flexible demographic simulations in both ancestral parental populations and hybrid populations. Species-specific genetic parameters, including base composition and local recombination rates, can be incorporated into simulations. Users also specify a number of technical parameters that impact accuracy, including sequencing depth, sequencing error, and cross-contamination rates.

In addition to simulating admixed genomes with *mixnmatch*, we provide a paired pipeline called *ancestryinfer* that implements local ancestry inference on real or simulated data. This pipeline builds off of a previously published HMM (AncestryHMM; Corbett-Detig & Nielsen, 2017) and includes several features that seamlessly integrate its use with raw sequence data, automating and parallelizing steps from mapping Illumina reads to the output of posterior probabilities for ancestry states. Moreover, both pipelines are user-friendly, with parameters defined in a text-editable configuration file and automated and docker-based installation options. Together, *mixnmatch* and *ancestryinfer* will make it feasible for users to perform sophisticated simulations to predict accuracy and apply the same approaches when analyzing their data.

Methods & Results

Overview

The overall structure of *mixnmatch* is described in Figure 1 and in more detail in Figure S1. First, the pipeline simulates parental haplotypes, either using user-provided genomes or the coalescent simulator *macs* (Chen, Marjoram, & Wall, 2009), which allows for simulations of complex population history (Figure S1; Supporting Information 1). With *macs*-based simulations, users can still provide one of the parental genomes for use as a base sequence (Figure S1).

Next, ancestry tract lengths for hybrid individuals are generated with SELAM (Corbett-Detig & Jones, 2016), using information about the admixture proportion, demographic history of the hybrid population, and number of generations since initial admixture. Based on these tract lengths, hybrid genomes are constructed from previously simulated parental haplotypes. For each individual, reads are generated at user-specified lengths and depth using the *wgsim* program (H. Li, 2011). If desired, users can simulate cross-contamination between samples during read generation. Together, these features of *mixnmatch* allow for simulation of experimental design choices and the history of both the parental and hybrid population populations, making it a powerful tool for studies of hybridization.

Output files produced by the *mixnmatch* simulator include fastq and fasta files for each individual, a bed file indicating true ancestry along the simulated chromosome, and all necessary files for running an efficient local ancestry inference HMM, implemented through our associated pipeline, *ancestryinfer* (Table S1). After running local ancestry inference with *ancestryinfer*, users can evaluate the performance of local ancestry inference with a provided script that summarizes accuracy by comparing inferred versus true ancestry at each ancestry informative site.

Installation

Users can install the two pipelines using step-by-step instructions provided with *mixnmatch* and *ancestryinfer* or by loading a docker image with all dependencies for both

pipelines pre-installed (see Appendix 1; user manual). Parameters for each pipeline are set in a text-editable configuration file (examples available on github: https://github.com/Schumerlab/mixnmatch; https://github.com/Schumerlab/ancestryinfer). Instructions for setting parameters can be found in the *mixnmatch* and *ancestryinfer* user manuals (see Appendix 1-2). *mixnmatch* and *ancestryinfer* can be parallelized with a SLURM resource management system (Appendix 1-2) and the non-parallel version can be run any Linux system or on Mac operating systems.

Simulation of parental haplotypes and definition of ancestry informative sites

Generating parental haplotypes is the first step in *mixnmatch* simulations (Figure S1). An important feature of *mixnmatch* is the ability to simulate demographic history. This is crucial because the population history of each parental species will influence genetic diversity and the extent of background linkage disequilibrium within species, as well as divergence between species, all of which can impact accuracy in local ancestry inference. To model this, users provide a *macs* command (Chen, Marjoram, & Wall, 2009) describing the demographic history of the two parental species in the configuration file. *mixnmatch* executes this command and converts the *macs* output to nucleotide sequences using the *seq-gen* program (Rambaut & Grassly, 1997). Users can optionally provide species-specific base composition and transition/transversion ratios in the configuration file, as well as a local recombination map. If provided, this recombination map will be used in the *macs* simulations of parental haplotypes and in generating hybrid ancestry tracts (see *Simulation of hybrid genomes*).

One of these simulated sequences from each parental population is set aside as to be used as the reference sequence. The remaining haplotypes are then used to define ancestry informative markers and are later used to generate hybrid chromosomes (see *Simulation of hybrid genomes* below). Ancestry informative markers are defined as highly differentiated markers among a randomly selected subset of the simulated parental haplotypes. Users specify the required frequency difference between species and the number of parental haplotypes to use to evaluate this in the configuration file. We note that the appropriate choices for these values will depend on the level of divergence and shared polymorphisms between species; users can rely on predictions from population genetic theory (e.g. Wakeley & Hey, 1997) or *mixnmatch*

simulations to explore these parameters. Together, these steps model the impacts of demographic history on the number and distribution of ancestry informative sites, as well as the process researchers typically follow in identifying them.

Other options for parental haplotype generation

Although the method of haplotype generation described above incorporates demography, recombination, and incomplete lineage sorting in the parental species, it lacks other complexities of real genome sequences such as repetitive elements and local variation in base composition. To accommodate these additional challenges, we allow users to provide an ancestral sequence to which simulated mutations are added. This option incorporates features of real sequences while modeling mutations and ancestral recombination events using a coalescent framework.

Another biological variable that can impact accuracy in ancestry HMMs is drift between the reference panel used to define ancestry informative sites and the populations that actually contributed to the hybridization event. Users can tell *mixnmatch* to allow drift between the source population and the population used to define ancestry informative sites (using *macs*). If users choose this option, ancestry informative sites are defined based on the drifted parental population *instead* of the hybridizing parental population. This generates realistic allele frequency differences between the reference and source parental populations, as well as covariance in allele frequency differences due to linkage. In real data, this could contribute to errors in downstream analysis.

We also provide an option for using *mixnmatch* with the exact reference genomes and ancestry informative sites that users plan to rely upon in their experiments. Instead of a *macs* command describing parental population history, this option takes the two parental reference genomes and a number of population genetic parameters as input (Figure S1). This approach is described in Supporting Information 1 (see also Schumer et al., 2016). Although this option has some limitations that should be considered (Supporting Information 1), it allows users to evaluate performance on species-specific reference genomes and ancestry informative sites.

192 **Simulation of hybrid genomes**

193

194 In addition to allowing users to simulate the demographic history of the parental species,
195 *mixnmatch* models the demographic history of the hybrid population (Figure S1). Any process
196 that influences the distribution of ancestry tract lengths, from bottlenecks to assortative mating,
197 could impact the accuracy of local ancestry inference. To incorporate this, *mixmatch* uses a
198 previously developed tool (SELAM, Corbett-Detig & Jones, 2016) to model the effects of
199 demography on ancestry tracts. Users specify the mixture proportions contributed from each
200 parent species to the hybrid population and the number of generations since initial admixture in
201 the *mixnmatch* configuration file. In addition, users can choose to provide a parameter file
202 describing the demographic history of the hybrid population (Appendix 1) as well as selection on
203 hybrids.

204 Using these tract lengths, *mixnmatch* next generates hybrid chromosomes. For each
205 ancestry tract, *mixnmatch* extracts the focal region from a randomly selected parental haplotype
206 of the appropriate ancestry. If users have provided a local recombination map, *mixnmatch* uses
207 this in converting tract coordinates from genetic to physical distance; otherwise a global
208 recombination rate is used. This process is repeated until an entire haplotype is generated, and
209 two such haplotypes are combined to generate both chromosomes within a diploid hybrid
210 individual. Importantly, this approach introduces variation to the simulation from processes such
211 as incomplete lineage and sampling of a reference panel, both of which can impact downstream
212 accuracy.

213 The pipeline next simulates reads uniformly from these hybrid chromosomes using the
214 *wgsim* program (H. Li, 2011), with user specified read lengths, read mate type, coverage, indel
215 and error rates. At the same time contamination can be simulated. During this step *mixnmatch*
216 writes out the true ancestry for each individual at every position along the chromosome,
217 facilitating analysis of accuracy downstream.

218 The final output of *mixnmatch* includes all of the files needed for running our paired
219 pipeline for local ancestry inference, *ancestryinfer,* as well as files needed for other local
220 ancestry inference tools. These include simulated reference genomes, ancestry informative sites
221 and counts for each allele in the parental reference panel, simulated Illumina reads, and bed
222 formatted files containing the true ancestry for each individual (Table S1).

One possible shortcoming of our approach for generating hybrid haplotypes is that it does not model coalescence among samples after admixture, which could generate errors in local ancestry inference not captured by our simulation approach. This is most likely to impact simulations of very small populations or ancient admixture (Corbett-Detig & Nielsen, 2017). We also note that the number of parental haplotypes used to generate the hybrid chromosomes is determined by the total number of parental haplotypes users choose to simulate. Simulating fewer parental haplotypes will decrease *mixnmatch* runtime, but users should ensure that the total number of parental haplotypes simulated captures most of the genetic variation within the parental populations (e.g. Figure S2; Watterson, 1975).

A versatile ancestry inference pipeline

To facilitate local ancestry inference analysis of real and simulated data, we developed a paired pipeline called *ancestryinfer*. This pipeline automates steps from read mapping to local ancestry inference, and is easy-to-use and parallelizable (Supporting Information 2). Briefly, the work flow of this pipeline (Figure 1) begins with mapping reads from a hybrid individual to both parental references independently with *bwa mem* (H. Li & Durbin, 2009) and identifying reads that do not map uniquely to either of the parental genomes. These reads are then excluded from the hybrid individual's bam file using *ngsutils* (Breese & Liu, 2013). Such reads may fall within repetitive regions of the parental genomes, be impacted by mapping bias, incompleteness of one parental reference, or insertions/deletions that disrupt mapping. These technical issues have received less attention as it relates to their impact on local ancestry inference (Supporting Information 3) but have been shown to have major impacts in other types of analyses such as allele-specific expression (Degner et al., 2009; Stevenson, Coolon, & Wittkopp, 2013).

Next, reads matching each parental allele at ancestry informative sites are counted from a *samtools* mpileup file (H. Li, 2011) generated for each hybrid individual. There are two options in the pipeline for identifying ancestry informative sites. If the genomes provided by the user are co-linear, users can direct *ancestryinfer* to automatically identify sites that differ between them. Alternately, users can provide the locations of ancestry informative sites and their estimated frequencies in the parental species. The latter option allows users to take advantage of reference assemblies for both species if they are available.

Counts for each parental allele at ancestry informative sites are subsampled to thin to one ancestry informative site per read if multiple sites occur within one read. We implement this thinning because mismapping can generate clusters of errors and non-independence between sites is not modeled in the HMM. Finally, Ancestry_HMM (Corbett-Detig & Nielsen, 2017) is applied to infer posterior probabilities of each ancestry state at ancestry informative sites along the genome. In addition, *ancestryinfer* summarizes the intervals over which ancestry transitions occur (Figure S3). With the exception of false switches in ancestry generated by errors, these intervals reflect observed crossover events in hybrids. These crossover intervals can be used to generate recombination maps (see below, *Generating a hybrid recombination map using observed ancestry transitions*).

If users have run the *ancestryinfer* pipeline on data simulated by *mixnmatch*, the accuracy of local ancestry inference can be summarized by running a script provided with the *mixnmatch* pipeline (Appendix 2). Briefly, the script generates hard-calls at a user-specific posterior probability threshold and compares true and inferred ancestry at each ancestry informative site along the chromosome. The output of this script includes plots summarizing individual-level accuracy, accuracy as a function of tract length (Figure S4), and a file tabulating all accurate and inaccurate calls in individual tracts as well as mean posterior error.

Predicted accuracy of local ancestry inference in simulated data

Basic simulation setup

Using *mixnmatch* and *ancestryinfer*, we next tested the accuracy of local ancestry inference with simulated data under a range of biological and technical scenarios. For these simulations, we started with a base parameter set (Table S2) and then systematically modified parameters in individual simulations. For this base parameter set we simulated 200 generations since initial admixture, 50-50 mixture proportions between the two parental species, per site polymorphism rates in each of the parental species of 0.1%, and pairwise sequence divergence between the parental species of 0.5%. We used the first 10 Mb of chromosome 1 from the swordtail fish species *Xiphophorus birchmanni* as the ancestral sequence, and provided *mixnmatch* with an inferred recombination map (Schumer et al., 2018) from that same region.

We simulated 100 parental haplotypes for each species, which will capture the majority of parental polymorphisms segregating in these populations (Figure S2; Watterson, 1975). We sampled 20 parental haplotypes from both species to define ancestry informative sites and required a frequency difference of 95% between species for a site to be treated as ancestry informative. A complete description of these simulations can be found in Supporting Information 4. For this set of parameters, the pipeline required a total of 0.66 CPU hours and was run with 96 Gb of memory for the sequence generation step and 64 Gb of memory for the hybrid genome simulation step. Simulations were performed on Dell C6420 servers on Stanford's Sherlock High Performance Computing cluster.

In simulations with this parameter set, accuracy of local ancestry inference was high, with estimated error rates of <0.5% per ancestry informative site (Figure S4). As expected, shorter ancestry tracts have higher per-basepair error rates (Figure S4). Because tract lengths often differ systematically across the genome and in particular around selected sites (Sedghifar, Brandvain, Ralph, & Coop, 2015; Shchur, Svedberg, Medina, Corbett-Detig, & Nielsen, 2019), this highlights the importance of considering local error rates throughout the genome when analyzing local ancestry data (Supporting Information 4, Figure S5).

Simulations under a range of scenarios

To understand what biological and technical variables impact accuracy in local ancestry inference, we next modified individual parameters in turn. Below we summarize the scenarios we tested that had the greatest impact on accuracy. A full description of our simulations can be found in Supporting Information 4.

Intuitively, with increasing divergence between species, there will be more ancestry informative sites. This is predicted to result in more accurate local ancestry inference. To evaluate this, we performed simulations varying pairwise divergence between the hybridizing species from 0.25% to 1%. We note that these simulations focus on deeper divergence than what has been considered in previous work (Maples et al., 2013; Medina et al., 2018) but span realistic levels of divergence found in naturally hybridizing species (Brandvain, Kenney, Flagel, Coop, & Sweigart, 2014; Schumer et al., 2014; Teeter et al., 2008; Turissini & Matute, 2017). As expected, accuracy increased with higher divergence between the hybridizing species (Figure 2),

as did the precision with which the locations of ancestry transitions were identified (Figure S3; consistent with previous results, Medina et al., 2018).

As the time since initial admixture increases, recombination events in each generation split haplotypes of a given ancestry into smaller and smaller pieces. Since these short tracts will contain fewer ancestry informative sites, this leads to the prediction that local ancestry inference will be more accurate in populations that have hybridized recently, which is indeed what we observed (Figure 2).

Following a similar logic, skewed admixture proportions are expected to reduce accuracy of ancestry inference in tracts derived from the "minor" parental species (i.e. the parental species that contributed less to the initial hybridization event). This is because only recombination events that occur in regions heterozygous for ancestry from the two parent species are detectable with ancestry HMMs, and minor parent haplotypes are more frequently found in this state (Gravel, 2012). As expected, we observe that the accuracy of local ancestry inference within minor parent tracts is reduced in simulations with skewed initial admixture proportions (Figure S6).

Ideally, reference panels for defining ancestry informative sites should be derived from the same parental populations that contributed to the admixture event. In practice, this is often not possible since source populations may no longer exist, may be unknown, or may themselves be admixed, making it more sensible to use allopatric populations for a reference panel. However, such populations are also expected to have some level of genetic drift from the admixing populations, which could impact accuracy. In *mixnmatch* this can be modeled by adding drift to the simulation and specifying which populations to use in defining ancestry informative sites.

To investigate the impact of using reference panels with drift from the hybridizing populations, we used *mixnmatch* to simulate two additional populations that split from the parental source populations before hybridization and treated these populations as the reference panel (0.4-3*Ne* generations ago in different simulations, with initial divergence between species occurring 8*Ne* generations ago; Supporting Information 4). We found that accuracy substantially decreased with increasing drift between the reference population and the source parental populations (Figure 2). Notably, this can be partially remediated by increasing the required frequency difference between the parental populations when defining ancestry informative sites (Supporting Information 4).

A common decision that researchers make is how much coverage to collect per sample. Intuitively, early generation hybrids will require less data to accurately infer local ancestry than later generation hybrids because of differences in the distribution of ancestry tract lengths. We simulated genome-wide coverage between 0.05-0.5X with *mixnmatch* (Supporting information 4). As expected, increased coverage improved accuracy but our results also suggested that beyond a certain level of coverage, improvements in accuracy plateau (Figure 2). However, higher coverage continued to improve the resolution of the locations of ancestry transitions (Figure S7).

In general, we find that the HMM implemented in *ancestryinfer* (Corbett-Detig & Nielsen, 2017) is not particularly sensitive to user-provided priors for admixture time or admixture proportion, but is somewhat sensitive to recombination rate priors (Supporting Information 4). Providing a local recombination prior in *ancestryinfer* modestly increases accuracy when the recombination map does not contain errors (Figure S8). However, in practice recombination maps will contain errors that depend on the method used for map construction among other factors (Supporting Information 4). With moderate levels of error in map inference our simulations suggest that users may benefit from providing a uniform recombination prior (Figure S8).

In recent years there has been substantial interest in the ecological and evolutionary genomics community in restriction site associated sequencing (or RAD-seq; Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012; Van Tassell et al., 2008) as a low-cost option for generating genomic data. However, RAD-seq data may be suboptiomal for local ancestry inference applications. This is because overdispersion in the spacing between sampled sites, coverage variation, and genealogical biases generated by variants in restriction enzyme cut-sites introduced by this method could all reduce the accuracy of local ancestry inference. To explore this, we generated reads *in silico* associated with a commonly used enzyme in RAD (*EcoRI*) but otherwise performed simulations as described above (Supporting Information 5). We found that in the case of *ancestryinfer* performance with RAD data is poor (Figure S9), likely due to the reliance on fewer ancestry informative sites for inference (Supporting Information 5).

Finally, although we focus on modeling accuracy under neutral demographic scenarios, *mixnmatch* can also be used to simulate selection on hybrids. The SELAM program that is used

to simulate ancestry tract lengths in *mixnmatch* accommodates selection on hybrid populations (Corbett-Detig & Jones, 2016; Figure S10, Supporting Information 6). This allows users to implement versatile selection scenarios in *mixnmatch* (Appendix 1), and use it to explore signatures of selection on hybrids. For example, recent work has described the impact of selection against hybrid incompatibilities on the number and distribution of ancestry junctions (Hvala, Frayer, & Payseur, 2018), the impact of selection on ancestry tract lengths (Sedghifar et al., 2015; Shchur et al., 2019), and the local frequency of haplotypes derived from the minor parental species (Sankararaman et al., 2014; Schumer et al., 2018; Vernot & Akey, 2014). We demonstrate the use of this feature of *mixnmatch* in Supporting Information 6 and Figure S10.

Application to natural and artificial swordtail hybrids

To demonstrate the utility of *ancestryinfer* with real data, we applied it to data from F_1 and F_2 hybrids generated from crosses between the swordtail fish species *X. birchmanni* and *X. malinche*. We constructed libraries using a tagmentation based library preparation protocol and collected low coverage whole-genome sequence data for these libraries (~0.2X coverage per individual, Supporting Information 7, Appendix 3). We used previously collected low-coverage sequence data from 60 individuals of each parental species to estimate allele frequencies at ancestry informative sites (Schumer et al., 2018).

Illumina sequencing data from lab-generated hybrids was input into the *ancestryinfer* pipeline to infer local ancestry along the 24 swordtail chromosomes (here 150 x 2 paired-end data was used). We used the appropriate parameters for mixture proportion, generations since mixture, and global recombination rate based on known information about the cross (see details in Supporting Information 7). We converted posterior probabilities for a given ancestry state into hard calls using a threshold of 0.9 and examined local patterns of ancestry in F_1 and F_2 hybrids. We also summarized expected ancestry proportions, heterozygosity in ancestry, and the numbers of observed ancestry transitions genome-wide (Figure 3).

We find that local and global ancestry patterns in F_1 and F_2 hybrids mirror expectations for each cross type (Figure 3), and that the results are consistent with extremely low error rates in ancestry inference. For example, estimated homozygosity at ancestry informative sites in F_1 hybrids is <0.1% (Figure 3). Importantly, this high level of accuracy is predicted from

409 simulations of early generation hybrids with *mixnmatch* (Supporting Information 7), suggesting
410 that *mixnmatch* simulations are capturing important properties of real data.
411
412 *Generating a hybrid recombination map using observed ancestry transitions*
413
414 Accurate local ancestry inference has a large number of downstream applications. One
415 such application is inferring the locations of crossovers for the construction of genetic maps
416 (Amores et al., 2014; Rastas, Calboli, Guo, Shikano, & Merilä, 2015; Salomé et al., 2012). As
417 discussed previously, if users specify a posterior probability threshold in the *ancestryinfer*
418 configuration file, the program will output a bed file containing recombination intervals inferred
419 from observed ancestry transitions in hybrids.
420 We used the locations of observed ancestry transitions in 139 F_2 hybrids that we
421 generated between *X. birchmanni* and *X. malinche* (at a posterior probability threshold of 0.9;
422 Supporting Information 7-8) to estimate the recombination rate in 5 Mb windows. We used a
423 large window size due to the spatial scale over which ancestry transitions were localized (lower
424 and upper 5% quantile of intervals genome-wide: 23 kb - 667 kb) and because we expected the
425 resulting map to be relatively coarse, given a total of 4038 inferred crossovers genome-wide
426 (average 1.2 per individual per chromosome).
427 We compared inferred recombination rates in this F_2 map to a linkage disequilibrium
428 based recombination map for *X. birchmanni* that we had previously generated (Schumer et al.,
429 2018). As expected, we observed a strong correlation in estimated recombination rate between
430 the linkage disequilibrium based and crossover maps (R=0.82, Figure 4, Supporting Information
431 8). Simulations suggest that the observed correlation is consistent with the two recombination
432 maps being indistinguishable, given the low resolution of the F_2 map (Supporting Information 8).
433
434 **Discussion**
435
436 With an increasing appreciation that hybridization is a common evolutionary process,
437 there has been renewed interest in local ancestry inference in the fields of genetics and
438 evolutionary biology. Accurate local ancestry information is important for applications from
439 admixture mapping to studying genome evolution after hybridization. Despite this, there are few

simulation tools that have been developed to model the impacts of biological and technical variables on the accuracy of local ancestry inference.

We demonstrated the use of *mixnmatch* as a flexible tool to predict the accuracy of local ancestry inference under a range of biological scenarios. As expected *a prori*, our simulations show that the factors with the strongest impact on accuracy include the number of ancestry informative sites that distinguish the hybridizing species, the length of the ancestry tracts containing these sites, and the frequency at which sites are erroneously defined as ancestry informative (either due to genetic drift or high levels of shared polymorphisms; Figure 2). We show how *mixnmatch* can also help users make important decisions about their projects, such as how much coverage to collect per hybrid individual and how many parental individuals to sequence to define ancestry informative sites.

mixnmatch is primarily designed to allow users to explore demographic and technical parameters that may influence the accuracy of local ancestry inference. However, because it uses the SELAM program to generate ancestry tract lengths (Corbett-Detig & Jones, 2016), it is also possible to implement natural selection during admixture (Figure S10). This will allow users to study the impacts of selection on local ancestry, ancestry junctions, and ancestry tract lengths. We predict that this will be a useful feature of *mixnmatch* for the many research groups studying selection after hybridization.

Simulated data from *mixnmatch* can be used to evaluate the accuracy of any ancestry inference program. However, it is designed to pair seamlessly with the *ancestryinfer* pipeline we describe here, which automates steps from read mapping to local ancestry inference. *ancestryinfer* has excellent accuracy under a broad range of biological conditions, and is fast and easy to use.

ancestryinfer is also intended to be an easy-to-use pipeline for local ancestry inference in real data. We demonstrate an application of the *ancestryinfer* pipeline to real data by using it to identify the locations of crossover events in *X. birchmanni* x *X. malinche* F_2 hybrids (Figure 3) and constructing a recombination map (Figure 4). Other possible uses of *ancestryinfer* include generating ancestry probabilities for QTL mapping or for studying genome evolution in natural hybrid populations, highlighting the versatile applications of the *mixnmatch* and *ancestryinfer* pipelines.

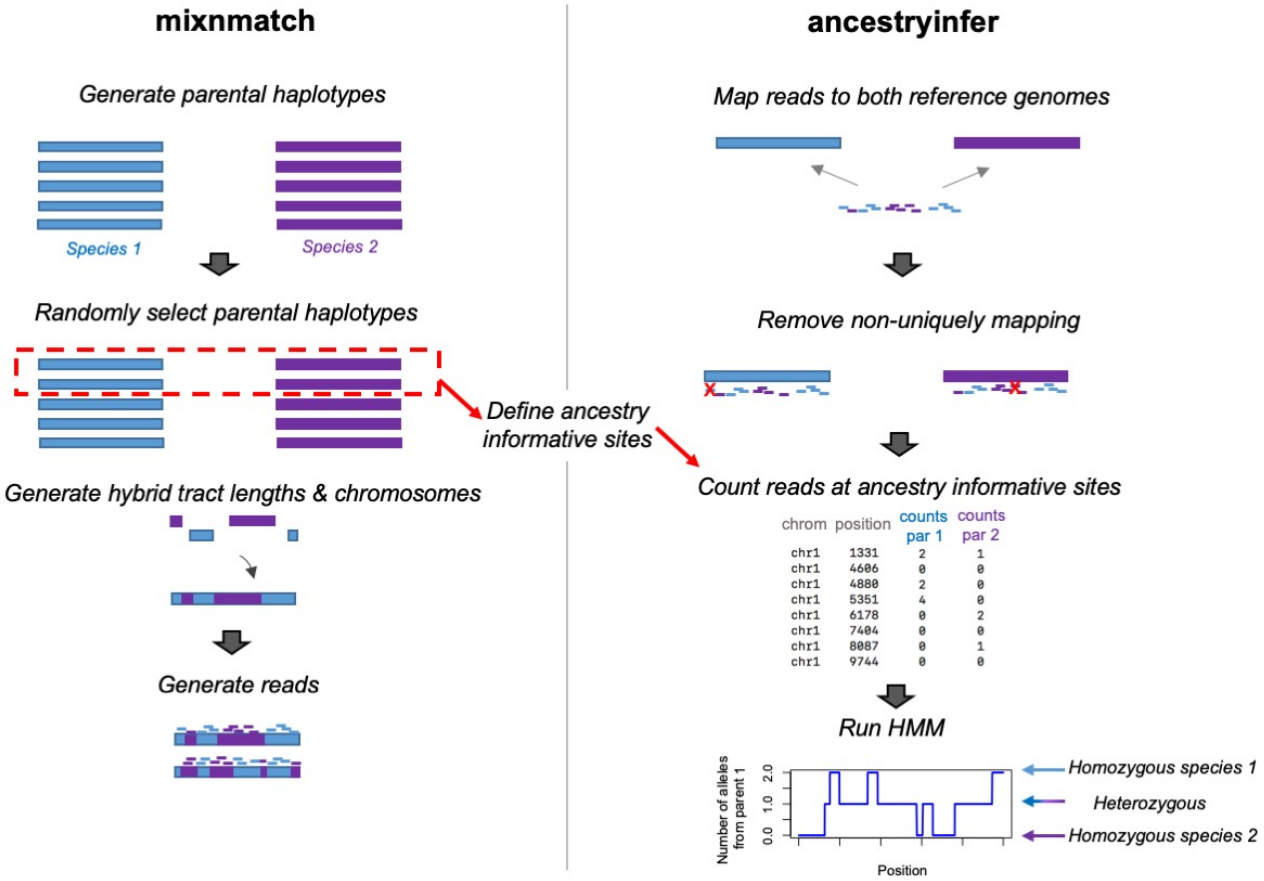
Data accessibility

Data associated with this manuscript is available on Dryad (doi:XXXX; Schumer, Powell, & Corbett-Detig, 2019) and *mixnmatch* and *ancestryinfer* pipelines are available on github (https://github.com/Schumerlab/mixnmatch; https://github.com/Schumerlab/ancestryinfer) and dockerhub (https://hub.docker.com/repository/docker/schumer/mixnmatch-ancestryinfer-image)

Acknowledgements

We thank Peter Andolfatto, Andrés Bendesky, Quinn Langdon, Ben Moran, David Reich, Alisa Sedghifar, and members of the Schumer and Corbett-Detig labs for helpful discussions and feedback on this work. Stanford University and the Stanford Research Computing Center provided computational resources and support for this project. This work was supported by a Hanna H. Gray fellowship and NIH 1R35GM133774 grant to MS and by NIH 1R35GM128932 and an Alfred P. Sloan Fellowship to RC-D.

485 **Figures**
486
487



488
489 **Figure 1.** Schematic showing the major steps in the *mixnmatch* (left) and *ancestryinfer* (right)
490 pipelines. *mixnmatch* can be used to simulate hybrid data under user-specified parental and
491 hybrid demographic scenarios and under a range of technical parameters. A more detailed
492 description of the *mixnmatch* pipeline and its options is outlined in Figure S1. Simulated
493 Illumina data output by *mixnmatch* can be input into our automated ancestry inference pipeline,
494 *ancestryinfer*, as can data from natural and artificial hybrids. The red box indicates the
495 haplotypes chosen from the parental species during *mixnmatch* to identify ancestry informative
496 sites that can later be used in *ancestryinfer*. Blue shading indicates tracts and reads derived from
497 parent species 1 and purple indicates tracts and reads derived from parent species 2. Red X's
498 mark reads that will be removed for not mapping uniquely to both parental genomes.
499
500

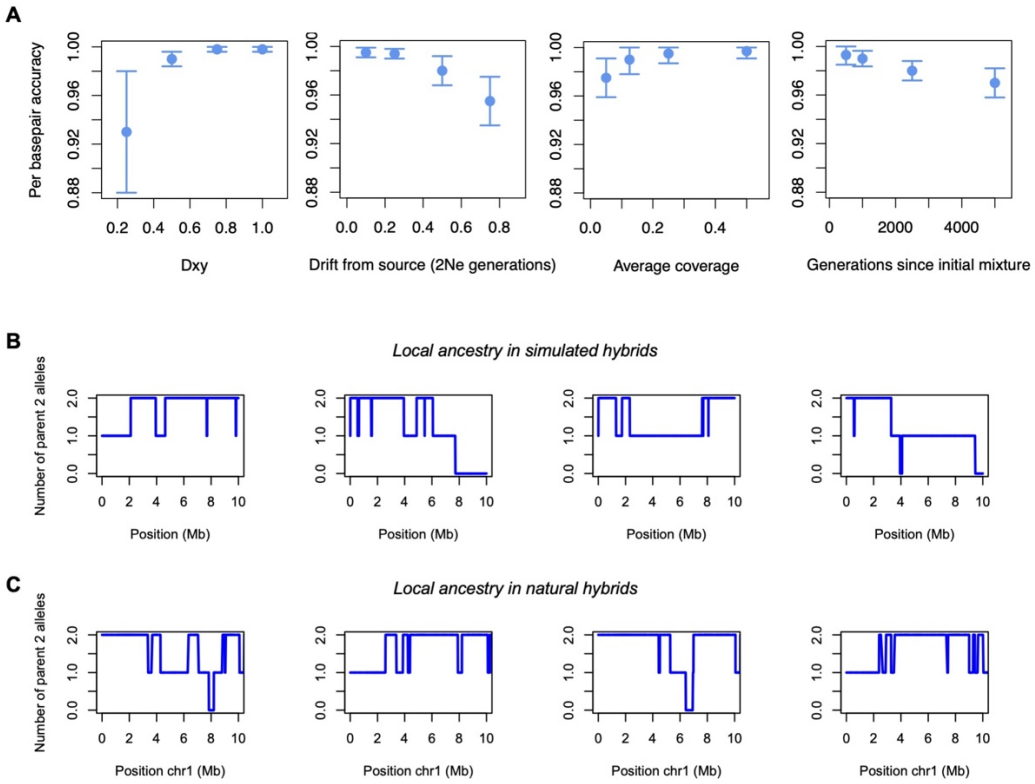


Figure 2. A) Results of *mixnmatch* simulations evaluating accuracy under different biological and technical parameters. All simulations start with the same basic parameter set (Table S2) and systematically vary the focal parameter (see Supporting Information 2). Points indicate the mean of individual-level accuracy and whiskers indicate two standard deviations. B) Example local ancestry results for simulated hybrids. Parameter used in this simulation were ~1X average coverage, sequence divergence of 0.5%, within species polymorphism rates of 0.1%, 20 generations since initial admixture, and 35% of the genome derived from parent species 1. C) Example local ancestry results for the first 10 Mb of chromosome 1 for *X. birchmanni* x *X. malinche* hybrid individuals from the Chahuaco falls hybrid population, where per-base coverage and inferred parameter values match those simulated in B. Note the qualitative similarities between C and B in the number of ancestry transitions and the size of the ancestry tracts. In C parent 2 alleles are those derived from the *X. malinche* parental species.

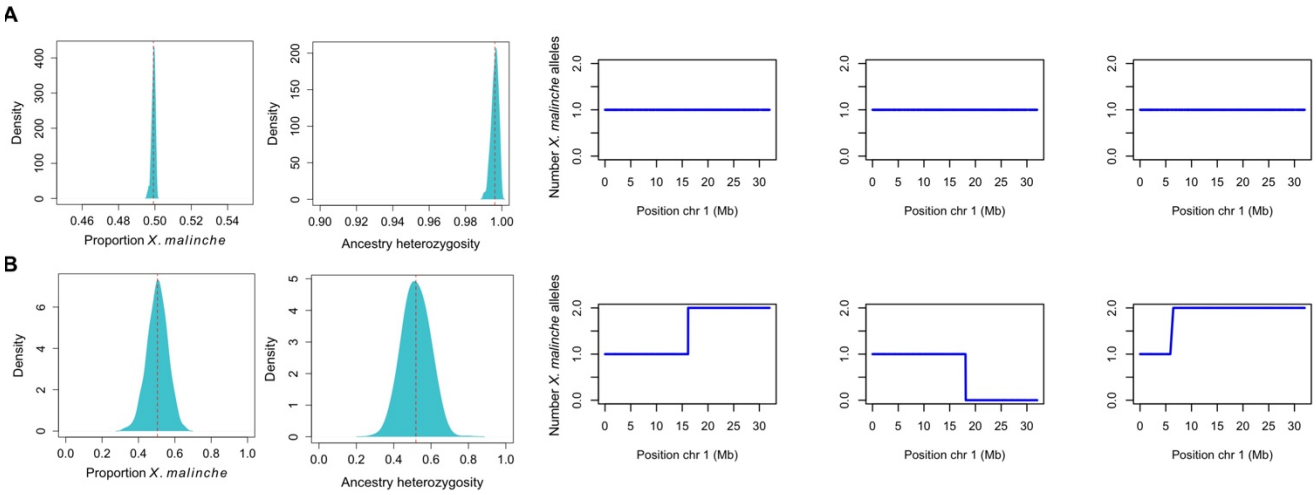


Figure 3. Results of the *ancestryinfer* pipeline run on F_1 (A) and F_2 (B) hybrids generated between *X. birchmanni* and *X. malinche*. A) As expected, we infer that F_1 hybrids have precisely 50% of their genomes derived from each parental species and infer nearly 100% heterozygosity at ancestry informative sites in these individuals. Example local ancestry plots for chromosome 1 for a subset of these F_1 hybrids are also shown (right). B) Similarly, genome-wide ancestry distributions and genome-wide ancestry heterozygosity in F_2 hybrids follows predicted distributions. Example local ancestry plots for chromosome 1 for a subset of these F_2 hybrids are also shown (right).

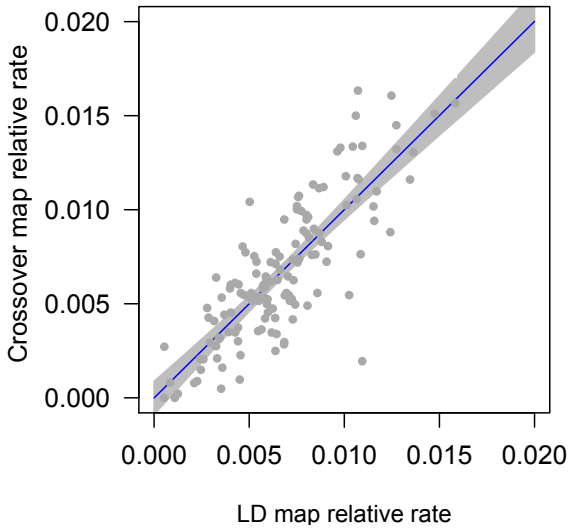


Figure 4. Comparison of F_2 crossover recombination map generated with *ancestryinfer* and previously published linkage disequilibrium map from the *X. birchmanni* parental species. Relative rates per 5 Mb window for both the linkage disequilibrium map (x-axis) and F_2 map (y-axis) are shown by gray dots. The blue line shows the best fit regression line between the maps ($R^2 = 0.67$) and the gray area shows the 95% confidence intervals. Simulations suggest that the observed correlation is consistent with recombination rates being identical across the two maps (Supporting Information 8).

534 **References**

535

536 Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in
537 unrelated individuals. *Genome Research*, *19*(9), 1655–1664. doi: 10.1101/gr.094052.109

538 Amores, A., Catchen, J., Nanda, I., Warren, W., Walter, R., Schartl, M., & Postlethwait, J. H.
539 (2014). A RAD-Tag Genetic Map for the Platyfish (*Xiphophorus maculatus*) Reveals
540 Mechanisms of Karyotype Evolution Among Teleost Fish. *Genetics*, *197*(2), 625–641.
541 doi: 10.1534/genetics.114.164293

542 Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern,
543 D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping.
544 *Genome Research*, *21*(4), 610–617. doi: 10.1101/gr.115402.110

545 Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing
546 the power of RADseq for ecological and evolutionary genomics. *Nature Reviews*
547 *Genetics*, *17*(2), 81–92. doi: 10.1038/nrg.2015.28

548 Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., …
549 Gravel, S. (2016). The Great Migration and African-American Genomic Diversity. *PLOS*
550 *Genetics*, *12*(5), e1006059. doi: 10.1371/journal.pgen.1006059

551 Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G., & Sweigart, A. L. (2014). Speciation and
552 Introgression between Mimulus nasutus and Mimulus guttatus. *PLOS Genetics*, *10*(6),
553 e1004410. doi: 10.1371/journal.pgen.1004410

554 Breese, M. R., & Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating
555 next-generation sequencing datasets. *Bioinformatics (Oxford, England)*, *29*(4), 494–496.
556 doi: 10.1093/bioinformatics/bts731

557 Cande, J., Andolfatto, P., Prud'homme, B., Stern, D. L., & Gompel, N. (2012). Evolution of
558 multiple additive loci caused divergence between *Drosophila yakuba* and *D. santomea* in
559 wing rowing during male courtship. *PloS One*, *7*(8), e43888. doi:
560 10.1371/journal.pone.0043888
561 Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence
562 data. *Genome Research*, *19*(1), 136–142. doi: 10.1101/gr.083634.108
563 Corbett-Detig, R., & Jones, M. (2016). SELAM: simulation of epistasis and local adaptation
564 during admixture with mate choice. *Bioinformatics*, *32*(19), 3035–3037. doi:
565 10.1093/bioinformatics/btw365
566 Corbett-Detig, R., & Nielsen, R. (2017). A Hidden Markov Model Approach for Simultaneously
567 Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data
568 in Samples of Arbitrary Ploidy. *PLOS Genetics*, *13*(1), e1006529. doi:
569 10.1371/journal.pgen.1006529
570 Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K.
571 (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-
572 sequencing data. *Bioinformatics*, *25*(24), 3207–3212. doi: 10.1093/bioinformatics/btp579
573 Gravel, S. (2012). Population Genetics Models of Local Ancestry. *Genetics*, *191*(2), 607–619.
574 doi: 10.1534/genetics.112.139808
575 Heliconius Genome, C. (2012). Butterfly genome reveals promiscuous exchange of mimicry
576 adaptations among species. *Nature*, *487*(7405), 94–98.
577 Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design
578 and Analysis of Admixture Mapping Studies. *The American Journal of Human Genetics*,
579 *74*(5), 965–978. doi: 10.1086/420855

580 Hvala, J. A., Frayer, M. E., & Payseur, B. A. (2018). Signatures of hybridization and speciation
581 in genomic patterns of ancestry. *Evolution*, *72*(8), 1540–1552. doi: 10.1111/evo.13509
582 Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J. R., …
583 Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage
584 in snowshoe hares. *Science*, *360*(6395), 1355–1358. doi: 10.1126/science.aar5273
585 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping
586 and population genetical parameter estimation from sequencing data. *Bioinformatics*,
587 *27*(21), 2987–2993. doi: 10.1093/bioinformatics/btr509
588 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
589 transform. *Bioinformatics*, *25*(14). doi: 10.1093/bioinformatics/btp324
590 Li, R., Bitoun, E., Altemose, N., W Davies, R., Davies, B., & Myers, S. (2018). A high-
591 resolution map of non-crossover events in mice reveals impacts of genetic diversity on
592 meiotic recombination. doi: 10.1101/428987
593 Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative
594 Modeling Approach for Rapid and Robust Local-Ancestry Inference. *American Journal*
595 *of Human Genetics*, *93*(2), 278–288. doi: 10.1016/j.ajhg.2013.06.020
596 Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018). Estimating the Timing of
597 Multiple Admixture Pulses During Local Ancestry Inference. *Genetics*, *210*(3), 1089–
598 1107. doi: 10.1534/genetics.118.301411
599 Montana, G., & Pritchard, J. K. (2004). Statistical Tests for Admixture Mapping with Case-
600 Control and Cases-Only Data. *The American Journal of Human Genetics*, *75*(5), 771–
601 789. doi: 10.1086/425281

602 Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., VerPloeg, S. G., Bruns, P. C., … Matson, C.
603 W. (2019). Adaptive introgression enables evolutionary rescue from extreme
604 environmental pollution. *Science*, *364*(6439), 455–457. doi: 10.1126/science.aav4155
605 Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., …
606 Reich, D. (2004). Methods for High-Density Admixture Mapping of Disease Genes. *The*
607 *American Journal of Human Genetics*, *74*(5), 979–1000. doi: 10.1086/420871
608 Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012).
609 Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093. doi:
610 10.1534/genetics.112.145037
611 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest
612 RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model
613 and Non-Model Species. *PLoS One*, *7*(5), e37135. doi: 10.1371/journal.pone.0037135
614 Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using
615 multilocus genotype data. *Genetics*, *155*(2), 945–959.
616 Rambaut, A., & Grassly, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation
617 of DNA sequence evolution along phylogenetic frees. *Computer Applications in the*
618 *Biosciences*, *13*(3).
619 Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., & Merilä, J. (2015). Construction of
620 Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an
621 Example. *Genome Biology and Evolution*, *8*(1), 78–93. doi: 10.1093/gbe/evv250
622 Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., & Weigel, D.
623 (2012). The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*,
624 *108*(4), 447–455. doi: 10.1038/hdy.2011.95

625 Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., … Reich, D.
626 (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*,
627 *507*(7492), 354–357. doi: 10.1038/nature12961
628 Schumer, M., Cui, R., Powell, D. L., Dresner, R., Rosenthal, G. G., & Andolfatto, P. (2014).
629 High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish
630 species. *ELife*, *3*, e02535. doi: 10.7554/eLife.02535
631 Schumer, M., Cui, R., Rosenthal, G. G., & Andolfatto, P. (2016). simMSG: an experimental
632 design tool for high-throughput genotyping of hybrids. *Molecular Ecology Resources*,
633 *16*(1), 183–192. doi: 10.1111/1755-0998.12434
634 Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., … Przeworski, M.
635 (2018). Natural selection interacts with recombination to shape the evolution of hybrid
636 genomes. *Science*, *360*(6389), 656. doi: 10.1126/science.aar3684
637 Schumer, M., Powell, D. L., & Corbett-Detig, R., 2019. Data availability - Dryad, doi:XXXXX.
638
639 Sedghifar, A., Brandvain, Y., Ralph, P., & Coop, G. (2015). The Spatial Mixing of Genomes in
640 Secondary Contact Zones. *Genetics*, doi: 10.1534/genetics.115.179838
641 Shchur, V., Svedberg, J., Medina, P., Corbett-Detig, R., & Nielsen, R. (2019). On the
642 distribution of tract lengths during adaptive introgression. *BioRxiv*, 724815. doi:
643 10.1101/724815
644 Slotte, T., Hazzouri, K. M., Ågren, J. A., Koenig, D., Maumus, F., Guo, Y.-L., … Wright, S. I.
645 (2013). The Capsella rubella genome and the genomic consequences of rapid mating
646 system evolution. *Nature Genetics*, *45*, 831.

647 Stevenson, K. R., Coolon, J. D., & Wittkopp, P. J. (2013). Sources of bias in measures of allele-
648 specific expression derived from RNA-seq data aligned to a single reference genome.
649 *BMC Genomics*, *14*(1), 536. doi: 10.1186/1471-2164-14-536
650 Teeter, K. C., Payseur, B. A., Harris, L. W., Bakewell, M. A., Thibodeau, L. M., O'Brien, J. E.,
651 … Tucker, P. K. (2008). Genome-wide patterns of gene flow across a house mouse
652 hybrid zone. *Genome Research*, *18*(1), 67–76. doi: 10.1101/gr.6757907
653 Turissini, D. A., & Matute, D. R. (2017). Fine scale mapping of genomic introgressions within
654 the Drosophila yakuba clade. *PLOS Genetics*, *13*(9), e1006971. doi:
655 10.1371/journal.pgen.1006971
656 Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C.
657 T., … Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep
658 sequencing of reduced representation libraries. *Nature Methods*, *5*(3), 247–252. doi:
659 10.1038/nmeth.1185
660 Vernot, B., & Akey, J. M. (2014). Resurrecting Surviving Neandertal Lineages from Modern
661 Human Genomes. *Science*, *343*(6174), 1017–1021. doi: 10.1126/science.1245938
662 Wakeley, J., & Hey, J. (1997). Estimating Ancestral Population Parameters. *Genetics*, *145*(3),
663 847–855.
664 Watterson, G. A. (1975). On the number of segregating sites in genetical models without
665 recombination. *Theoretical Population Biology*, *7*(2), 256–276.
666 Winkler, C. A., Nelson, G. W., & Smith, M. W. (2010). Admixture Mapping Comes of Age.
667 *Annual Review of Genomics and Human Genetics*, *11*(1), 65–89. doi: 10.1146/annurev-
668 genom-082509-141523
669