Supplementary Materials

A beginner's guide to low-coverage whole genome sequencing for population genomics

Runyang Nicolas Lou[1]*, Arne Jacobs[1,2], Aryn Wilder[3], Nina O. Therkildsen[1]*

[1]Department of Natural Resources and the Environment, Cornell University, Ithaca, NY 14853, USA
[2]Current address: Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, G12 8QQ, UK
[3]San Diego Zoo Institute for Conservation Research, Escondido, CA 92027, USA

*Corresponding authors: RNL (rl683@cornell.edu), NOT (nt246@cornell.edu)

TABLE OF CONTENTS

Supplementary methods

Section 4: Population genomic inference from lcWGS data under different experimental designs

In short, we created in silico populations under a diploid Wright-Fisher model using forward genetic simulation, and then simulated the low-coverage whole genome sequencing (lcWGS) process with a subset of individuals in these populations. We performed genotype-likelihood-based analyses on these simulated sequencing reads, and tested their power in population genetic inference. In addition, we simulated other high-throughput sequencing strategies, including pool-seq and RAD-seq, and compared their performance with that of lcWGS. Our entire simulation and analysis pipeline is available on GitHub (https://github.com/therkildsen-lab/lcwgs-simulation).

First, we tested the accuracy of lcWGS in allele frequency estimation with different sequencing strategies in a single simulated population with stable population size and no selection. We used SLiM3 (Haller & Messer, 2019) to randomly generate a starting nucleotide sequence on a 30Mb chromosome, and then created a diploid population with all individuals initially having this same starting sequence. We aimed to simulate a large population with effective population size (*Ne*) on the order of 10^5. However, it is computationally expensive to directly simulate large population sizes with forward genetic simulation methods, since all individuals in the population need to be tracked in every generation, and more time is required to reach mutation-drift equilibrium. Therefore, we chose to scale down our simulated population size (*N*) by a factor of 100, and scale up the mutation rate (μ) and recombination rate (*r*) by a factor of 100. Because the most important parameters of the simulated population (e.g. nucleotide diversity, linkage disequilibrium, site frequency spectrum) depends on products in the form of $N\mu$, Nr, and etc., this scaling approach can generate a realistic population with a reasonable computational cost. Specifically, we set *N* to be 1,000, and ran the simulation with $\mu = 1\text{x}10^{-6}$ per bp per generation and r = 250 cM/Mb for 10,000 generations, resulting in a population that has achieved mutation-drift equilibrium with population genetic parameters similar to what we find in natural diploid animal populations with *Ne* on the order of 10^5 (Allio, Donega, Galtier, & Nabholz, 2017; Stapley, Feulner, Johnston, Santure, & Smadja, 2017). All mutations are neutral in this simulation. We outputted the entire haplotype sequences at the last generation in fasta format. We also output the true allele frequency for each site. Next, for each haplotype sequence, we used ART-MountRainier (W. Huang, Li, Myers, & Marth, 2012) to simulate the sequencing process on an Illumina platform with 150-base paired-end reads and 10x coverage for each haplotype. We then sorted the resulting bam files and merged the two bam files originating from the two haplotypes of each individual. We selected a combination of sample size (5, 10, 20, 40, 60, 80, 160) and coverage per sample (0.25x, 0.5x, 1x, 2x, 4x, 8x) by randomly subsampling these merged bam files. For each of these different combinations of sample size and coverage, we called SNPs and performed genotype likelihoods (using Samtools's genotype likelihood model) and allele frequency estimation using ANGSD-0.931 with the following options `-GL 1 -doGlf 2 -doMaf 1 -doMajorMinor 5 -doCounts 1 -doDepth 1 -dumpCounts 3 -SNP_pval 1e-6 -rmTriallelic 1e-6 -setMinDepth 2 -minInd 1 -minMaf 0.0005 -minQ 20`. We were then able to compare the inferred allele frequencies with the true allele frequencies in the simulated population, and quantify the accuracy in allele frequency estimation by calculating the coefficient of determination (R^2) and root-mean-square error (RMSE) using custom R scripts. We also estimated the sample allele frequency likelihoods (SAF) and subsequently the site frequency spectrum (SFS) using ANGSD. For SAF, we found that a more stringent depth filter has better performance, so we used the following options `-doSaf 1 -GL 1 -doCounts 1 -setMinDepth sample_size*coverage`. For SFS, we found that extending the number of iterations can improve its performance, and thus used realSFS with the following options `-tole 1e-08 -maxIter 1000`. From the estimated SFS, we calculated different estimators of theta (e.g. Watterson's estimator, Tajima's estimator) and performed some neutrality tests (e.g. Tajima's D), also using ANGSD with the following options: `-GL 1 -doSaf 1 -doThetas 1 -doCounts 1 -setMinDepth sample_size*coverage`. Lastly, to compare the performance between different genotype likelihood models, we replicated the entire analysis pipeline using GATK's genotype likelihood model (`-GL 2`).

Then, we tested the power of lcWGS in resolving the genetic structure of spatially distributed populations. Again, we began by randomly creating a starting sequence on a 30Mb chromosome, but this time we created nine populations, each with N of 500. These nine populations are distributed on a three-by-three grid, with a constant bidirectional migration rate (m) equal to 0.002 connecting each pair of adjacent populations (Figure x). Similarly, we scaled up the neutral mutation rate (μ) to 2x10^{-7} per bp per generation, and recombination rate (r) to 50cM/Mb. We ran the simulation for 10,000 generations, resulting in a metapopulation that has achieved mutation-drift-migration equilibrium. This metapopulation consists of nine populations, each with population genetic parameters resembling a diploid animal population with effective population size (Ne) on the order of 10^{4}. We used ART to simulate the sequencing process, and subsampled the bam files to create different combinations of sample size (5, 10, 20, 40, 60, 80) and coverage per sample (0.125x, 0.25x, 0.5x, 1x, 2x, 4x). We called SNPs and estimated genotype likelihoods with the nine populations combined using `-GL 1 -doGlf 2 -doMaf 1 -doMajorMinor 5 -doCounts 1 -doDepth 1 -dumpCounts 1 -doIBS 2 -makematrix 1 -doCov 1 -P 6 -SNP_pval 1e-6 -rmTriallelic 1e-6 -setMinDepth 2 -minInd 1 -minMaf 0.05 -minQ 20` in ANGSD. This step outputs a covariance matrix (`-doCov 1`) and a distance matrix (`-doIBS 2`) among individuals, and in addition to these, we also used PCAngsd (Meisner & Albrechtsen, 2018) to generate another covariance matrix using the estimated genotype likelihoods. Using the `eigen()` function and the `cmdscale()` function in R, we conducted principal component analysis (PCA) and principal coordinate analysis (PCoA) with these covariances matrices and distance matrix, respectively, plotted the samples on the first two principal components / principal coordinates, and compared these with the true spatial structure that was simulated. Also, we performed PCA with the true sample genotypes using PLINK2 as an additional comparison.

Lastly, we tested the power of lcWGS in detecting signatures of divergent selection between two populations connected by gene flow. This simulation consists of two stages: a neutral burn-in stage and a selection stage. Two populations under mutation-drift-migration equilibrium are created in the burn-in stage, and then selection is imposed on these populations in the selection stage. In the burn-in stage, we began by randomly creating a starting sequence on a 30Mb chromosome and two populations, each with a population size (N) of 500, and with a constant bidirectional migration rate (m) between them. We used a scaled-up recombination rate (r) and neutral mutation rate (μ), ran the simulation for 5,000 generations, and outputted the entire populations. In the first generation of the selection stage, we read the output from the burn-in stage into SLiM, selected 11 evenly distributed positions on the chromosome, and at each of these positions we added a non-neutral mutation to one randomly sampled genome in the first population. These mutations were set to be beneficial in the first population with a certain selection coefficient (s) and deleterious in the second population with a selection coefficient of (1/*s*). Despite this, since these non-neutral mutations each exist in a single copy, a majority of them are likely to get lost in the first few generations of the selection due to drift, in which case the simulation needs to be reset. To avoid resetting the simulation too many times (which can take a long time), we instantly expanded the population size by a factor of 10 (to 5,000) in each population after introducing the non-neutral mutations, which would then exist in multiple copies. Correspondingly, we scaled down the original *m*, *r*, and *μ* by a factor of 10, in order to preserve the key population genomic parameters of the simulated populations. We ran the simulation for an additional 200 generations. If more than half of the selected alleles become lost due to drift or Hill-Robertson interference during the process, we restart from the beginning of the selection stage with a different random seed (the same burn-in is always used). After the selection stage is completed, the SNP density is mainly determined by the mutation rate (*μ*), the background level of differentiation between the two populations is mainly determined by the migration rate (*m*), the level of differentiation at the selected locus is mainly determined by both the selection coefficient (*s*) and the migration rate (*m*), and the width of the genomic region that shows high differentiation between the two populations is mainly determined by the recombination rate (*r*). We were therefore able to create population pairs with different genomic landscapes of differentiation by reiterating this process with different combinations of mutation rate (*μ*), selection coefficients (*s*), migration rates (*m*), and recombination rates (*r*) (Table S2). Then, we again subsampled each population, and used ART to simulate the sequencing process with the same combinations of sample size (5, 10, 20, 40, 60, 80, 160) and coverage per sample (0.25x, 0.5x, 1x, 2x, 4x, 8x) as in our neutral model. Using ANGSD, we called SNPs with the two populations combined through `-dosaf 1 -GL 1 -doGlf 2 -doMaf 1 -doMajorMinor 5 -doCounts 1 -doDepth 1 -dumpCounts 1 -SNP_pval 1e-6 -rmTriallelic 1e-6 -setMinDepth 2 -minInd 1 -minMaf 0.0005 -minQ 20`, estimated genotype likelihoods and allele frequencies for each population through `-dosaf 1 -GL 1 -doGlf 2 -doMaf 1 -doMajorMinor 5 -doCounts 1 -doDepth 1 -dumpCounts 1 -setMinDepth 1 -minInd 1 -minQ 20`, and finally estimated per-SNP Fst between the population pair from the two-dimensional site frequency spectrum estimated from realSFS using the default option. Using custom R scripts, we visualized and compared the Fst landscape under different simulation scenarios and sequencing strategies.

In addition to these investigations on different sequencing designs of lcWGS, we have also compared lcWGS with two other commonly used high-throughput sequencing strategies, namely Pool-seq and RAD-seq. With Pool-seq, we were mainly interested in its accuracy in allele frequency estimation (in comparison to the estimation with individually barcoded lcWGS samples), particularly when the sequencing yield from different individuals in the pool is uneven. Therefore, we simulated Pool-seq with our neutral model under two different scenarios. In the first scenario, we assumed that the sequencing yield is equal among individuals. In this case, the simulation and analysis is exactly the same as in lcWGS until the last step, where instead of using the allele frequency estimates outputted by ANGSD, we calculated allele frequencies based on the allele counts in the population instead (this was generated by `-doCounts 1 -dumpCounts 1`). In the second scenario, we kept the total sequencing yield to be the same, but added variation in the contribution of each individual to the pool. To do this, we sampled each individual's sequencing yield from an empirical distribution, which we obtained by subsampling and rescaling the individual sequencing yield from three of our lcWGS projects where we tried our best effort to generate even yield among samples by pooling by DNA molarity. These empirical sequencing yields have a right-skewed distribution with a standard deviation that is 60% of the mean (Figure

S7). We subsampled each individual bam file according to its target yield, and inputted these subsampled bam files to the same ANGSD pipeline for SNP calling, genotype likelihoods estimation, and allele frequency estimation. Allele frequency estimates outputted by the pipeline would represent the result from lcWGS, and allele frequencies calculated from allele counts would represent the estimates from Pool-seq. We again calculated R^2 and RMSE from these allele frequency estimates as a measure of their accuracy.

With RAD-seq, we were mainly interested in its power in identifying genomic islands of differentiation. Therefore, we simulated RAD-seq with our divergent selection model. We assumed that with the high coverage of RAD-seq, genotypes can always be called correctly, so we used true genotypes instead of simulating the sequencing process. We used R to randomly sample 150-bp fragments on our 30Mb genome as our RAD tags at a range of different densities (4, 8, 16, 32, 64, and 128 per Mb), obtained each sample's true genotype at these fragments, and calculated sample allele frequencies. We used these allele frequencies to estimate per-SNP Fst (Fst = 1 - H_S / H_T), visualized and then compared these Fst results with those from lcWGS simulation.

Section 5: Analysis of down-sampled *Heliconius* data

To determine the effect of sequencing coverage on our ability to detect local signatures of differentiation and global population structure we re-analysed *Heliconius spp*. whole-genome data from (Van Belleghem et al., 2017). Raw whole-genome data for 70 H. erato individuals were downloaded from NCBI (Supplementary Table Sx) and mapped to the H. erato demophoon reference genome (*Heliconius_erato_demophoon_v1*) using BOWTIE2 (Langmead & Salzberg, 2013) using the --very-sensitive setting. Reads with mapping qualities (MAPQ) below 20 were filtered out and the remaining reads sorted using SAMTOOLS v.1.9 (Heng Li et al., 2009). Duplicated reads were removed using MARKDUPLICATES v.2.9.0 from PICARD TOOLS and reads realigned around indels using PICARD.

Subsequently, we subsampled each filtered bam file based on the fraction of reads to an approximated coverage of 8x (30M reads per individual), 4x (15M reads), 2x (7.5M reads), 1x(3.75M reads) and 0.5x(1.625M reads) using SAMTOOLS. Individuals with insufficient coverage for a mean of 8x were filtered out (2 individuals).

To determine how the ability to detect local signatures of differentiation differs with coverage, we estimated Fst between individuals with red-bar and no red-bar along the genomic scaffold containing the underlying gene optix (scaffold Herato1801:) (Van Belleghem et al., 2017). Individuals with the same phenotypes were pooled across sampling sites and subspecies to achieve sample sizes of 23 red-barred individuals (*H. e. demophoon*, *H. e. favorinus*; *H. e. hydara* and *H. e. notabilis*) and 28 non-barred individuals (*H. e. amalfreda*, *H. e. emma*; *H. e. erato*; *H. e. lativitta* and *H. e. etylus*). Using each set of subsampled bam file, we identified variant sites across scaffold Herato1801 using ANGSD v.0.28 with the following criteria: SNP_p-val=1e-6; minDepth = Number of individuals * 0.1x; maxDepth = coverage * N.ind + (2 * coverage *N.ind); minInd=75% of individuals (= 40); minQ = 30; and minMAF=0.05 (Korneliussen, Albrechtsen, & Nielsen, 2014). Fst values were estimated based on these variant sites (-sites option) in ANGSD based on genotype likelihoods in 50kb sliding windows with a 20kb step size to make them comparable to results in (Van Belleghem et al., 2017).

To understand how the sequencing coverage affects the ability to detect global population structure in Heliconius, we performed a principal components analysis for all individuals at each coverage based on covariance matrices estimated in ANGSD. Covariance matrices were estimated using a random-read sampling procedure in ANGSD and PCA was performed using the eigen function in R. All results were plotted in R using ggplot.

Section 6: Genotype imputation analysis

To explore imputation performance under different scenarios, we simulated a 30Mb chromosome for three neutrally evolving populations that have reached mutation-drift equilibrium, as described above. We set the mutation rate (μ) to be 1x10-8/bp/generation for all three populations, and altered their effective population size (Ne) and recombination rate (r) to create three different scenarios with different levels of genetic diversity and linkage disequilibrium (LD). (Note that in a neutral population, genetic diversity is proportional to the product of effective population size and mutation rate, whereas LD is inversely proportional to the product of effective population size and recombination rate.) The three scenarios include the following: 1) a low diversity, high LD scenario (r = 0.5 cM/Mb, Ne = 1,000); 2) a medium diversity, medium LD scenario (r = 0.5 cM/Mb, Ne = 10,000); and 3) a medium diversity, low LD scenario (r = 2.5, Ne = 10,000). In order to generate a large sample from a single, neutrally evolving population of stable size, we sampled with replacement 2n haplotypes (n diploid individuals) from the offspring of the final generation of the simulation, for sample sizes n=25, 100, 250, N=500 and N=1000. Bamfiles were simulated using ART-MountRainier as described above to average depths of 1x, 2x and 4x. For each dataset (five sample sizes x three depths x three population scenarios = 45 datasets), we first called SNPs in ANGSD using the settings (-GL 1 -doGlf 2 -doMaf 1 -doMajorMinor 5 -doCounts 1 -doDepth 1 -dumpCounts 3 -P 6 -SNP_pval 1e-6 -rmTriallelic 1e-6 -setMinDepth 2 -minInd 1 -minMaf 0.0005 -minQ 20).

For each scenario, we evaluated the accuracy of genotype dosages and genotypes called using imputation without a reference panel in the programs Beagle v.3.3.2 and STITCH v.3.6.2. We compared the imputed results with called genotypes and estimated genotype dosages without imputation in ANGSD. (Although ANGSD recommends basing downstream analyses on genotype likelihoods rather than called genotypes, we use it as a baseline for evaluating any improvement of genotype calls by imputation.)

We called genotypes at variable sites using the three methods (no imputation, imputation in Beagle, and imputation in STITCH). We called non-imputed genotypes directly from the posterior genotype probability in ANGSD, using minor allele frequencies as a prior and a posterior probability cutoff of 0.90 (-postCutoff 0.90 -doPost 1 -doMaf 1 -GL 2 -dogeno 5 -doMajorMinor 3). Because ANGSD does not directly output genotype dosages, we converted posterior genotype probabilities using the formula genotype dosage=P(AA | data)*0 + P(AB | data)*1 + P(BB | data)*2. For imputation in STITCH, we explored performance under varying settings of the parameter K (K=25, 30 and 35), and examined output plots as well as r2 values between simulated genotypes and imputation dosages. In most cases K=30 performed best or very close to best; thus, we used the settings K=30, nGen=10, and S=4, and called genotypes with posterior probability ≥ 0.90. For imputation in Beagle, we passed genotype likelihoods estimated in ANGSD directly to Beagle for imputation under default settings. We called genotypes from posterior genotype probability threshold of 0.9 using the script gprobs2beagle.jar (https://faculty.washington.edu/browning/beagle_utilities/utilities.html). We evaluated the performance of each method in the following ways, by the proportion of correct genotype calls (genotype concordance), the proportion of genotypes actually called, and by the r2 between allelic dosage and true genotypes within allele frequency bins of size 0.05. We report average values for all sites with MAF>0.05, excluding variant sites that were not identified (false negatives) or non-variant sites called as SNPs (false positives) in the ANGSD SNP-calling step.

Sensitivity of population genomic inference power to simulation assumptions

In Section 4 of this paper, we have tested the performance of different types of population genomic inference under different lcWGS experimental designs using forward genetic simulation. We found that for most of these analyses, distributing the same amount of sequencing effort across more samples can consistently improve inference power. This conclusion should be relatively robust regardless of the parameter settings in our simulation model, although the power of inference under each combination of sample size and coverage can be strongly affected by these model assumptions. Here, we briefly present a qualitative discussion on how the power of different types of population genomic inference could be impacted by different parameter choices in the simulation.

Section 4.1: Given the same true allele frequency, the accuracy of allele frequency estimation at a single SNP should be largely independent of simulation parameters other than sample size and coverage. The values of RMSE and r^2 genome-wide, however, will be sensitive to the site frequency spectrum (SFS) in the simulated data, since errors are strongly affected by the true allele frequencies (Figure 2). As a result, any processes that can skew the SFS (e.g. demographic expansion and contraction, selection) could affect the values of RMSE and r^2, although the directionality of the change is context dependent.

Section 4.2: For the inference of spatial structure, higher migration rate is an obvious driver for lower inference power (Figure 4). We have also shown that with more SNPS (which can result from a larger genome, larger population size, or higher mutation rate), inference power can improve (Figure S9). On the other hand, stronger LD (caused by lower population size or lower recombination rate) should decrease the power of inference, since SNPs can become highly correlated with each other, resulting in fewer independent SNPs that are informative.

Section 4.3: Similarly, a larger number of SNPs in the dataset due to higher mutation rate can also lead to higher power to locate the region under divergent selection, as a window-based approach can have more information to work with. Stronger LD due to lower recombination rate generates more distinct patterns of linked selection, therefore also enhances the power to locate the general region of interest. Both factors, however, have a more complex effect on the power to locate the causal SNPs due to the higher number of linked neutral SNPs that potentially become false positives. Stronger divergent selection should be able to more reliably increase the detection power of both the general region of interest and the causal SNPs. Lastly, the effects of population size and migration rate is also complex. On the one hand, higher population size leads to more SNPs in the dataset. On the other hand, it can result in narrower peaks that are more difficult to detect due to reduced LD. Lower migration rate increases the Fst values of the selected SNPs, but also increases the background noise. A more quantitative power analysis is therefore warranted to better understand the effect of these simulation parameters.

References for software in Table 2 of the main text

Angsd (Korneliussen et al., 2014)
Atlas (Link et al., 2017)
MAPGD (Maruki & Lynch, 2015)
GPAT (Domyan et al., 2016)
ngsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014)
PCAngsd (Meisner & Albrechtsen, 2018)
GATK (McKenna et al., 2010)
Freebayes (Garrison & Marth, 2012)
Reveel (L. Huang, Wang, Chen, Bercovici, & Batzoglou, 2016)
EBG (Blischak, Kubatko, & Wolfe, 2018)
BaseVar (Liu et al., 2018)
Heterozygosity-em (Bryc, Patterson, & Reich, 2013) (https://github.com/kasia1/heterozygosity-em)
ngsF (Vieira, Fumagalli, Albrechtsen, & Nielsen, 2013)
ngsRelate (Korneliussen & Moltke, 2015)
ngsF-HMM (Vieira, Albrechtsen, & Nielsen, 2016)
Bcftools/ROH (Narasimhan et al., 2016)
skmer (Sarmashghi, Bohmann, P Gilbert, Bafna, & Mirarab, 2019)
ngsDist (Vieira, Lassalle, Korneliussen, & Fumagalli, 2016)
lostruct (Han Li & Ralph, 2019)
ngsAdmix (Skotte, Korneliussen, & Albrechtsen, 2013)
Entropy (Gompert et al., 2014)
Ohana (Cheng, Mailund, & Nielsen, 2017; Cheng, Racimo, & Nielsen, 2019)
evalAdmix (Garcia-Erill & Albrechtsen, 2020)
AlphaAssign (Whalen, Gorjanc, & Hickey, 2019)
WHODAD (Snyder-Mackler et al., 2016)
ngsLD (Fox, Wright, Fumagalli, & Vieira, 2019)
GUS-LD (Bilton et al., 2018)
PopLD (Maruki & Lynch, 2014)
SNPTEST (Marchini, Howie, Myers, McVean, & Donnelly, 2007)
svgem (Lucas-Lledó, Vicente-Salvador, Aguado, & Cáceres, 2014)
HMMploidy (https://github.com/SamueleSoraggi/HMMploidy)
loimpute (Wasik et al., 2019)
STITCH (Davies, Flint, Myers, & Mott, 2016)
NOISYmputer (Lorieux, Gkanogiannis, Fragoso, & Rami, 2019)
LB-Impute (https://github.com/dellaporta-laboratory/LB-Impute)
LinkImpute (Money et al., 2015)
LepMap3 (Rastas 2017)

Supplementary tables

Table S1. Heliconius erato short read archive (SRA) IDs. Individuals used for the subsampling and genotype-likelihood-based analysis of H. erato subspecies, with SRA ID and subspecies names. Samples from (Van Belleghem et al., 2017).

SRA ID	***H. erato* subspecies**
SRS1618075	amalfreda
SRS1618086	amalfreda
SRS1618008	amalfreda
SRS1618009	amalfreda
SRS1618010	amalfreda
SRS1618033	emma
SRS1618034	emma
SRS1618062	emma
SRS1618063	emma
SRS1618065	emma
SRS1618066	emma
SRS1618067	emma
SRS1618069	erato
SRS1618070	erato
SRS1618071	erato
SRS1618072	erato
SRS1618073	erato
SRS1618084	erato
SRS1618014	etylus
SRS1618015	etylus
SRS1618016	etylus
SRS1618017	etylus
SRS1618018	etylus
SRS1618053	lativitta
SRS1618044	lativitta

SRS1618045	lativitta
SRS1618046	lativitta
SRS1618047	lativitta
SRS1618002	demophoon
SRS1618093	demophoon
SRS1618094	demophoon
SRS1618098	demophoon
SRS1618100	demophoon
SRS1617995	demophoon
SRS1618032	favorinus
SRS1618057	favorinus
SRS1618056	favorinus
SRS1618058	favorinus
SRS1618059	favorinus
SRS1618060	favorinus
SRS1618083	favorinus
SRS1618102	hydara
SRS1617999	hydara
SRS1618068	hydara
SRS1618074	hydara
SRS1618087	hydara
SRS1618101	hydara
SRS1618005	notabilis
SRS1618012	notabilis
SRS1618090	notabilis
SRS1618091	notabilis

Table S2. Parameter settings for the simulation of divergent selection

Scenario	*N*	*μ* (per bp per generation)	*r* (cM/Mb)	*m* (per generation)	*s*
High Ne, high gene flow	500	10^{-6}	250	0.01	0.08
Low Ne, low gene flow	500	$2*10^{-7}$	50	0.005	0.08

Supplementary figures

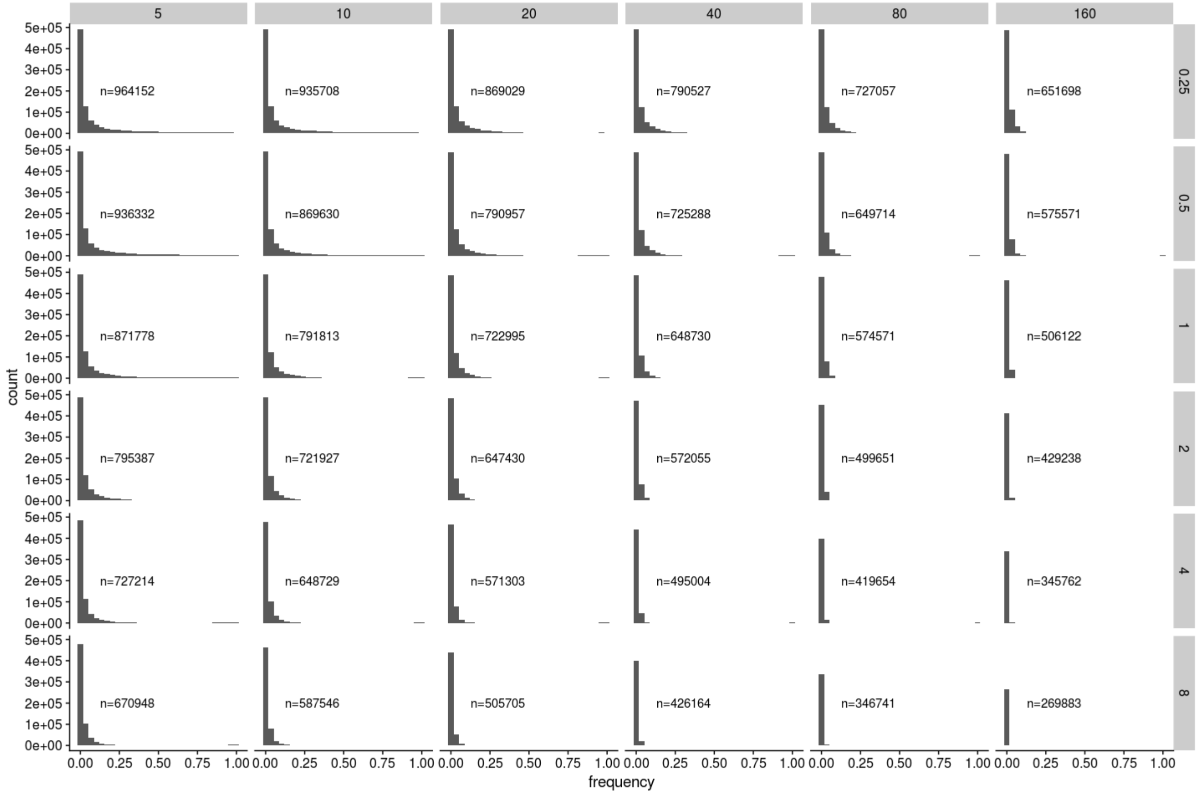

Figure S1. Histogram of the allele frequencies of false negative SNPs with lcWGS. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right.

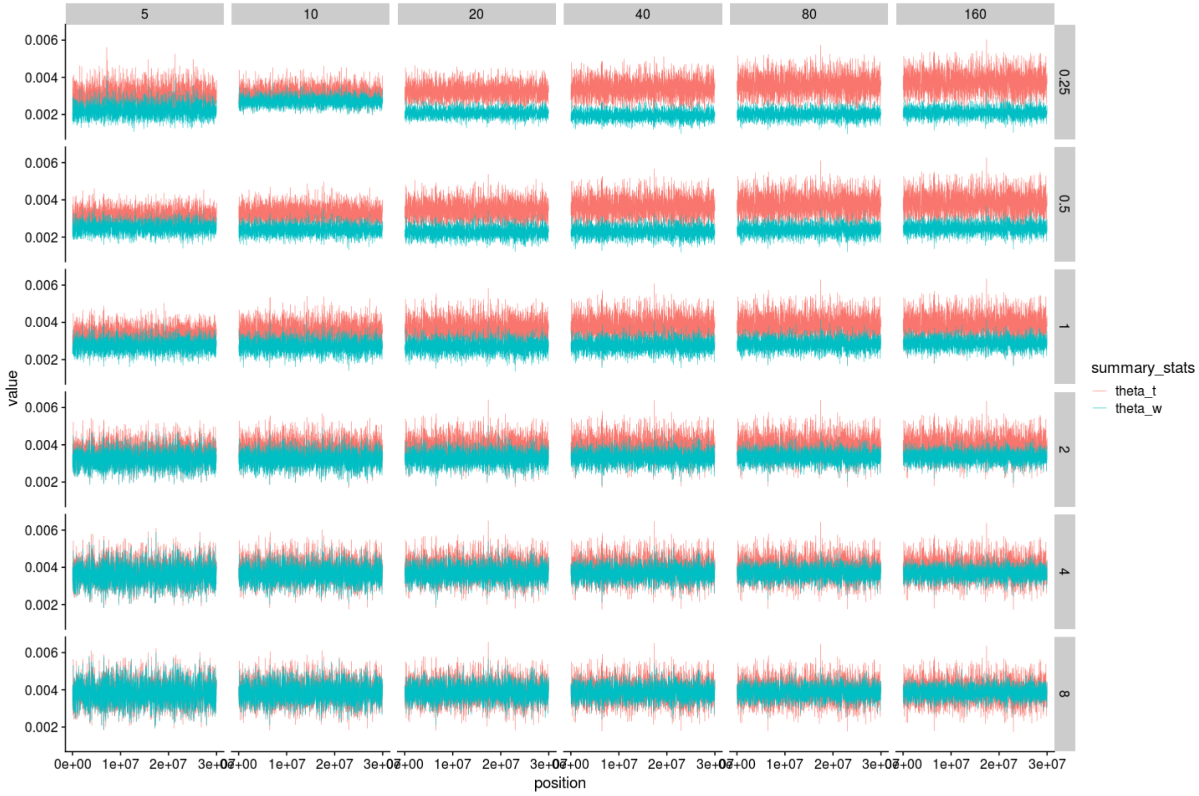

Figure S2. Tajima's θ (aka π) and Watterson's θ estimated using Samtool's genotype likelihood model with ANGSD in 10kb windows. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. The true chromosome-average values for both statistics should be 0.004.

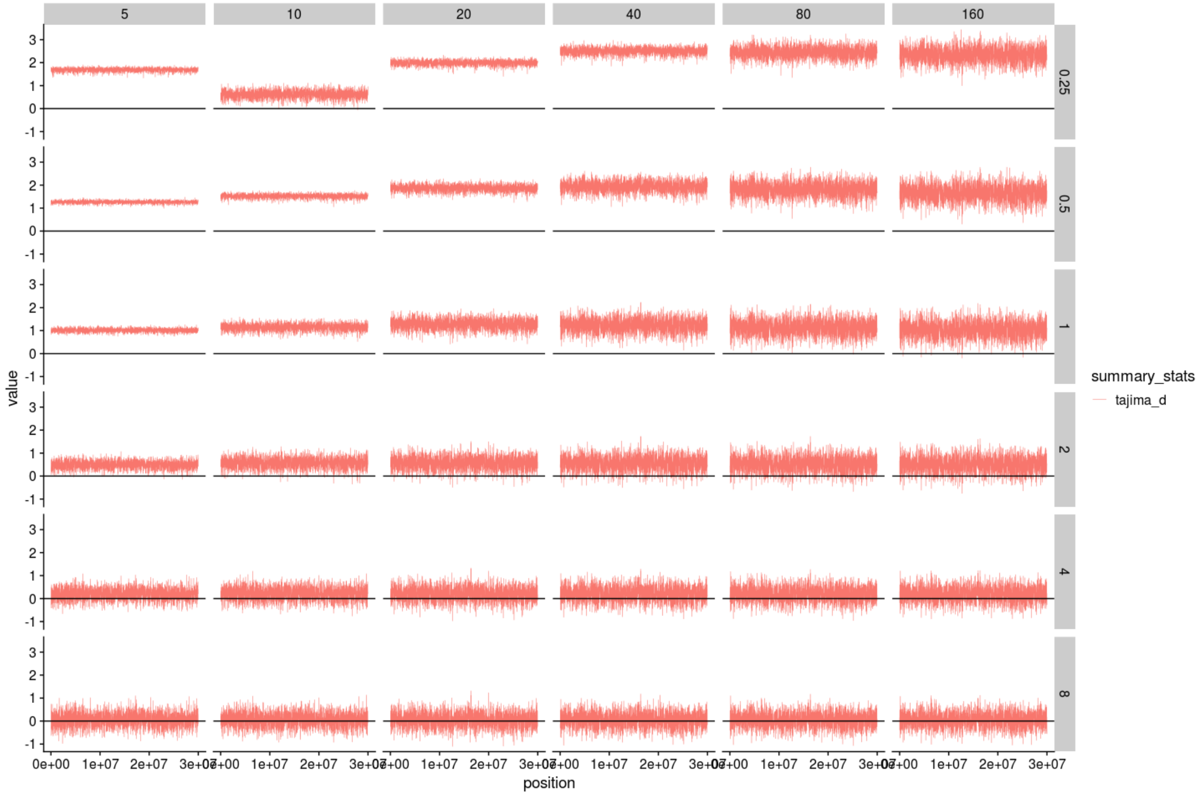

Figure S3. Tajima's D estimated using Samtool's genotype likelihood model with ANGSD in 10kb windows. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. The true chromosome-average Tajima's D should be 0, which is marked by a black line.

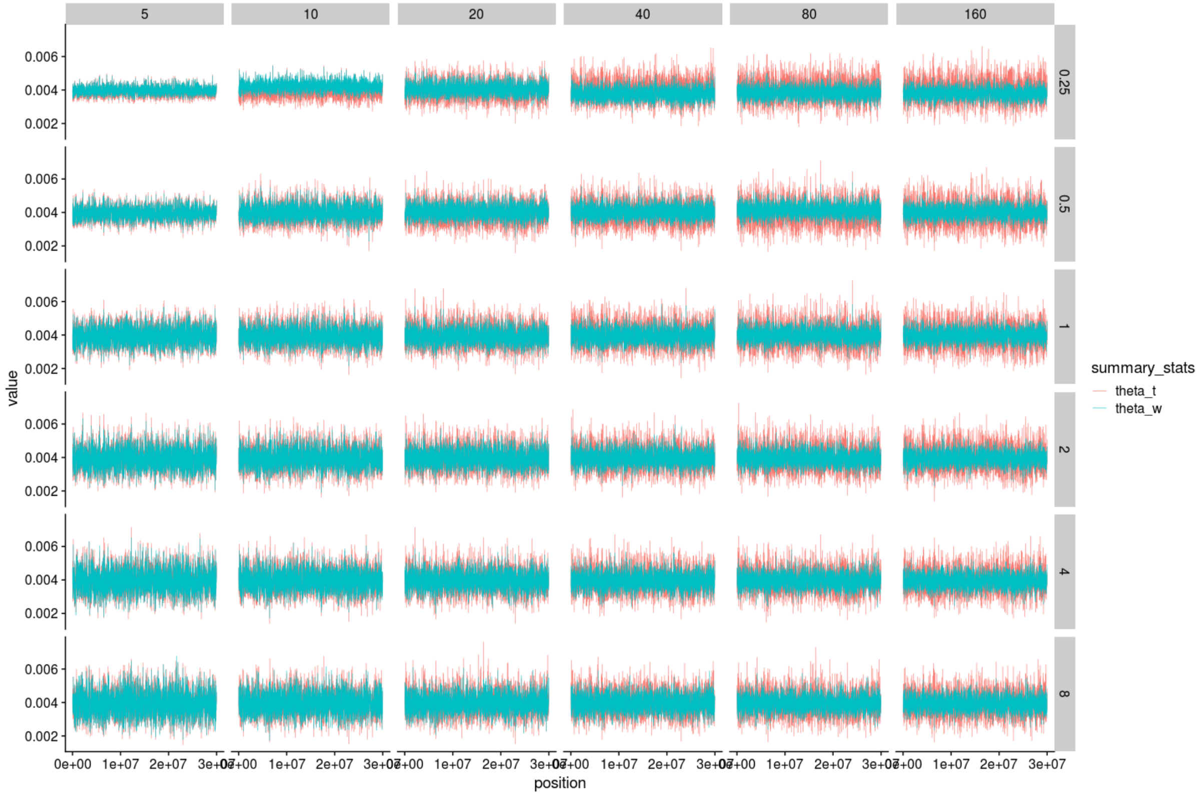

Figure S4. Tajima's θ (aka π) and Watterson's θ estimated using GATK's genotype likelihood model with ANGSD in 10kb windows. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. The true chromosome-average values for both statistics should be 0.004.

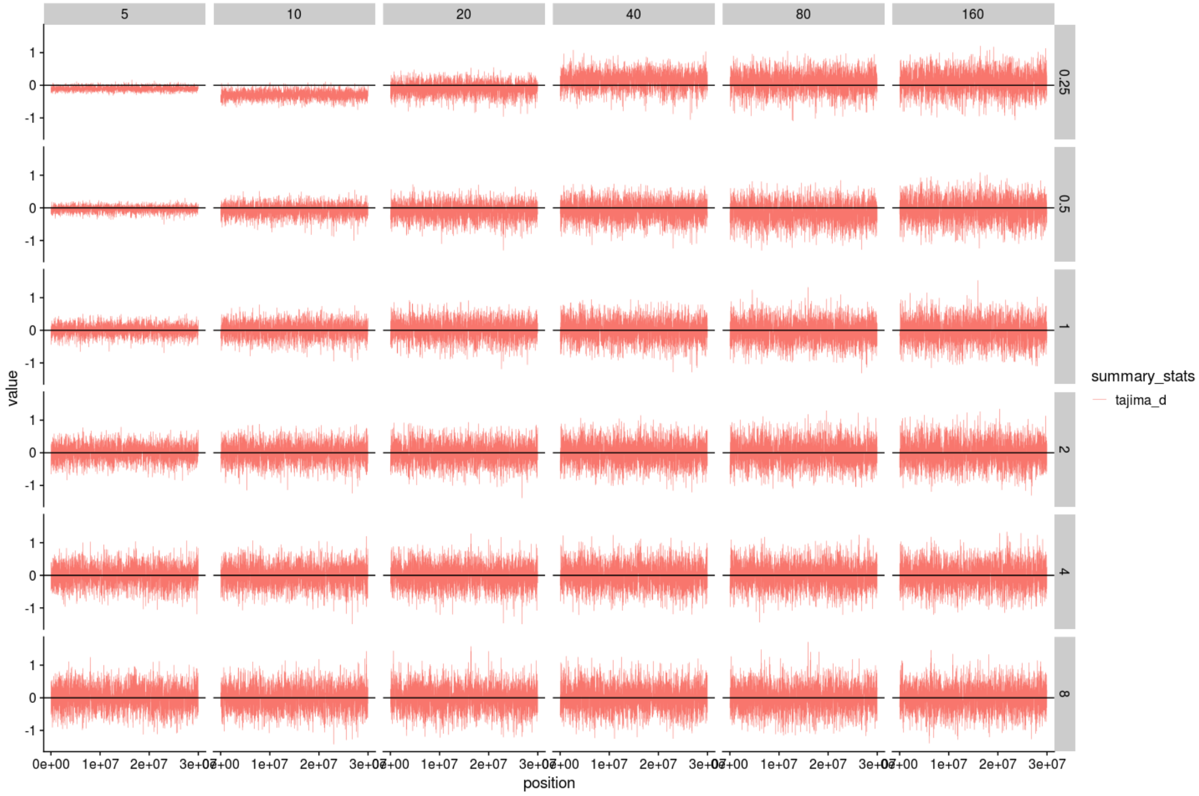

Figure S5. Tajima’s D estimated using GATK’s genotype likelihood model with ANGSD in 10kb windows. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. The true chromosome-average Tajima’s D should be 0, which is marked by a black line.

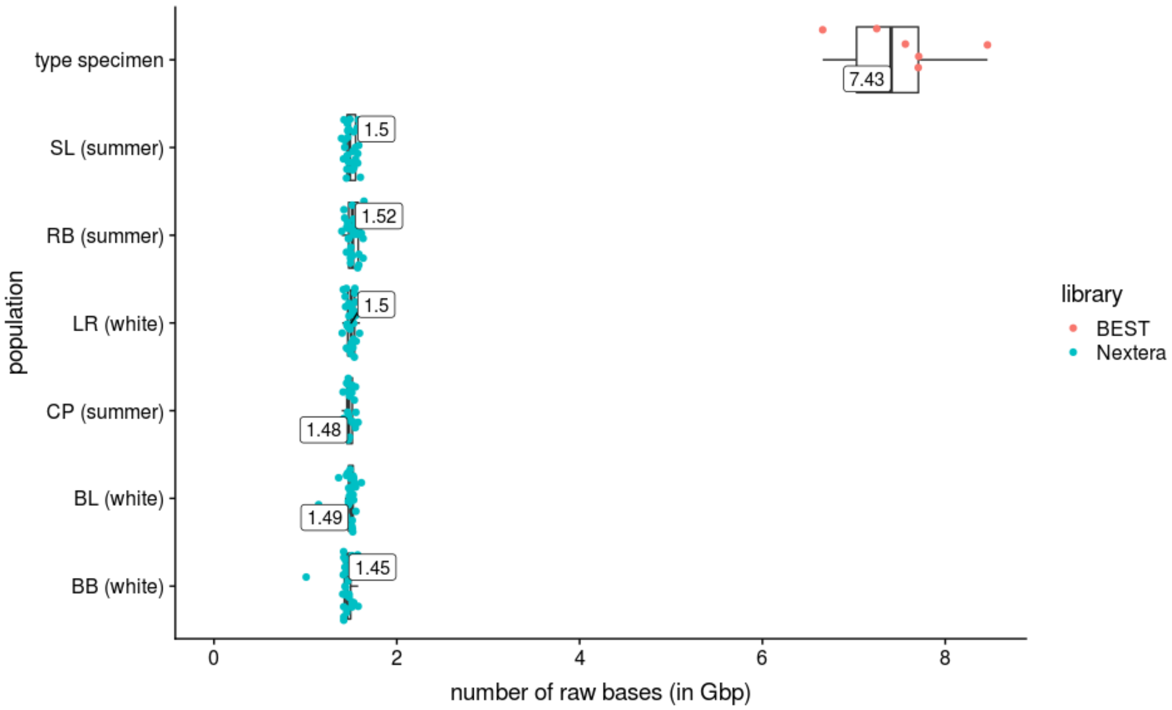

Figure S6. An empirical example from one of our lcWGS projects of the distribution of raw sequencing yield from individual samples when they are repooled based on the first round of sequencing. This is to demonstrate that equal distribution of sequencing effort can be approximated by such a sequencing design. (The type specimens were designed to have higher sequencing yield then other samples.)

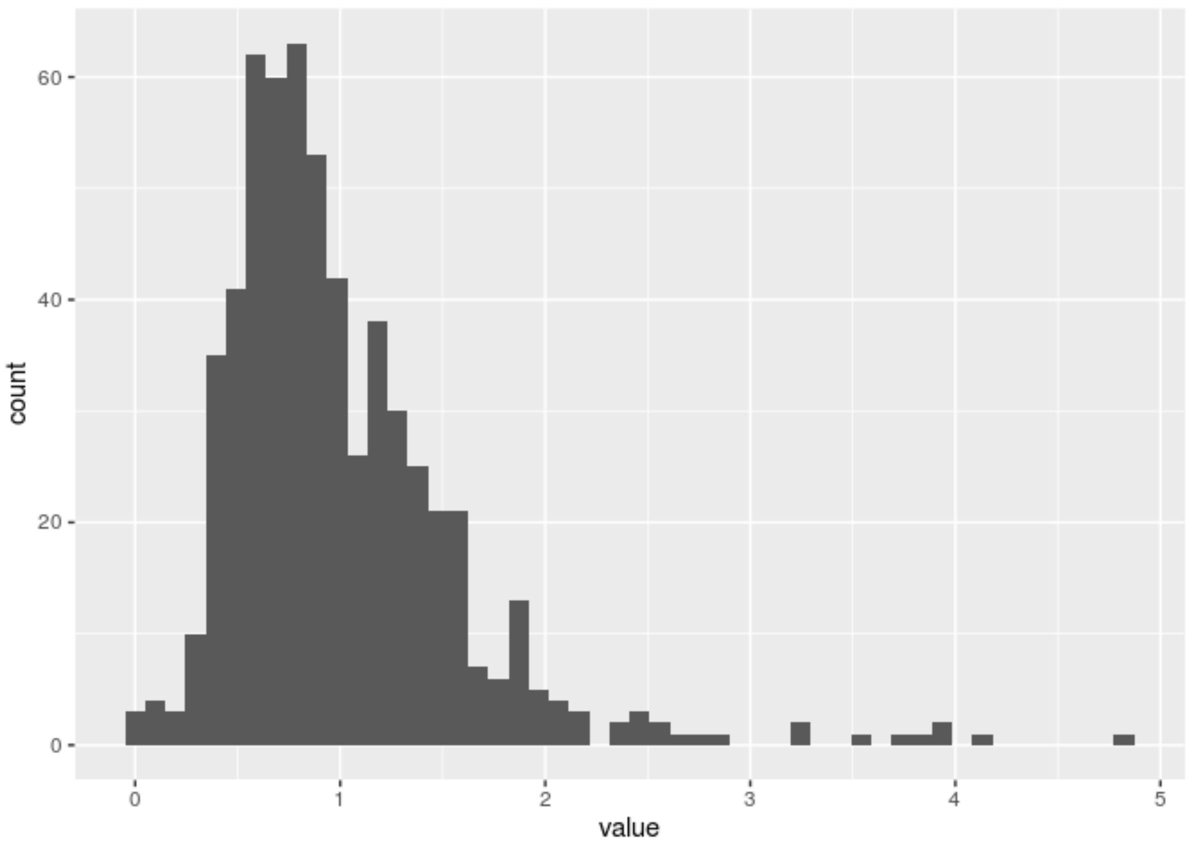

Figure S7. The sequencing coverage distribution that we sampled from when simulating uneven sequencing coverage among samples. This distribution is obtained by merging the distributions of coverage among samples from three of our lcWGS projects where we pooled samples by molarity.

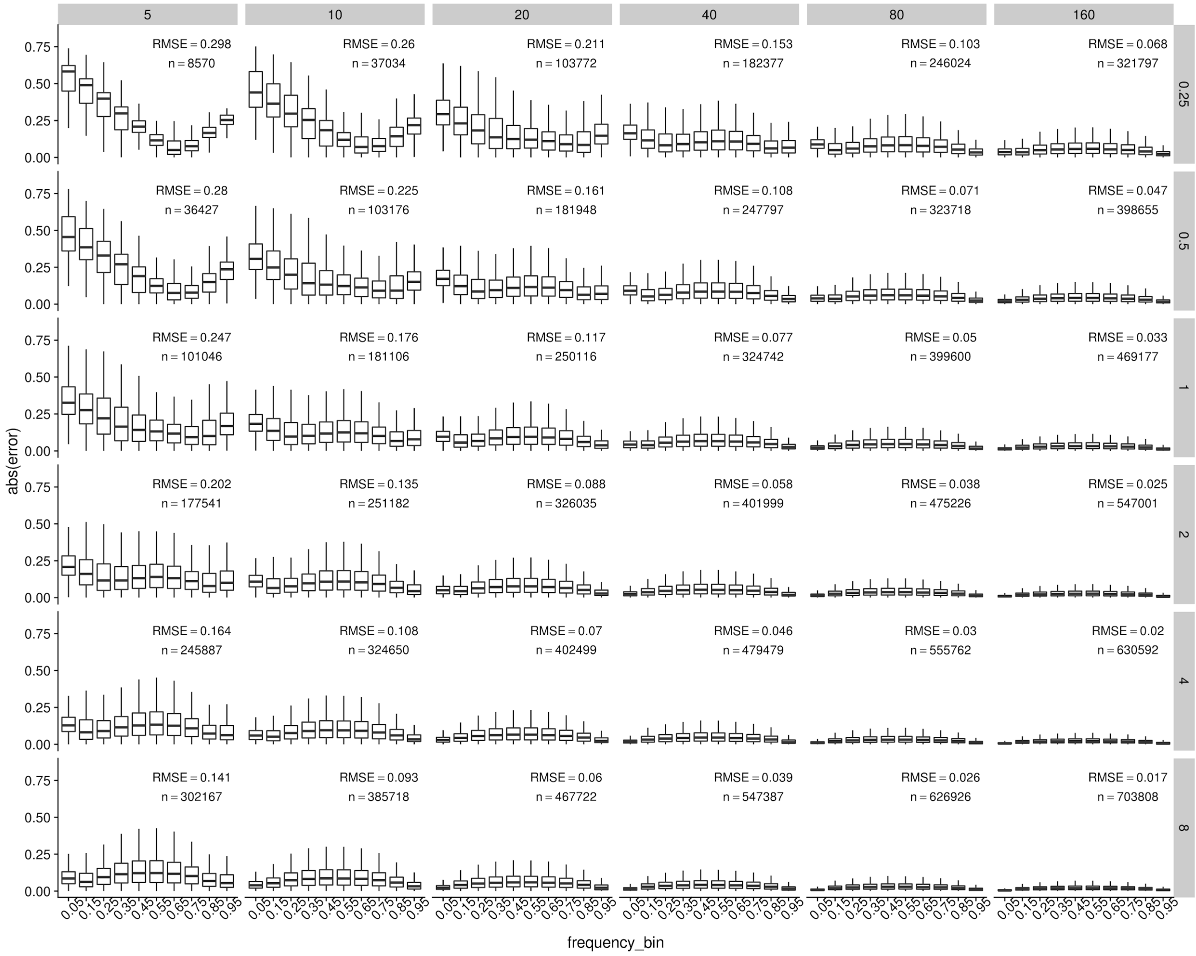

Figure S8. The error in allele frequency estimation with Pool-seq when sequencing effort is evenly distributed among samples. Derived alleles are binned according to their true frequencies on the x axis, and their absolute errors (|estimated frequency - true frequency|) are shown on the y-axis. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. The root mean squared error (RMSE) and the number of SNPs called (SNP count; this includes the true positives and the false positives) are shown in each facet.

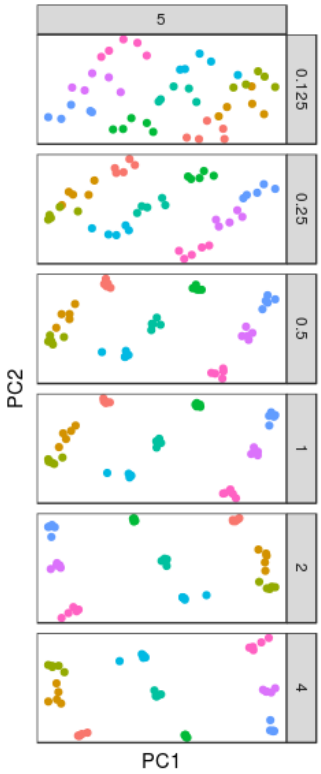

Figure S9. The spatial population structures inferred through principal component analysis (PCA) with lcWGS data using PCA. The first two principal components are shown. This result is from our higher gene flow scenario (an average of 1 effective migrant from one population to another every generation), but a longer chromosome is simulated (300Mb, or 10 times longer than the scenarios shown in Figure 4). Sample size remains five per sample, and coverage increases from top to bottom.

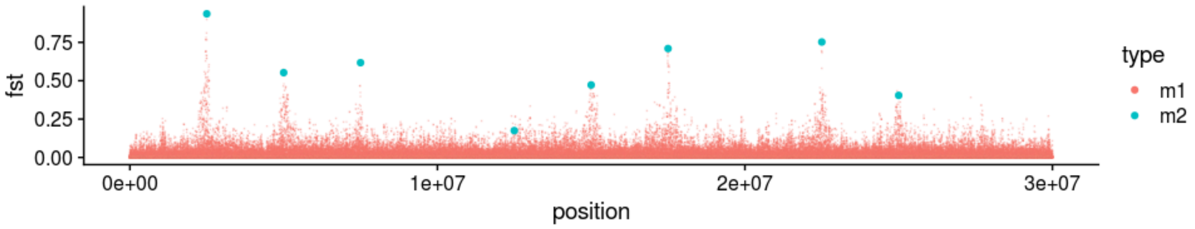

Figure S10. The true per-SNP F_{ST} values along the chromosome between the two simulated populations in a scenario with smaller N_e ($N_e = 10^4$) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). Neutral SNPs (m1) are shown in red and selected SNPs (m2) are shown in blue.

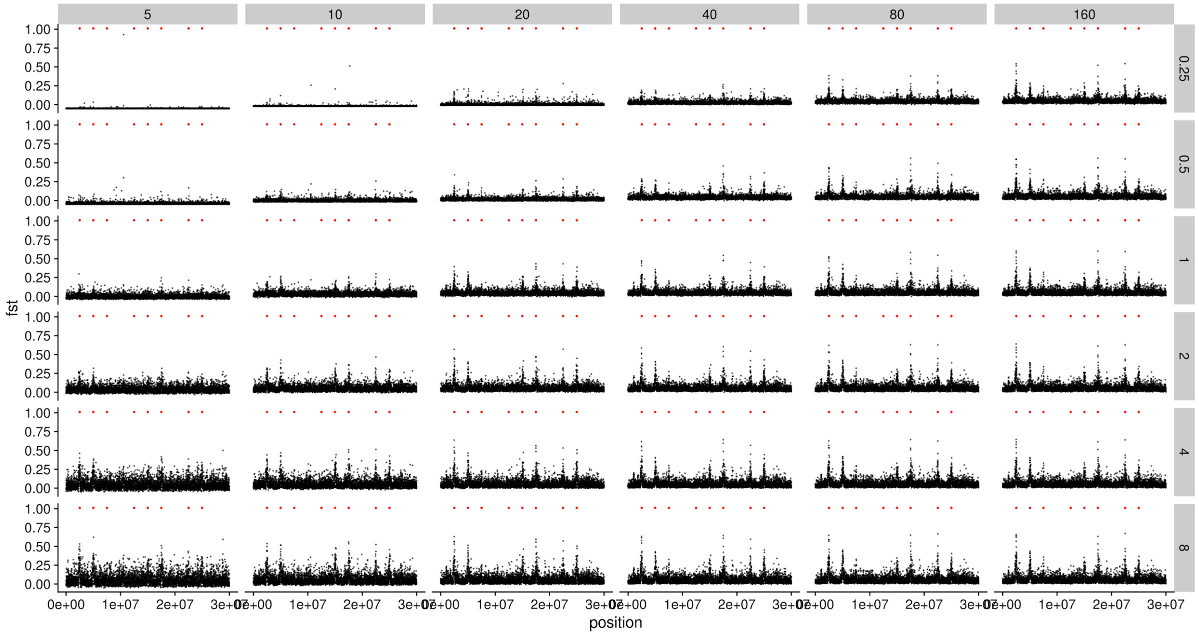

Figure S11. Genome-wide scan for divergent selection with lcWGS data in a scenario with smaller N_e ($N_e = 10^4$) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). The F_{ST} values inferred from lcWGS data in 1kb windows along the chromosome are shown on the y axis. Sample size increases from left to right, and coverage increases from top to bottom. The black points mark both the selected and neutral SNPs, and the red points only mark the positions of the selected SNPs (not their inferred Fst values).

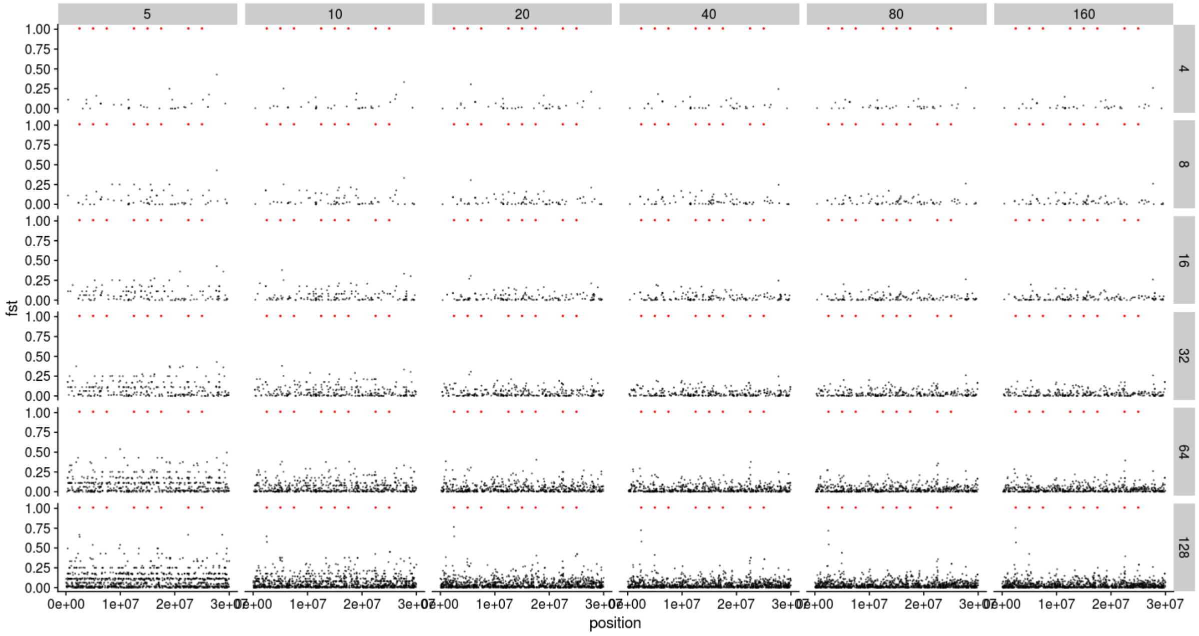

Figure S12. Genome-wide scan for divergent selection with RAD-seq data in a scenario with smaller N_e ($N_e = 10^4$) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). The per-SNP F_{ST} values inferred from RAD-seq data are shown on the y axis and the SNP positions are shown on the x axis. Sample size increases from left to right, and RAD-tag density increases from top to bottom. The black points mark both the selected and neutral SNPs, and the red points only mark the positions of the selected SNPs (not their inferred Fst values).

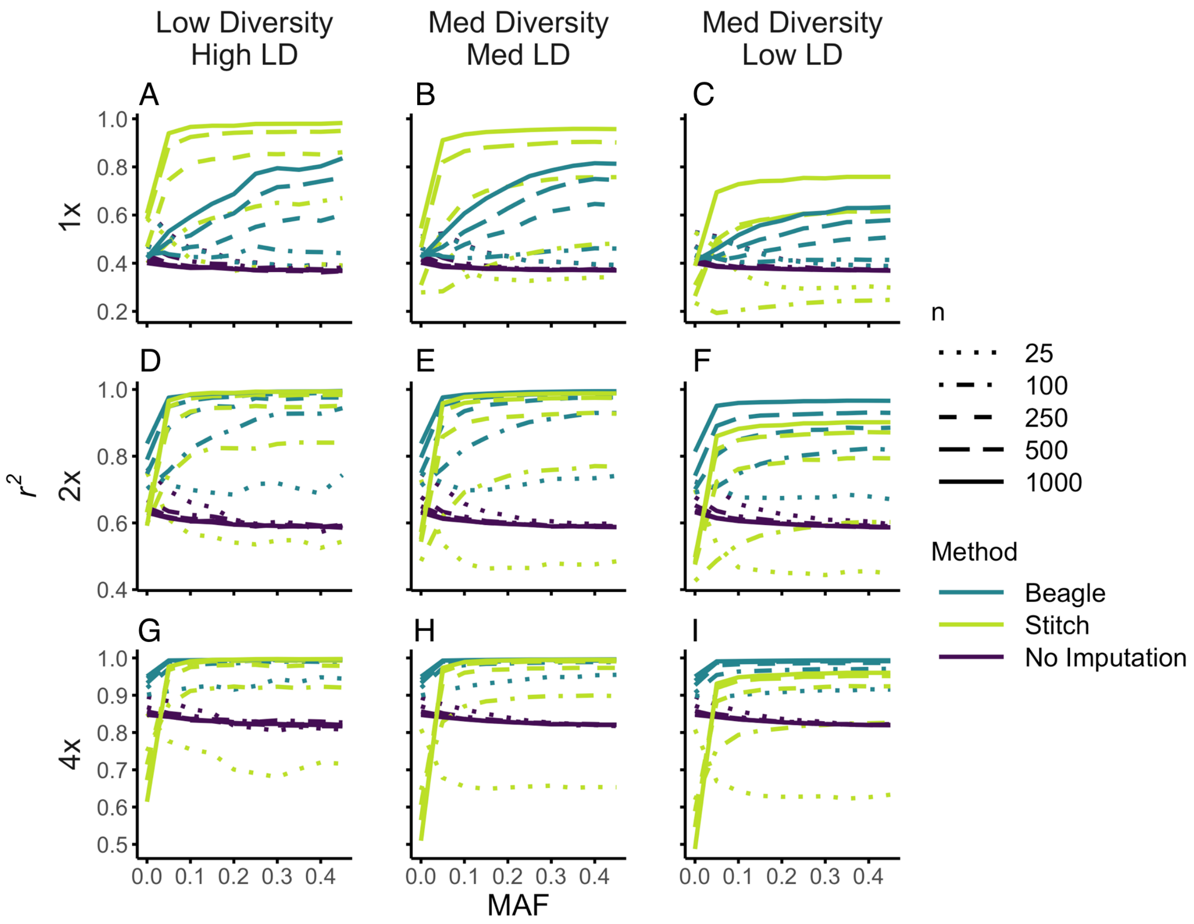

Figure S13. Genotype estimation accuracy (r^2) by minor allele frequency (MAF) for imputation in STITCH and Beagle compared to posterior genotypes estimated without imputation. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (plots in rows correspond to 1x, 2x and 4x coverage) were tested for each method (line colors) under different diversity and linkage disequilibrium scenarios. Note the different y-axis scales.

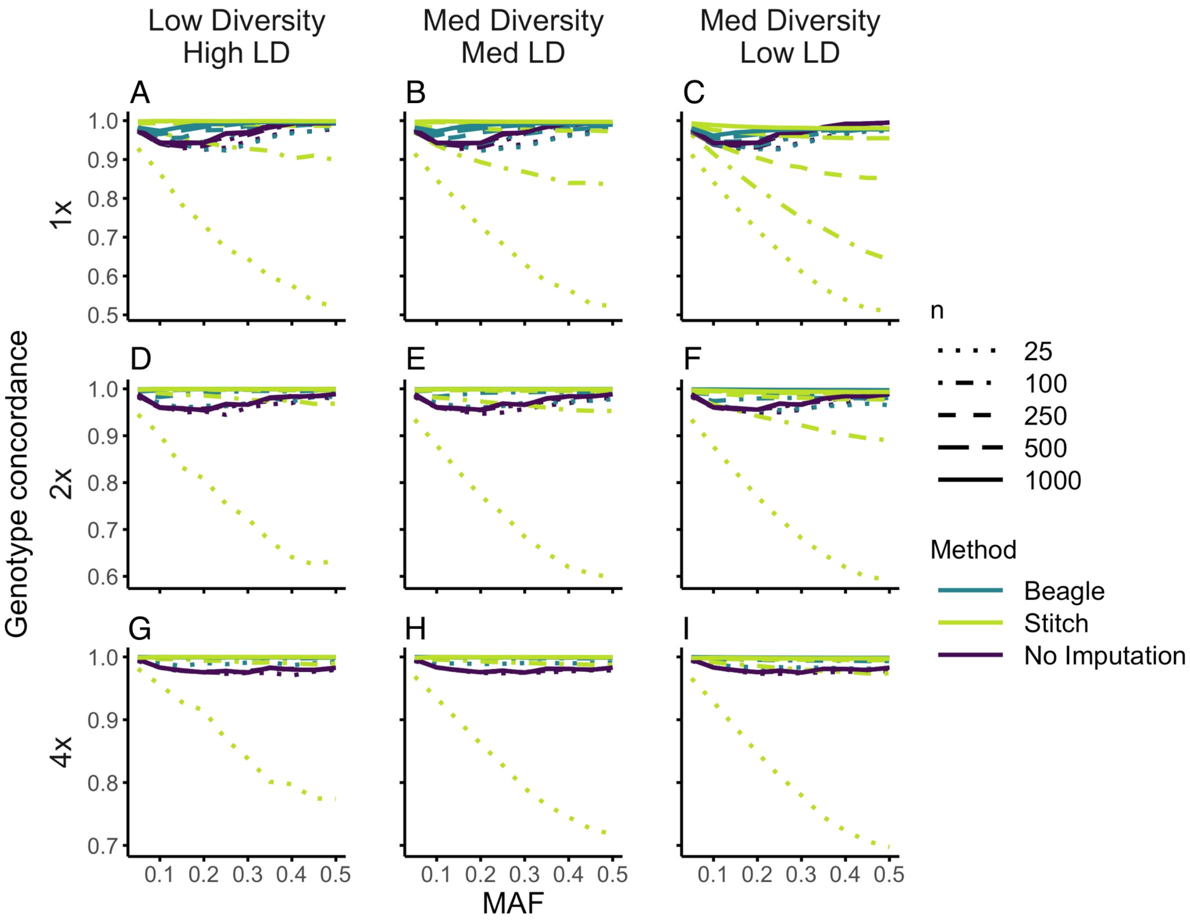

Figure S14. Genotype concordance by minor allele frequency (MAF) for imputation in STITCH and Beagle and without imputation. Genotypes were called with minimum posterior genotype probability of 0.9. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (plots in rows correspond to 1x, 2x and 4x coverage) were tested for each method (line colors) under different diversity and linkage disequilibrium scenarios. Note the different y-axis scales.

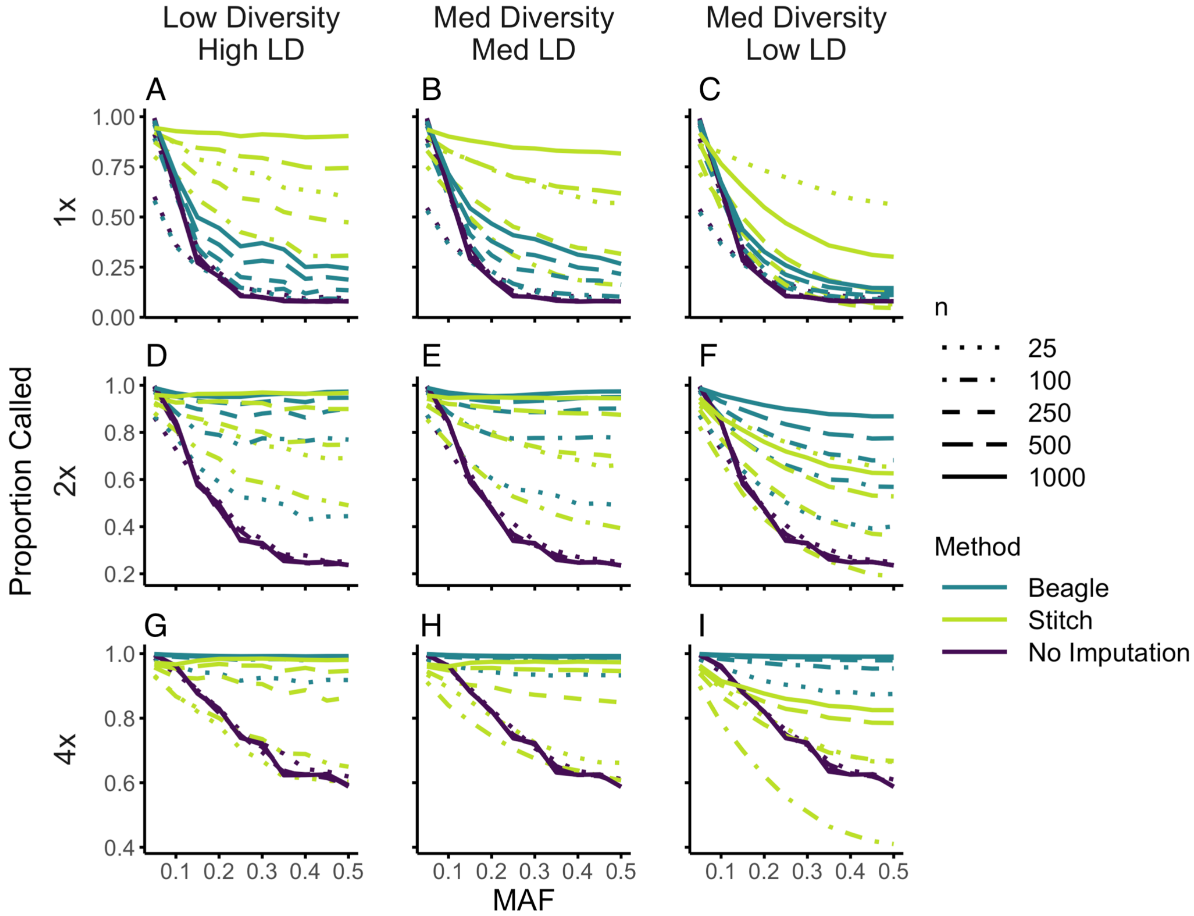

Figure S15. Proportion of genotypes called by minor allele frequency (MAF) for imputation in STITCH and Beagle and without imputation. Genotypes were called with minimum posterior genotype probability of 0.9. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (plots in rows correspond to 1x, 2x and 4x coverage) were tested for each method (line colors) under different diversity and linkage disequilibrium scenarios. Note the different y-axis scales.

Supplementary References

Allio, R., Donega, S., Galtier, N., & Nabholz, B. (2017). Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Molecular Biology and Evolution*, *34*(11), 2762–2772.

Bilton, T. P., McEwan, J. C., Clarke, S. M., Brauning, R., van Stijn, T. C., Rowe, S. J., & Dodds, K. G. (2018). Linkage Disequilibrium Estimation in Low Coverage High-Throughput Sequencing Data. *Genetics*, *209*(2), 389–400.

Blischak, P. D., Kubatko, L. S., & Wolfe, A. D. (2018). SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics*, Vol. 34, pp. 407–415. doi: 10.1093/bioinformatics/btx587

Bryc, K., Patterson, N., & Reich, D. (2013). A novel approach to estimating heterozygosity from low-coverage genome sequence. *Genetics*, *195*(2), 553–561.

Cheng, J. Y., Mailund, T., & Nielsen, R. (2017). Fast admixture analysis and population tree estimation for SNP and NGS data. *Bioinformatics* , *33*(14), 2148–2155.

Cheng, J. Y., Racimo, F., & Nielsen, R. (2019). Ohana: detecting selection in multiple populations by modelling ancestral admixture components. doi: 10.1101/546408

Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, *48*(8), 965–969.

Domyan, E. T., Kronenberg, Z., Infante, C. R., Vickrey, A. I., Stringham, S. A., Bruders, R., … Shapiro, M. D. (2016). Molecular shifts in limb identity underlie development of feathered feet in two domestic avian species. *eLife*, *5*, e12115.

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics* , *35*(19), 3855–3856.

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* , *30*(10), 1486–1487.

Garcia-Erill, G., & Albrechtsen, A. (2020). Evaluation of model fit of inferred admixture proportions. *Molecular Ecology Resources*, *20*(4), 936–949.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv. Retrieved from http://arxiv.org/abs/1207.3907

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, *23*(18), 4555–4573.

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637.

Huang, L., Wang, B., Chen, R., Bercovici, S., & Batzoglou, S. (2016). Reveel: large-scale population genotyping using low-coverage sequencing data. *Bioinformatics* , *32*(11), 1686–1696.

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

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*, 356.

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* , *31*(24), 4009–4011.

Langmead, B., & Salzberg, S. L. (2013). Langmead. 2013. Bowtie2. *Nature Methods*, *9*, 357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Li, H., & Ralph, P. (2019). Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics*, *211*(1), 289–304.

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017).

ATLAS: Analysis Tools for Low-depth and Ancient Samples. *bioRxiv*. doi: 10.1101/105346

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., … Xu, X. (2018). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell*, *175*(2), 347–359.e14.

Lorieux, M., Gkanogiannis, A., Fragoso, C., & Rami, J.-F. (2019). NOISYmputer: genotype imputation in bi-parental populations for noisy low-coverage next-generation sequencing data. *bioRxiv*. doi: 10.1101/658237

Lucas-Lledó, J. I., Vicente-Salvador, D., Aguado, C., & Cáceres, M. (2014). Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm. *BMC Bioinformatics*, *15*, 163.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913.

Maruki, T., & Lynch, M. (2014). Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics*, *197*(4), 1303–1313.

Maruki, T., & Lynch, M. (2015). Genotype-Frequency Estimation from High-Throughput Sequencing Data. *Genetics*, *201*(2), 473–486.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, *210*(2), 719–731.

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., & Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3* , *5*(11), 2383–2390.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* , *32*(11), 1749–1751.

Rastas, P. (2017). Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* , *33*(23), 3726–3732.

Sarmashghi, S., Bohmann, K., P Gilbert, M. T., Bafna, V., & Mirarab, S. (2019). Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, *20*(1), 34.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702.

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples. *Genetics*, *203*(2), 699–714.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1736). doi: 10.1098/rstb.2016.0455

Van Belleghem, S. M., Rastas, P., Papanicolaou, A., Martin, S. H., Arias, C. F., Supple, M. A., … Papa, R. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Ecology & Evolution*, *1*(3), 52.

Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics* , *32*(14), 2096–2102.

Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, *23*(11), 1852–1861.

Vieira, F. G., Lassalle, F., Korneliussen, T. S., & Fumagalli, M. (2016). Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biological Journal of the Linnean Society. Linnean Society of London*, *117*(1), 139–149.

Wasik, K., Berisa, T., Pickrell, J. K., Li, J. H., Fraser, D. J., King, K., & Cox, C. (2019). Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics (p. 632141). doi: 10.1101/632141

Whalen, A., Gorjanc, G., & Hickey, J. M. (2019). Parentage assignment with genotyping-by-sequencing data. *Journal of Animal Breeding and Genetics*, *136*(2), 102–112.