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 Wildlife Alliance, Escondido, CA 92027, USA

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

TABLE OF CONTENTS

Supplementary methods

Section 2: Estimation of the cost of lcWGS

The cost estimates presented in Table 1 assume a per library cost of 8 USD (details in Therkildsen and Palumbi 2017). This is the pro-rated cost of the reagents needed for a single library. An important consideration for researchers adopting lcWGS for the first time, is that many of the reagents needed are only available in relatively large batches, requiring a substantial upfront investment. One of the most expensive reagents to acquire is often a sufficiently large set of indexed (barcoded) adapter oligos needed to individually label each library. To avoid misassigned reads due to index hopping, we recommend a unique dual index strategy (i.e. two unique oligos per sample for the P5 and P7 ends of the library construct (MacConaill et al., 2018)). With November 2020 pricing, custom synthesis of each adapter oligo pair would cost ~44 USD, bringing the initial investment for oligos for 50 uniquely barcoded samples (which can then be pooled in a single sequencing lane) to ~2,200 USD. Several commercial barcoding adapter kits are also available and may be a cheaper option if a relatively small total number of samples are to be processed. The investment in indexed adapters is for most users a one-time investment in a resource that can split among laboratories.

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

In short, we used SLiM3 (Haller & Messer, 2019) to generate forward genetic simulations of a 30Mbp chromosome within *in silico* populations under a diploid Wright-Fisher model. The simulated populations had an effective population size (N_e) of 10^5 (unless otherwise noted), a mutation rate of 10^{-8} per base per generation, and a recombination rate of 2.5 cM/Mbp. These parameters were set to resemble a typical metazoan species with a relatively large population size (Allio, Donega, Galtier, & Nabholz, 2017; Stapley, Feulner, Johnston, Santure, & Smadja, 2017), and see a discussion of how different parameter choices can affect our results in the supplementary materials). We then sampled a subset of individuals in these populations and used ART-MountRainier (Huang, Li, Myers, & Marth, 2012) to simulate different lcWGS experimental designs with different combinations of sample size and coverage per sample. We performed genotype-likelihood-based analyses of these simulated sequencing reads with ANGSD, and tested the power of different experimental designs in population genetic inference. We used the Samtools genotype likelihood model implemented in ANGSD (-GL 1) and only report the results from GATK model (-GL 2) when the two show significant discrepancies. In addition, we simulated other high-throughput sequencing strategies, including Pool-seq and RAD-seq, and compared their performance with that of lcWGS (detailed methods in the supplementary materials).

To examine the performance for different types of population genomic inference, we generated three separate sets of simulations. First, we simulated an isolated population to test the accuracy of lcWGS in estimating key population genetic parameters in a single population. Second, we simulated two different metapopulations to test the ability of lcWGS to infer spatial structure among subpopulations under different levels of connectivity. Lastly, we simulated two populations closely connected by gene flow under divergent selection, and tested the power of lcWGS to identify the genetic loci under selection. The key model parameters used in our simulations are summarized in Table S2, and our entire simulation and analysis pipeline is available on GitHub (https://github.com/therkildsen-lab/lcwgs-simulation).

Population genomic inference for single populations: First, we tested the accuracy of low-coverage sequencing 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 30Mbp 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/Mbp 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 et al., 2017; Stapley et al., 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 et al., 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 the Samtools 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 (Figure 2). 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 run the realSFS module in ANGSD 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 neutrality tests (e.g. Tajima's D) in 10kb windows, using ANGSD with the following options: `-GL 1 -doSaf 1 -doThetas 1 -doCounts 1 -setMinDepth sample_size*coverage`, and the thetaStat module in ANGSD with the following options: `do_stat -win 10000 -step 10000` (Figure S2, S3). To compare the performance between different genotype likelihood models, we replicated the entire analysis pipeline above using the GATK genotype likelihood model (`-GL 2`) (Figure S2, S3). Lastly, from the genotype likelihoods calculated using the Samtools model, we estimated linkage disequilibrium (LD) between intermediate frequency SNPs (minimum minor allele frequency = 0.1) within 5kb of each other using ngsLD (Fox et al. 2019) with the following options: `--probs --rnd_sample 1  --max_kb_dist 5  --min_maf 0.1` (Figure S4). We then fitted the estimated r^2 values with the LD decay model described by Hill and Weir (1988) using the `fit_LDdecay.R` script in ngsLD with the following options: `--fit_level 2 --n_ind $SAMPLE_SIZE --fit_boot 1000` (Figure S5). We also computed the theoretical expectation of LD decay curve using the effective population size and recombination rate used in our simulation, also based on the model described by Hill and Weir (1988) (Figure S4, S5).

Inference of spatial structure: Then, we tested the power of low-coverage sequencing in resolving the genetic structure of spatially distributed populations. Again, we began by randomly creating a starting sequence on a 30Mbp 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 0.0005 (or 0.002 in the high migration rate scenario) connecting each pair of adjacent populations (Figure 4). Similarly, we scaled up the neutral mutation rate (μ) to $2x10^{-7}$ per bp per generation, and recombination rate (r) to 50cM/Mbp. 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 (Figure 4, S10, S11). Also, we performed PCA with the true sample genotypes using PLINK2 as an additional comparison (Figure 4). Lastly, to test whether performance improves with genome-wide data instead of a single chromosome, we simulated a longer chromosome of 300Mbp under the high migration rate scenario, and repeated the entire pipeline but only with 5 samples per population (Figure S9).

Scans for divergent selection in the face of gene flow: Lastly, we tested the power of low-coverage sequencing 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 30Mbp 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 complete, 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 (Figure 5, S12, S13).

Comparison with Pool-seq: In addition to these investigations on different sequencing designs of low-coverage whole genome sequencing, we have also compared low-coverage whole genome sequencing 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 low-coverage samples), particularly when the sequencing yield from different individuals in the pool is uneven, which is avoidable with a lcWGS design by repooling (Figure S6) but is almost inevitable with pool-seq. 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 low-coverage whole genome sequencing 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 `-minQ 20 -doCounts 1 -dumpCounts 1`) (Figure 3). 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 low-coverage whole genome sequencing projects where we tried our best effort to generate even yield among samples by pooling by DNA molarity. These empirical sequencing 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 low-coverage whole genome sequencing, 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 (Figure S8).

Comparison with RAD-seq: 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 low-coverage whole genome sequencing simulation (Figure 6, S14).

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 S3) 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.

Box 4: Using imputation to bolster genotype estimation from lcWGS

Simulations: To explore imputation performance under different scenarios, we used the same forward simulation framework as in section 4.1 to simulate a 30MB chromosome for three neutrally evolving populations that have reached mutation-drift equilibrium. 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), creating three different scenarios with different levels of genetic diversity and linkage disequilibrium (LD). Genetic diversity and LD are known to affect imputation performance(Pasaniuc et al., 2012). 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, and accordingly, our three scenarios were characterized by 1) a low diversity, high LD scenario (r = 0.5 cM/Mbp, Ne = 1,000); 2) a medium diversity, medium LD scenario (r = 0.5 cM/Mbp, Ne = 10,000); and 3) a medium diversity, low LD scenario (r = 2.5, Ne = 10,000).

We generated sample sizes of 25, 100, 250, 500 or 1000 individuals from a single, neutrally evolving population of stable size for each simulated scenario. We sampled with replacement 2n haplotypes (n diploid individuals) from the offspring of the final generation of the simulation. Similar to our approach in Section 4, we used ART-MountRainier(W. Huang et al., 2012) to simulate bam files of sequence reads to average depths of 1x, 2x and 4x per individual for each sample size, for a total of five sample sizes x three depths x three population scenarios = 45 datasets.

SNP calling and genotype estimation with and without imputation: For each dataset, 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. For comparison, we called genotypes and estimated genotype dosages without imputation in ANGSD v.0.931. (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.) For all downstream analyses, we first identified 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).

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.

Before running the full 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 the imputation in Beagle, we passed genotype likelihoods estimated in ANGSD directly to Beagle and ran the 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 r^2 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.

Genotype calling rates and genotype concordance with imputation: At the smallest sample size tested (n=25), there was little to no improvement in accuracy using Beagle, and accuracy actually decreased when imputation was performed in STITCH with 25 samples (Figures S16-S18), suggesting that such small sample sizes are inadequate for reliable imputation; thus we focused our results on n≥100. For all sample sizes and sequencing depths across scenarios, the accuracy of genotype estimates varied with allele frequency. The correlation (r^2) between imputed allelic dosage and true genotypes was low for sites with minor allele frequency (MAF) < 0.05 to 0.10, but increased and was relatively consistent across higher MAF bins (Figure S16). Genotype concordance (GC), by contrast, had the opposite relationship with MAF; GC was higher for sites with low MAF and decreased with higher MAF (Figure S16). This is because it is easy to achieve high accuracy by calling the homozygous major genotype when the minor allele is rare. In order to summarize overall imputation performance, we averaged r^2, GC and the proportion of called genotypes across sites with MAF>0.05 for each combination of method, scenario and study design (Figure S16-18).

Genotype concordance (GC) was universally high for all methods and sequencing strategies (GC>0.9), except for imputation of 100 samples from the medium diversity, high LD scenario in STITCH (Figure S19D-F). At 1x coverage, fewer than half of genotypes were called by Beagle and without imputation, especially for sites with higher MAF (Figure S18). GC was similar under the medium diversity, medium LD scenario compared to the low diversity, high LD scenario (Figure S19D-E), except GC was somewhat lower for genotypes imputed in STITCH at 1x coverage. The least improvement in GC using imputation was seen under medium diversity, low LD scenario (Figure S19F). For n≤250 samples sequenced at 1x and 2x coverage, GC for genotypes imputed in STITCH were less accurate than those estimated without imputation.

Overall, imputation accuracy required larger sample sizes or was reduced altogether as genetic diversity and recombination rates increased. This was particularly true for the program STITCH, which estimates distinct haplotype probabilities within a given region across a mosaic of ancestral haplotypes(Davies, Flint, Myers, & Mott, 2016), a problem that becomes increasingly complex under high recombination. Imputation showed larger improvements with increasing sample size in STITCH than in Beagle, especially at low coverage (1x), whereas Beagle improved more with increasing sequence read depth (Figure 9).

Allele frequency estimation from imputed genotype probabilities: Because imputation increased the accuracy of posterior genotype probabilities under most of the tested scenarios and study designs, we asked whether allele frequency estimation was improved by using imputed genotype probabilities compared to MAF estimation without imputation. To estimate MAF from imputed genotype probabilities, we summed over the posterior genotype probabilities (-domaf 4 in ANGSD), and compared the results to MAF estimated from genotype likelihoods using the EM algorithm implemented in ANGSD (-domaf 1). Under some scenarios and study designs, imputation resulted in small improvements in accuracy of allele frequency estimation (Figure S20). Imputation yielded the largest improvements in allele frequency estimation for large sample sizes (N≥250) sequenced at 1x coverage from the low diversity, high LD population, and from the medium diversity, medium LD population. For small sample sizes from the medium diversity, low LD population, MAF estimated from genotype probabilities imputed in STITCH were less accurate. Beagle showed more consistent, modest improvements, increasing MAF estimation accuracy when coverage was ≥2x for all sample sizes and scenarios.

Under the low diversity, high LD scenario, allele frequency estimates based on genotype probabilities imputed in STITCH from 1000 samples at 1x coverage were slightly more accurate (r^2=0.999) than for 500 samples at 2x coverage (r^2=0.998) and 250 samples at 4x coverage (r^2=0.997). However, given that smaller sample sizes are already sufficient for estimating allele frequencies with high accuracy without imputation (r^2=0.990 for MAF estimated from 250 samples sequenced at 1x coverage; Figure S20), imputation is not likely to contribute to analyses of these types of population-level statistics as much as it would for individual-level and genotype-level analyses like GWAS.

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.

Additional details on software packages for the analysis of low-coverage data

In this section, we include some additional details about the software packages that we introduced in Section 4 of the main text. When applicable, we highlight the methodological differences between the different packages for solving the same problem.

Genotype likelihood models: Four different genotype likelihood models are currently implemented in ANGSD. The GATK model (McKenna et al., 2010) assumes that base quality scores at the same site from different sequencing reads are each an independent and unbiased representation of the probabilities of sequencing error, whereas the Samtools model (Li, 2011) assumes that these quality scores are not completely independent. Both the SYK model (Kim et al., 2011) and the SOAPsnp model (Li et al., 2009) assume that the quality scores could be biased and thus implement a quality score recalibration step. In the SKY mode, type specific error rates (e.g. the probability of an A being called a T) are estimated and accounted for in GL calculation. In the SOAPsnp model, in addition to the type specific errors, strand and read position specific errors can be accounted for as well, but a set of invariant loci should be provided to minimize biases. Additional genotype likelihood models are adopted by other software packages and they can be useful alternatives to ANGSD for specific types of data. For example, the program Atlas (Kousathanas et al., 2017) explicitly incorporates post-mortem DNA damage in addition to sequencing error in its genotype likelihood model, making it well-suited for ancient DNA studies. EBG (Blischak, Kubatko, & Wolfe, 2018) uses a simplified version of the SAMtools model but relaxes ANGSD's assumption of diploidy, allowing the analysis of polyploid samples.

SNP identification: In ANGSD, SNPs are inferred by first estimating allele frequencies at each site (including the presumably invariable loci) and then testing whether its minor allele frequency is significantly larger than zero (Korneliussen et al., 2014). Accordingly, the first step is to restrict the number of alleles that can possibly occur at each site to two: a major allele, and a minor allele. The identities of these alleles can be determined through a maximum likelihood approach (Jørsboe & Albrechtsen, 2019; Skotte, Korneliussen, & Albrechtsen, 2012) or by user specification. Next, the likelihood of the minor allele frequency at each site can be formulated as a function of genotype likelihoods across all individuals (see Equation 2 in (Kim et al., 2011)), and these minor allele frequencies can be estimated using a maximum likelihood approach. In this way, all possible genotypes for each individual can be considered, effectively avoiding explicitly calling genotypes. Then, polymorphic sites will be identified through a likelihood ratio test (Kim et al., 2011). The list of polymorphic sites (i.e. SNPs) can then be exported and used for downstream analyses, along with the genotype likelihoods at each of these sites for each individual. Other software programs address SNP calling in similar ways. Atlas, for example, follows the same general framework as ANGSD, but has made modifications (Kousathanas et al., 2017) to accommodate cases where the sample size is very small and neither the major nor the minor alleles is specified by users, which is often the case for ancient DNA studies (Kousathanas et al., 2017).

Dimensionality reduction methods for population structure inference: The random read sampling method employed by ANGSD does not take full advantage of the entire dataset. In contrast, ngsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014) uses a more sophisticated method where posterior genotype probabilities are first calculated with an empirical Bayes approach. This approach is valid under the assumption of Hardy-Weinberg equilibrium across the entire sample set, but for most structured populations, this assumption will not hold, which can lead to inaccurate PCA results (e.g. population clusters can have long tails, see Meisner & Albrechtsen, 2018). PCAngsd (Meisner & Albrechtsen, 2018) therefore takes one step further and uses an iterative approach to correct for potential violation of the HWE assumption by updating prior genotype probabilities based on the PCA result in each previous iteration, since these PCA results can represent the population structure that exists in the data (Meisner & Albrechtsen, 2018).

Model-based clustering for population structure inference: NGSAdmix (Skotte, Korneliussen, & Albrechtsen, 2013) adopts a maximum likelihood implementation of the classic STRUCTURE model (Tang, Peng, Wang, & Risch, 2005)(Pritchard, Stephens, & Donnelly, 2000), (Tang et al., 2005), but formulates a likelihood function with sequencing data as its observed data and uses genotype likelihoods to consider all possible genotypes for each individual (see Equation 6 in Skotte et al., 2013). It then uses an expectation-maximization (EM) algorithm to estimate model parameters. Because of the more complex formulation of the likelihood function, however, NGSAdmix tends to be computationally demanding. As an alternative, Ohana (Cheng, Racimo, & Nielsen, 2019) adopts the same likelihood function as NGSAdmix but uses a sequential quadratic programming (QP) method instead of EM for optimization, which should speed up computation. No formal comparison between the performance of the two methods is available to date, but separate evaluations on simulated and real data have shown that both methods deliver great accuracy even at very low coverage (Cheng et al., 2019; Skotte et al., 2013). Distinct from both NGSAdmix and Ohana, PCAngsd uses individual allele frequencies, an intermediate output from its PCA analysis, as input for a non-negative matrix factorization (NMF) algorithm to infer admixture proportions.

Genome-wide association analysis: In Kim et al. (2011), case / control association is tested by first estimating allele frequencies within case and control individuals with the approach as described in the “SNP identification” section, and then using a likelihood ratio test for differences between case and control individuals at each locus (see equations 6-7 in Kim et al. 2011). The first step in Skotte et al. (2012) and Jørsboe & Albrechtsen (2019) is to calculate the posterior genotype probability using an empirical Bayes approach, with priors informed by either population allele frequencies or the SFS. Skotte et al. (2012) then used a score statistics approach to test for significant associations with the phenotype at each site. This approach is computationally efficient, but cannot estimate the effect size of the loci. In contrast, (Jørsboe & Albrechtsen, 2019) employs a maximum likelihood approach to explicitly estimate the effect size of each locus. As expected, this approach is slower than the score statistics method. To take advantage of both methods, ANGSD also implements a hybrid approach, first using the score statistic to identify significant loci, and then using the maximum-likelihood approach to estimate effect sizes of these significant loci.

Linkage disequilibrium: GUS-LD (Bilton et al., 2018) constructs a likelihood function of the LD coefficient D and uses a numerical method to optimize the likelihood function. In contrast, ngsLD (Fox, Wright, Fumagalli, & Vieira, 2019) constructs a likelihood function of the haplotype frequencies between each pair of SNPs instead, and uses an EM algorithm to optimize it (Fox et al., 2019). Different LD statistics, such as D, D’ and r^2, can then be derived from the inferred haplotype frequencies. Furthermore, ngsLD incorporates several other helpful features, such as LD pruning and the fitting of an LD decay model.

Allele frequency estimation: As mentioned in the SNP identification section, ANGSD takes a maximum-likelihood approach to estimate allele frequencies among all samples (Kim et al., 2011). It then uses the same algorithm to estimate the frequencies of the minor alleles in each population separately for each site identified as polymorphic (based on the selected filtering and confidence threshold). It is important to note that a SNP significance filter or a minimum minor allele frequency filter should not be applied in population-specific allele frequency estimation, because sites fixed for the major allele in a subset of populations (which would be removed by these filters) are typically of interest. Other programs that can estimate allele frequencies from genotype likelihoods follow the same general workflow. Atlas (Kousathanas et al., 2017), for example, adopts a similar maximum likelihood framework, but also provides a Bayesian inference option.

Genetic diversity and neutrality test statistics within a single population: To estimate θ in different parts of the genome, ANGSD adopts an empirical Bayes approach, where the SFS within a window (posterior) can be formulated and solved as the product of the SAF likelihoods within the window (likelihood) and the genome-wide or chromosome-wide SFS (prior) (see the equation in the “Empirical Bayes” section in Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013). Different θ estimators can then be extracted from the SFS in each window.

Genetic differentiation between populations: ANGSD implements the method-of-moment estimator of F_{ST} developed by (Reynolds, Weir, & Cockerham, 1983). While different estimators of θ depend on the local SFS within a single population, Reynolds et al.’s estimator of pairwise Fst can be formulated as a function of the local two-dimensional SFS (the matrix with the joint distribution of allele counts in two populations). Therefore, ANGSD again takes an empirical Bayes approach, using the maximum likelihood method to estimate a genome-wide two-dimensional SFS, which it then uses as a prior to calculate the two-dimensional SFS at each genomic locus. Fst at each locus can then be derived from these locus-specific SFS. GPAT (http://www.yandell-lab.org/software/gpat.html) implements two additional methods to estimate Fst using genotype likelihoods as its input. In the first method (wcFst), GPAT estimates allele frequencies from genotype likelihoods and directly plugs the estimated allele frequencies into Weir and Cockerham's Fst estimator. This method is computationally efficient but may not account for the uncertainties in the estimated allele frequencies as well as ANGSD does. In the second method (bFst), GPAT implements a Bayesian framework as described by (Holsinger, Lewis, & Dey, 2002). This Bayesian approach has the advantage of being able to provide a confidence interval for Fst, but it is computationally expensive.

References for software in Table 2 of the main text

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

Supplementary tables

Table S1. List of population genomics studies using lcWGS.

Cayuela, Hugo, Clément Rougeux, Martin Laporte, Claire Mérot, Eric Normandeau, Maëva Leitwein, Yann Dorant, et al. 2021. "Genome-Wide DNA Methylation Predicts Environmentally-Driven Life History Variation in a Marine Fish." bioRxiv. https://doi.org/10.1101/2021.01.28.428603.
Clucas, Gemma V., R. Nicolas Lou, Nina O. Therkildsen, and Adrienne I. Kovach. 2019. "Novel Signals of Adaptive Genetic Variation in Northwestern Atlantic Cod Revealed by Whole-genome Sequencing." Evolutionary Applications 134 (4): 1289.
Crawford, Jacob E., Ricardo Amaru, Jihyun Song, Colleen G. Julian, Fernando Racimo, Jade Yu Cheng, Xiuqing Guo, et al. 2017. "Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans." American Journal of Human Genetics 101 (5): 752–67.
Foote, Andrew D., Michael D. Martin, Marie Louis, George Pacheco, Kelly M. Robertson, Mikkel Holger S. Sinding, Ana R. Amaral, et al. 2019. "Killer Whale Genomes Reveal a Complex History of Recurrent Admixture and Vicariance." Molecular Ecology 8 (7): e1002837–18.
Foote, Andrew D., Nagarjun Vijay, María C. Ávila-Arcos, Robin W. Baird, John W. Durban, Matteo Fumagalli, Richard A. Gibbs, et al. 2016. "Genome-Culture Coevolution Promotes Rapid Divergence of Killer Whale Ecotypes." Nature Communications 7 (January): 11693.
Fuller, Zachary L., Veronique J. L. Mocellin, Luke A. Morris, Neal Cantin, Jihanne Shepherd, Luke Sarre, Julie Peng, et al. 2020. "Population Genetics of the Coral Acropora Millepora: Toward Genomic Prediction of Bleaching." Science 369 (6501). https://doi.org/10.1126/science.aba4674.
Gignoux-Wolfsohn, Sarah A., Malin L. Pinsky, Kathleen Kerwin, Carl Herzog, Mackenzie Hall, Alyssa B. Bennett, Nina H. Fefferman, and Brooke Maslo. 2021. "Genomic Signatures of Selection in Bats Surviving White-Nose Syndrome." Molecular Ecology. https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15813?casa_token=erFTF_VQsOkAAAAA:eptQO8I66AX6xJVdRd6iv90BA9avfwJVDnQyG9sb_lWAd0aoJNYK6mI7P_qFrAMo-fBbCKAOD2DWrg.
Ilardo, Melissa A., Ida Moltke, Thorfinn S. Korneliussen, Jade Cheng, Aaron J. Stern, Fernando Racimo, Peter de Barros Damgaard, et al. 2018. "Physiological and Genetic Adaptations to Diving in Sea Nomads." Cell 173 (3): 569–80.e15.
Jones, Felicity C., Manfred G. Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, et al. 2012. "The Genomic Basis of Adaptive Evolution in Threespine Sticklebacks." Nature 484 (7392): 55–61.
Liu, Shiping, Eline D. Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, et al. 2014. "Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears." Cell 157 (4): 785–94.
Mérot, Claire, Emma Berdan, Hugo Cayuela, Haig Djambazian, Anne-Laure Ferchaud, Martin Laporte, Eric Normandeau, Jiannis Ragoussis, Maren Wellenreuther, and Louis Bernatchez. 2021. "Locally-Adaptive Inversions Modulate Genetic Variation at Different Geographic Scales in a Seaweed Fly." bioRxiv. https://doi.org/10.1101/2020.12.28.424584.
Oziolor, Elias M., Noah M. Reid, Sivan Yair, Kristin M. Lee, Sarah Guberman VerPloeg, Peter C. Bruns, Joseph R. Shaw, Andrew Whitehead, and Cole W. Matson. 2019. "Adaptive Introgression Enables Evolutionary Rescue from Extreme Environmental Pollution." Science 364 (6439): 455–57.
Powell, Daniel L., Mateo García-Olazábal, Mackenzie Keegan, Patrick Reilly, Kang Du, Alejandra P. Díaz-Loyo, Shreya Banerjee, et al. 2020. "Natural Hybridization Reveals Incompatible Alleles That Cause Melanoma in Swordtail Fish." Science 368 (6492): 731–36.
Reid, Noah M., Dina A. Proestou, Bryan W. Clark, Wesley C. Warren, John K. Colbourne, Joseph R.

Shaw, Sibel I. Karchner, et al. 2016. “The Genomic Landscape of Rapid Repeated Evolutionary Adaptation to Toxic Pollution in Wild Fish.” Science 354 (6317): 1305–8.

Rowan, Beth A., Darren Heavens, Tatiana R. Feuerborn, Andrew J. Tock, Ian R. Henderson, and Detlef Weigel. 2019. “An Ultra High-Density Arabidopsis Thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features.” Genetics 213 (3): 771–87.

Szarmach, Stephanie J., Alan Brelsford, Christopher Witt, and David P. L. Toews. 2021. “Comparing Divergence Landscapes from Reduced-Representation and Whole-Genome Re-Sequencing in the Yellow-Rumped Warbler (Setophaga Coronata) Species Complex.” bioRxiv. https://doi.org/10.1101/2021.03.23.436663.

Therkildsen, Nina O., Aryn P. Wilder, David O. Conover, Stephan B. Munch, Hannes Baumann, and Stephen R. Palumbi. 2019. “Contrasting Genomic Shifts Underlie Parallel Phenotypic Evolution in Response to Fishing.” Science 365 (6452): 487–90.

Wang, Hongru, Filipe G. Vieira, Jacob E. Crawford, Chengcai Chu, and Rasmus Nielsen. 2017. “Asian Wild Rice Is a Hybrid Swarm with Extensive Gene Flow and Feralization from Domesticated Rice.” Genome Research 27 (6): 1029–38.

Wilder, Aryn P., Stephen R. Palumbi, David O. Conover, and Nina Overgaard Therkildsen. 2020. “Footprints of Local Adaptation Span Hundreds of Linked Genes in the Atlantic Silverside Genome.” Evolution Letters 4 (5): 430–43.

Table S2. Model parameters used for forward genetic simulation.

Scenario*	Chromosome length (in Mb)	Number of populations	Population size (*N*)†	Mutation rate (μ)	Recombination rate (*r*)	Migration rate (*m*)	Selection coefficient (*s*)	Corresponding figures
Single population	30	1	1000	10^{-6}	$2.5x10^{-6}$	NA	NA	3-4, S1-7
Spatial structure (low migration)	30	9	500	$2x10^{-7}$	$5x10^{-7}$	0.0005	NA	5A, S10
Spatial structure (high migration)	30	9	500	$2x10^{-7}$	$5x10^{-7}$	0.002	NA	5B, S11
Spatial structure (high migration, longer chromosome)	300	9	500	$2x10^{-7}$	$5x10^{-7}$	0.002	NA	S9
Divergent selection‡ (large Ne, high migration)	30	2	5000	10^{-7}	$2.5x10^{-7}$	0.001	0.08	6-7
Divergent selection‡ (small Ne, low migration)	30	2	5000	$2x10^{-8}$	$5x10^{-8}$	0.0005	0.08	S12-14
Imputation test (low diversity, high LD)	30	1	1000	10^{-8}	$5x10^{-9}$	NA	NA	9, S16-20
Imputation test (medium diversity, medium LD)	30	1	1000	10^{-7}	$5x10^{-8}$	NA	NA	9, S16-20
Imputation test (medium diversity, low LD)	30	1	1000	10^{-7}	$2.5x10^{-7}$	NA	NA	9, S16-20

* Each entry is linked to its corresponding simulation pipeline on GitHub.

† Note that since we scaled down population size and scaled up mutation rate, recombination rate, migration rate, and selection coefficient in order to speed up computation, these population sizes do not represent the effective population size of our simulated populations.

‡ These parameters are the ones used in the selection stage of the simulation. Prior to the selection stage, a burn-in stage was first performed, during which the population size was further scaled down, whereas mutation rate and recombination rate were scaled up, all by a factor of 10. See supplementary methods for details.

Table S3. 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

Supplementary figures

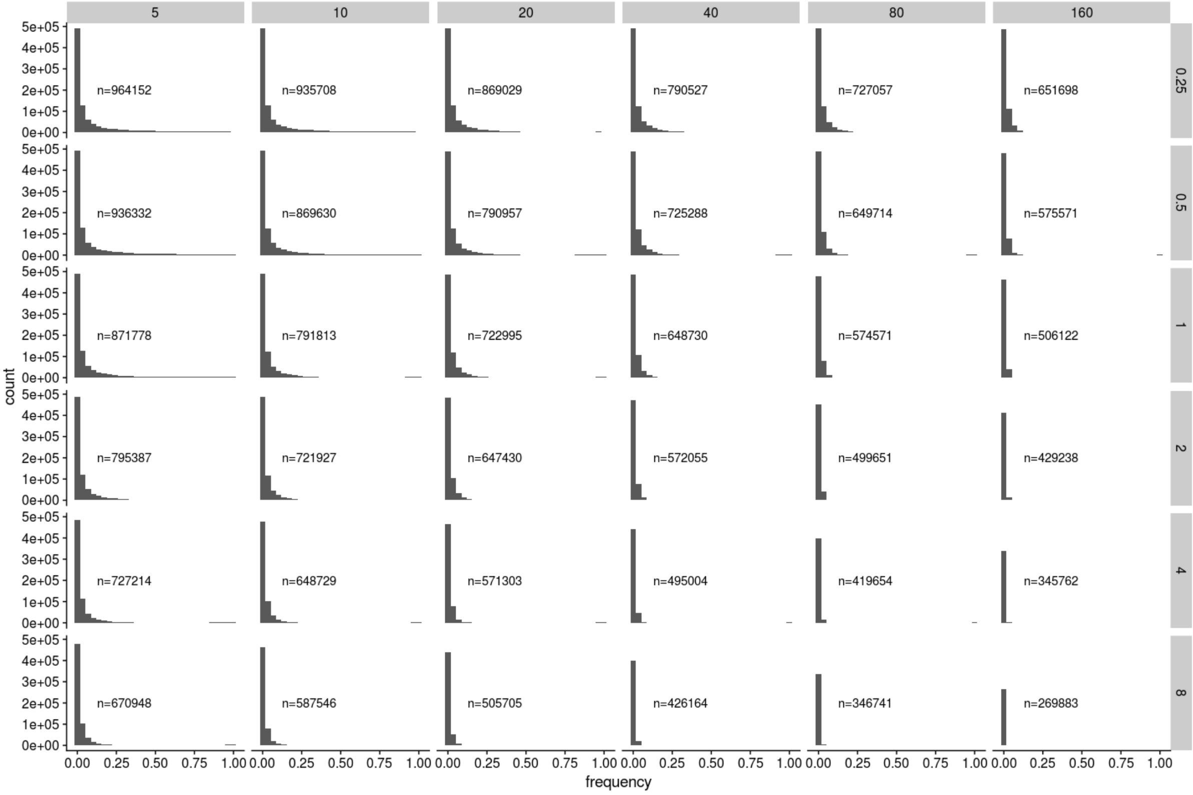


Figure S1. Histogram of the allele frequencies of false negative SNPs with lcWGS. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right.

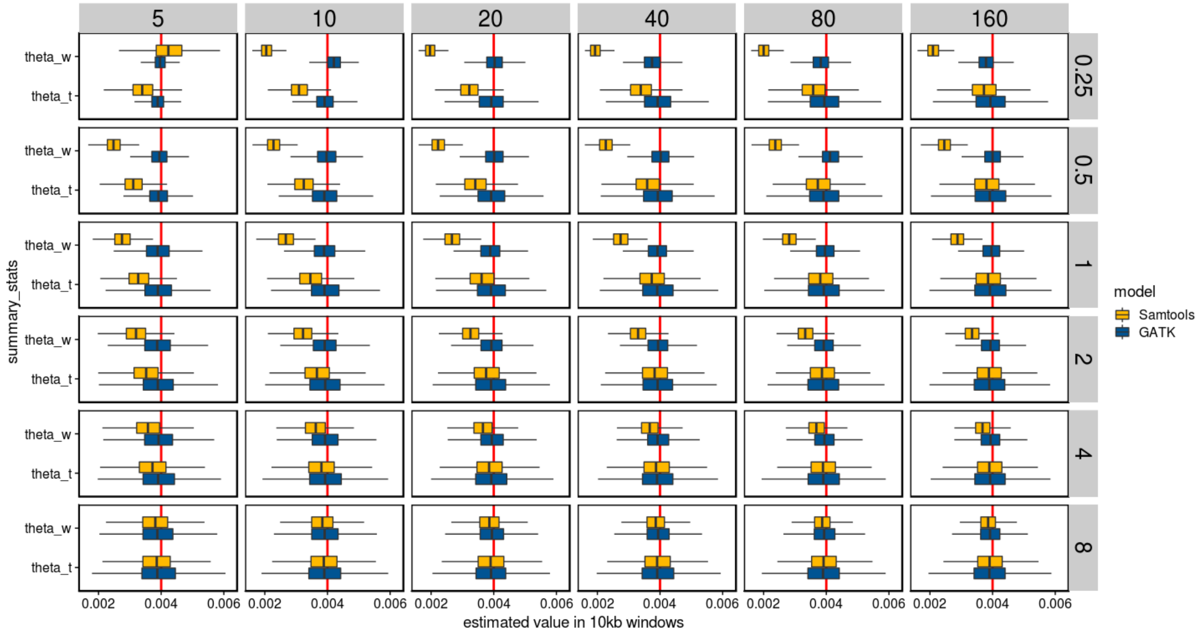


Figure S2. Distribution of Tajima's θ (aka π) and Watterson's θ estimated using the Samtools genotype likelihood model and the GATK genotype likelihood model in 10kb windows. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing 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, which is marked with a read line.

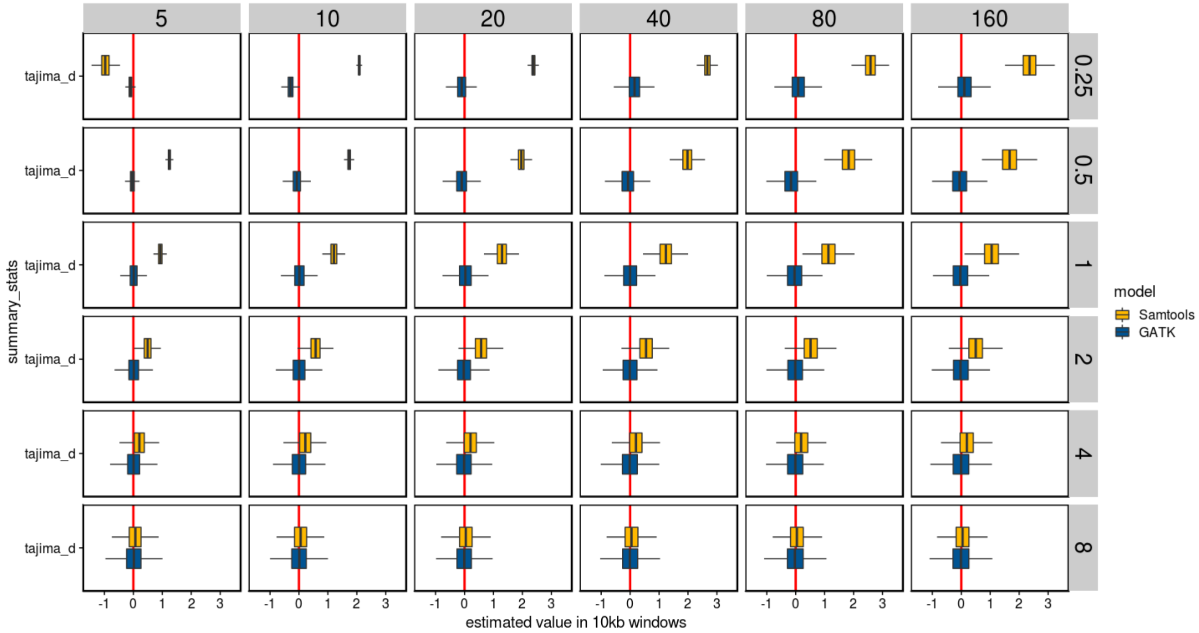


Figure S3. Tajima's D estimated using the Samtools genotype likelihood model and the GATK genotype likelihood model in 10kb windows. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the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

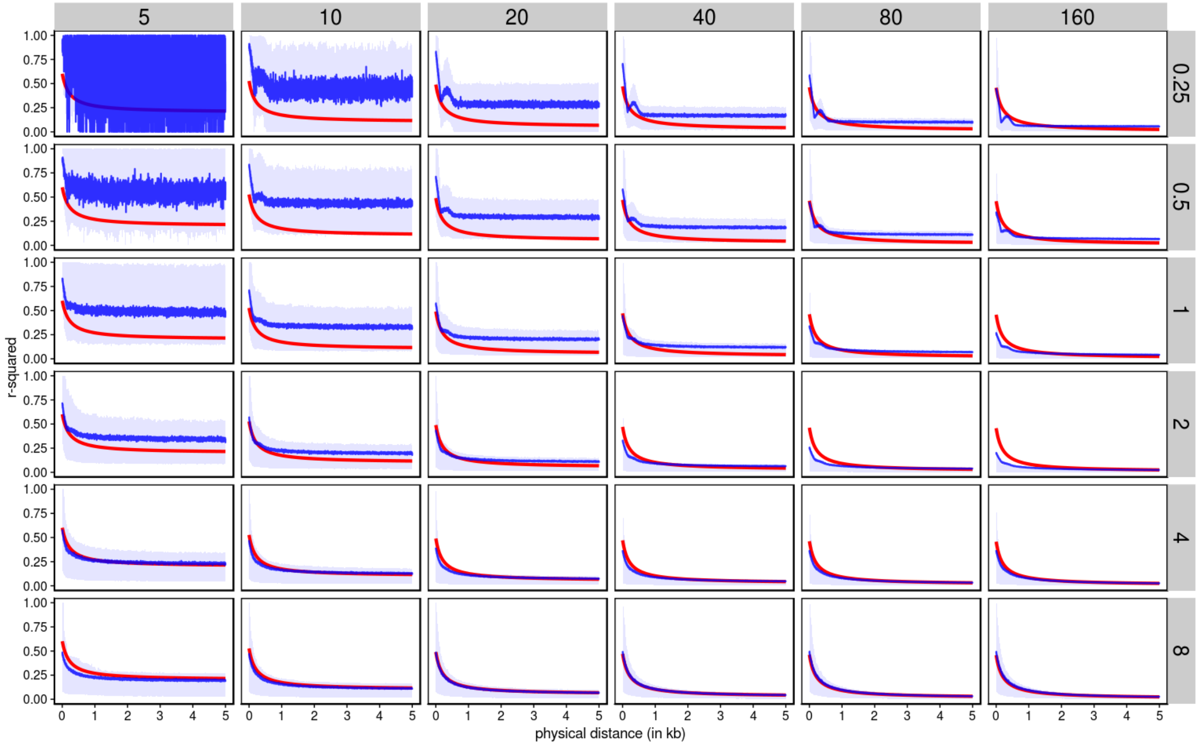


Figure S4. Linkage disequilibrium (LD) estimated using ngsLD from simulated data. LD, shown on the y axis, is measured as r^2 between pairs of SNPs, and the physical distance between these SNP pairs is shown on the x axis. The blue line shows the mean of the estimated r^2 for each distance value, and the lighter blue area shows its interquartile range. The red line marks the theoretical expectation of r^2 under mutation-drift equilibrium. 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.

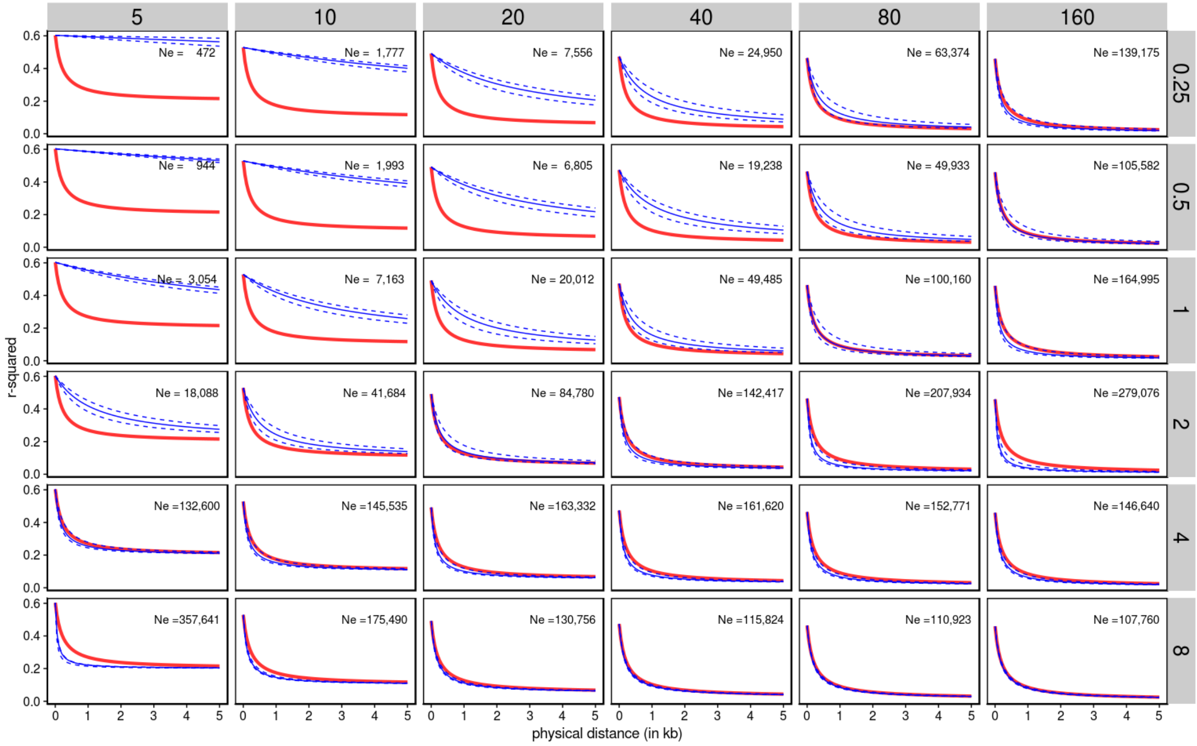


Figure S5. Estimated linkage disequilibrium (LD) fitted to a linkage decay model using ngsLD. The solid blue line shows the best fitted model, and the dashed blue lines represent its 95% confidence interval. When the true recombination rate is known, the effective population size (Ne) can be calculated from the estimated LD decay rate and is shown on the top right corner in each facet. The true effective population size used in the simulation is 100,000. The red line marks the theoretical expectation of r^2 under mutation-drift equilibrium, given by (Hill & Weir, 1988). 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.

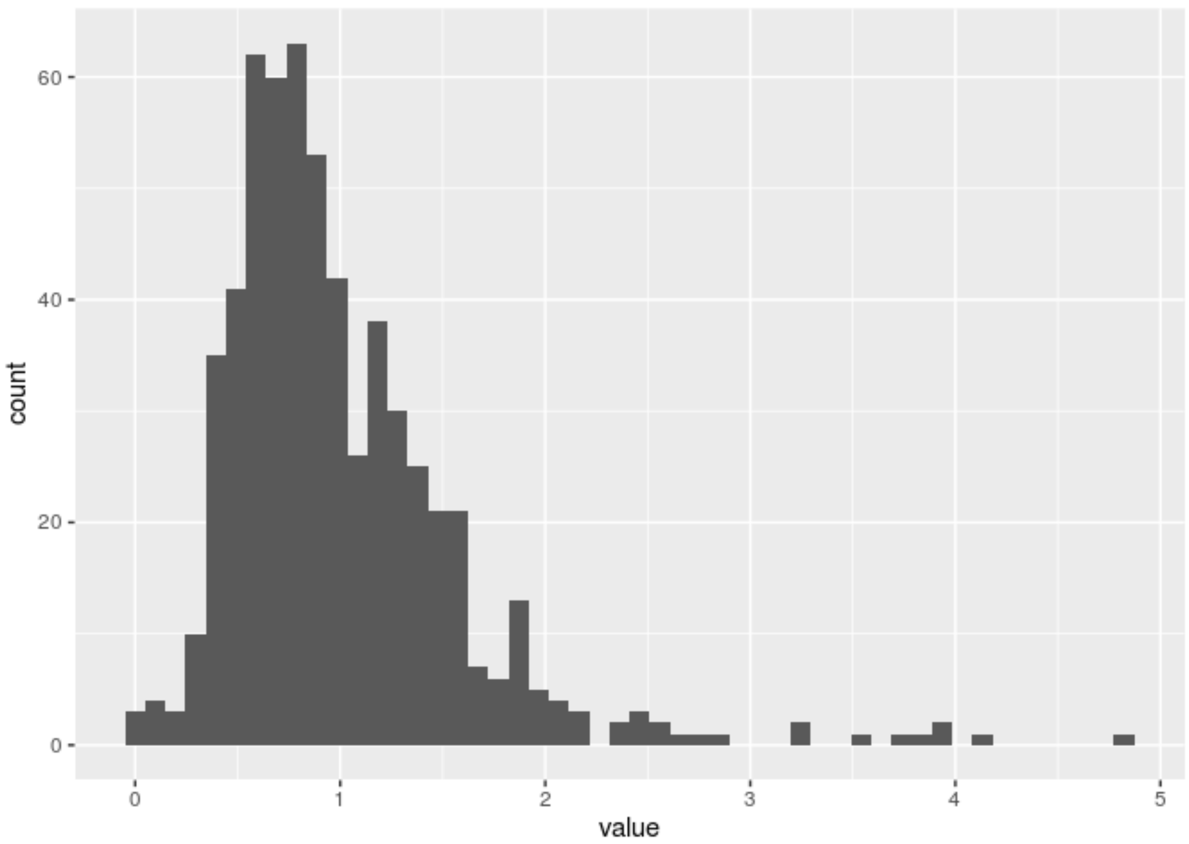


Figure S6. The sequencing coverage distribution that we sampled from when simulating uneven sequencing coverage among samples. This distribution is obtained by merging the distributions of coverage among samples from three of our lcWGS projects where we pooled samples by molarity.

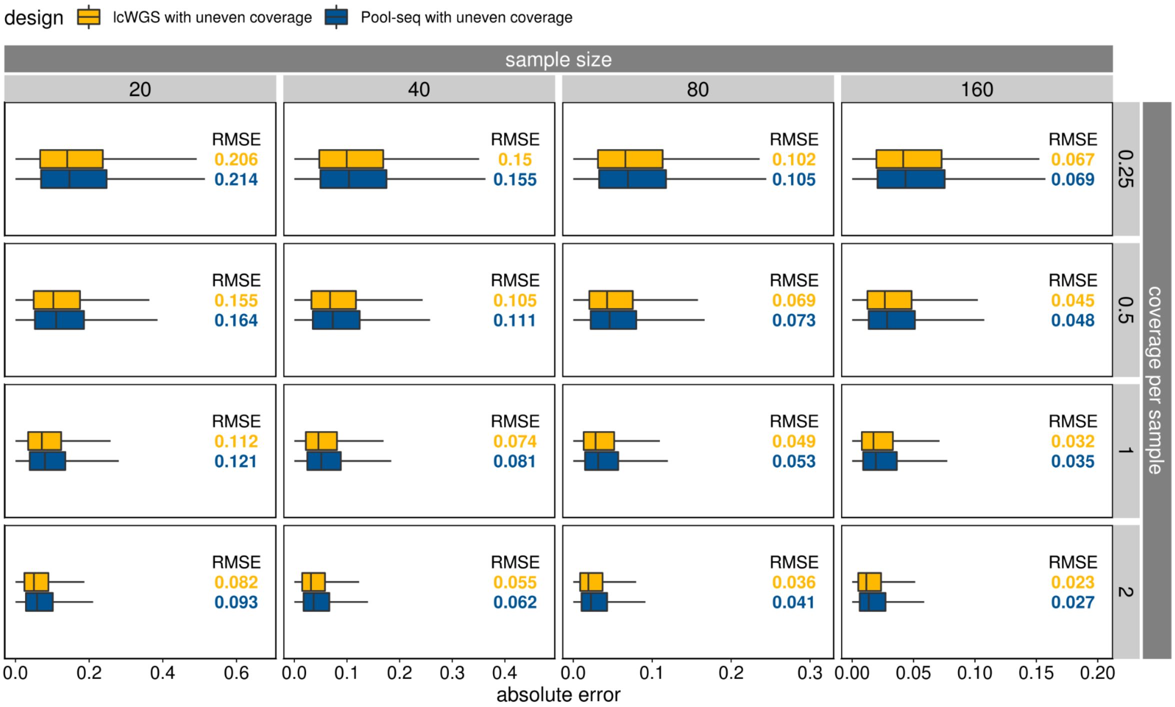


Figure S7. The error in allele frequency estimation with lcWGS (yellow) and Pool-seq (blue) data, both with uneven coverage among individual samples. The distribution of absolute errors (|estimated frequency - true frequency|) is shown with the box plots along the x-axis. The lower and upper hinges of the box plots show 25th and 75th percentile of the absolute errors, and the whiskers extend to the largest or smallest values no further than 1.5 times the interquartile range. Outlier points are hidden. 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) for the two sequencing designs are shown in each facet. False negative SNPs are not included in this figure. See supplementary methods and Figure S7 for how uneven coverage was simulated.

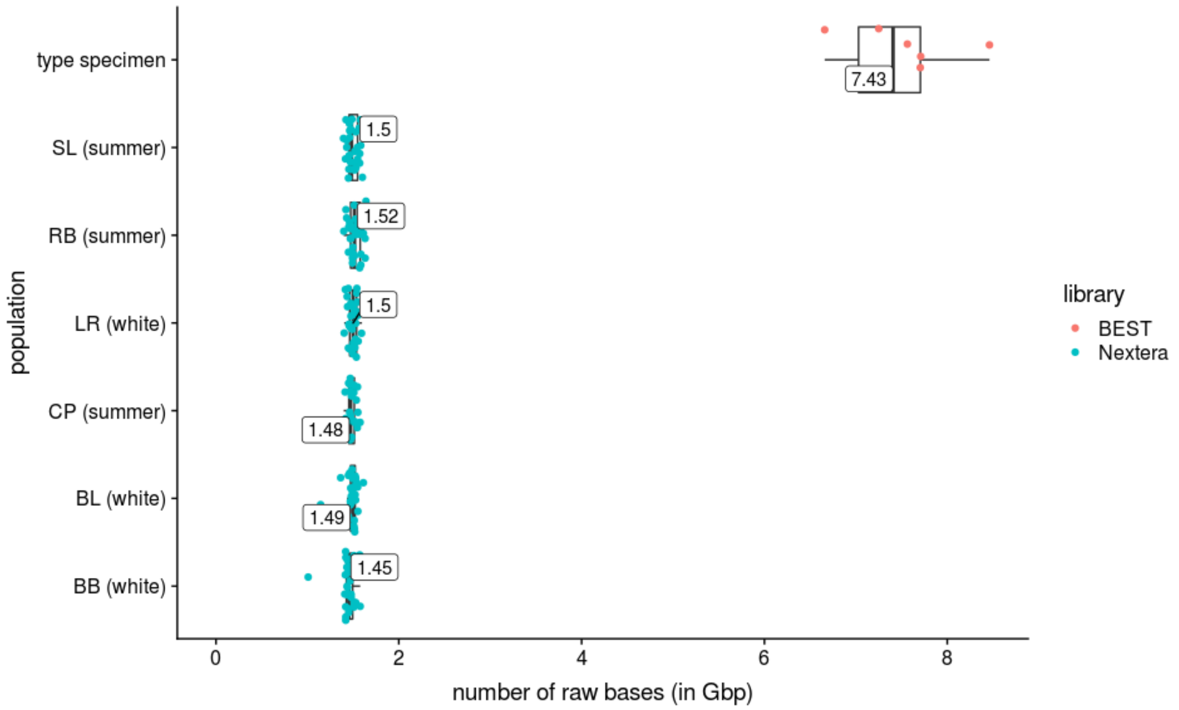


Figure S8. An empirical example from one of our lcWGS projects of the distribution of raw sequencing yield from individual samples when they are repooled based on the first round of sequencing. This is to demonstrate that equal distribution of sequencing effort can be approximated by such a sequencing design. (The type specimens were designed to have higher sequencing yield then other samples.)

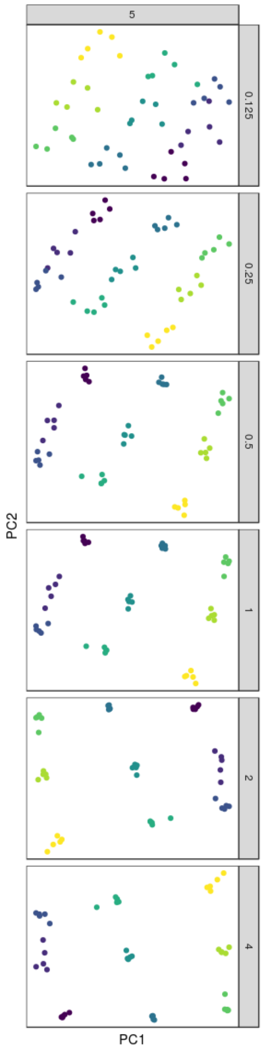


Figure S9. The spatial population structures inferred through principal component analysis (PCA) with lcWGS data using PCA. The first two principal components are shown. This result is from our higher gene flow scenario (an average of 1 effective migrant from one population to another every generation),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

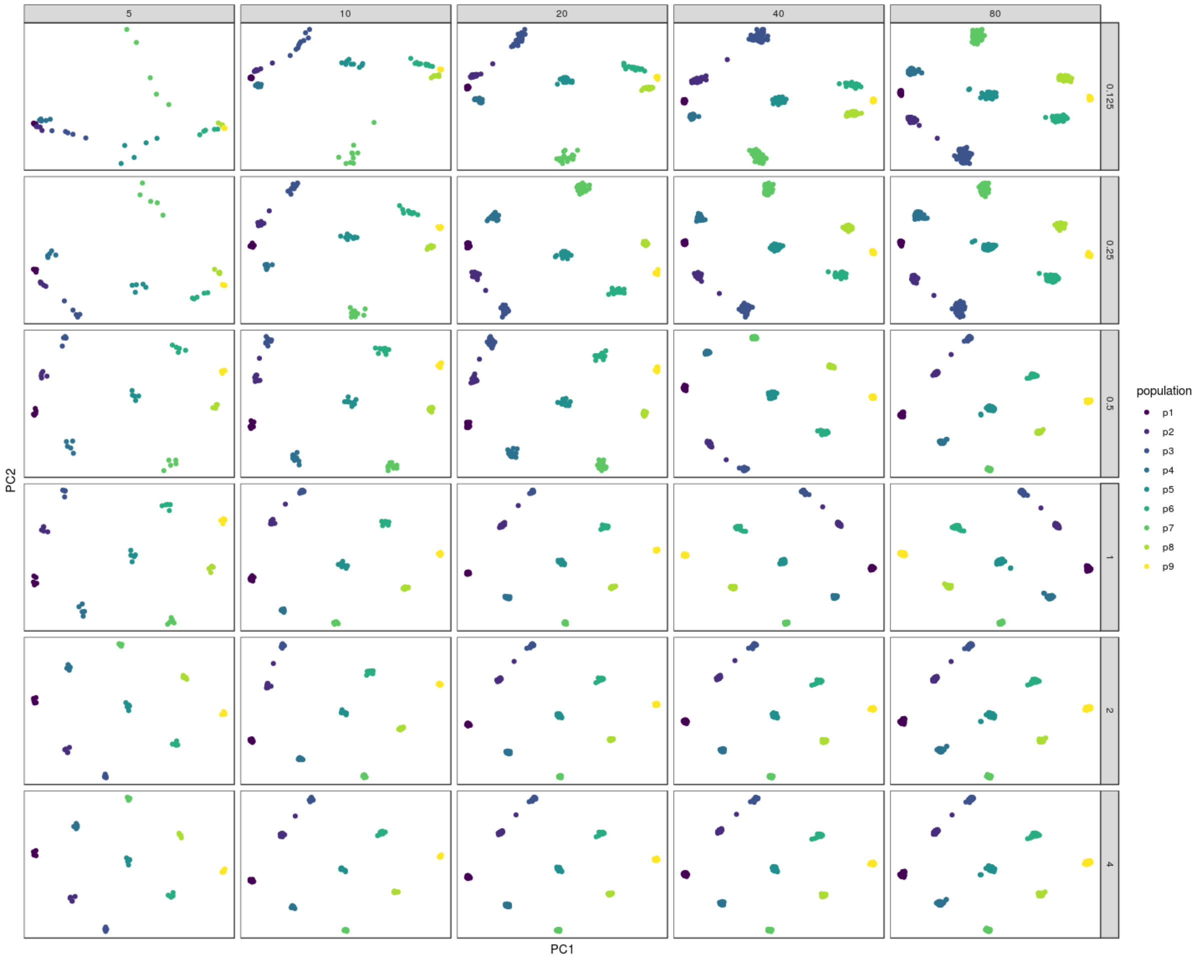


Figure S10. Patterns of spatial population structure inferred through principal component analysis (PCA) with lcWGS data using PCAngsd, in a scenario with lower gene flow (an average of 0.25 effective migrants per generation). Sample size per population increases across panels from left to right, and coverage per sample increases from top to bottom. This figure is based on the same dataset as Figure 5A, in which case ANGSD was used instead of PCAngsd to perform the PCA.

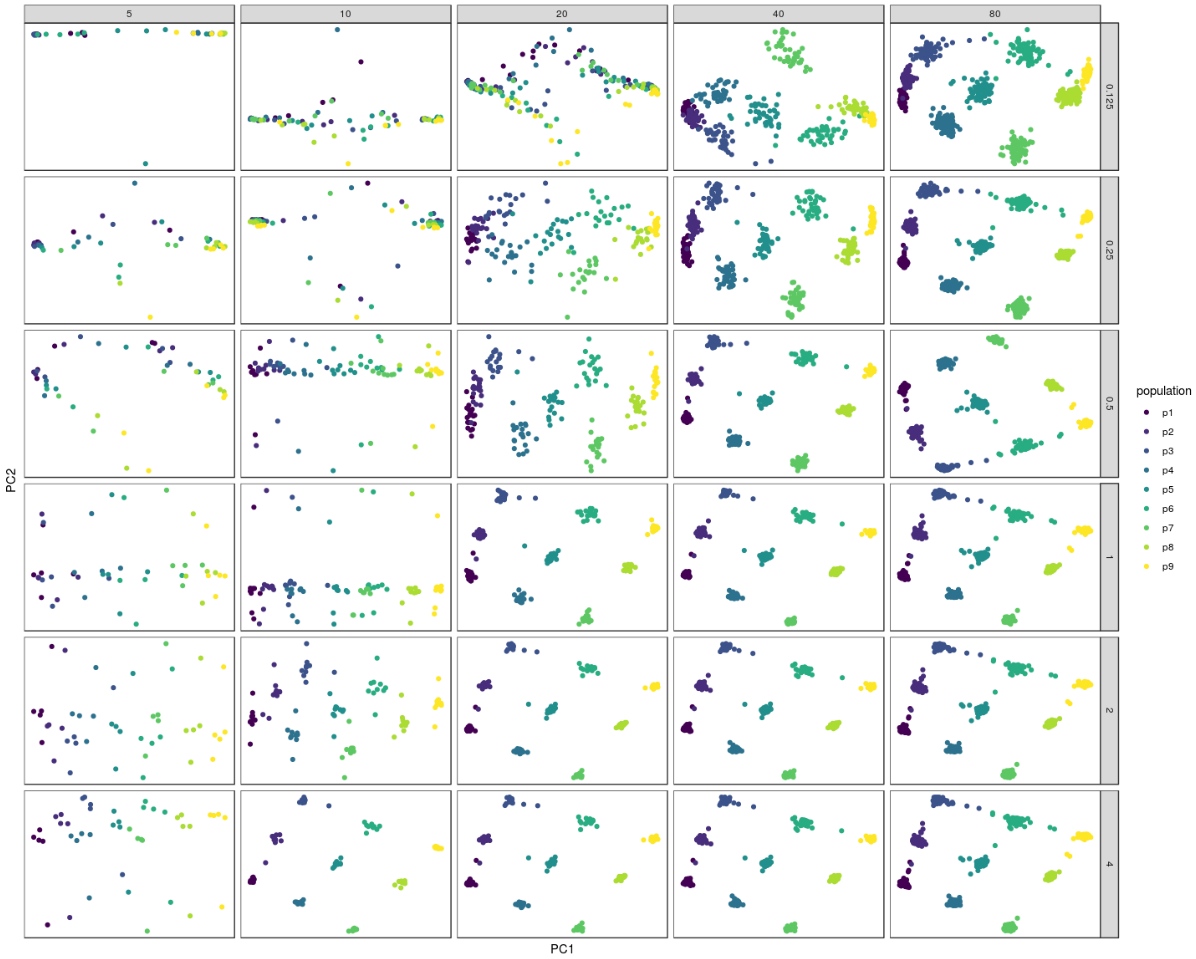


Figure S11. Patterns of spatial population structure inferred through principal component analysis (PCA) with lcWGS data using PCAngsd, in a scenario with higher gene flow (an average of 1 effective migrants per generation). Sample size per population increases across panels from left to right, and coverage per sample increases from top to bottom. This figure is based on the same dataset as Figure 5B, in which case ANGSD was used instead of PCAngsd to perform the PCA.

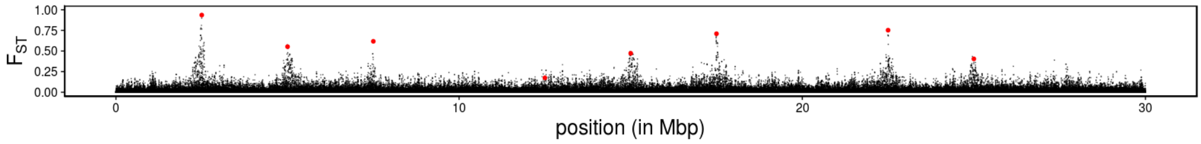


Figure S12. The true per-SNP F_{ST} values along the chromosome between the two simulated populations in a scenario with smaller N_e ($N_e = 10^4$) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). Neutral SNPs are shown in black and selected SNPs are shown in black.

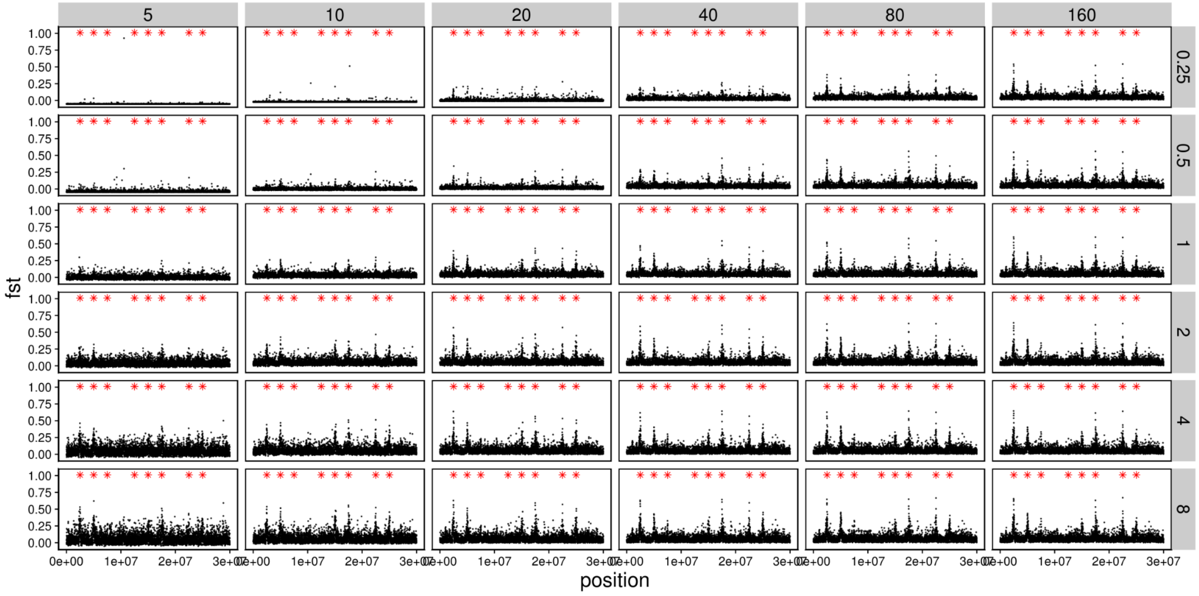


Figure S13. Genome-wide scan for divergent selection with lcWGS data in a scenario with smaller N_e ($N_e = 10^4$) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). The F_{ST} values inferred from lcWGS data in 5kb windows along the chromosome 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 asterisks only mark the positions of the selected SNPs (not their inferred Fst values).

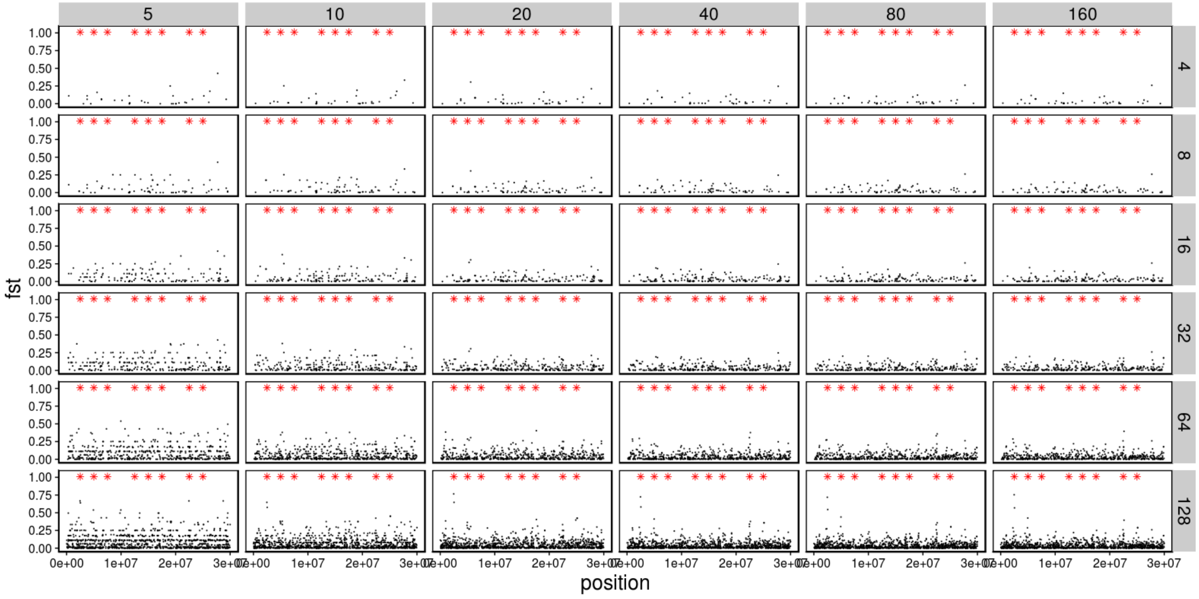
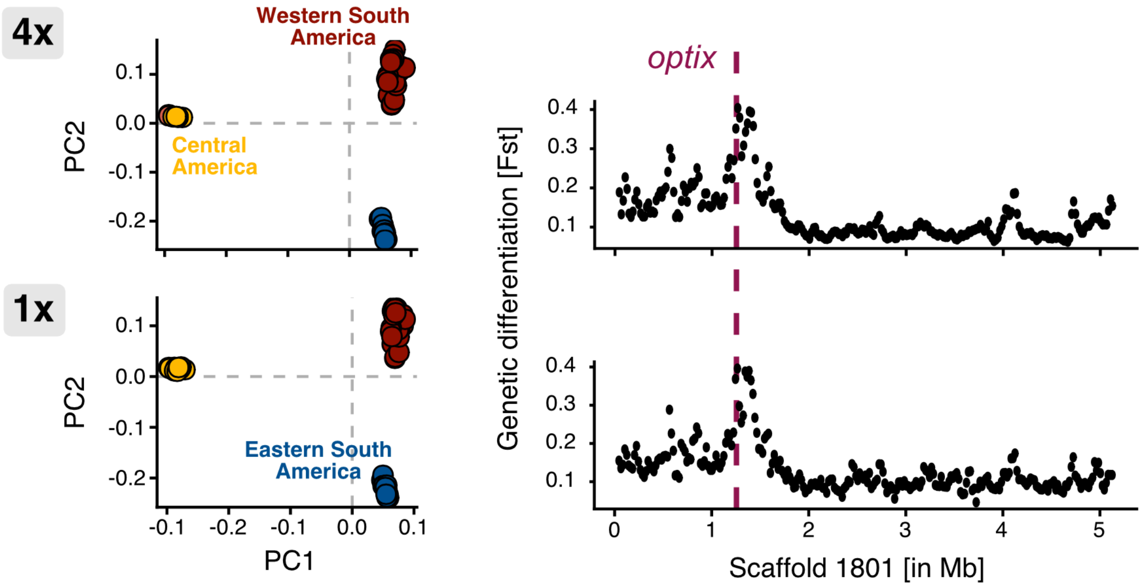


Figure S14. Genome-wide scan for divergent selection with RADseq data in a scenario with smaller N_e ($N_e = 10^4$) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). The per-SNP F_{ST} values inferred from RAD-seq data are shown on the y axis 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 asterisks only mark the positions of the selected SNPs (not their inferred Fst values).



958

959 **Figure S15.** Principal components plot and estimates of genetic differentiation around the
960 optix gene for the *Heliconius* dataset at 4x (top) and 1x coverage (bottom), respectively.

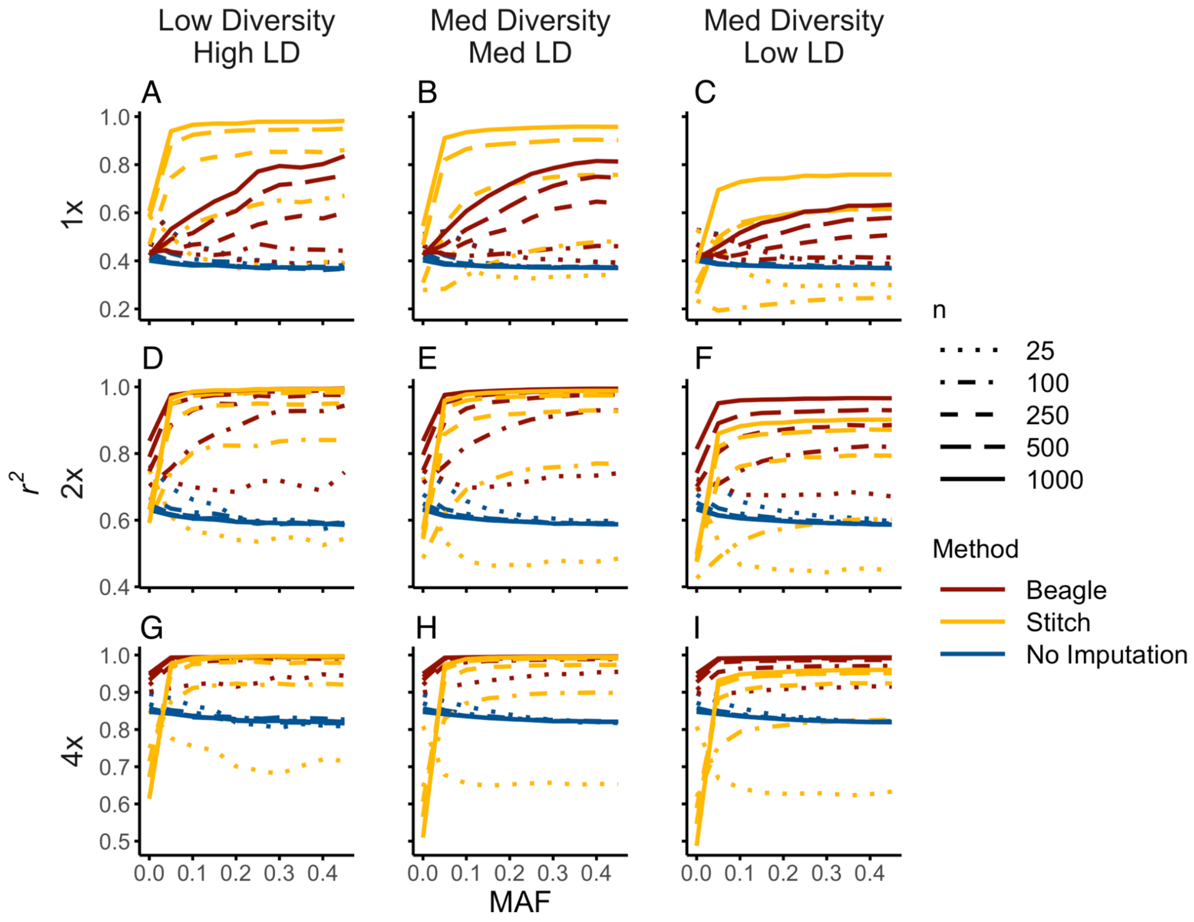


Figure S16. Genotype estimation accuracy (r^2) by minor allele frequency (MAF) for imputation in STITCH and Beagle compared to posterior genotypes estimated without imputation. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (plots in rows 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.

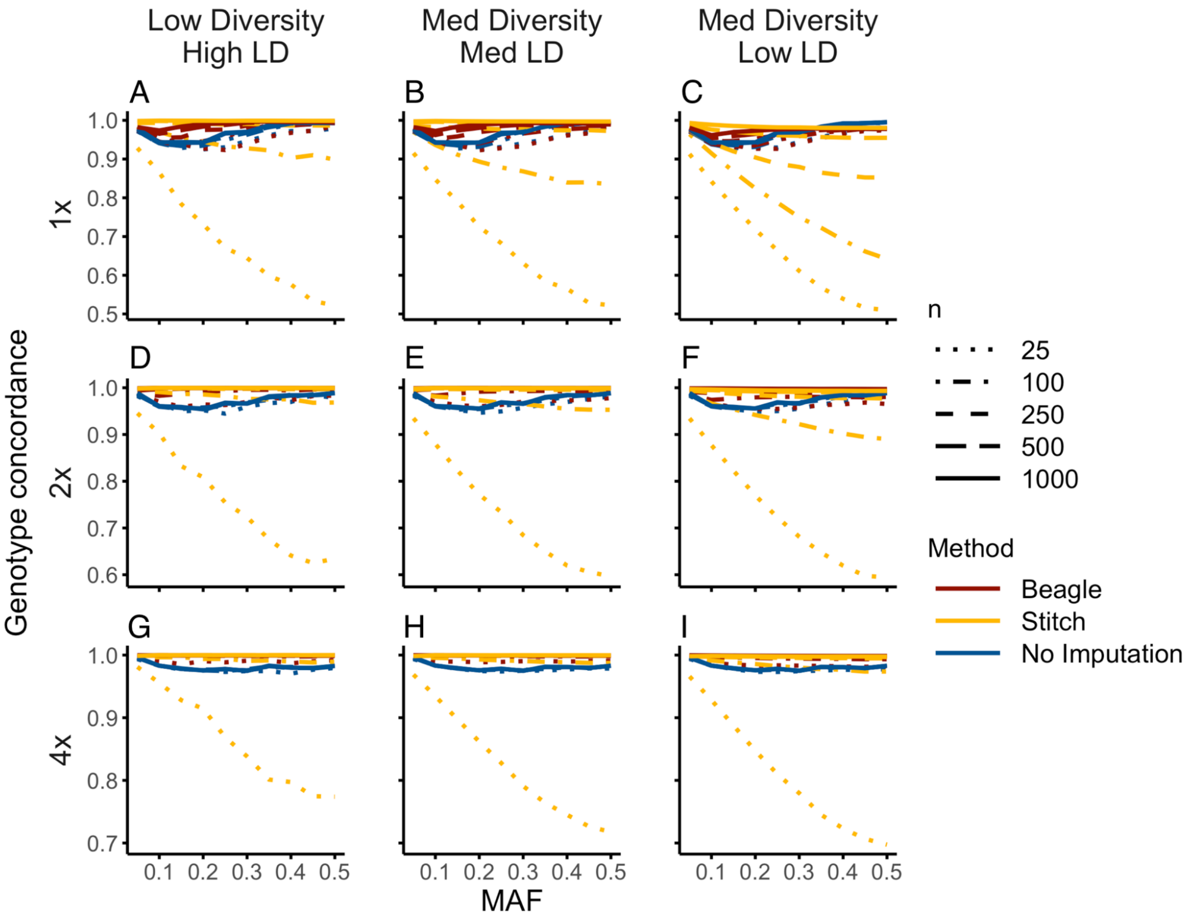


Figure S17. Genotype concordance by minor allele frequency (MAF) for imputation in STITCH and Beagle and without imputation. Genotypes were called with minimum posterior genotype probability of 0.9. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (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.

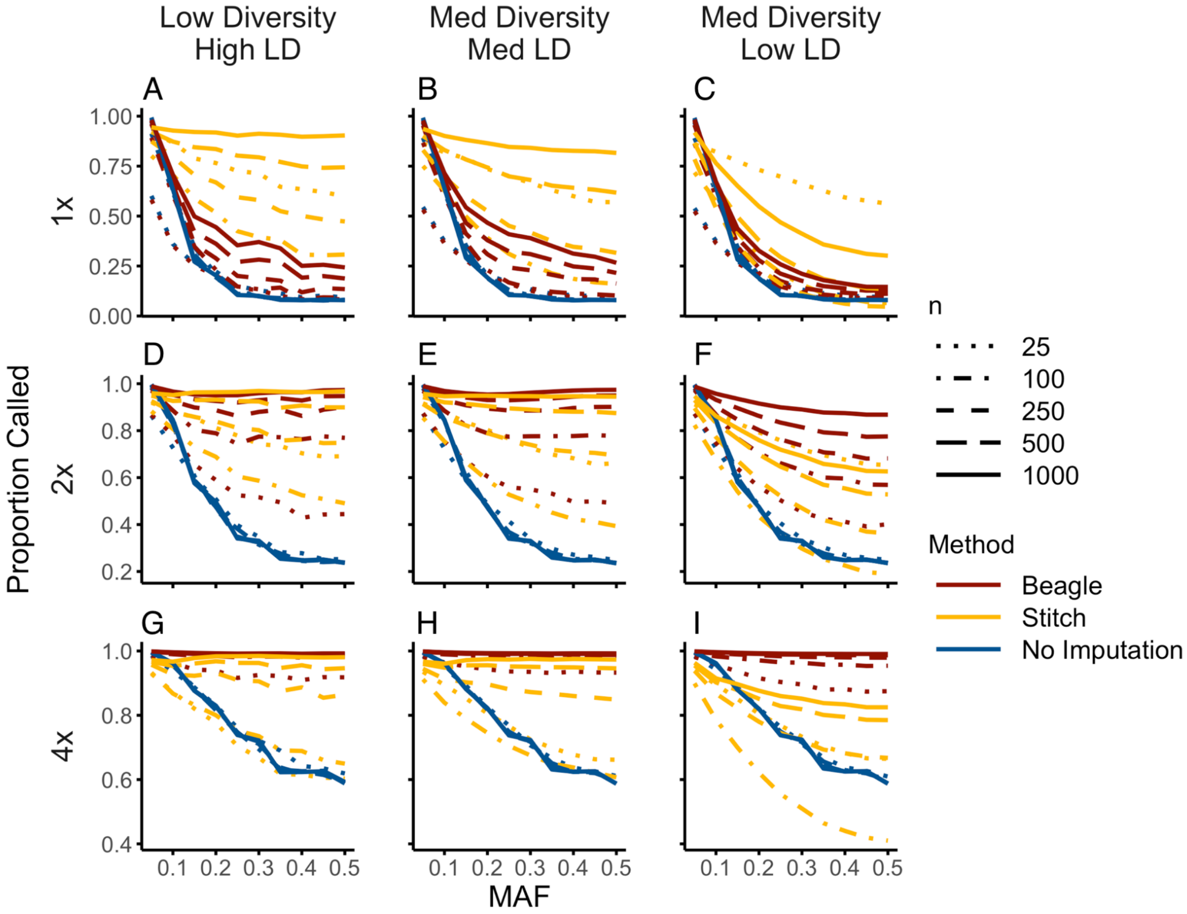


Figure S18. Proportion of genotypes called by minor allele frequency (MAF) for imputation in STITCH and Beagle and without imputation. Genotypes were called with minimum posterior genotype probability of 0.9. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing 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.

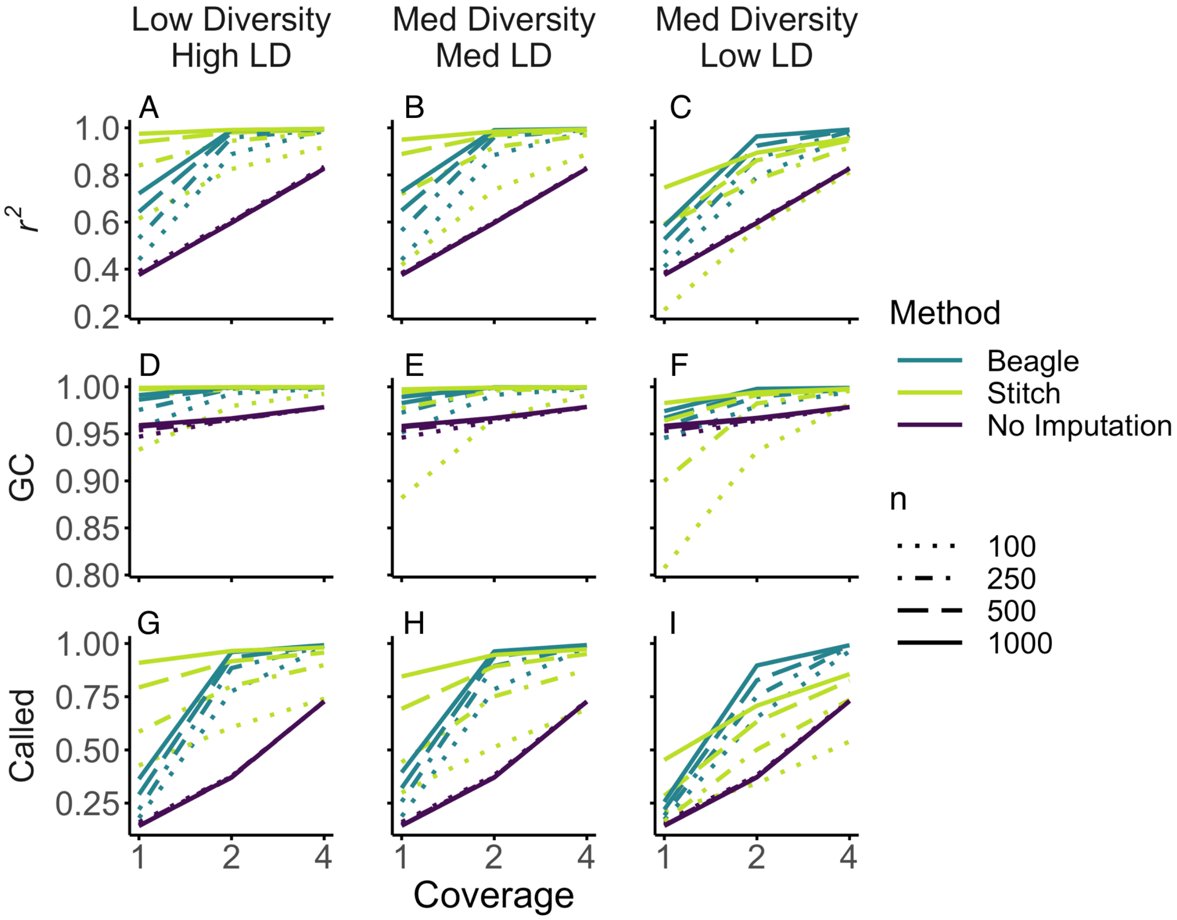


Figure S19. Genotype estimation by imputation in STITCH and Beagle compared to posterior genotypes estimated without imputation for sites with MAF>0.05. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (x-axis) were tested for each method (line colors) under different diversity and linkage disequilibrium scenarios. **(A)-(C)** Mean r^2 between true genotypes and estimated genotype dosage. **(D)-(F)** Genotype concordance (GC) between true and called genotypes with posterior genotype probability>0.9. G-I) Proportion of genotypes called with posterior genotype probability>0.9.

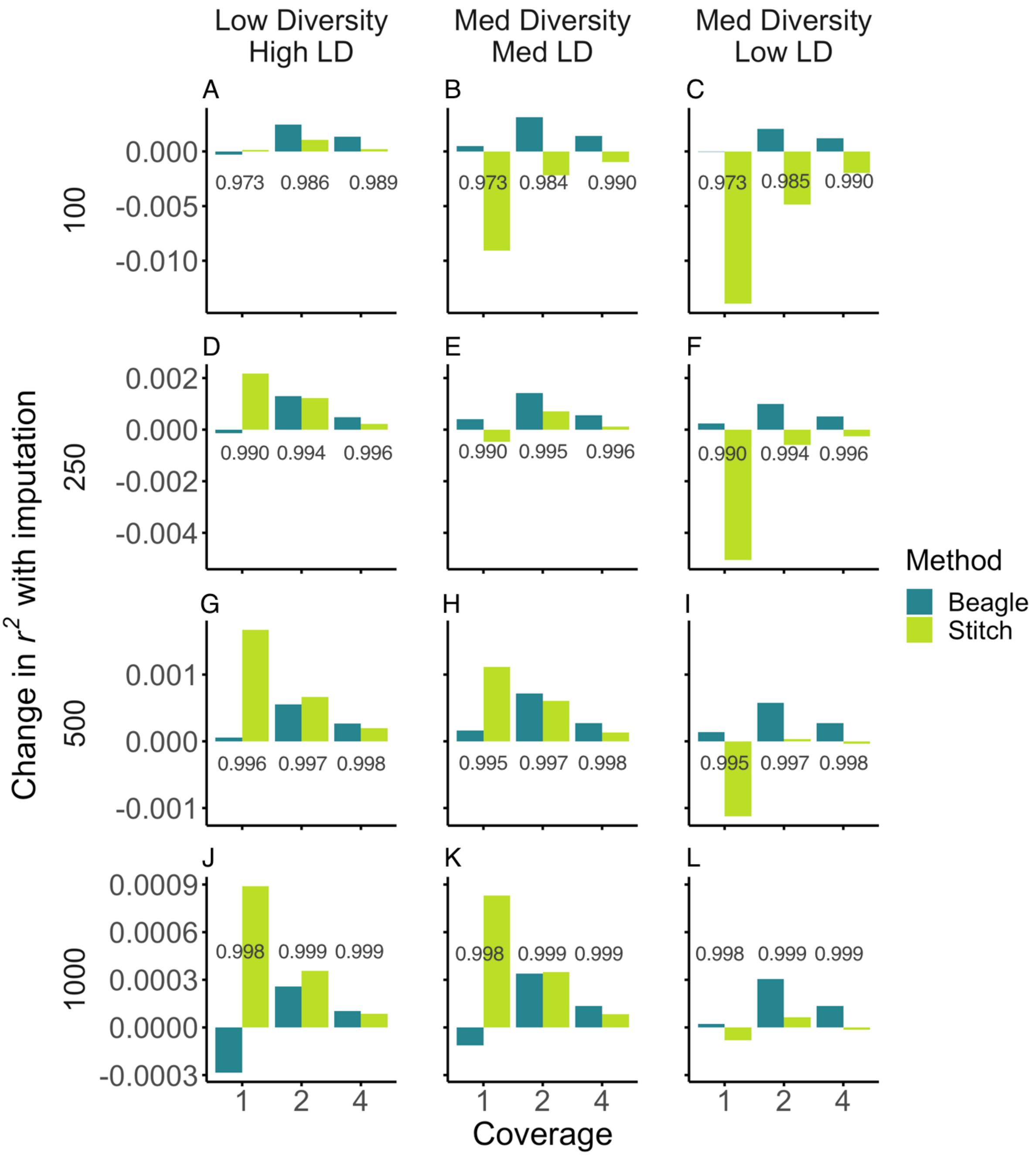


Figure S20. Change in accuracy (r^2) of minor allele frequencies (MAF) estimation using imputed genotype probabilities from STITCH and Beagle, relative to non-imputed genotype likelihoods. Values above the x-axis show r^2 for MAF estimated without imputation. The three diversity/LD scenarios are arranged in columns, sample sizes (n=100, 250, 500 and 1000) are arranged in rows, and sequencing depths are shown on the x-axis. 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*, *34*(3), 407–415.

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.

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, *33*(1), 54–78.

Holsinger, K. E., Lewis, P. O., & Dey, D. K. (2002). A Bayesian approach to inferring population structure from dominant markers. *Molecular Ecology*, *11*(7), 1157–1164.

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.

Jørsboe, E., & Albrechtsen, A. (2019). A Genotype Likelihood Framework for GWAS with Low Depth Sequencing Data from Admixed Individuals. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/10.1101/786384v1.full-text

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., … Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, *12*, 231.

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.

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*, 289.

Kousathanas, A., Leuenberger, C., Link, V., Sell, C., Burger, J., & Wegmann, D. (2017). Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*, *205*(1), 317–332.

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

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* , *27*(21), 2987–2993.

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 (p. 105346). doi: 10.1101/105346

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., & Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124–1132.

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 (p. 658237). 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.

MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., … Thorner, A. R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, *19*(1), 30.

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.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., … Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, *44*(6), 631–635.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, *105*(3), 767–779.

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. (2012). Association testing for next-generation sequencing data using score statistics. *Genetic Epidemiology*, *36*(5), 430–437.

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

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, *28*(4), 289–301.

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 = Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie*, *136*(2), 102–112.