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)

Abstract

Low-coverage whole genome sequencing (lcWGS) has emerged as a powerful and cost-effective approach for population genomic studies in both model and non-model species. However, with read depths too low to confidently call individual genotypes, lcWGS requires specialized analysis tools that explicitly account for genotype uncertainty. A growing number of such tools have become available, but it can be difficult to get an overview of what types of analyses can be performed reliably with lcWGS data, and how the distribution of sequencing effort between the number of samples analyzed and per-sample sequencing depths affects inference accuracy. In this introductory guide to lcWGS, we first illustrate how the per-sample cost for lcWGS is now comparable to RAD-seq and Pool-seq in many systems. We then provide an overview of software packages that explicitly account for genotype uncertainty in different types of population genomic inference. Next, we use both simulated and empirical data to assess the accuracy of allele frequency and genetic diversity estimation, detection of population structure, and selection scans under different sequencing strategies. Our results show that spreading a given amount of sequencing effort across more samples with lower depth per sample consistently improves the accuracy of most types of inference, with a few notable exceptions. Finally, we assess the potential for using imputation to bolster inference from lcWGS data in non-model species, and discuss current limitations and future perspectives for lcWGS-based population genomics research. With this overview, we hope to make lcWGS more approachable and stimulate its broader adoption.

Keywords: genotype likelihoods, bioinformatics, allele frequencies, population structure, selection scan, genotype imputation

1. Introduction

Despite massive reductions in the cost of DNA sequencing over the past decades, researchers remain faced with decisions about how to distribute sequencing effort along three dimensions: 1) how much of the genome to sequence (breadth of coverage), 2) how deeply to sequence each sample (depth of coverage), and 3) the total number of samples to sequence. Until recently, reduced-representation sequencing (e.g. RAD-seq), through which a small random portion of the genome can be sequenced deeply in many individuals to allow for simultaneous variant discovery and high-confidence genotyping, has been the most popular approach for population genomics of non-model organisms (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Davey et al., 2011; McKinney, Larson, Seeb, & Seeb, 2017). While this approach undoubtedly has led to a breakthrough in our ability to examine genome-wide patterns of variation, an important limitation is that large stretches of the genome between markers remain unsampled (Figure 1A). Accordingly, RAD-seq data may miss signatures of selection and adaptive divergence that are highly localized in the genome (Lowry et al., 2017; Tiffin & Ross-Ibarra, 2014).

In a growing number of cases, whole genome sequencing has identified striking peaks of differentiation or strong associations with phenotypes that went completely undetected with RAD-seq data [see e.g. 1) Aguillon, Walsh, & Lovette, 2020 vs. Aguillon, Campagna, Harrison, & Lovette, 2018; 2) Campagna et al., 2017 vs. Campagna, Gronau, Silveira, Siepel, & Lovette, 2015; 3) Clucas, Lou, Therkildsen, & Kovach, 2019 vs. Clucas et al., 2019; and 4) Szarmach, Brelsford, Witt, & Toews, 2021], suggesting that full genome coverage often is needed to understand mechanisms of adaptation. However, whole genome sequencing at sufficient depths to confidently call individual genotypes is still prohibitively expensive on a population scale for many projects. A popular cost-effective alternative is to sequence pools of individuals (Pool-seq; Schlötterer, Tobler, Kofler, & Nolte, 2014; Figure 1B). When the number of individuals pooled and sequencing depth are sufficient, Pool-seq is a powerful approach for obtaining reliable estimates of population-level parameters (Futschik & Schlötterer, 2010; Zhu, Bergland, González, & Petrov, 2012). However, all information about individuals is lost, making it difficult to control for uneven contribution to the pool and precluding individual-level analyses as well as detection of cryptic substructure among sampled individuals (Anderson, Skaug, & Barshis, 2014; Fuentes-Pardo & Ruzzante, 2017).

Low-coverage whole genome sequencing (lcWGS) is emerging as a cost-effective alternative that allows population-scale screening of the entire genome while retaining individual information for - in many cases - a comparable cost to RAD-seq and Pool-seq. The underlying strategy is to maximize the information content in the sequence data by spreading it across the entire genomes of many separately barcoded individuals (Figure 1C). This way, we sacrifice depth of coverage (repeated sequencing of the same locus in the same individual), and therefore confidence in individual genotypes, in return for much greater breadth of coverage and potentially also larger sample sizes.

At low depth of coverage, individual genotypes cannot reliably be inferred (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012; Nielsen, Paul, Albrechtsen, & Song, 2011). However, for most population-level questions, it is not the specific genotype of any particular individual that matters, but rather the overall population characteristics (e.g. allele frequencies, linkage disequilibrium (LD) patterns, etc). Similarly, for questions about genetic relationships between individuals, it is not the genotype at any particular single nucleotide polymorphism (SNP) that matters, but rather patterns of variation across SNPs genome-wide. Accordingly, probabilistic analysis frameworks that account for uncertainty about the true genotype (instead of assuming that any one genotype is correct) can integrate over the uncertainty about individual genotypes for population-level inference of variation at particular SNPs (e.g. allele frequencies, population differentiation), and integrate over the uncertainty about an individual's genotype at each particular SNP to make inference about that individual's overall genetic signature (e.g. level of inbreeding, admixture proportions; Buerkle & Gompert, 2013; Nielsen et al., 2012, 2011).

Simulation studies have demonstrated that when sequencing data are analyzed within this type of probabilistic statistical framework that accounts for genotype uncertainty, sampling many individuals each at low read depth actually provides more accurate estimates of many population parameters than higher read depth for fewer individuals (Buerkle & Gompert, 2013; Fumagalli, 2013; Nevado, Ramos-Onsins, & Perez-Enciso, 2014). In fact, these studies have suggested that spreading sequencing depth to 1–2 reads per locus and individual (1–2x coverage or less) - and increasing sample sizes accordingly - maximizes the information gained about a population. Many recent empirical studies have demonstrated the power of this approach (examples are listed in Table S1). Some of the first applications included identification of genomic regions repeatedly associated with marine-freshwater adaptation in stickleback (Jones et al., 2012), adaptation to an Arctic lifestyle in polar bears (Liu et al., 2014), and divergence among killer whale ecotypes (Foote et al., 2016). More recently, lcWGS was used e.g. to identify genes involved in rapid adaptation to fisheries-induced size selection in experimental populations of Atlantic silversides (Therkildsen et al., 2019), map hybrid incompatibility genes in swordtail fish (Powell et al., 2020), scan for soft sweeps in response to white-nose syndrome in bats (Gignoux-Wolfsohn et al., 2021), build ultra-dense crossover maps in Arabidopsis (Rowan et al., 2019), and assess admixture patterns and elevated differentiation across massive linkage blocks along environmental gradients in several non-model organisms (Clucas et al., 2019; Mérot et al., 2021; Wilder, Palumbi, Conover, & Therkildsen, 2020).

Yet, despite the clear promise, adopting a lcWGS approach can seem daunting because working with genomic data in a probabilistic framework requires both a shift in the way we think about our data and a different toolbox that incorporates genotype uncertainty in downstream analysis. In recent years, there has been a proliferation of programs that can explicitly account for genotype uncertainty in population genomic inference. But for newcomers, it can be difficult to get an overview of what types of analyses can reliably be performed with this data type and what experimental designs will provide the most robust results for a particular system and question, e.g. how to best divide a given sequencing effort between the number of samples vs. the depth of sequencing per sample.

The goal of this paper is to provide a practical “field guide” for researchers considering a lcWGS approach. We first illustrate that lcWGS is now a feasible option for many research projects by comparing the current costs and requirements of lcWGS to alternative sequencing strategies (Section 2). Next, we introduce the basic statistical framework used to account for genotype uncertainty inherent to lcWGS data, and provide an overview of current analytical tools built under a probabilistic framework to help readers identify software that can robustly perform common types of population genomics inference with lcWGS data (Section 3). To guide experimental design, we then use both genetic simulations (Section 4) and down-sampling of empirical data (Section 5) to assess the accuracy of population genomic inference under different sequencing strategies. We evaluate trade-offs between sample size and depth of coverage, compare the power of lcWGS to RAD-seq and Pool-seq, and explore the potential of genotype imputation for bolstering inference with lcWGS data. Finally, in Sections 6 and 7, we review challenges and limitations associated with lcWGS data and discuss future perspectives. With this practitioner-centered overview, we hope to make lcWGS more accessible to a wider group of researchers and stimulate broader adoption of this powerful approach, while inspiring future development of population genomic inference methods for lcWGS data.

2. Feasibility: What does lcWGS cost and what resources are required?

2.1 Current sequencing costs

It is a widespread assumption that whole genome sequencing approaches are still too expensive for researchers working on modest budgets. Yet, due to spectacular drops in sequencing costs over the past decades (the cost per Mb of sequencing is today >600,000 times cheaper than in 2000; Wetterstrand, 2021), lcWGS can now - in many cases - be performed at similar per-sample costs as reduced-representation techniques. Table 1 provides estimates of the total per-sample cost for both library preparation and sequencing (based on May 2021 pricing) for organisms of different genome sizes. The cost of lcWGS inevitably scales with genome size (because more sequence data are needed to provide a target coverage level of a large vs. a small genome), and this approach therefore may remain impractical for organisms with extremely large genome sizes. However, even for organisms with sizeable genomes around 1 Gb (e.g. many birds, fish, invertebrates, and plants), the per-sample cost with 1-2x sequencing coverage (20-35 USD) is now on par with the 30 USD recently reported for genotyping 20,000 variable RAD-seq loci, the 15 USD for a custom sequence capture approach for 500 - 10,000 loci (Meek & Larson, 2019) and 48 USD for custom exome capture (Puritz & Lotterhos, 2018). For organisms with smaller genome sizes, lcWGS can be cheaper than reduced-representation approaches, and prices are likely to drop further as sequencing costs continue to decrease.

2.2. Library preparation

In most cases, Pool-seq approaches remain the most cost-effective way to obtain genome-wide population-level data because it only requires preparation of a single sequencing library per population. The obvious downside is that all individual-level information is lost, precluding many types of analysis. Despite this limitation, Pool-seq has gained popularity because preparation of individually indexed libraries for hundreds of samples used to be labor-intensive and costly (the costs for preparing hundreds of libraries could easily outweigh the cost of sequencing). LcWGS has now become a viable alternative because of the development of cheap library preparation methods with efficient workflows that make it both practical and affordable to process hundreds of samples (see Table S1 for an overview of methods used in recent lcWGS studies). Therkildsen & Palumbi (2017), for example, describe a robust easy-to-implement protocol based on reduced reaction volumes of Illumina's Nextera kit, which brings per-sample reagent costs down to ~8 USD (based on current reagent pricing). Several other protocols that stretch reagents in commercial kits reach similar price points (e.g. Gaio et al., 2019; Li et al., 2019). An advantage of commercial kit-based protocols is that they often work "straight out of the box" or require only limited optimization. Substantial further cost savings can be achieved with protocols based on in-house expression and purification of tn5 transposase (the enzyme used in Illumina's Nextera tagmentation approach), such as described by Picelli et al., (2014) and Hennig et al., (2018). With those protocols, per-sample library costs can be brought to <<1 USD, substantially reducing overall project costs when analyzing hundreds of samples and essentially eliminating the added cost of individually indexed libraries, making total costs for lcWGS equivalent to Pool-seq for similar total sequencing effort per population.

LcWGS library preparation methods also tend to be very efficient and scalable. For example, tn5 (tagmentation)-based protocols (like the one used by Therkildsen & Palumbi 2017) make it possible to prepare 96 libraries in <5 hours (with <3 hours hands-on time) - substantially less time than needed for most RAD-seq protocols (Meek & Larson, 2019). The Therkildsen and Palumbi (2017) protocol also works well for relatively degraded DNA and requires only very small amounts of input DNA (~2.5 ng). For highly degraded DNA, we have had great success with the Carøe et al. (2018) single-tube method. Other cost-effective protocols produce successful lcWGS libraries even from picogram-levels of input DNA (Hennig et al., 2018; Meier, Salazar, Kučka, Davies, & Dréau, 2020; Picelli et al., 2014), for example enabling high throughput production of libraries from individual zooplankters (Beninde, Möst, & Meyer, 2020). Methods that sidestep DNA extraction with tagmentation directly on cells or tissue may lead to additional efficiencies for lcWGS library preparation in the future (Vonesch et al., 2020).

2.3. The need for a reference genome

For non-model organisms, a key constraint associated with lcWGS is the need for a reference genome to map the short-read sequence data generated from each individual. If a reference genome is not already available for the species of interest, a common solution is to map to a reference genome of a related species. While this can work well in some contexts, increasing phylogenetic divergence between the re-sequenced species and the reference genome can restrict mapping to the genomic regions that are most conserved between the two taxa and bias

estimates of population genomic parameters (Bohling, 2020; Nevado et al., 2014). Major differences in genome organization (e.g. structural and copy number variants) can also exist even between closely related species (Ekblom & Wolf, 2014). For these reasons, a species-specific reference sequence is preferable where it can be obtained.

As a shortcut to obtaining species-specific reference sequence without de novo assembling a full genome, Therkildsen and Palumbi (2017) mapped lcWGS reads to a reference transcriptome, in practice performing 'in-silico' exome capture. However, major advances in affordable long-read sequencing, powerful genome scaffolding techniques, and improved assembly algorithms now enable chromosome-scale assemblies at a much lower cost and faster speed than earlier approaches (reviewed by Rice & Green (2019)), facilitating high-quality assemblies of mammalian-sized genomes (several Gb) with chromosome-length scaffolds for as little as 1,000 USD (Dudchenko et al., 2018; Gatter, von Löhneysen, Drozdova, Hartmann, & Stadler, 2020). Therefore, at this point, it probably makes sense to start most new lcWGS studies with a de novo genome assembly or upgrade, if a reference sequence of sufficient quality is not available.

BOX 1: Glossary

Base quality score: A metric associated with each base (nucleotide) in a sequencing read that indicates the probability that the base is called incorrectly.

Bayesian inference: A statistical inference strategy that estimates model parameters by characterizing its posterior probability distribution (i.e. P(parameter | data)). By the Bayes theorem, the posterior probability is formulated as a product of the likelihood function and the prior probability distribution (probability distribution of model parameters before considering the data) divided by the marginal probability of the data (which is a constant), i.e. P(parameter | data) = P(data | parameter) * P(parameter) / P(data)

Genotype dosage: The expected count of an allele in an individual. For a diploid individual, the genotype dosage of the B allele = P(AA | data)*0 + P(AB | data)*1 + P(BB | data)*2, where A and B represent the two alleles at the site, and e.g. P(AB | data) represents the posterior probability of the heterozygous genotype.

Genotype imputation: A method to infer missing genotypes and bolster genotype likelihood estimation by identifying stretches of haplotypes shared between individuals.

Genotype likelihoods (GLs): The probability of observing the sequencing data at a certain site in an individual given that the individual has each of the possible genotypes at this site (e.g. for diploids there are 10 possible genotypes, which can be reduced to 3 if the major and minor alleles are known), i.e. P(data | genotype), or L(genotype).

Genotype likelihood (GL) model: The mathematical model used to compute GLs. Different GL models are built under different assumptions about the data, in particular about the sequencing error profile. For example, the GATK model assumes that the sequencing quality scores accurately capture the probability of sequencing error, and that all errors are independent. In comparison, the Samtools model assumes that once a first error occurs at a certain site in an individual, subsequent errors are more likely to occur at this site.

Low-coverage whole genome sequencing (lcWGS): We use this term to refer to whole genome re-sequencing of individuals (i.e. labeled with unique barcodes) with depth too low to reliably call genotypes without imputation (<5x). Note, however, that even for medium sequencing depth (5-20x), inference accuracy may improve under a probabilistic analysis framework based on GLs, rather than working with called genotypes (Nielsen et al., 2011).

Mapping quality score: A metric associated with each sequencing read aligned to the reference genome that indicates the probability that the read is aligned to the correct position in the reference sequence.

Maximum likelihood inference: A statistical inference strategy that estimates model parameters by choosing the parameters that maximize the likelihood of the data. In other words, the maximum likelihood estimators of model parameters = argmax(L(parameter))

Posterior genotype probability: The probability of an individual having one of the possible genotypes at a certain site given the sequencing data, i.e. P(genotype | data).

Prior genotype probability: The probability of an individual having one of the possible genotypes at a certain site before considering the sequencing data for this individual at this site, i.e. P(genotype). The prior genotype probability can be uniform (i.e. all genotypes are equally likely to occur), or can be informed by the allele frequency or the site frequency spectrum (SFS) at this site for all individual samples. It is often used for the estimation of posterior genotype probability in Bayesian inference.

Restriction site-associated DNA sequencing (RAD-seq): A group of techniques for sequencing short flanking regions around restriction enzyme cut sites to obtain random samples of genetic markers across the entire genome. These markers are typically sequenced at high depth (e.g. >20x) for each individual so that individual genotypes can be confidently determined.

Sample allele frequency (SAF) likelihood: The probability of observing sequencing data at a certain site across all individual samples given each possible sample allele frequency at this site (e.g. for diploids, the possible sample allele frequencies range from 0 to 2n; n=sample size), i.e. P(data | sample allele frequency).

Whole genome sequencing of pools of individuals (Pool-seq): A whole genome sequencing strategy in which unlabeled DNA from multiple individuals is pooled before sequencing. The sequencing depth is typically low on a per-individual level but high for each pool (e.g. >50x).

Due to the absence of individual barcodes, all individual-level information is lost in the sequencing data.

3. The toolbox: What types of analysis can we do with low-coverage data?

The major challenge in working with lcWGS data is that individual genotypes cannot be accurately inferred (Li, Sidore, Kang, Boehnke, & Abecasis, 2011; Nielsen et al., 2012, 2011). Many analytical tools that incorporate uncertainty about individual genotype calls have therefore been developed in recent years, covering a broad diversity of common types of population genomic inference. We here briefly introduce the most widely used applications (see Table 2 for a more comprehensive list and the Supplementary Text Part 3 for additional details) and also provide a tutorial with example data as a starting point for exploration: https://github.com/nt246/lcwgs-guide-tutorial.

Currently, the most widely used program for lcWGS analysis is ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014), a comprehensive package that implements an extensive variety of analysis options. Because of its broad use and versatility, ANGSD will feature prominently in this section's overview of available tools. However, we also seek to highlight that a variety of alternative programs are available for most types of analysis (Table 2).

3.1. Accounting for genotype uncertainty

The most common way to incorporate uncertainty about true genotypes is to use genotype likelihoods (GLs) rather than genotype calls in downstream analyses. A GL reflects the probability of observing the sequencing reads that cover a specific site in an individual if said individual has a particular genotype at this site. GLs refer to the set of likelihoods computed for each of all the possible genotypes that individual could hold at that site (e.g. for diploids there are ten possible genotypes: AA, AC, AG, AT, CC, CG, CT, GG, GT, and TT, which can be reduced to three possible genotypes if the major and minor allele at a site are known, i.e. major-major, major-minor, minor-minor).

The key factors that prevent us from confidently identifying the true genotype with lcWGS data are uncertainties about 1) whether both alleles of a diploid individual have been sampled in the stochastic sequencing process, 2) whether the base call (A, C, G, or T) at each position of a sequencing read is correct, and 3) whether sequencing reads have been mapped to the correct position in the genome. Several different models have been proposed for how the first two sources of uncertainty should be accounted for in estimation of GLs (to our knowledge, no current models directly factor in mapping accuracy). Currently, the most commonly used GL models are probably the GATK model (McKenna et al., 2010) and the Samtools model (Li, 2011; Li et al., 2009) implemented in ANGSD (Korneliussen et al., 2014). The key difference between these two GL models is that the GATK model assumes sequencing errors are independent, whereas the Samtools model assumes a correlated error structure.

Unfortunately, the effects of GL model choice on downstream analyses are still incompletely understood and likely depend on a diversity of factors including the accuracy of base calling and base quality scores, the sequencing depth, and the type of inference sought. In our comparative assessment (Section 4), we found that many types of analysis gave nearly identical results with the GATK and the Samtools models, (Figure S4-S7), but that GL model choice can strongly influence the number of rare alleles estimated from simulated low-coverage (≤2x) data (Section 4.1, see also Korneliussen et al., 2014 for a similar finding). However, more research is needed to compare the performance of different GL models, and in the meantime, it may be prudent to compare inference with several different GL models with a subset of the data for each new dataset, particularly for analyses that rely on rare alleles.

On a related note, while base quality scores should reflect the probability of each called base in a sequencing read being incorrect, it is widely recognized that instrument-reported values can sometimes be inaccurate (i.e. poorly predicting the true frequency of sequencing errors; Callahan et al., 2016; Ni & Stoneking, 2016). Given the central importance that base quality scores typically play in estimating GLs when coverage is low, miscalibrated scores can potentially bias inference, especially related to rare alleles (Kousathana et al. 2017). It may therefore be advantageous to recalibrate base quality scores by first identifying putative sequencing errors in the data and then adjusting the base quality scores based on the observed error rates and patterns. This type of recalibration can be performed as an extra data preprocessing step but is also implemented in some GL models (e.g. the SOAPsnp model in ANGSD). Unfortunately, some of the most widely used methods (e.g. those implemented in GATK and ANGSD; Auwera & O'Connor, 2020; Li et al., 2009) require a database of known variable sites, which is not available for most organisms, and inputting an inaccurate variant database can sometimes inadvertently result in further miscalibration of quality scores (Orr 2020). For non-model species, there may be more promise in approaches based on synthetic spike-ins (e.g. PhiX; Zook et al., 2012; Ni & Stoneking, 2016) or monomorphic genomic regions (e.g. sex chromosomes, ultra-conserved elements, or organellar DNA; Kousathana et al. 2017) for which no true genetic variation is expected and sequencing errors can more readily be identified. Other recently proposed techniques based on k-mer analysis (Orr, 2020) or comparison of quality score profiles (Chung & Chen, 2017) also sidestep the need for a variant database. However, none of these methods have yet been extensively validated for low-coverage data. For now, a conservative approach may be to filter out bases with low quality scores, but that results in data loss and does not fully address the issue of potential miscalibration, so more research in this area is needed.

3.2. From raw reads to SNP identification

The initial steps in processing lcWGS data are similar to those used in many other NGS approaches, such as high-coverage whole genome sequencing and Pool-Seq (Figure 2). These include trimming adapter sequence and bases with low quality scores, mapping (aligning) reads to a suitable reference genome, removing poorly mapped and duplicated reads, and - depending on requirements of downstream tools - potentially clipping overlapping sections of read pairs and realigning reads that span indels (see e.g. Therkildsen & Palumbi 2017). It is in

the downstream processing of the resulting filtered bam files that high-coverage and low-coverage workflows diverge and where a probabilistic framework based on GLs becomes central for low-coverage data.

The optimal approach in a GL-based framework would arguably be to always compute GLs for every site in the genome, including sites that appear to be invariant in a sample (because with lcWGS data we cannot be completely confident that we have not missed an alternative allele in one or more of our samples). While this approach is required for some types of analysis (e.g. all estimates of genetic diversity and the site frequency spectrum), other types of analysis (e.g. analysis of population structure or outlier scans) are more tractable and computationally efficient if only polymorphic sites are considered. Thus, a more practical solution for those types of analysis is to initially identify likely polymorphic sites and restrict downstream GL-based inference to those sites.

Although many types of genetic variants exist, lcWGS analysis has so far typically only considered bi-allelic single-nucleotide polymorphisms (SNPs). A range of different programs can identify SNPs from lcWGS data (Table 2). Because of built-in integration of a broad variety of downstream analysis tools, ANGSD is often a convenient option. ANGSD identifies SNPs as sites with minor allele frequencies significantly larger than zero. In this case, the number of alleles at each site is restricted to two (major and minor allele), with the identities of these alleles either determined through a maximum likelihood approach, setting the more common allele as the major allele (Jørsboe & Albrechtsen, 2019; Skotte, Korneliussen, & Albrechtsen, 2012) or by user specification (e.g. setting the reference or ancestral allele as the major allele). ANGSD currently does not allow for identification of indels or multi-nucleotide polymorphisms, but users could potentially identify bi-allelic indels with a different tool, such as Freebayes (Garrison & Marth, 2012) or GATK (McKenna et al., 2010), and import estimated GLs into ANGSD for use in downstream analysis. Regardless of the program used, quality control filters can be crucial to ensure data reliability. Table 3 provides an overview of key filters that should be considered for different types of analysis with lcWGS data.

3.3. Individual-level analyses

Despite the lack of called genotypes, lcWGS data can be used for a wide range of individual-level analyses, which we define as those that do not require a priori grouping of individual samples. It should be noted that the input formats for the different approaches differ between programs and that in some cases SNP identification can be performed as part of the analyses (see specific manuals). Note also that none of the analyses listed in this subsection are possible with Pool-seq data.

Population structure: A key component of many population genomic studies is to characterize population structure, using dimensionality reduction (e.g. PCA and PCoA) and/or model-based clustering (e.g. admixture analysis). Dimensionality reduction methods are based on a covariance matrix (PCA) or distance matrix (PCoA). Several methods for computing these matrices while accounting for genotype uncertainty have been implemented. ANGSD, for example, can either randomly sample one read per individual per site or use the most common

allele to represent the individual's allele frequency at this site (as either 0 or 1) and then calculate the covariance and distance between every pair of individuals from these allele frequencies. This simple approach has been shown to work well for datasets with very low sequencing depth and uneven coverage across samples (see Section 4.2 and the ANGSD manual). PCAngsd (Meisner & Albrechtsen, 2018), in contrast, estimates the covariance matrix from GLs while taking population structure into account.

Model-based clustering methods that estimate admixture proportions of each sample assuming a model of discrete ancestral populations are also implemented in several software programs using GLs as input. These include NGSAdmix (Skotte, Korneliussen, & Albrechtsen, 2013) and Ohana (Cheng, Racimo, & Nielsen, 2019) that both adopt a maximum likelihood implementation of the classic STRUCTURE model (Pritchard, Stephens, & Donnelly, 2000; Tang, Peng, Wang, & Risch, 2005), but differ in their optimization approaches. PCAngsd implements admixture analysis with a different approach, which uses an intermediate output from its PCA analysis as a starting point for model-based clustering. PCAngsd has been shown to outperform NGSAdmix in runtime without strongly compromising its inference accuracy, making it potentially more suitable for larger datasets (Meisner & Albrechtsen, 2018).

Selection scans: Several of the mentioned clustering programs also implement selection scan approaches that do not require a priori grouping of individuals, as their general strategy is to locate outlier loci that exhibit patterns of genetic variation among individuals that are highly different from the genome-wide average. For example, PCAngsd (Meisner & Albrechtsen, 2018; Meisner, Albrechtsen, & Hanghøj, 2021) implements the FastPCA method by Galinsky et al. (2016) in a GL framework and in Ohana, SNPs that exhibit a significantly different covariance structure can be identified as potentially under selection.

Genome-wide association studies (GWAS): Multiple statistical frameworks have been developed to take genotype uncertainty into account in scans for genotype-phenotype associations. GWAS often require large sample sizes to gain sufficient power, and a lcWGS/GL-based approach provides an opportunity to maximize the number of individuals studied in a cost-efficient way. Several GL-based GWAS approaches implemented in ANGSD have shown power to discover meaningful associations, including in the presence of population structure (Jørsboe & Albrechtsen, 2019; Skotte et al., 2012). These methods range from simple case / control associations for identifying variants associated with binary phenotypes (Kim et al., 2011) to analysis of quantitative traits with incorporation of covariates (Skotte et al. 2012; Jørsboe & Albrechtsen 2019). The maximum likelihood approach recently developed by Jørsboe & Albrechtsen (2019) also explicitly estimates the effect size of each locus.

Linkage disequilibrium (LD): LD estimation has many important applications, for example relating to inference of population size, demographic history, selection, and discovery of structural variants (Slatkin, 2008). In addition, since many downstream analyses make assumptions about the independence of genomic loci, LD estimation is essential for excluding strongly linked loci from a dataset (i.e. LD pruning). Several approaches have been developed to estimate LD from GLs (i.e. taking genotype uncertainty into account), implemented e.g. in

GUS-LD (Bilton et al., 2018) and ngsLD (Fox, Wright, Fumagalli, & Vieira, 2019). Unfortunately, the computational complexity of GUS-LD is too high for it to be practical for whole genome data, but ngsLD has a more efficient algorithm and has different built-in functionalities to reduce its computational complexity (e.g. restricting LD estimation between SNPs within a set distance, setting a minor allele frequency filter, etc.), and comparative evaluation has indicated that ngsLD tends to show less bias at low read depths (1-2x) than GUS-LD (Bilton et al., 2018; Fox et al., 2019).

Other types of analyses: In addition to the examples discussed above, many other specialized software packages have been developed to account for genotype uncertainty in various types of inference, including estimation of relatedness among individuals (Korneliussen & Moltke, 2015; Link et al., 2017), parentage inference (Whalen, Gorjanc, & Hickey, 2019) and pedigree analysis (Snyder-Mackler et al., 2016), estimation of individual inbreeding coefficients (Link et al., 2017; Vieira, Fumagalli, Albrechtsen, & Nielsen, 2013) and identity-by-descent tracts (Vieira, Albrechtsen, & Nielsen, 2016), tests for introgression such as computation of ABBA-BABA/D-statistics (Korneliussen et al., 2014), and construction of linkage maps (Rastas, 2017). More examples are listed in Table 2. It is also important to note that samples sequenced at low-coverage of the nuclear genome typically have very high sequencing depth across the mitochondrial genome due to its much higher copy number in each cell, enabling recovery of high-confidence full mitochondrial genome sequences for each individual (see e.g. Therkildsen & Palumbi 2017). LcWGS thus provides a cost-effective way to generate full mitochondrial genome sequences for hundreds of individuals, enabling unprecedented resolution for phylogeographic analysis (Lou et al., 2018; Margaryan et al., 2020).

3.4. Population-level analyses

When individual samples can be grouped into discrete populations or categories based on either prior information (e.g. sampling location or experimental treatment) or results from individual-level population structure analyses (e.g. model-based clustering), analyses can be conducted at the population level.

Allele frequency estimation: The estimation of population-specific allele frequencies is essential for most population genomic studies as it is a required input for many downstream analyses. Many programs, such as ANGSD (implementing the method of Kim et al., 2011) or ATLAS (Link et al., 2017), can estimate minor allele frequencies for each site using a maximum-likelihood or Bayesian approach. In programs where population-specific estimates are obtained by running the program on each population separately (e.g. ANGSD), it is crucial for users to explicitly define the same alleles as major and minor in all populations to avoid inadvertently computing the frequency of opposite alleles in different populations.

Site frequency spectrum (SFS): The population-specific SFS is another population genomic parameter essential for many downstream analyses. A challenge in estimating the SFS with low-coverage data is that low-frequency SNPs are less likely to be identified as polymorphic and therefore an SFS directly estimated from allele frequencies at identified SNP positions can be

biased towards intermediate frequencies. To get around this issue, ANGSD estimates the SFS by using the sample allele frequency (SAF) likelihoods to formulate the likelihood function of the SFS, which the program then optimizes (Nielsen et al., 2012). Depending on the availability of an outgroup or ancestral reference genome, the inferred SFS can either be folded or unfolded and ANGSD can estimate the SFS jointly for up to four populations (Korneliussen et al., 2014). This probabilistic approach can correct for the bias caused by low-coverage data, but its performance can be sensitive to the choice of underlying GL model (Korneliussen et al., 2014, see also Section 4.1). Another important limitation is that the runtime of the SFS estimation algorithm currently implemented in ANGSD grows quadratically with the number of samples and it can become impractical to run across the whole genome if the sample size is very large. One strategy is to estimate SFS by chromosome or in smaller windows and sum them up in the end. Implementation of a faster algorithm (Han, Sinsheimer, & Novembre, 2015) may also be included in future ANGSD releases (Fumagalli, personal communication).

Genetic diversity and neutrality test statistics within a single population: Derived estimators of genome-wide genetic diversity θ, such as nucleotide diversity π and Watterson's estimator, can be directly calculated from the population-specific SFS. These estimators of θ can also be computed within genomic windows from window-specific SFS and subsequently, different neutrality test statistics (e.g. Tajima's D) that evaluate the skewness of SFS in each genomic window can be calculated. Individual heterozygosity estimates can be obtained by estimating the SFS for individuals (rather than populations). All these diversity statistics can be computed based on an infinite sites model implemented in ANGSD. In contrast, ATLAS (Link et al., 2017) bases its θ estimation on a model that allows for back mutations (Felsenstein, 1981), which can be more appropriate when working with ancient samples. Regardless of the method used, it is important to note that when generating diversity estimates, non-variable sites should be included in the calculation, and therefore minimum minor allele frequency filters or SNP p-value filters should not be used.

Genetic differentiation between populations: In addition to estimates of *within*-population diversity, the genetic differentiation *between* populations can be estimated with a variety of different statistics, from simply quantifying the allele frequency difference to more complex statistics such as relative genetic differentiation (F_{ST}) and absolute genetic divergence (d_{xy}). Various estimators of F_{ST} can be computed from GL data using ANGSD, ngsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014), or vcflib (Garrison et al., 2021; see the Supplementary Text Part 3 for more detail). vcflib can also estimate pF_{ST}, which, contrary to what the name suggests, is not an F_{ST} estimator, but a statistic that quantifies the significance of allele frequency differences between populations in face of genotype uncertainty (Domyan et al., 2016). In contrast to F_{ST}, no established method to estimate d_{xy} from GLs has, to our knowledge, been included in major software packages. Various custom scripts have been shared (see e.g. https://github.com/mfumagalli/ngsPopGen/tree/master/scripts, https://github.com/marqueda/PopGenCode/blob/master/dxy_wsfs.py). Note, however, that d_{xy} may be over-estimated with these scripts so they should be used only for inspecting the relative distribution of d_{xy} across the genome (Foote et al., 2016) and not to make inferences based on its absolute values.

Other analyses based on derived statistics: In addition to the methods that work directly with GLs, many other types of population-level analysis can be conducted based on the derived statistics mentioned above. For example, several commonly used software tools can use allele frequency matrices as input to infer population relationships and potential gene flow (e.g. Treemix (Bradburd, Coop, & Ralph, 2018; Pickrell & Pritchard, 2012) and conStruct (Bradburd et al., 2018; Pickrell & Pritchard, 2012)), perform selection scans (e.g. BayPass (Gautier, 2015), WFABC (Foll & Gaggiotti, 2008; Foll, Shim, & Jensen, 2015)), association analyses (e.g. BayPass, LFMM2 (Caye, Jumentier, Lepeule, & François, 2019)), or variance partitioning analyses (e.g. RDA (Forester, Lasky, Wagner, & Urban, 2018)). To run these programs, population-level allele frequencies are estimated as explained above (e.g. using ANGSD), but have to be transformed into the appropriate input format using custom scripts. Similarly, the population-specific or multi-dimensional SFS estimated from ANGSD can be used to infer demographic history (e.g. with δaδi (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) or fastsimcoal2 (Excoffier & Foll, 2011)), or to explicitly control for the effect of demography in selection scans (e.g. SweepFinder2 (DeGiorgio, Huber, Hubisz, Hellmann, & Nielsen, 2016)). Both locus-specific neutrality test statistics and F_{ST} values can be used in selection scans (e.g. OutFLANK (Whitlock & Lotterhos, 2015)), and genome-wide F_{ST} estimates can be used, for example, to test for isolation by distance (Mantel test) or to estimate effective migration surfaces (e.g. EEMS (Petkova, Novembre, & Stephens, 2016)). Furthermore, Ancestry_HMM (Medina, Thornlow, Nielsen, & Corbett-Detig, 2018) and ancestryinfer (Schumer, Powell, & Corbett-Detig, 2020) can infer local ancestry across the genome without called genotypes, although they require detailed SNP information for reference populations. Using derived statistics as input data can be a powerful approach to expand the available toolbox for lcWGS. However, unlike the GL-based programs listed in the rest of this section and Table 2, this approach does not carry uncertainty about parameter estimation downstream. Accordingly, if summary statistics rather than GLs are used as input for analysis, p-values etc. should be interpreted with caution and in light of the expected precision given the sample size and sequencing depth (see Section 4).

4. Experimental design: The tradeoffs between sequencing depth per sample and total number of samples analyzed

With a finite sequencing budget, do we learn more about a population from adding more sequencing depth to each individual or stretching the sequencing effort over more individuals? Several previous studies have used simulated data to address this question (e.g. Buerkle & Gompert, 2013; Fumagalli, 2013; Nevado et al., 2014). In general, these studies have found that sampling many individuals at 1x or 2x read depth provides more accurate estimates of many population parameters than higher read depth for fewer individuals. However, both the simulation (e.g. Haller & Messer, 2019; Huang, Li, Myers, & Marth, 2012) and the GL-based data analysis toolboxes (e.g. Fumagalli, Vieira, Linderoth, & Nielsen, 2014; Korneliussen et al., 2014; Meisner & Albrechtsen, 2018) have evolved rapidly since these studies were conducted,

and a more up-to-date evaluation is now needed. Here, we used simulated data to compare common types of population genomic inference under a wide range of sample size and sequencing depth combinations, including depths <1x, which were not explicitly evaluated in earlier studies. Full details about all the simulations and analyses can be found in the Supplementary Methods (Part 1) and Table S2, and our entire simulation and analysis pipeline is available on GitHub (https://github.com/therkildsen-lab/lcwgs-simulation).

4.1. Population genomic inference for single populations

We used SLiM (Haller & Messer, 2019) to simulate an isolated population that has reached mutation-drift equilibrium, and evaluated the accuracy of lcWGS (reads simulated with ART (Huang et al., 2012)) for inferring key population genomic parameters, including allele frequencies, the SFS, θ, Tajima's D (estimated with ANGSD (Korneliussen et al., 2014)), and linkage disequilibrium (estimated with ngsLD (Fox et al., 2019)) under different experimental designs.

As expected, the accuracy of allele frequency estimation consistently increases with both higher sample size and depth of coverage per individual (as measured by the r^2 values in Figure 3). The number of false negative SNPs (i.e. true SNPs in the population that fail to be identified) similarly decreases with higher sample size and higher coverage per individual (Figure S1). Importantly, however, distributing the same total sequencing effort (i.e. the product of sample size and coverage per individual) across more samples, with each sample receiving lower coverage (i.e. going from bottom left to top right in Figure 3) also consistently improves allele frequency estimation, even when each sample is sequenced at a coverage as low as 0.25x. The increased accuracy arises because each allele is less likely to be sequenced more than once with lower per-sample coverage, and thus the effective sample size gets higher.

Consistent with what the authors of ANGSD have previously shown (Korneliussen et al., 2014), we found that for SFS-based inference, the choice of GL model used can strongly influence its result. With the Samtools GL model, Watterson's θ is systematically underestimated when the average coverage is low (≤4x), although Tajima's θ (π) estimates are more robust (Figure S2). Consequently, Tajima's D tends to be overestimated (Figure S3). In contrast, when the GATK GL model is used, Watterson's θ, Tajima's θ, and Tajima's D can all be accurately estimated even at coverage as low as 0.5x (Figure S2, S3). The two GL models differ in performance because both the GATK model and our simulation model assume that each base quality score reflects an independent and unbiased measurement of the probability of sequencing error (Huang et al., 2012; McKenna et al., 2010), whereas the Samtools model assumes that if one sequencing error occurs at a certain locus, subsequent errors are more likely (Li, 2011; Li et al., 2009). As a result, with the Samtools model, lower-frequency mutations are less likely to be identified as polymorphic sites and more likely to be interpreted as sequencing errors when the coverage is low. This leads to an underestimation of the number of singleton mutations when using the Samtools model, and therefore Watterson's θ tends to be underestimated, at least for our simulated data. We note, however, that these low-frequency SNPs have minimal impact on many other types of population genomic analyses and, in fact, are often filtered out. Consistent with this, we did not observe any strong discrepancies between the two GL models in other

types of analysis that we performed in this study (Figure S4-S7). We also stress that the sequencing errors modeled in our simulations may not accurately represent the sequencing error profile in real life, so our result should not be interpreted as a recommendation of one GL model over the other.

For LD estimation, we found that relative estimates (which may be adequate for many uses, e.g. for the identification of LD blocks or LD pruning) could be reliably obtained with per-sample coverage as low as 1-2x. However, higher per-sample coverage (e.g. ≥4x) is required to get precise and accurate absolute estimates of LD (e.g. for use in demographic inference) even with sample size as large as 160 (Figure S4, S5, Fox et al. 2019).

Box 2. Performance of lcWGS vs. Pool-seq in allele frequency estimation

A key advantage of lcWGS over Pool-seq is that each sequencing read can be assigned to an individual so we can detect uneven sequencing coverage and account for it in parameter estimation. But does that matter in practice if the contribution of each individual to the sequencing pool is roughly equal? With our simulated data, we found that a lcWGS analysis approach that accounts for individual-level GLs consistently provides slightly more accurate allele frequency estimates than Pool-seq analysis (which ignores individual-level information) even when all individuals contribute equally to the sequencing pool (Figure 4). This is because the sampling variance inherent to next-generation sequencing creates stochastic variation in the sequencing depth for each individual at each locus (so some by chance will be overrepresented while others will be underrepresented). In practice, inaccuracies due to measurement and pipetting errors, variation in DNA quality, and sequencing biases make it almost impossible to ensure the optimal scenario of even amounts of sequences among samples (Figure S8, see also Schlötterer, Tobler, Kofler, & Nolte, 2014), further enhancing the value of being able to account for sample overrepresentation with individually barcoded reads (Figure S9, S10).

4.2. Inference of spatial structure

To evaluate the power of different lcWGS sampling designs in detecting population structure, we simulated a metapopulation consisting of nine subpopulations located on a three-by-three grid that have reached mutation-drift-migration equilibrium. We first examined a scenario in which gene flow among subpopulations is low (0.25 effective migrants between neighboring subpopulations per generation). In this scenario, the spatial structure among subpopulations can be correctly inferred from PCA even with extremely low sample size (5 samples per subpopulation) and coverage (0.125x coverage per sample; Figure 5A). In addition, migrant individuals and hybrids, when included in the sample, can be identified in the PCA (Figure 5A), which would not be possible with a Pool-seq design.

We then increased the level of gene flow (1 effective migrant between subpopulations every generation). As expected, the power of PCA to resolve the weaker spatial structure slightly declines, but interestingly, small sample size causes a greater loss of power than low coverage

does (Figure 5B). Subpopulations fail to form discrete clusters in PCA space when the sample size per population is 5, unless the coverage is 2x or higher per sample. On the other hand, with a sample size of 10, the correct spatial structure can be inferred with a coverage as low as 0.125x (i.e. a per-population coverage of only 1.25x; Figure 5B). The reason why we can push the per-sample coverage so low is that PCA depends on reliable covariance estimation between some, but not all pairs, of samples in the dataset. To get reliable covariance estimates in a sample pair, both samples need to have at least 1x coverage at some informative SNPs. As sample size increases, the number of all available sample pairs increases quadratically, and the number of sample pairs for which enough informative SNPs are shared also increases quadratically. Therefore, the overall population structure is more likely to be correctly extrapolated from these sample pairs. We also note that, due to computational limitations, our simulations are based on only a single 30Mb chromosome. Since the power of PCA depends on the number of informative SNPs shared between pairs of samples, with a larger genome size, even lower sequencing depth and/or sample size would be required to resolve the spatial structure among subpopulations, given the same SNP density as simulated here (see Figure S11 for an example of this). Lastly, we found that the read sampling method implemented in ANGSD (Korneliussen et al., 2014), the results of which are presented here, outperforms PCAngsd (Meisner & Albrechtsen, 2018) in scenarios with low sample size (e.g. ≤10 samples per population) or very low coverage (e.g. ≤0.25x per sample; Figure S12, S13).

4.3. Scans for divergent selection in the face of gene flow

A primary advantage of lcWGS compared to reduced-representation sequencing approaches is the increased resolution in genome scans for signatures of selection, for example in the form of outlier loci that show elevated levels of differentiation between populations or elevated/depressed neutrality test statics within a single population. To evaluate how experimental design affects our ability to detect outliers, we simulated two populations connected by gene flow that are strongly affected by divergent selection on a subset of loci. We estimated per-SNP F_{ST} between the two populations, as well as Tajima's D and Fay and Wu's H within one of the populations, from lcWGS data to identify the loci under selection (details in the Supplementary Methods).

We first examined a scenario where the size of each population is large (the effective population size (N_e) = $5x10^4$) and gene flow is high (5 effective migrants per generation). In this scenario, seven out of eight SNPs under divergent selection, along with their neighboring neutral SNPs, show highly elevated F_{ST} values compared to the genome-wide background, creating a distinct pattern of narrow genomic islands of divergence (Figure 6; Turner, Hahn, & Nuzhdin, 2005). This F_{ST} landscape can be recovered from lcWGS data with a total sequencing coverage per population as low as 10x (e.g. 40 samples per population and 0.25x coverage per sample, Figure 6). For a given total sequencing effort, however, we observe an increase in background F_{ST} when the sequencing is spread over fewer individuals (e.g. 5 samples per population and 2x coverage per sample give more background noise than 40 samples each at 0.25x coverage), which can lead to overestimated genome-average F_{ST} (Figure S7) and more false positive signals in the outlier detection (Figure 6). With similar sequencing effort, most of these loci

under selection can also be identified by signals of decreased Tajima's D and Fay and Wu's H, although the absolute values of these estimates are sensitive to both sample size and coverage (Figure S14, S15). Unlike for F_{ST}, spreading the same sequencing effort across more samples does not consistently improve the accuracy of these neutrality test statistics (as some require higher coverage for accurate estimation). We also estimated F_{ST} and neutrality test statistics in a scenario with smaller population size ($N_e = 10^4$) and lower gene flow (2.5 effective migrants per generation), and the same general conclusions hold (Figure S16, S17).

4.4. The optimal experimental design depends on study goals

Perhaps unsurprisingly, our simulation results suggest that there is not a single lcWGS experimental design that is ideal for all purposes. Instead, the optimal design depends on the goals, system, and budget of a study. For many common types of population genomic inference (e.g. allele frequency estimation, population structure analysis, genetic differentiation between populations), higher accuracy can be achieved by spreading a given sequencing effort thinly across more samples (Figures 3, 5, 6). There are, however, some notable exceptions. For example, inference that depends heavily on low-frequency alleles (e.g. Watterson's θ, Tajima's D) can be very sensitive to the chosen GL model when per-sample sequencing coverage is low, so until we have a better understanding of which GL models best fit the empirical data, sequencing each sample with relatively higher coverage (e.g. >4x) might generate more robust results for these types of analyses (Figure S2, S3). Nevertheless, if relative measures of these statistics are of interest rather than their absolute values (e.g. for outlier detection), lower coverage of each sample may be adequate (Figure S14, S15). Similarly, the methods that are currently available for LD estimation with lcWGS data can generate biased absolute estimates when the coverage is lower than 4x (Figure S4, S5), but note that reliable relative estimates of LD can be obtained at lower coverage.

It is important to keep in mind that tradeoffs exist between sample size and per-sample depth: with a given budget, the higher per-sample sequencing depth needed for robust estimation of the SFS (e.g. for demographic inference using δaδi) or absolute values of e.g. Tajima's D or LD will likely compromise the accuracy for other estimates, e.g. of allele frequencies or F_{ST} outliers (unless sample sizes are large even with the higher per-sample coverage). Accordingly, researchers must carefully consider what types of inference are most essential to their study goals and strike an appropriate balance. Based on our results here and those from previous studies, we provide some general guidelines to lcWGS experimental design in Table 4. For more targeted guidance, we also encourage researchers to build on our simulation pipeline (https://github.com/therkildsen-lab/lcwgs-simulation) to optimize the experimental design for their specific studies.

Box 3. Performance of lcWGS vs. RAD-seq in selection scans

Compared to lcWGS, RAD-seq has the advantage of generating high-confidence genotype calls, but suffers from a sparser coverage of the genome, which can result in missed signals in selection scans (Lowry et al., 2017). Here, we simulated RAD-seq data for our two divergent selection scenarios with a range of realistic sample sizes and RAD tag densities. In the scenario with larger population size and higher gene flow, we found that even with a large sample size

and a much higher marker density than typically used (128 RAD tag SNPs per Mb, i.e. ~128,000 SNPs in a 1Gb genome), RAD-seq picked up some, but tended to miss several of the narrow F_{ST} peaks. With a lower, much more commonly used marker density (e.g. 8 tags per Mb, or ~8,000 SNPs in a 1Gb genome), the majority of the selection-induced peaks would be missed, regardless of sample size (Figure 7). In the scenario where the population size is smaller and gene flow is lower, RAD-seq is more likely to sample SNPs within the true F_{ST} peaks due to the stronger linked selection, but because of the higher background noise in these scenarios, it still struggles to detect distinct F_{ST} peaks (Figure S18). These findings are consistent with a growing number of empirical examples where RAD-seq missed signatures of selection clearly detected with WGS data (see the Introduction).

5. Application to empirical data

To supplement our simulation-based evaluation of lcWGS inference with an exploration of how sequencing depth affects the identification of polymorphic sites, population structure analysis and detection of outlier loci in empirical data, we subsampled and re-analysed previously published whole genome sequencing data from the Neotropical butterfly *Heliconius erato* (Van Belleghem et al., 2017). The *H. erato* radiation comprises several subspecies that show a vast visual diversity in Müllerian mimicry related to wing patterns, and many of the underlying candidate genes have been identified (Reed et al., 2011; Van Belleghem et al., 2017). For example, the *optix* gene has been shown to control the red band phenotype in multiple *Heliconius* species and accordingly shows strong differentiation among subspecies with different band patterns (Reed et al., 2011; Van Belleghem et al., 2017). We subsampled resequencing data (originally average coverage of 11x ± 2.3x per individual) mapped to the *H. erato demophoon* (v1) to coverage depths of 8x, 4x, 2x, 1x, 0.5x and 0.25x (see Supplementary Methods (Part 1)) and analysed them in a GL framework. For simplicity, we focus on results for 8x, 2x and 0.5x coverage, as results from 4x and 1x are very similar to 8x and 2x, respectively (see supplementary Figure S19).

First, we found a positive correlation between the number of variable sites identified during SNP identification in ANGSD and the mean genome-wide sequencing coverage (Figure 8a; quadratic function: $r^2 = 0.98$, $p=0.00099$). Across all 51 individuals used in the final analyses, the number of SNPs identified with a p-value threshold of 10-6 ranged from 12,266 at 0.5x coverage to 14,851,731 at a mean coverage depth of 8x. It has to be noted though, that the number of detected SNPs depends on the p-value threshold, and for a dataset with a mean per-individual coverage of 0.25x a lower p-value threshold would have to be used to identify any SNPs at all (Figure 8).

Second, we reconstructed the population structure using PCA, performed on covariance matrices estimated using random read sampling in ANGSD (see Supplementary Methods). The PCA showed a very similar clustering pattern for all datasets regardless of coverage level, with populations grouping into three distinct clusters corresponding to the geographic origin of samples (Central America, East of Andes, West of Andes; Figure 8b). One subspecies (*H. erato*

hydara) sampled from two geographic regions was split over two clusters. On a finer population structure scale, we observed a slightly wider spread of data points at the lowest coverage (0.5x), although the general clustering was comparable to higher coverages.

Lastly, comparing the genetic differentiation between *H. erato* subspecies with (n=28) and without (n=23) the red bar phenotype (Van Belleghem et al., 2017), we recovered the well-characterized F_{ST} peak around the *optix* gene at per-individual coverages as low as 1x (Figure 8c; Van Belleghem et al., 2017). At 0.5x coverage, we were restricted to estimating F_{ST} within fewer genomic windows compared to higher coverages (112 50kb windows at 0.5x vs 255 50kb windows at >1x along scaffold 1801), leading to much sparser window coverage across the scaffold and therefore a noisier signal (Figure 8c). However, even at this low resolution, we detected one differentiated genomic window in the optix region, albeit the estimated F_{ST} was elevated at 0.5x (F_{ST} ~0.6) compared to higher coverages (F_{ST} ~0.4).

Overall, these results suggest that even at a comparatively low individual sequencing coverage of 0.5-1x and moderate sample sizes of 20-30 per population, we can detect population structure and recover distinct peaks of differentiation across the genome in empirical data.

Box 4. Using imputation to bolster genotype estimation from lcWGS

The majority of current population genomic inference methods, including all the lcWGS methods discussed in this paper so far, consider data on a SNP-by-SNP basis and accordingly ignore all the information contained in the surrounding haplotype structure. Imputation can be used to boost genotyping accuracy by leveraging LD patterns between variants to identify shared stretches of chromosome and incorporate information from flanking alleles to infer missing or low-confidence genotypes (Li et al., 2011; Pasaniuc et al., 2012). Imputation has been used extensively to obtain genotype calls from low-coverage data in humans and agricultural species, but has seen limited application in non-model species because most imputation methods, such as Beagle (Browning & Yu, 2009) and findhap (VanRaden, Sun, & O'Connell, 2015), rely on externally generated haplotype reference panels, which are unavailable for most species. In contrast, the more recently developed program STITCH imputes directly from sequence read data without reference panels, and has been shown to perform well when sample sizes are large (n>2000; Davies, Flint, Myers, & Mott, 2016). However, sample sizes of this magnitude are not achievable in many studies, especially for rare or elusive species. To evaluate the utility of imputation without reference panels with sample sizes more typical of studies of non-model species, we simulated three populations with varying levels of genetic diversity and LD, tested combinations of sequencing depths and sample sizes, and identified the conditions under which reference panel-free imputation is likely to bolster genomic analyses of lcWGS data.

Imputed genotype accuracy

We simulated three populations characterized by 1) low diversity, high LD (N_e = 1,000, r = 0.5 cM/Mb); 2) medium diversity, medium LD (N_e = 10,000, r = 0.5 cM/Mb); and 3) medium

diversity, low LD (N_e = 10,000, r = 2.5). For each population, we subsampled 25, 100, 250, 500 or 1000 individuals and simulated sequencing reads to average depths of 1x, 2x and 4x per sampled individual. We compared genotype dosages for all SNPs with minor allele frequency >0.05 imputed without reference panels in Beagle and STITCH, to those estimated without imputation in ANGSD (see the Supplementary Text Part 1 and Table S2 for details on simulations, genotype dosage estimation and imputation).

Our analysis suggests that using imputation without reference panels does improve population genomic inference under certain circumstances. Imputation was most effective under the low diversity, high LD scenario (Figure 9A). Under this scenario, genotype dosages imputed in STITCH from large sample sizes (n≥500) sequenced at 1x coverage were highly correlated with true genotypes (r^2>0.94), and all experimental designs with sample sizes ≥100 showed a substantial improvement in genotype estimation (Figure 9A). In the medium diversity and medium LD scenario, larger sample sizes were necessary to achieve similar imputation accuracy (e.g., n=1000 was needed for r^2=0.95; Figure 9B). Performance was markedly worse in the scenario with medium diversity and low LD, but there was nonetheless an improvement when imputing from large sample sizes (n≥250) or greater sequencing depths (≥2x) compared to genotypes called without imputation (Figure 9C).

Considerations for using imputation in non-model systems

Choosing whether to apply imputation to real-world data will depend on the details of the study system and the experimental design. In general, imputation accuracy increases with SNP density and LD between SNPs (de Bakker, Neale, & Daly, 2010; Shi et al., 2018), and our results suggest that populations with lower LD (even those with greater SNP density) require greater sample sizes and/or coverage to achieve the same imputation accuracy. For populations with higher LD, STITCH can substantially boost genotype accuracy for samples sequenced at 1x coverage, provided sample sizes are adequate (n≥100). When coverage is higher (≥2x), Beagle tends to perform similarly to or even outperform STITCH. However, for populations with lower LD, the improvement in genotype accuracy by imputation may be small unless sample sizes are ≥1000 and/or coverage is ≥2x for the conditions tested here; at smaller sample sizes or lower coverage, the potential benefit of imputation for low LD populations may not warrant the computational time.

Imputation provides another potential benefit for spreading sequencing effort thinly among many individuals in some circumstances. As our results have shown, by leveraging LD information from all samples, imputation can to some extent make up for the genotype uncertainty inherent in lcWGS data. For example, in the high LD population, genotypes imputed in STITCH from 1000 samples sequenced at 1x coverage were only slightly lower in accuracy (r^2=0.975) than for 500 samples at 2x coverage (r^2=0.981) and 250 samples at 4x coverage (r^2=0.982). For many questions where a large sample size is necessary to achieve adequate power, such as GWAS, what can be gained from increased sample size could readily outweigh the minimal loss in genotype accuracy. In addition, for some GWAS methods, the remaining genotype uncertainty can be incorporated directly into the analysis (Skotte et al., 2012; Jørsboe & Albrechtsen, 2019).

Because the performance of imputation varies with the LD and diversity of populations, a priori information on population history may help researchers anticipate how well imputation will perform. A set of “true genotypes” (e.g. from high-depth samples) and quality metrics output by the imputation programs (Browning & Yu, 2009; Davies et al., 2016) can also be used. Populations with small effective population size or that have experienced recent bottlenecks, such as threatened or endangered species, will have higher genome-wide LD (Hayes, Visscher, McPartlan, & Goddard, 2003; Waples & Do, 2010), making them potentially good systems for applying imputation if relatively large sample sizes (e.g. ≥100 for the scenarios simulated here) can be obtained. Where pedigree information is available, methods that incorporate the pedigree into imputation can be used (e.g. Ros-Freixedes, Whalen, Gorjanc, Mileham, & Hickey, 2020; Whalen, Ros-Freixedes, Wilson, Gorjanc, & Hickey, 2018). Finally, although imputation has been mainly applied to regular short-read data, the haplotype reconstruction step could be greatly simplified by long-read or linked-read data that is becoming increasingly available (see Section 7).

6. Current limitations and future developments

Despite the many strengths of lcWGS, there are also clear limitations to this data type. Here, we outline key constraints that researchers should consider before adopting the approach and discuss prospects for overcoming these constraints in the future.

Not suitable for analyses requiring genotype calls: It is important to stress that the potential for improved inference accuracy by spreading sequencing effort thinly over many individuals is only realized if the resulting uncertainty about individual genotypes is accounted for statistically in downstream analyses, with approaches such as those reviewed in Section 3. As discussed, hard-calling genotypes from lcWGS data remains likely to bias inference regardless of how large the sample size is, so lcWGS data is not well-suited for analysis types or downstream software that require genotypes as input, unless imputation can provide more accurate genotype calls (see Box 4 for details). However, as outlined in Section 3, GL-based inference frameworks are available for most major types of population genomic analysis and many additional approaches are under development. Alternatively, many researchers are now embracing a hybrid approach, where they sequence a few samples at higher coverage in order to conduct some analyses that require confident genotype calls, and perform the rest of their analyses using lcWGS data from more samples (e.g. Foote et al., 2016; Liu et al., 2014; Pečnerová et al., 2021; Westbury et al., 2018). Furthermore, another promising strategy with a hybrid dataset is to form a reference panel using a subset of high-coverage samples, and impute the genotypes of low-coverage samples (e.g. Fuller et al., 2020).

Lack of user-friendly software interfaces and documentation: Unfortunately, a key barrier to the wider adoption of lcWGS has been a lack of user-friendly interfaces and sparse documentation for programs that handle GL data. Accordingly, these tools are only accessible to users with prior expertise in bioinformatics, and the development of workflows often requires a substantial time investment. We hope that this beginner’s guide can be part of the effort to

increase the accessibility of lcWGS. We are also aware that efforts are underway to develop a more user-friendly version of ANGSD, which should make this powerful and versatile software package accessible to a broader set of researchers (Altinkaya and Fumagalli, personal communication).

Computational demands: Another practical limitation is the often much greater computational cost of probabilistic GL-based methods compared to methods based on called genotype. For example, SFS estimation from GLs in ANGSD is computationally intensive with very large sample sizes, which may be prohibitive for researchers without access to high-performance computational resources. New, more efficient algorithms (e.g. Han et al., 2015) and strategies for analyzing smaller sections of the genome in turn (see Section 3) may alleviate some of these constraints, but the computational demands for analysis should definitely be considered, especially for researchers transitioning to lcWGS after working with much smaller datasets such as RAD-seq.

Limitations and gaps in the current toolbox: Although tremendous progress has been made in the development of methods and tools for the analysis of lcWGS data over the past decade, some key analytical challenges remain. One important issue is the potential sensitivity to the choice of GL model in some types of analyses (see Section 4.1 and Box 4 in Fuentes-Pardo & Ruzzante, 2017). A better understanding of which GL models best match the real error structures generated by different sequencing platforms and more well-established methods for base quality score recalibration is essential for more robust inference from low-coverage data. In addition, alignment error is not taken into account in any of the current GL models, which could be problematic for genomes with high repeat content or for poor-quality reference genomes. The current analysis framework implemented in most software packages is also centered on the analysis of diploid organisms; extension to an arbitrary ploidy level would expand its usefulness for working with haploid and polyploid organisms, and key parts of this framework have already been developed (Blischak, Kubatko, & Wolfe, 2018). There also remain types of analysis for which GL-based methods are not yet available. However, new analytical approaches for lcWGS data continue to emerge. For example, GL-based equivalents to some established approaches, such as implementation of the Pairwise Sequentially Markovian Coalescent (PSMC) model, are currently under development (ngsPSMC [https://github.com/ANGSD/ngsPSMC]).

Analysis susceptible to batch artifacts: LcWGS data have great potential for reusability because the possibility to combine different datasets does not depend on the selection of the same restriction enzyme or markers. However, because of the low level of redundancy in the data, lcWGS could be particularly susceptible to batch effects when different datasets are combined. As mentioned earlier, some GL-based approaches are heavily dependent on accurate modeling of the error structure in the data, which can vary between sequencing batches. For example, the sequencing error rate could be overestimated in one batch and underestimated in another (Lou & Therkildsen, 2021), leading to artificial differences between batches that could confound real biological signals. Many of these batch effects can be

mitigated with simple bioinformatic approaches, although extra care needs to be taken (Lou & Therkildsen 2021).

Limited ability to phase lcWGS data: A major limitation is that no bioinformatic solution is yet available to allow accurate phasing of lcWGS data without a reference panel, therefore prohibiting haplotype-based analyses. Haplotype data are a rich source of information, e.g. for inference of local ancestry tracks across the genome, demographic histories, or ongoing selective sweeps (see Leitwein, Duranton, Rougemont, Gagnaire, & Bernatchez, 2020) for a detailed overview). Despite major technological advances, long-read sequencing that can recover haplotype information remains too costly for typical population genomic studies. However, the recent development of an affordable linked-read low-coverage sequencing approach (Meier et al., 2021) promises to open many new opportunities for haplotype-based inference on a population scale by enabling efficient phasing and imputation of low-coverage linked-read data without a reference panel. Phased haplotype data will provide substantial improvement in imputation performance compared to the short-insert lcWGS data explored in Box 4, and make completely new types of analysis possible with lcWGS data.

Limitations for small sample sizes and very large genomes: LcWGS will not be an optimal solution for all study systems. In particular, for species that are rare or difficult to collect (e.g. endangered or elusive species), it may be impossible to obtain adequate sample sizes for accurately estimating population genomic parameters with lcWGS (see Section 4). In these cases, many types of analysis, such as demographic history, diversity, selective sweeps and inbreeding levels, can be performed based on deep sequencing of the genome of a few or even just a single individual (e.g. Li & Durbin, 2011). For species with extremely large genomes (e.g. many amphibians and pine species), whole genome sequencing may also remain impractical at any sequencing depth from a cost or data storage/handling perspective, and reduced representation approaches such as RAD-seq or targeted sequence capture may be preferable (Burgon et al., 2020; McCartney-Melstad, Mount, & Shaffer, 2016). Of note, however, for targeted methods like sequence capture, low-coverage sequencing of larger sample sizes and associated GL-based analysis can, similar to lcWGS, confer distinct advantages over sequencing fewer individuals at higher depth (e.g. Snyder-Mackler et al., 2016; Therkildsen et al., 2019; Warmuth & Ellegren, 2019; Wilder et al., 2020).

7. Conclusion

In conclusion, although some limitations still exist for the use of lcWGS, this approach offers many advantages over reduced-representation sequencing or pooled WGS approaches and is ripe for broader implementation. We are excited about how its cost-effectiveness democratizes population-scale whole genome analysis, which until recently was only available to well-funded research groups working on model species. The ability to obtain full genome data for hundreds of individuals even on modest research budgets, and the rapidly expanding toolbox for versatile analysis of lcWGS data now makes it an increasingly promising approach for molecular ecology, conservation and evolutionary biology research. We hope this guide will inspire broader adoption to expedite the exploration of genomic variation across the tree of life.

Acknowledgements

We would like to thank Philipp Messer and Robbie Davies for advice on analysis, the science Twitter community for help compiling the list of studies using lcWGS (Table S1), and Matt Hare, Matteo Fumagalli, Andy Foote, Mats Pettersson, Daniel Wegman, Jonas Meisner, Anders Albrechtsen, the Therkildsen Lab at Cornell University, and the Editor for very helpful feedback that has substantially improved this manuscript. A very special thanks to Claire Mérot for generously sharing her perspective and providing extensive suggestions that really helped make this guide more relevant, focused, and user-friendly. This study was funded through a National Science Foundation grant to NOT (OCE-1756316).

Data availability

All scripts used to generate the analysis presented in this manuscript will be available in a GitHub repository release deposited in Zenodo (DOI: 10.5281/zenodo.5037406). The NCBI SRA accession numbers for the Heliconius data re-analyzed in this project is available in Table S3.

Author contributions

NOT conceived of the project. All the authors designed the research jointly and collaborated to compile the overview of available methods. RNL simulated the test data and performed the comparative analysis for different experimental designs, AJ performed the analysis of the empirical data and designed the graphics, and APW performed the imputation analysis. All the authors provided input on all analyses and wrote the manuscript together.

References

Aguillon, S. M., Campagna, L., Harrison, R. G., & Lovette, I. J. (2018). A flicker of hope: Genomic data distinguish Northern Flicker taxa despite low levels of divergence. *The Auk*, *135*(3), 748–766.

Aguillon, S. M., Walsh, J., & Lovette, I. J. (2020). Extensive hybridization reveals multiple coloration genes underlying a complex plumage phenotype. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/10.1101/2020.07.10.197715v1.abstract

Alex Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, *22*(11), 3028–3035.

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, *23*(3), 502–512.

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

Auwera, G. A. V. der, & O'Connor, B. D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. ***(1st Edition).*** O'Reilly Media.

Beninde, J., Möst, M., & Meyer, A. (2020). Optimized and affordable high-throughput sequencing workflow for preserved and nonpreserved small zooplankton specimens. *Molecular Ecology Resources*. 20(6), 1632-1646

Berner, D. (2019). Allele Frequency Difference AFD–An Intuitive Alternative to FST for Quantifying Genetic Population Differentiation. Genes, 10(4), 308. doi: 10.3390/genes10040308

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.

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution*, *10*(14), 7585–7601.

Bradburd, G. S., Coop, G. M., & Ralph, P. L. (2018). Inferring Continuous and Discrete Population Genetic Structure Across Space. *Genetics*, *210*(1), 33–52.

Browning, B. L., & Yu, Z. (2009). Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *American Journal of Human Genetics*, *85*(6), 847–861.

Burgon, J. D., Vieites, D. R., Jacobs, A., Weidt, S. K., Gunter, H. M., Steinfartz, S., … Elmer, K. R. (2020). Functional colour genes and signals of selection in colour-polymorphic salamanders. *Molecular Ecology*, *29*(7), 1284–1299.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583.

Campagna, L., Gronau, I., Silveira, L. F., Siepel, A., & Lovette, I. J. (2015). Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular Ecology*, *24*(16), 4238–4251.

Campagna, L., Repenning, M., Silveira, L. F., Fontana, C. S., Tubaro, P. L., & Lovette, I. J. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. *Science Advances*, *3*(5), e1602404.

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., … Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in*

Ecology and Evolution / British Ecological Society, *9*(2), 410–419.
Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution*, 36(4), 852–860. doi: 10.1093/molbev/msz008
Cheng, J. Y., Racimo, F., & Nielsen, R. (2019). Ohana: detecting selection in multiple populations by modelling ancestral admixture components. bioRxiv doi: 10.1101/546408
Chung, J. C. S., & Chen, S. L. (2017). Lacer: Accurate base quality score recalibration for improving variant calling from next-generation sequencing data in any organism. BioRxiv, 130732. doi: 10.1101/130732
Clucas, G. V., Kerr, L. A., Cadrin, S. X., Zemeckis, D. R., Sherwood, G. D., Goethel, D., … Kovach, A. I. (2019). Adaptive genetic variation underlies biocomplexity of Atlantic Cod in the Gulf of Maine and on Georges Bank. *PLOS ONE*, *14*(5), e0216992.
Clucas, G. V., Lou, R. N., Therkildsen, N. O., & Kovach, A. I. (2019). Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing. *Evolutionary Applications*, *12*(10), 1971–1987.
Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, *12*(7), 499–510.
Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, *48*(8), 965–969.
de Bakker, P. I. W., Neale, B. M., & Daly, M. J. (2010). Meta-analysis of genome-wide association studies. *Cold Spring Harbor Protocols*, *2010*(6), db.top81.
DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics, 32*(12), 1895–1897.
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.
Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., … Aiden, E. L. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000 *bioRxiv* doi: 10.1101/254797
Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042.
Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics, 27*(9), 1332–1334.
Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376.
Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*(2), 977–993.
Foll, M., Shim, H., & Jensen, J. D. (2015). WFABC: a Wright Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, *15*(1), 87–98.
Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, *7*, 11693.
Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Molecular Ecology*, *27*(9), 2215–2233.
Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage

disequilibrium using genotype likelihoods. *Bioinformatics, 35*(19), 3855–3856.
Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, *26*(20), 5369–5406.
Fuller, Z. L., Mocellin, V. J. L., Morris, L. A., Cantin, N., Shepherd, J., Sarre, L., … Przeworski, M. (2020). Population genetics of the coral Acropora millepora: Toward genomic prediction of bleaching. *Science*, 369(6501). doi: 10.1126/science.aba4674
Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS One*, *8*(11), e79667.
Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, *195*(3), 979–992.
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.
Futschik, A., & Schlötterer, C. (2010). The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, *186*(1), 207–218.
Gaio, D., To, J., Liu, M., Monahan, L., Anantanawat, K., & Darling, A. E. (2019). Hackflex: low cost Illumina sequencing library construction for high sample counts *bioRxiv* doi: https://doi.org/10.1101/779215
Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics*, *98*(3), 456–472.
Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *bioRxiv*. Retrieved from http://arxiv.org/abs/1207.3907
Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. bioRxiv doi: 10.1101/2021.05.21.445151
Gatter, T., von Löhneysen, S., Drozdova, P., Hartmann, T., & Stadler, P. F. (2020). Economic Genome Assembly from Low Coverage Illumina and Nanopore Data. *bioRxiv* doi: 10.1101/2020.02.07.939454
Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, *201*(4), 1555–1579.
Gignoux-Wolfsohn, S. A., Pinsky, M. L., Kerwin, K., Herzog, C., Hall, M., Bennett, A. B., … Maslo, B. (2021). Genomic signatures of selection in bats surviving white-nose syndrome. *Molecular Ecology*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15813
Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10), e1000695.
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.
Han, E., Sinsheimer, J. S., & Novembre, J. (2015). Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics, 31*(5), 720–727.
Hayes, B. J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, *13*(4), 635–643.
Hennig, B. P., Velten, L., Racke, I., Tu, C. S., Thoms, M., Rybin, V., … Steinmetz, L. M. (2018). Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3, 8*(1), 79–89.
Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics, 28*(4), 593–594.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61.

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.

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.

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P.-A., & Bernatchez, L. (2020). Using Haplotype Information for Conservation Genomics. *Trends in Ecology & Evolution*, *35*(3), 245–258.

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., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496.

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., Wu, K., Ruan, C., Pan, J., Wang, Y., & Long, H. (2019). Cost-reduction strategies in massive genomics experiments. *Marine Life Science & Technology*, *1*(1), 15–21.

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. doi: 10.1101/gr.088013.108

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research*, 21(6), 940–951.

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Molecular Ecology Resources, 19(3), 639–647. doi: 10.1111/1755-0998.12995

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017). ATLAS: Analysis Tools for Low-depth and Ancient Samples. *bioRxiv*. https://doi.org/10.1101/105346.

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., … Wang, J. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*(4), 785–794.

Lou, R. N., Fletcher, N. K., Wilder, A. P., Conover, D. O., Therkildsen, N. O., & Searle, J. B. (2018). Full mitochondrial genome sequences reveal new insights about post-glacial expansion and regional phylogeographic structure in the Atlantic silverside (Menidia menidia). *Marine Biology*, *165*(8), 124.

Lou, R. N. and Therkildsen, N. O. Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, identification, and mitigation. *bioRxiv*.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A.

(2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152.

Margaryan, A., Lawson, D. J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., … Willerslev, E. (2020). Population genomics of the Viking world. *Nature*, *585*(7825), 390–396.

McCartney-Melstad, E., Mount, G. G., & Shaffer, H. B. (2016). Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, *16*(5), 1084–1094.

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.

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al . (2016). *Molecular Ecology Resources*, *17*(3), 356–361.

Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018). Estimating the Timing of Multiple Admixture Pulses During Local Ancestry Inference. *Genetics*, *210*(3), 1089–1107.

Meek, M. H., & Larson, W. A. (2019). The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molecular Ecology Resources*, *19*(4), 795–803.

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., Aldás, I., … Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25), 461–441.

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

Meisner, J., Albrechtsen, A., & Hanghøj, K. (2021). Detecting Selection in Low-Coverage High-Throughput Sequencing Data using Principal Component Analysis *bioRxiv* doi: 10.1101/2021.03.01.432540

Mérot, C., Berdan, E., Cayuela, H., Djambazian, H., Ferchaud, A.-L., Laporte, M., … Bernatchez, L. (2021). Locally adaptive inversions modulate genetic variation at different geographic scales in a seaweed fly. *Molecular Biology and Evolution*, msab143. doi: 10.1093/molbev/msab143

Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, *23*(7), 1764–1779.

Ni, S., & Stoneking, M. (2016). Improvement in detection of minor alleles in next generation sequencing by base quality recalibration. *BMC Genomics*, 17(1), 139. doi: 10.1186/s12864-016-2463-2

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PloS One*, *7*(7), e37558.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451.

Orr, A. J. (2020). Methods for Detecting Mutations in Non-Model Organisms (Ph.D. Thesis, Arizona State University). Arizona State University, United States -- Arizona. Retrieved from https://www.proquest.com/docview/2476130546/abstract/12747DA614FF4224PQ/1

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.

Pečnerová, P., Garcia-Erill, G., Liu, X., Nursyifa, C., Waples, R. K., Santander, C. G., … Hanghøj, K. (2021). High genetic diversity and low differentiation reflect the ecological

versatility of the African leopard. *Current Biology*, 31(9), 1862-1871.e5. doi: 10.1016/j.cub.2021.01.064

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100.

Picelli, S., Björklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, *24*(12), 2033–2040.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967.

Powell, D. L., García-Olazábal, M., Keegan, M., Reilly, P., Du, K., Díaz-Loyo, A. P., … Schumer, M. (2020). Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science*, *368*(6492), 731–736.

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

Puritz, J. B., & Lotterhos, K. E. (2018). Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Molecular Ecology Resources*, *18*(6), 1209–1222.

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

Reed, R. D., Papa, R., Martin, A., Hines, H. M., Counterman, B. A., Pardo-Diaz, C., … McMillan, W. O. (2011). optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, *333*(6046), 1137–1141.

Rice, E. S., & Green, R. E. (2019). New Approaches for Genome Assembly and Scaffolding. *Annual Review of Animal Biosciences*, *7*, 17–40.

Roesti, M., Salzburger, W., & Berner, D. (2012). Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, 12(1), 94. doi: 10.1186/1471-2148-12-94

Ros-Freixedes, R., Whalen, A., Gorjanc, G., Mileham, A. J., & Hickey, J. M. (2020). Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. *Genetics, Selection, Evolution: GSE*, *52*(1), 18.

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R., & Weigel, D. (2019). An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics*, *213*(3), 771–787.

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. *Nature Reviews. Genetics*, *15*(11), 749–763.

Schumer, M., Powell, D. L., & Corbett-Detig, R. (2020). Versatile simulations of admixture and accurate local ancestry inference with mixnmatch and ancestryinfer. *Molecular Ecology Resources*, *20*(4), 1141–1151.

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., … Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. *Human Heredity*, *83*(3), 107–116.

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.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477–485.

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.

Szarmach, S., Brelsford, A., Witt, C. C., & Toews, D. (2021). Comparing divergence landscapes

from reduced-representation and whole-genome re-sequencing in the yellow-rumped warbler (Setophaga coronata) species *bioRxiv*. Retrieved from https://www.biorxiv.org/content/10.1101/2021.03.23.436663v1.abstract

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.

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208.

Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H., & Palumbi, S. R. (2019). Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science*, *365*, 487–490.

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, *29*(12), 673–680.

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in Anopheles gambiae. *PLoS Biology*, *3*(9), e285.

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.

VanRaden, P. M., Sun, C., & O'Connell, J. R. (2015). Fast imputation using medium or low-coverage sequence data. *BMC Genetics*, *16*, 82.

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.

Vonesch, S. C., Li, S., Tu, C. S., Hennig, B. P., Dobrev, N., & Steinmetz, L. M. (2020). Fast and inexpensive whole genome sequencing library preparation from intact yeast cells bioRxiv doi: 10.1101/2020.09.03.280990

Waples, R. S., & Do, C. (2010). Linkage disequilibrium estimates of contemporary N e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, *3*(3), 244–262.

Warmuth, V. M., & Ellegren, H. (2019). Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data. *Molecular Ecology Resources*, *19*(3), 586–596.

Westbury, M. V., Hartmann, S., Barlow, A., Wiesel, I., Leo, V., Welch, R., … Hofreiter, M. (2018). Extended and Continuous Decline in Effective Population Size Results in Low Genomic Diversity in the World's Rarest Hyena Species, the Brown Hyena. *Molecular Biology and Evolution*, 35(5), 1225–1237. doi: 10.1093/molbev/msy037

Wetterstrand, K. A. (2021). DNA sequencing costs: data from the NHGRI genome sequencing program (GSP) Available at www.genome.gov/sequencingcostsdata. Accessed Jan 15 2021

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.

Whalen, A., Ros-Freixedes, R., Wilson, D. L., Gorjanc, G., & Hickey, J. M. (2018). Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. *Genetics, Selection, Evolution: GSE*, *50*(1), 67.

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST. *The American Naturalist*, *186*(S1), S24–S36.

Wilder, A. P., Palumbi, S. R., Conover, D. O., & Therkildsen, N. O. (2020). Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. *Evolution*

Letters, *n/a*(n/a). doi: 10.1002/evl3.189
Zhu, Y., Bergland, A. O., González, J., & Petrov, D. A. (2012). Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. *PloS One*, *7*(7), e41901.
Zook, Justin M., Daniel Samarov, Jennifer McDaniel, Shurjo K. Sen, and Marc Salit. 2012. "Synthetic Spike-in Standards Improve Run-Specific Systematic Error Analysis for DNA and RNA Sequencing." PloS One 7 (7): e41356.

Table 1. Total cost per sample for both library preparation and sequencing based on May 2021 price levels (rounded up to nearest dollar)

Genome size (Gb)	Cost per sample (USD)*		Example organisms
	1x coverage	2x coverage	
0.2	11(3)	13(5)	Fruit fly, Honeybee, Arabidopsis
0.65	16(8)	25(17)	Atlantic silverside, Stickleback, Eastern oyster
1	21(13)	34(26)	Zebra finch, Chicken, Purple sea urchin
3	47(39)	86(78)	Human, Atlantic salmon, African clawed frog

*Cost estimates do not include labor and assume that samples are sequenced efficiently on an Illumina NovaSeq instrument. The assumed costs break down to 8 USD per library (Therkildsen & Palumbi, 2017) and ~13 USD per Gb sequence data in a shared S4 lane (see supplementary methods for estimates of initial investment costs). The numbers in brackets show the cost of sequencing only (i.e. the approximate total cost with a cheap homebrew library preparation method (see section 2.2)).

Table 2. List of published software for the analysis of lcWGS data. References for each software can be found in the main text (Section 3) or in the Supplementary Material Part 4.

Analysis type		Software						
Analysis	**Method**	**ANGSD**	**ATLASs**	**MAPGD**	**vcflib**	**ngsTools***	**PCAngsd**	**Specialised software**
SNP identification		✓	✓					BaseVar, EBG, Freebayes, GATK, Reveel, etc.
Population structure	PCA	✓				✓	✓	
	Individual genetic distance	✓	✓			✓		skmer
	Local PCA†							lostruct
	Admixture						✓	Entropy, evalAdmix, ngsAdmix, Ohana
Selection scan	PCA-based; ancestry-corrected						✓	Ohana
Association analysis		✓						SNPTEST
Linkage disequilibrium				✓		✓		GUS-LD, PopLD
Individual relatedness	Relatedness			✓			✓	ngsRelate
	Parentage							AlphaAssign
	Pedigree analysis							WHODAD
Inbreeding	Inbreeding coefficient		✓		✓	✓	✓	ngsRelate
	IBD tracts					✓		
	Runs of homozygosity							bcftools roh
Ancestry relationships	D-statistics/ABBA-BABA	✓			✓			
Linkage map construction								Lep-MAP3
Allele frequency estimation		✓	✓	✓				
Site frequency spectrum		✓				✓		
Within population genetic diversity	θ estimators (e.g. Watterson's, π)	✓	✓			✓		
Within population neutrality stats	e.g. Tajima's D, Fay & Wu's H	✓						
Individual level genetic diversity	Individual heterozygosity	✓	✓	✓				heterozygosity-em
Population differentiation	F_{ST}	✓			✓	✓		
	dxy					✓		
Allele frequency differentiation‡		✓			✓			
Hardy-Weinberg equilibrium		✓	✓	✓			✓	
Structural variants								svgem

Quality score recalibration		✓	✓					
Ploidy inference								HMMploidy
Genotype imputation								Beagle, LB-Impute, LinkImput, loimpute, NOISYmputer, STITCH, etc.

* ngsTools is a collection of loosely-connected programs including ngsSim, ngsF, ngsPopGen, ngsUtils, ngsDist, ngs-HMM, and ngsLD

† LocalPCA can be conducted by using lostruct together with custom scripts that perform the PCA with low-coverage data (e.g. with PCAngsd).

‡ ANGSD can be used to test for statistical significance of allele frequency differentiation between two groups with the option -doAsso 1, and vcflib implements the estimation of pFst (Domyan et al. 2016)

Table 3. Key data filters to consider in the analysis of lcWGS data

Category	Filter	Recommendation
General filters	Base quality	Base quality scores are factored into the calculation of genotype likelihoods, so if they accurately reflect the probability of sequencing error, bases with low scores also carry useful information. However, base quality scores are sometimes miscalibrated, so noise may be reduced if bases with scores below a threshold, e.g. 20, are either trimmed off prior to analysis or ignored. Alternatively, all base quality scores can be recalibrated based on estimated error profiles in the data (see Section 3.1).
	Mapping quality	Mapping quality is **not** considered in genotype likelihood estimation in currently available tools, so it is often advisable to remove low-confidence and/or non-uniquely mapped reads prior to analysis (e.g. reads with mapping quality <20). Filtering out reads that do not map in proper pairs should also further increase confidence in reads being mapped to the correct location, but could cause biases in regions with structural variation.
	Minimum depth and/or number of individuals	To avoid sites with low or confounding data support in downstream analysis, minimum depth and/or minimum number of individual filters can be used to exclude sites with much reduced sequencing coverage compared to the rest of the dataset (e.g. regions with low unique mapping rates, such as repetitive sequences). Appropriate thresholds will vary between data sets, but could e.g. be to exclude sites with read data for <50% of individuals (globally or within each population), or with <0.8x average depth across individuals (after filtering on mapping quality)
	Maximum depth	Maximum depth filters are used to exclude sites with exceptionally high coverage (e.g. regions that are susceptible to dubious mapping, such as copy number variants). Common maximum depth thresholds could be one or two standard deviations above the median genome-wide depth.
	Duplicate reads	PCR and optical duplicates can give inflated impressions of how many unique molecules have been sequenced, which - particularly in the presence of preferential amplification of one allele - could bias genotype likelihood estimation. We therefore recommend removing duplicate reads prior to any analysis.

	Indels	Reads mapped to indels are frequently misaligned, especially if the ends of reads span an indel. To avoid false SNPs, we recommend either using dedicated tools to realign reads covering indels, using a haplotype-based variant caller (e.g. Freebayes or GATK) to estimate genotype likelihoods, or excluding bases flanking indels.
	Overlapping sections of paired-end reads	If the DNA insert in a library fragment is shorter than the combined length of paired reads, there will be a section of overlap between the forward and reverse reads. While some variant callers (e.g. GATK) account for the pseudo-replication in overlapping ends of read pairs, the current implementation of ANGSD treats each end of a read pair as independent (this may change in a future release (Korneliussen, personal communication)). When treated as independent, read support for overlapping sections will be "double counted", which may bias genotype likelihoods. A conservative approach is to soft-clip one of the overlapping read ends.
Filters on polymorphic sites*	p-value	The significance threshold (often in the form of maximum p-value) can be adjusted to fine-tune the sensitivity of polymorphism detection, with lower p-values leading to fewer, but higher-confidence, SNP calls. A commonly used cut-off is 10^{-6}.
	SNPs with more than two alleles	Most software programs for downstream analyses assume that all SNPs are biallelic, so SNPs with more than two alleles can be filtered out in the SNP identification step to avoid violation of such assumptions.
	Minimum minor allele frequency (MAF)	For many types of analysis, e.g. PCA, admixture analysis, detection of F_{ST} outliers and estimation of LD, low-frequency SNPs are uninformative and can even bias results (e.g. Linck & Battey, 2019; Roesti, Salzburger, & Berner, 2012). For those types of analysis, imposing a minimum MAF filter of 1-10% can substantially speed up computation time. Appropriate thresholds depend on coverage, sample size (how many copies does a MAF threshold correspond to) and the type of downstream analysis.
Restricting analysis to a predefined site list	List of global SNPs	For comparison of parameter estimates for multiple populations, it is important to ensure that data are obtained for a shared set of sites and that SNP polarization (which allele we track the frequency of) is consistent. For programs like ANGSD where population-specific estimates are obtained by analyzing the data from each population separately, a good strategy is to first conduct a global SNP calling with all samples and then restrict population-specific analysis to those SNPs with consistent major and

		minor allele designations and no MAF or SNP p-value filter (because that would incorrectly generate “missing data” if a site is fixed in a particular population).

* Note that no SNP significance or minimum MAF threshold should be used when estimating genetic diversity (e.g. theta and the SFS) as all sites contain relevant information. This also applies to the estimation of the absolute values of d_{xy}.

Table 4. Experimental design recommendations for different types of population genomic analyses using lcWGS data

Type of analyses	Examples	Recommendations on experimental design
Allele frequency and differentiation	Population allele frequencies, most genotype-environment association analysis (GEA) methods, F_{ST} (as implemented e.g. in vcflib), pFst	Prioritize larger sample sizes, ≥10 samples per population, ≥10x coverage per population, (Figure 3, 4). Avoid uneven sample size for estimation of F_{ST} (Berner, 2019)
SFS-based analyses (absolute estimation of rare-allele-dependent metrics)	Absolute estimation of Watterson's θ, Tajima's D, individual heterozygosity Reconstruction of demographic history (e.g. δaδi)	Prioritize higher coverage per sample, >4x coverage per sample, ≥5 samples per population, (Figure S2, S3).
SFS-based analyses (relative estimation of rare-allele-dependent metrics, or non-rare-allele-dependent metrics)	Relative estimation of Watterson's θ and Tajima's D (e.g. for outlier scans) π, d_{xy}, F_{ST} (as implemented in ANGSD)	Prioritize larger sample sizes, ≥10 samples per population, ≥10x coverage per population, (Figure 6, S2, S3, S7, S14-S17). Avoid uneven sample size for estimation of F_{ST}, (Figure S7, see also Berner, 2019).
Population structure	PCA, admixture analysis	Prioritize larger sample sizes, ≥10 samples per population, extremely low per-sample coverage (e.g. 0.125x, Figure 5, S6, S11) or highly uneven per-sample coverage (e.g. ranging from 0.5x to 6x, Skotte et al. 2013) can be viable.
Absolute estimation of linkage disequilibrium	LD decay rate, demographic inference	Prioritize higher coverage per sample, ≥4x coverage per sample, ≥20 samples per population, (Figure S4, S5; Bilton et al., 2018; Fox et al., 2019; Maruki & Lynch, 2014).
Relative estimation of linkage disequilibrium	LD pruning, LD block identification	Per-sample coverage as low as 1x could be viable, ≥20x coverage per population, (Figure S4, S5).
Genotype imputation without reference panels	STITCH, Beagle	STITCH: prioritize larger sample size (≥500) over per-sample coverage (1x could be sufficient), Beagle: prioritize higher per-sample coverage (≥2x) over sample sizes (≤250 could be sufficient), (Figure 9).

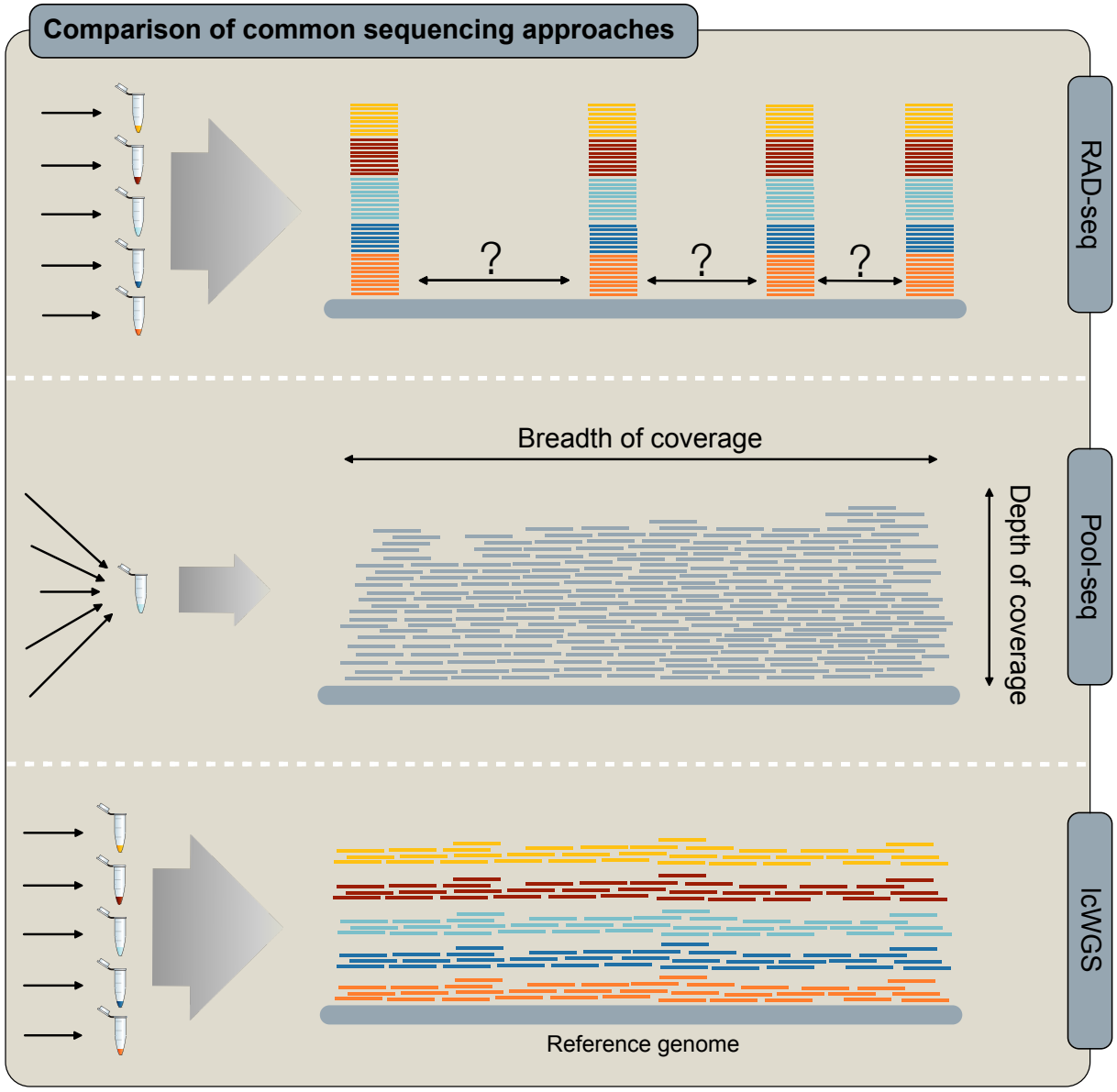

Figure 1. Diagram showing the distribution of sequencing reads mapped to a reference genome under **(A)** a RAD-seq, **(B)** a Pool-seq, and **(C)** a lcWGS design.

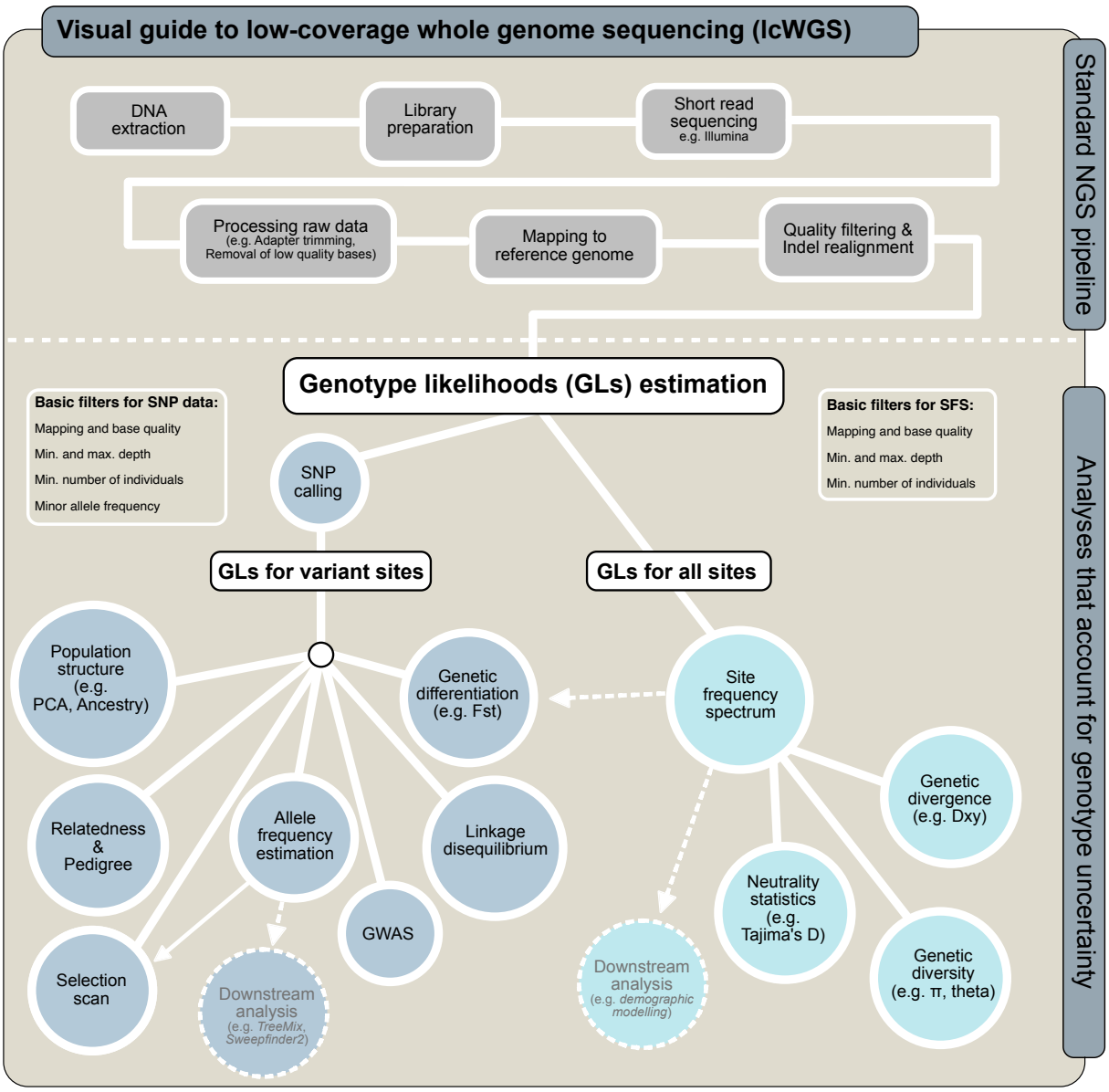

Figure 2. Diagram showing a typical computational pipeline for lcWGS data. **Top:** The data pre-processing part of the pipeline, which is similar to pipelines used for other types of NGS data. **Bottom:** The data analysis part of the pipeline, which is based on a probabilistic framework using genotype likelihoods to account for genotype uncertainty. The path through the SFS to diversity statistics and F_{ST} illustrated here reflects the workflow implemented in ANGSD. Other tools (e.g ATLAS) can infer these statistics directly from GLs without an SFS prior.

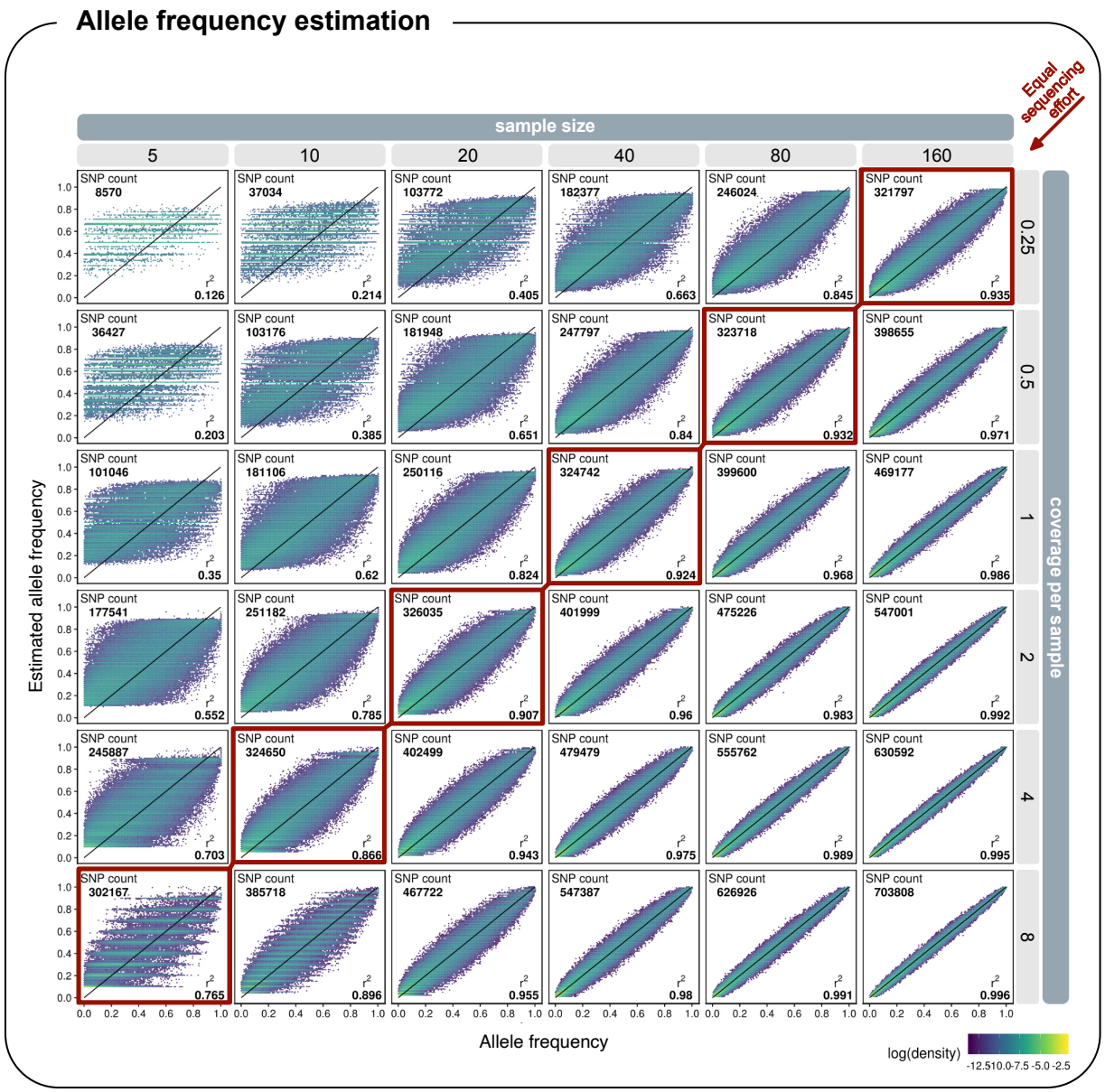

Figure 3. The estimated vs. true allele frequencies at all called SNPs (i.e. true positives + false positives) with lcWGS. Across the different facets, sample size increases from left to right, and coverage per sample increases from top to bottom. The total sequencing effort remains the same along the diagonals from bottom left to top right (one example highlighted with red boxes). The color in the plot area indicates the local density of points, with yellow corresponding to the highest density and dark blue corresponding to the lowest density. r^2 and the number of SNPs called (SNP count) are shown in each facet. The black 1:1 line in each facet indicates the positions where the estimated allele frequency is equal to the true allele frequency. False negative SNPs are not included in this figure; their distribution is shown in Figure S1.

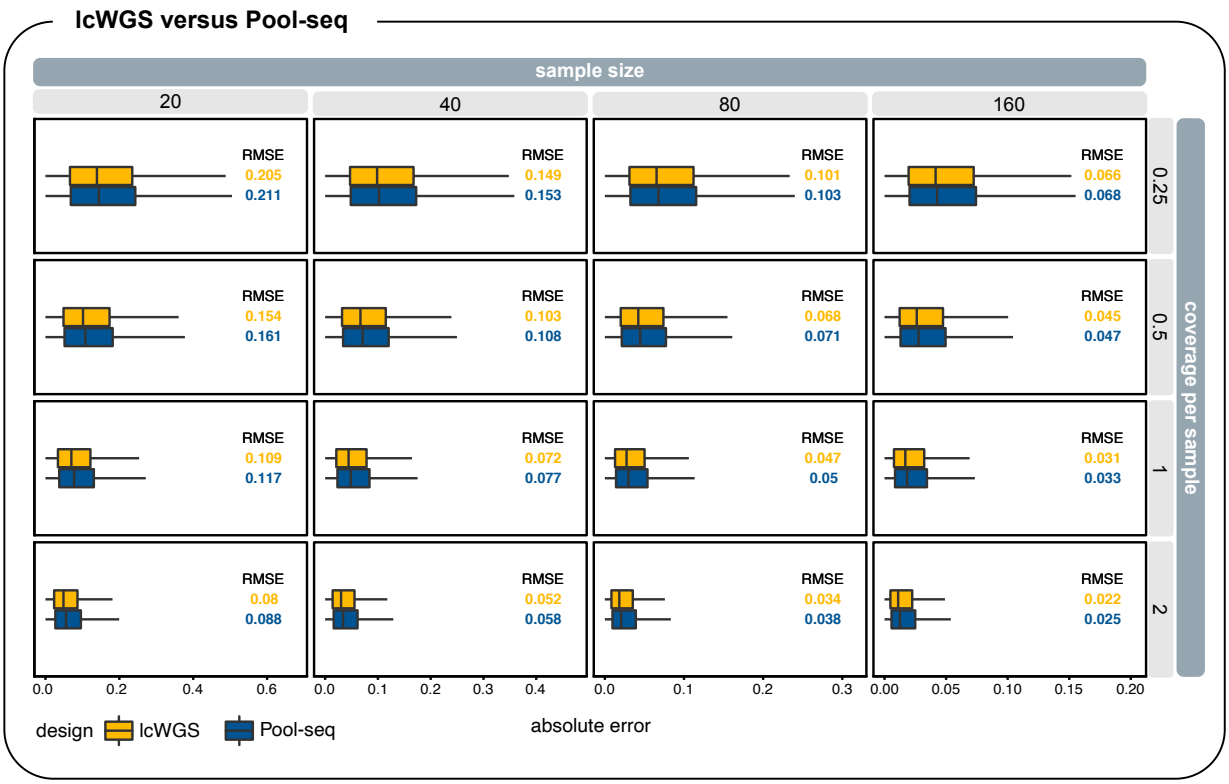

Figure 4. A comparison of the error in allele frequency estimation with lcWGS (yellow) and Pool-seq (blue) data. The distribution of absolute errors (|estimated frequency - true frequency|) is shown with the box plots along the x-axis. The left and right hinges of the box plots show the interquartile ranges of 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 per sample 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; note the differences in scale of the x-axes. False negative SNPs are not included in this figure; their distribution is shown in Figure S1.

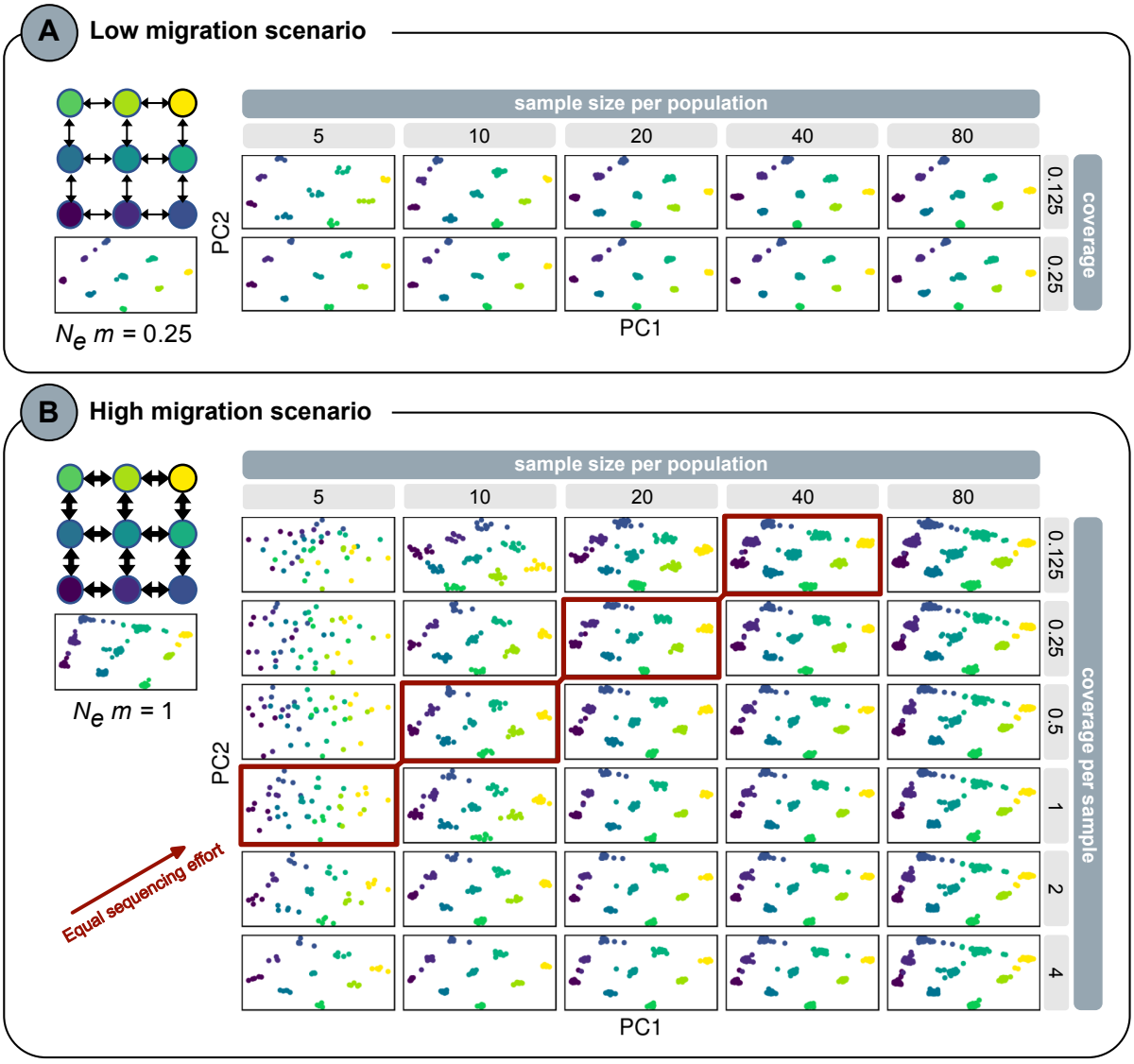

Figure 5. Patterns of spatial population structure inferred through principal component analysis (PCA) with lcWGS data. **(A)** A scenario with lower gene flow (an average of 0.25 effective migrants per generation). **(B)** A scenario with higher gene flow (an average of 1 effective migrant from one population to another every generation). Left: schematics of the scenario that was simulated (each node corresponds to a simulated population) and a PCA based on the true genotypes. Right: the first two principal components from the PCA with simulated lcWGS data; each point corresponds to an individual sample and its color corresponds to the population it is sampled from. The sample size per population increases across panels from left to right, and the coverage per sample increases from top to bottom.

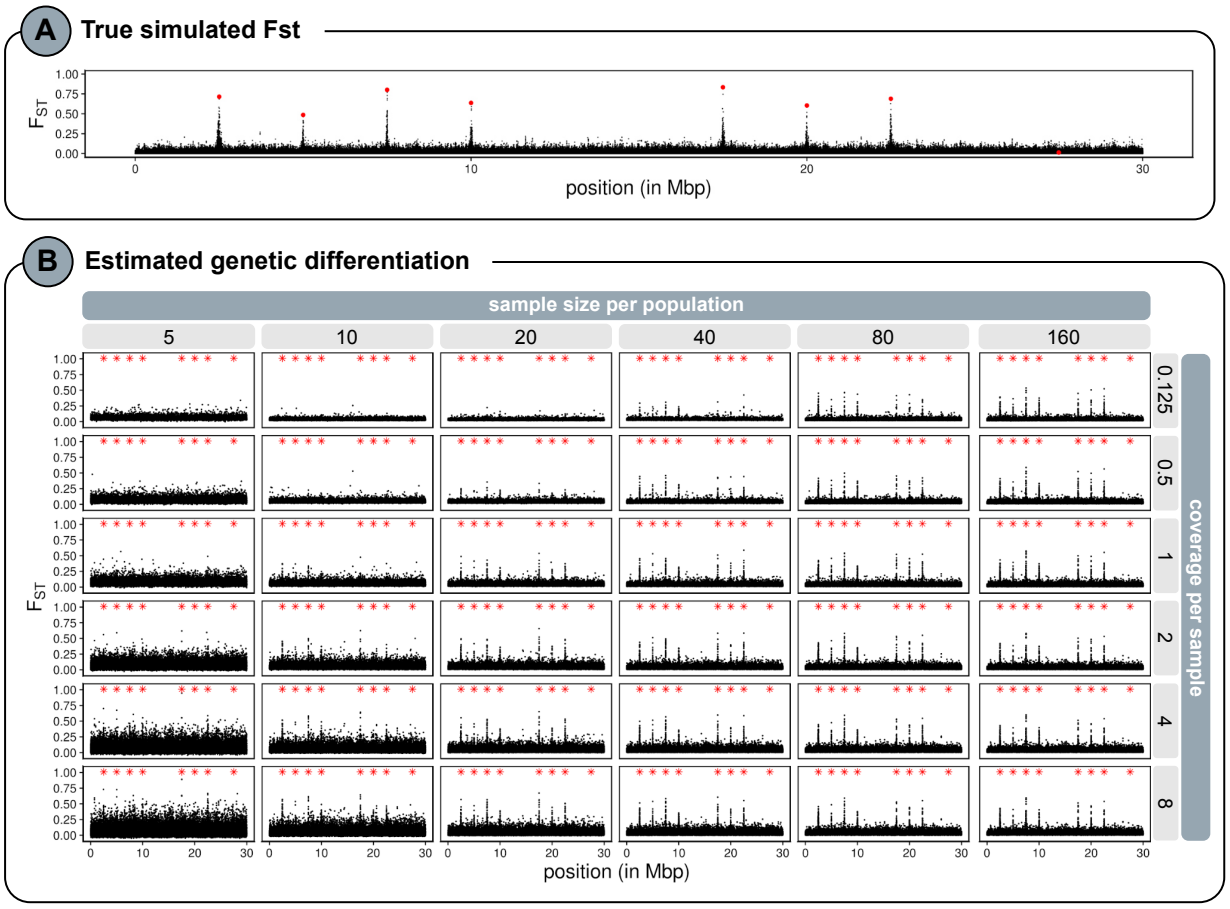

Figure 6. Genome-wide scans for divergent selection with lcWGS data. **(A)** The true per-SNP F_{ST} values along the chromosome between the two simulated populations. **(B)** The F_{ST} values inferred from lcWGS data in 1kb windows along the chromosome. The sample size per population increases from left to right, and the coverage per sample increases from top to bottom. In **(A)**, the red points mark the positions of SNPs under selection and the black points mark the neutral SNPs. In **(B)**, 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 F_{ST} values).

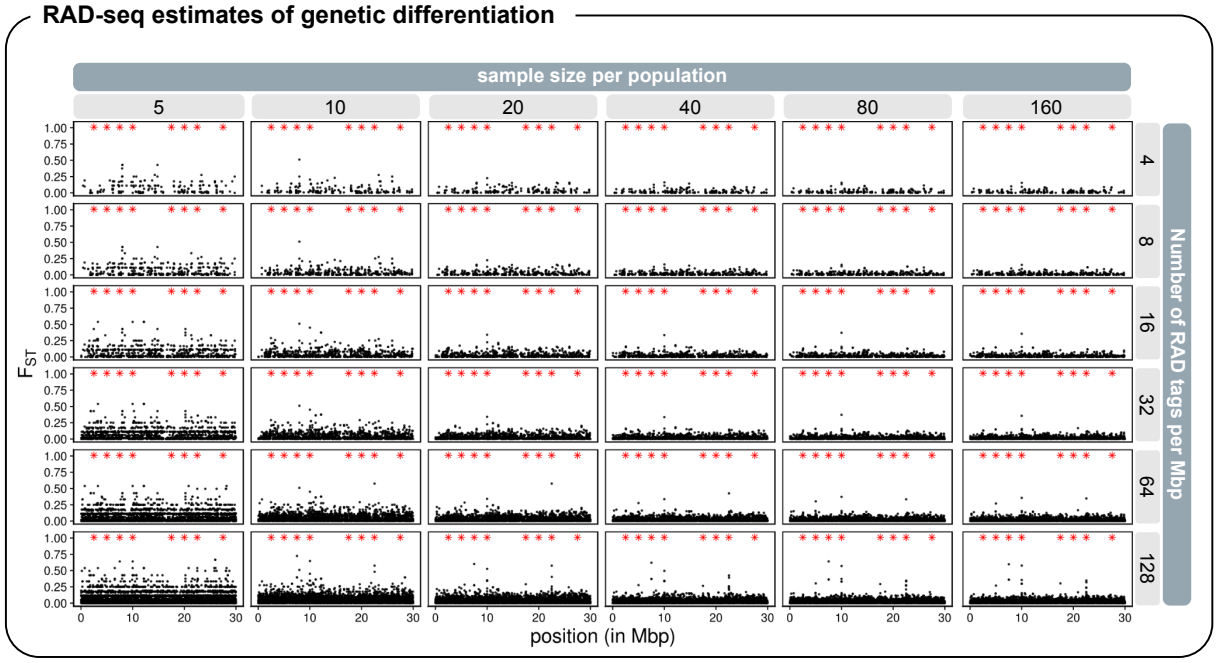

Figure 7. Genome-wide scans for divergent selection with RAD-seq data. The per-SNP F_{ST} values inferred from RAD-seq data are shown on the y axis and the SNP positions are shown on the x axis. The sample size per population increases from left to right, and the RAD-tag density increases from top to 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 F_{ST} values).

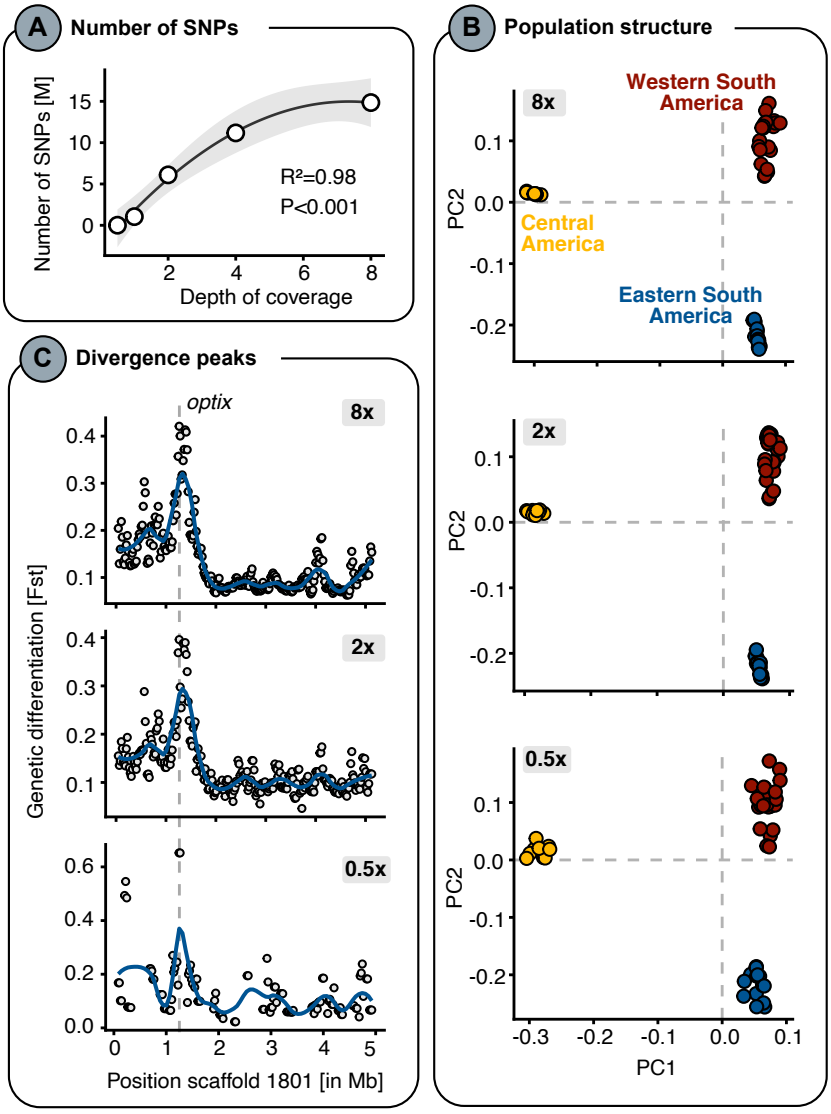

Figure 8. Application of genotype-likelihood-based analysis to downsampled empirical data. **(A)** Correlation between the number of identified SNPs (in millions) and depth of sequencing coverage in the downsampled *Heliconius* dataset. **(B)** Principal components analysis for three different coverages (8x, 2x and 0.5x) of 51 samples. Estimates of population structure are highly concordant across the coverage levels. Subspecies are pooled and colored by their broader region of origin. **(C)** Estimates of genetic differentiation (F_{ST}) between *Heliconius* subspecies with the red-bar phenotype (n=23) and without the red-bar phenotype (n=28) along the scaffold containing the causal *optix* candidate genes in 50kb sliding windows with 20kb steps. F_{ST} estimates are highly concordant between 8x and 2x coverage, but sparser at 0.5x due to the lower number of identified variant sites.

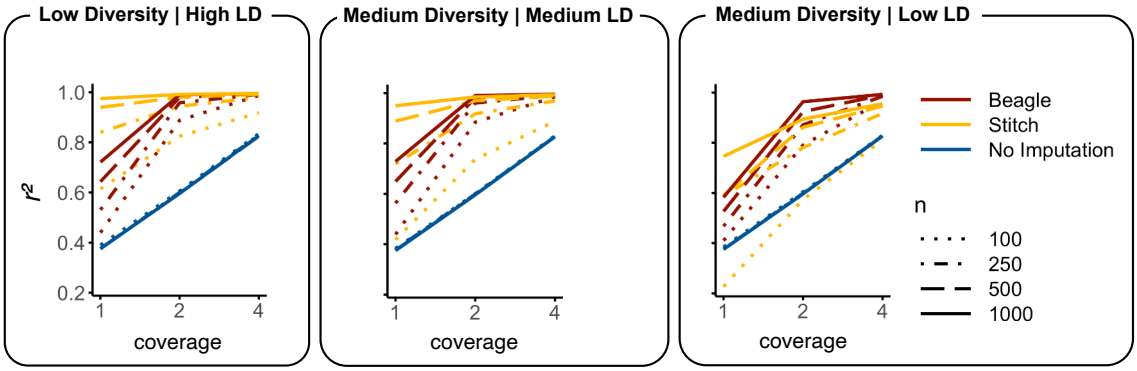

Figure 9. Genotype imputation in STITCH and Beagle compared to posterior genotypes estimated without imputation in three simulated populations with varying diversity and linkage disequilibrium. r^2 between true genotypes and estimated genotype dosages are shown for combinations of sample size (n; with increasing n indicated by more contiguous lines), sequencing coverage (x-axis) and method (line colors).