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

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

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

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

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 that 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 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 compared to sequencing fewer samples each at higher depth. 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 analysis. With this overview, we hope to make lcWGS more approachable and stimulate broader adoption.

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

1. Introduction

Despite massive drops 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 (breath of coverage), 2) how deeply to sequence each sample (depth of coverage), and 3) the total number of samples to sequence. Until recently, by far the most popular approach for population genomic studies of non-model species has been reduced-representation sequencing (e.g. RAD-seq), in which a small random portion of the genome can be sequenced deeply in many individuals to allow accurate genotype calls despite non-negligible error rates in individual sequence reads (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Davey et al., 2011; McKinney, Larson, Seeb, & Seeb, 2017). While RAD-seq and related approaches undoubtedly have 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 (Fig. 1A). Accordingly, RAD-seq data may completely miss important signatures of selection and adaptive divergence, which can be highly localized in the genome (Tiffin & Ross-Ibarra, 2014; Lowry et al., 2017).

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. Toews et al., 2016; Campagna et al., 2017 vs. Campagna, Gronau, Silveira, Siepel, & Lovette, 2015; Aguillon, Walsh, & Lovette, 2020 vs. Aguillon, Campagna, Harrison, & Lovette, 2018; Clucas, Lou, Therkildsen, & Kovach, 2019 vs. Clucas et al., 2019), 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 researchers. A popular cost-effective alternative is to sequence pools of individuals (Pool-seq; Schlötterer, Tobler, Kofler, & Nolte (2014)). When the number of individuals pooled and the sequencing depth is sufficient, Pool-seq represents a powerful approach to 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 all individual-level analysis as well as detection of cryptic substructure among sampled individuals (Fig 1B, Anderson, Skaug, & Barshis, 2014).

Low-coverage whole genome sequencing (lcWGS) is now 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 per-sample cost to RAD-seq and comparable per-population cost to 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 (Fig. 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 sample sizes.

At low depth of coverage, individual genotypes cannot reliably be inferred (Nielsen, Paul, Albrechtsen, & Song, 2011; Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012). 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 similarities or differences 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 take the uncertainty about true genotypes into account instead of assuming that any particular genotype call is correct, can integrate over the uncertainty about individual genotypes for population-level inference of variation at particular SNPs and integrate over the uncertainty about an individual's genotype at each particular SNP to make inference about that individual's overall genetic signature.

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 the number of individuals sequenced accordingly - maximizes the information gained about a population. Recent empirical studies have further demonstrated the power of this approach, for example in genome scans for regions of elevated differentiation between populations or differential admixture patterns, as well as analysis of LD patterns, genotype-phenotype associations, and fine-scale population structure (Ilardo et al., 2018; Clucas et al., 2019; Therkildsen et al., 2019; Wilder, Palumbi, Conover, & Therkildsen, 2020; Powell et al., 2020;)

Despite the clear promise, adopting a lcWGS approach can seem daunting because working with genomic data in a probabilistic framework rather than as called genotypes 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 the newcomer, 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 for their next population genomics project. We primarily use the term lcWGS to refer to whole genome re-sequencing with per-sample depths too low to reliably call genotypes without imputation (<5x), but note that even for medium sequencing depths (5-15x), inference accuracy may improve by adopting the probabilistic analysis frameworks discussed here, rather than working with hard-called genotypes (Nielsen et al., 2011). The paper is divided into seven sections. Following this introduction (Section 1), we first illustrate that lcWGS is now a feasible option for many research projects by comparing the current cost of lcWGS to alternative sequencing strategies and briefly reviewing practical considerations related to laboratory procedures, sample input requirements and the need for a reference sequence to map reads to (Section 2). Next, we introduce the basic statistical framework used to account for genotype uncertainty inherent in lcWGS data, and provide a comprehensive overview of existing analytical tools built under such a framework to help readers identify the software that can robustly perform common types of population genomics inference with lcWGS data (Section 3). We then expand on earlier work to guide experimental design by using 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 per sample, and compare the power of lcWGS to other sequencing strategies common in studies of non-model species, including RAD-seq and Pool-seq. Section 6 uses simulated data to explore the potential of genotype imputation for bolstering inference with lcWGS data in the absence of reference panels, and finally, in Section 7, we review challenges and limitations associated with lcWGS data and discuss future perspectives. With this practitioner-centered overview, we hope to make lcWGS seem more approachable 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, the cost of sequencing today is >600,000 times lower than in 2000 (Wetterstrand, 2020), and because of this spectacular price drop over the past decades, lcWGS can now - in many cases - be performed at similar per-sample costs as more widely used reduced-representation techniques. Table 1 provides estimates of the total per-sample cost for both library preparation and sequencing (based on November 2020 pricing) for organisms with different genome sizes. The cost of lcWGS inevitably scales with genome size (because more sequence data is needed to provide a target coverage level of a large vs. a small genome), and this approach therefore may remain an impractical solution for studies of organisms with extremely large genome sizes. However, even for organisms with sizeable genomes around 1 Gb (e.g. most birds and many fish, invertebrates, and plants), the per-sample cost with 1-2x sequencing coverage (20-32 USD) is now on par with the 30 USD recently reported as the typical cost for using RAD-seq to generate data for 20,000 variable loci (Meek & Larson, 2019), the 15 USD for a custom sequence capture approach to generate data for 500 - 10,000 loci (Meek & Larson, 2019), and the 48 USD reported for custom exome capture (Puritz & Lotterhos, 2018). For organisms with smaller genome sizes, lcWGS can end up cheaper than reduced-representation approaches, and prices are likely to drop further as sequencing costs continue to decrease.

2.2. Library preparation

Depending on target coverage levels, Pool-seq approaches remain the most cost-effective way to obtain genome-wide population-level data because they only require 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 separate indexed libraries for hundreds of individuals 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. 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, 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). Other cost-effective protocols produce successful lcWGS libraries even from picogram-levels of input DNA (Picelli et al., 2014; Hennig et al., 2018; Meier, Salazar, Kučka, Davies, & Dréau, 2020), 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 commonly used 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 (Nevado et al., 2014; Bohling, 2020). 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 1000 USD (Dudchenko et al., 2018; Gatter, von Löhneysen, Drozdova, Hartmann, & Stadler, 2020). At this point, it thus probably makes sense to start most new lcWGS with a de novo genome assembly or improvement, if a reference sequence of sufficient quality is not available.

BOX 1: Glossary

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 a constant, i.e. P(parameter | data) = P(data | parameter) * P(parameter) / P(data)

Empirical Bayes: a type of Bayesian inference method that differs from the classical Bayesian approach by having the prior probabilities estimated from the data.

Genotype dosage: the expected genotypic count. For diploid individuals, genotype dosage = 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 that identifies stretches of haplotypes shared between individuals so that missing genotypes or those with sequence read depth too low to confidently call can be more robustly estimated using information from shared haplotypes.

Genotype likelihoods: 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: AA, AC, AG, AT, CC, CG, CT, GG, GT, and TT), i.e. P(data | genotype), or L(genotype).

Genotype likelihood model: the mathematical model used to estimate genotype likelihoods. Different genotype likelihood models are built under different assumptions about the data, in particular about the 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, a second error is more likely to occur at the same site.

Likelihood ratio test: a hypothesis testing method that compares two competing hypotheses by evaluating the ratio of their likelihoods (i.e. probability of observing the data given each hypothesis), i.e. lIkelihood ratio test statistic = -2log(P(data | null hypothesis) / P(data | alternative hypothesis))

Maximum likelihood inference: a statistical inference strategy that estimates model parameters by choosing the parameters that maximize the likelihood of the data (i.e. p(data | parameter), or L(parameter)), i.e. maximum likelihood estimator of model parameter = 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.

Sample allele frequency 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 possibilities range from 0 to 2n; n=sample size), i.e. P(data | sample allele frequency).

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

3.1. Accounting for genotype uncertainty

Traditionally, most population genomic inference has been based on called genotypes. Yet, genotypes are not directly observable and must be inferred from sequencing data (or alternatively, targeted genotyping platforms). Because of the non-negligible error rates in sequencing data as well as the stochastic nature of allele sampling on high-throughput sequencing platforms that can result in uneven representation of the two chromosomes of a diploid individual, sequencing depths of at least 15-20x are typically required for confident genotype calls (Li, Sidore, Kang, Boehnke, & Abecasis, 2011; Nielsen et al., 2011). Many studies do call genotypes based on much lower sequencing depths, but while that may provide sufficient resolution for certain applications, low-depth genotype calls are likely to be highly error-prone and can substantially bias downstream analysis (Nielsen et al., 2012; Crawford & Lazarro, 2012; Han, Sinsheimer, & Novembre, 2014). Robust inference from lcWGS data, therefore, requires a new suite of analytical tools that instead of working with called genotypes, operate under a probabilistic framework, such that the uncertainty about individual genotypes can be incorporated in downstream analyses. Fortunately, such tools are now available to explicitly consider genotype uncertainty in most common types of population genomic inference. Here, we group these tools into three loosely defined categories: SNP discovery, individual-level analyses, and population-level analyses. We briefly introduce some of the most widely used software applications for each category and compile a more comprehensive list in Table 2. We also provide a tutorial with example data to provide a starting point for exploration at https://github.com/nt246/lcwgs-guide-tutorial. All discussion in this section, along with the next two sections, concerns genotype likelihoods-based inference on a SNP-by-SNP basis. In Section 6, we consider opportunities for further improving inference by leveraging population haplotype structures to impute missing or low confidence genotype information in species for which extensive reference panels are not available.

3.2. Genotype likelihoods

The most common way to incorporate uncertainty about true genotypes is to use genotype likelihoods (GLs) rather than genotype calls as input for downstream analyses, and genotype likelihoods thus form the foundation for the statistical framework used in most population genomic inference with lcWGS data. We note, however, that the use of genotype likelihoods is not exclusive to low-coverage data and can be an important step in to improve genotype calling pipelines for high and medium-coverage data as well, e.g. in GATK (McKenna et al. 2010). Genotype likelihoods are computed for each possible genotype held by each individual at each site of the genome. A genotype likelihood 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. Genotype likelihoods (plural) then refer to the set of likelihoods computed for each of all 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 is known (i.e. major-major, major-minor, minor-minor)). By basing downstream analyses on genotype likelihoods, genotype uncertainty caused by low coverage and sequencing errors (reflected by sequencing quality scores) can be explicitly taken into account.

Several models for computing genotype-likelihood-based on read data have been proposed. The main difference among them is their assumptions about how sequencing quality scores relate to the true probabilities of sequencing error. For example, the GATK model (McKenna et al., 2010) assumes that quality scores at the same site from different sequencing reads are each an independent and unbiased representation of the probabilities of sequencing error, whereas the Samtools model (Li, 2011) assumes that these quality scores are not completely independent. Both the SOAPsnp model (Li et al., 2009) and the SYK model (Kim et al., 2011) assume that the quality scores could be biased and thus implement a quality score recalibration step. All four of the above-mentioned models are implemented in ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014), which currently is the most widely used and versatile software package for the analysis of lcWGS data. Different genotype likelihood models adopted by other software packages can be useful alternatives to ANGSD for specific types of data. For example, the program Atlas (Kousathanas et al., 2017) explicitly incorporates post-mortem DNA damage in addition to sequencing error in its genotype likelihood model, making it well-suited for ancient DNA studies. EBG (Blischak, Kubatko, & Wolfe, 2018) uses a simplified version of the SAMtools model but relaxes ANGSD's assumption of diploidy, allowing the analysis of polyploid samples.

Unfortunately, the effects of genotype likelihood model choice on downstream analysis are still incompletely understood. Previous comparisons have suggested that while the genotype likelihood model choice seems to make little difference for some datasets, different models can give inconsistent results for other datasets, potentially biasing inference (Korneliussen et al., 2014). The sensitivity to genotype likelihood model choice may depend on the accuracy of base-calling and associated quality scores, the read coverage distribution and filtering, the sample size and particular individuals included in the sample, and how accurately data error structures match model assumptions (see Box 4 in Fuentes-Pardo & Ruzzante (2017)). More research is needed to compare the performance of genotype likelihood models, and in the meantime, it may be prudent to compare inference with several different models for each new dataset.

3.3. SNP identification and filtering

3.3.1. SNP identification: SNP calling is the procedure for identifying which sites in the genome are polymorphic within a sample or among a set of individuals. Arguably, the optimal solution in a genotype-likelihood-based framework is to avoid making hard calls about which sites are polymorphic and which are not, and instead use estimated genotype likelihoods for every site in downstream analysis. This approach is certainly appropriate for some types of analysis, e.g. for estimation of the site frequency spectrum (SFS) that require consideration of both polymorphic and non-polymorphic sites and low-frequency SNPs. Other types of analysis, however, are more tractable and computationally efficient when only considering sites that by some confidence criterion appear to be polymorphic.

Although some software tools are able to handle multi-allelic SNP data (e.g. GATK (McKenna et al., 2010) and Freebayes (Garrison & Marth, 2012)), biallelic SNPs are far more common and we will focus our discussion on those. In ANGSD, for example, SNPs are inferred by first estimating allele frequencies at each site (including the presumably invariable loci) and then testing whether its minor allele frequency is significantly larger than zero (Korneliussen et al., 2014). Accordingly, the first step is to determine the major and minor alleles at each site, either based on the genotype likelihoods of all individuals (Skotte, Korneliussen, & Albrechtsen, 2012), the provided reference or ancestral sequence, or by user specification, which can be useful when comparing with another dataset in which major and minor alleles are already determined. Next, the likelihood of the minor allele frequency at each site can be formulated as a function of genotype likelihoods across all individuals (see Equation 2 in Kim et al. (2011)), and these minor allele frequencies can be estimated using a maximum likelihood approach. In this way, all possible genotypes for each individual can be considered, effectively avoiding explicitly calling genotypes. Hardy Weinberg equilibrium (HWE) is assumed by default in this step in order to bridge allele frequency likelihoods with genotype likelihoods, although users can supply a table containing inbreeding coefficients for each individual to allow deviation from this assumption. Then, polymorphic sites will be identified through a likelihood ratio test, which evaluates whether the hypothesis that the minor allele frequency is equal to zero can be rejected based on a chosen significance threshold (Kim et al., 2011). The list of polymorphic sites (e.g. SNPs) can then be exported and used for downstream analyses, along with the genotype likelihoods at each of these sites for each individual.

Other software programs address SNP calling in different ways. Atlas (Kousathanas et al., 2017), for example, follows the same general framework as ANGSD, but has made slight modifications to accommodate cases where the sample size is very small and neither the major nor the minor alleles is specified by users, which is often the case for ancient DNA studies (Kousathanas et al., 2017). Furthermore, Atlas uses a different formulation of the likelihood ratio test such that allele frequencies are not required to be estimated before SNPs are called. The program Reveel (Huang, Wang, Chen, Bercovici, & Batzoglou, 2016), on the other hand, combines genotype likelihoods together with predefined prior genotype probabilities to calculate the posterior probability of genotypes for each sample at each site, and subsequently calculates the probability of alleles for a given sample at a given site. It then determines whether the site is polymorphic by integrating the probability of alleles from all samples using a monotonically increasing function and an arbitrary cutoff value. Reveel seems to work better with larger datasets (i.e. thousands of samples), and its computational time scales well with such sample sizes (Huang et al., 2016).

3.3.2. SNP and data filtering: Most software programs built for SNP discovery allow users to set specific quality control filters. Adjusting the SNP significance threshold (often in the form of maximum p-value), for example, can be used to finetune the sensitivity of the SNP calling step. A minimum depth filter and a minimum individual filter are used to exclude sites that are too difficult to map (e.g. tandem repeats), and a maximum depth filter is used to exclude sites that are susceptible to dubious mapping (e.g. due to copy number variation or homologs). Setting a minimum minor allele frequency filter excludes low-frequency SNPs that are uninformative for certain analyses. A minimum base quality and a minimum mapping quality filter can eliminate bases/alignments with very low levels of confidence. While the base quality score factors into genotype likelihoods estimation, the mapping quality score is not accounted for in any of the genotype likelihood models currently implemented in ANGSD and most other commonly used programs, so removal of reads with poor mapping quality prior to analysis can often reduce noise. It is important to note that different filters are warranted for different types of analyses. For example, as mentioned above, SFS estimation should not include any SNP significance or minimum minor allele frequency threshold, while various cut-off levels make sense for other types of analysis.

3.4. Individual-level analyses

We define individual-level analyses as those that do not require grouping individual samples into separate populations a priori. These analyses can typically be performed directly based on the genotype likelihoods estimated in the SNP discovery process. None of the analyses listed in this subsection are possible with Pool-seq data.

3.4.1. Population structure: A key component of many population genomic studies is to characterize the organization of genetic variation among individuals and populations (i.e. population structure). Two of the most widely used types of individual-based analyses for inferring population structure from lcWGS data are dimensionality reduction (e.g. PCA and PCoA) and model-based clustering (e.g. admixture analysis).

With dimensionality reduction methods, a metric to evaluate the genetic relationship between each pair of individual samples is often used as the input. In the case of PCA, an eigendecomposition is performed on a pairwise covariance matrix to find the principal components that can explain the highest proportions of variance in the data (Patterson, Price, & Reich, 2006), and in the case of PCoA, a multidimensional scaling (MDS) is performed on a pairwise distance matrix to achieve the same goal. Such covariance or distance matrices are typically generated from genotype matrices (e.g. Patterson et al. (2006)), but several different programs now can compute them while accounting for genotype uncertainty. For example, ANGSD 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). Covariance and distance between every pair of individuals can then be calculated from such allele frequencies. This method is computationally efficient, and despite its simplicity, it seems to perform well when a large number of individuals and polymorphic sites are included in the dataset even at extremely low coverage (<1x, see the ANGSD website and our evaluation below). However, this single read sampling process does not take full advantage of the entire dataset. In contrast, ngsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014) uses a more sophisticated method where posterior genotype probabilities are first calculated with an empirical Bayes approach from genotype likelihoods and prior genotype probabilities informed by the allele frequencies among all samples, and the covariance matrix can then be estimated from these posterior genotype probabilities. This approach is valid under the assumption of Hardy-Weinberg equilibrium across the entire sample set, but for most structured populations, this assumption will not hold, which can lead to inaccurate PCA results. PCAngsd (Meisner & Albrechtsen, 2018) therefore takes one step further and uses an iterative approach to correct for potential violation of the HWE assumption by updating prior genotype probabilities based on the PCA result in each previous iteration, since these PCA results can represent the population structure that exists in the data (Meisner & Albrechtsen, 2018).

Dimensionality reduction methods tend to be computationally efficient and do not rely on strong assumptions about the data, but oftentimes they can only provide a qualitative overview of the variation among individuals. In contrast, model-based clustering methods typically explicitly assume a model of discrete ancestral populations and aim to estimate the admixture proportion of each sample (i.e. the proportion of the sample's genome that originates from each discrete ancestral population). The most widely used programs for clustering analysis, including

STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) and the more computationally efficient ADMIXTURE (Alexander & Lange, 2011) and FRAPPE (Tang, Peng, Wang, & Risch, 2005), all require called genotypes as input. However, several specialized programs implement the same underlying model in a framework based on genotype likelihoods. For example, NGSAdmix (Skotte, Korneliussen, & Albrechtsen, 2013) adopts a maximum likelihood implementation of the classic STRUCTURE model (Tang et al., 2005), but formulates a likelihood function with sequencing data as its observed data and uses genotype likelihoods to consider all possible genotypes for each individual (see Equation 6 in Skotte, Korneliussen, & Albrechtsen, 2013)). It then uses an expectation-maximization (EM) algorithm to optimize the likelihood function and estimate model parameters such as admixture proportions. Because of the more complex formulation of the likelihood function, however, NGSAdmix tends to be computationally demanding. As an alternative, Ohana (Cheng, Racimo, & Nielsen, 2019) adopts the same likelihood function as NGSAdmix but uses a sequential quadratic programming (QP) method instead of EM for optimization, which should speed up computation. No formal comparison between the performance of the two methods is available to date, but separate evaluations on both simulated and real data have shown that both methods deliver great accuracy even for very low-depth data (Skotte, Korneliussen, & Albrechtsen, 2013; Cheng et al., 2019). Distinct from both NGSAdmix and Ohana, PCAngsd uses individual allele frequencies, an intermediate output from its PCA analysis, as input for a non-negative matrix factorization (NMF) algorithm to infer admixture proportions. This approach is shown to significantly outperform NGSAdmix in runtime without strongly compromising its inference accuracy, so it might be more suitable for larger datasets (Meisner & Albrechtsen, 2018).

3.4.2. Selection scans: Unlike the genomic signature of population structure that is mostly homogeneous across the entire genome, selection tends to leave its footprint only at its target loci and neighboring regions. In fact, a key advantage of lcWGS over reduced-representation sequencing techniques is its ability to more comprehensively uncover these localized signatures of selection. For selection scan methods that do not require a priori assignment of individuals into populations, the general strategy is to locate outlier loci that exhibit patterns of variation among all individual samples that are highly different from the genome-wide signal. For example, PCAngsd (Meisner & Albrechtsen, 2018) adopts the method by (Galinsky et al., 2016) and implements it for low-coverage data (i.e. in a genotype likelihood framework). This method measures the level of differentiation at each SNP along each of the top PC axes as its selection statistic. This statistic is expected to follow a chi-squared distribution if solely affected by population structure, so outlier SNPs (if there are any) may be affected by selection. Alternatively, in Ohana (Cheng et al., 2019), allele frequencies from K ancestral populations outputted from its genotype-likelihood-based admixture analysis are used to construct a covariance matrix that reflects the relationship among these ancestral populations. SNPs that exhibit a significantly different covariance structure can subsequently be identified using a likelihood ratio test as candidates for selection.

3.4.3. Genome-wide association analysis: Genome-wide association studies need a large number of individuals to detect significant genotype-phenotype associations. Using low-coverage whole-genome sequencing and genotype likelihoods allows one to maximise the number of individuals studied in a cost-efficient way. Several approaches that take genotype uncertainty into account for association analyses have been developed in recent years and have shown power to discover meaningful associations under a range of different scenarios, including the presence of population structure (Skotte et al., 2012; Jørsboe & Albrechtsen, 2019). Many of these approaches have been implemented in ANGSD. In Kim et al. (2011), for example, case / control association is tested by first estimating allele frequencies within case and control individuals with a genotype-likelihood-based maximum likelihood approach as described in the “SNP identification” section, and then using a likelihood ratio test for differences between case and control individuals at each locus (see equations 6-7 in Kim et al. 2011). In addition to binary phenotypes, genome-wide association with quantitative traits can be tested with the methods developed by Skotte et al. (2012) and Jørsboe & Albrechtsen (2019), and both approaches allow for incorporation of additional covariates. The first step in both methods is to calculate the posterior genotype probability using an empirical Bayes approach, with priors informed by either population allele frequencies or the SFS. Skotte et al. (2012) then used a score statistics approach to test for significant associations with the phenotype at each site. This approach is computationally efficient, but cannot estimate the effect size of the loci. In contrast, (Jørsboe & Albrechtsen, 2019) employs a maximum likelihood approach using an EM algorithm to explicitly estimate the effect size of each locus. As expected, this approach is slower than the score statistics method. To take advantage of both methods, ANGSD also implements a hybrid approach, first using the score statistic to identify significant loci, and then using the EM approach to estimate effect sizes of these significant loci.

3.4.4. Linkage disequilibrium: The estimation of linkage disequilibrium (LD) has many important applications, for example relating to inference of population size, demographic history, selection, and discovery of structural variants. In addition, since many downstream analyses make assumptions about the independence of genomic loci, LD estimates are essential for pruning lists of loci to be included in these analyses. Traditional methods to measure LD rely on resolving individual haplotypes from genotype data, but maximum likelihood approaches that account for genotype uncertainty in unphased sequencing data have been developed to enable LD estimation from lcWGS data. Simulations have suggested sampling more individuals each at lower coverage actually produces more accurate estimates of LD than higher coverage for fewer individuals and that a mean coverage of 2x appears to be the optimal allocation of resources for LD estimation (Maruki & Lynch, 2014; Bilton et al., 2018). The overall performance and dependence on read depth depends both on the underlying algorithm, the diversity levels of

LD patterns within the sampled populations, and the statistic used to summarize LD. GUS-LD (Bilton et al., 2018), for example, constructs a likelihood function of the LD coefficient and uses a numerical method (the Nelder–Mead method) to optimize the likelihood function. In contrast, ngsLD (Fox, Wright, Fumagalli, & Vieira, 2019) constructs a likelihood function of the haplotype frequencies between each pair of SNPs instead, and uses an EM algorithm to optimize it. Different LD statistics, such as D, D' and r^2, can then be derived from the inferred haplotype frequencies. These algorithmic differences make ngsLD less computationally demanding than GUS-LD, and comparative evaluation has indicated that ngsLD tends to show less bias at low read depths (1-2x) than GUS-LD (Fox et al., 2019). In addition to these maximum likelihood approaches, ngsLD also implements an alternative method where it simply calculates r^2 from genotype dosages between pairs of loci as a measurement of LD, and furthermore, it incorporates several other helpful features, such as LD pruning and the fitting of an LD decay model.

3.4.5. Other types of analysis: In addition to the examples discussed above, many other specialized software packages have been developed to account for genotype uncertainty in different 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 (Vieira, Fumagalli, Albrechtsen, & Nielsen, 2013; Link et al., 2017) and identity-by-descent tracts (Vieira, Albrechtsen, & Nielsen, 2016), tests for introgression such as computation of ABBA-BABA/D-statistic (Korneliussen et al., 2014), and construction of linkage maps (Rastas, 2017). More examples are listed in Table 2. We also note that samples sequenced to low-coverage of the nuclear genome typically have very high sequencing depth across the mitochondrial genome due to the much higher copy number in each cell. This enables recovery of high-confidence full mitochondrial genome sequences for each individual (see e.g. Therkildsen &

Palumbi, (2017)) that can be used for high-resolution phylogeographic analysis (Lou et al., 2018; Margaryan et al., 2020).

3.5. Population-level analyses

When individual samples can be grouped into discrete populations 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 on the population level. Two key population-level analyses are the estimation of allele frequencies and the SFS, both of which have been implemented for lcWGS data with genotype uncertainty taken into account. Numerous other population-level analyses can then be conducted directly using the estimated allele frequencies and/or the inferred SFS as their inputs, and we will focus our review on a few examples of these.

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

3.5.2. Site frequency spectrum: The population-specific SFS is another key population genomic parameter that is essential for many downstream analyses. It is possible to discretize the estimated allele frequency distribution and use it as the SFS, but a key issue with low coverage data is that low-frequency SNPs in the sample are less likely to be called and therefore an SFS directly estimated from allele frequencies can be biased towards intermediate frequencies. To get around this issue, ANGSD estimates the SFS by first calculating the sample allele frequency (SAF) likelihoods (the probability of data given each possible sample allele frequency) at each site from the genotype likelihoods of each individual using a dynamic programming algorithm. These SAF likelihoods can then be used to formulate the likelihood function of the SFS, which the program then optimizes (see equations 5-7 in Nielsen et al., (2012)). This method corrects for the bias caused by low-coverage data, and can be generalized to estimate the SFS jointly for up to four populations (Nielsen et al., 2012). Depending on the availability of an outgroup or ancestral reference genome, the inferred SFS can either be folded or unfolded.

The dynamic programming part of the ANGSD's workflow can greatly reduce the computational cost of the SAF calculation, but the runtime still grows quadratically with the number of samples and it can become impractical if the sample size is very large. Han et al. (2015) has thus proposed an alternative “score-limited dynamic programming” algorithm to speed up the SAF calculation with limited compromise on its accuracy (Han, Sinsheimer, & Novembre, 2015) .

3.5.3. Genetic diversity and neutrality test statistics within a single population: The genetic diversity within a population is often evaluated by the parameter $\theta = 4N_e\mu$. Different estimators of θ, such as Tajima's estimator (also known as nucleotide diversity or π) and Watterson's estimator, are essentially different linear combinations of the SFS, and therefore the genome-wide estimate of θ can be directly calculated from the population-specific SFS. However, population genomic studies often need to look beyond the average diversity across the genome. Particularly, genomic regions impacted by natural selection often leave a signature of reduced/increased θ and/or skewed SFS compared to the rest of the genome (Fay & Wu, 2000; Tajima, 1989). Although it is possible to use the maximum likelihood method described above to separately estimate the SFS in each window along the genome in order to calculate θ, it can be computationally intensive to do so. ANGSD therefore adopts an empirical Bayes approach, where the SFS within a window (posterior) can be formulated and solved as the product of the SAF likelihoods within the window (likelihood) and the genome-wide or chromosome-wide SFS (prior) (see the equation in the "Empirical Bayes" section in Korneliussen, Moltke, Albrechtsen, & Nielsen, (2013)). Different theta estimators can then be extracted from the SFS in each window. Subsequently, different neutrality test statistics (e.g. Tajima's D) can be calculated by taking the difference between different θ estimators to evaluate the skewness of SFS in each genomic window. If an unfolded SFS is available, additional θ estimators and neutrality test statistics can be estimated, such as Fay and Wu's H (Fay & Wu, 2000) and Zeng's E (Zeng, Fu, Shi, & Wu, 2006). This approach is shown to be computationally efficient and to give relatively accurate estimates with low-coverage data (Korneliussen et al., 2013). Lastly, when this same method of SFS estimation is applied to individual samples instead of populations, individual heterozygosity estimates can be obtained. Diversity statistics can also be estimated with other programs, e.g. Atlas (Kousathanas et al., 2017) that in contrast to the infinite sites model implemented in ANGSD, bases theta estimates on a model by Felsenstein (1981) that allows for back mutations.

3.5.4. Genetic differentiation between populations: Genetic differentiation between populations can be evaluated with a variety of different statistics, starting from simply quantifying the allele frequency difference to more complex statistics such as relative genetic divergence (F_{ST}), absolute genetic divergence (d_{xy}) and others (e.g. pFst). Many of these statistics can now be estimated within a genotype-likelihood based-framework. One of the oldest and most widely-used statistics among these is F_{ST} which evaluates the proportion of the total genetic variance that can be explained by population structure. ANGSD implements the method-of-moment estimator developed by Reynolds, Weir, & Cockerham (1983). While θ at each site in the genome depends on the local SFS within a single population, Reynolds et al.'s estimator of pairwise F_{ST} can be formulated as a function of the local two-dimensional SFS (the matrix with the joint distribution of allele counts in two populations). Therefore, ANGSD again takes an empirical Bayes approach, using the maximum likelihood method to estimate a genome-wide two-dimensional SFS, which it then uses as a prior to calculate SFS at each genomic locus. Fst at each locus can then be derived from these locus-specific SFS.

GPAT (http://www.yandell-lab.org/software/gpat.html) implements two additional methods to estimate F_{ST} using genotype likelihoods as its input. In the first method (wcFst), GPAT estimates allele frequencies from genotype likelihoods and directly plugs the estimated allele frequencies into Weir and Cockerham's F_{ST} estimator. This method is computationally efficient but may not account for the uncertainties in the estimated allele frequencies as well as ANGSD does. In the second method (bFst), GPAT implements a Bayesian framework as described by Holsinger, Lewis, & Dey (2002), with a modification in its original likelihood function such that genotype likelihoods can be used as input instead of called genotypes. This Bayesian approach has the advantage of being able to provide a confidence interval for F_{ST}, but it is computationally expensive.

In addition to these various F_{ST} estimators, GPAT can also estimate pFst, which quantifies the significance of allele frequency differences between populations, but is not an F_{ST} estimator itself (Domyan et al., 2016). In contrast, no established method to estimate d_{xy}, a measure of absolute divergence, has been included in major software packages to our knowledge. 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 distribution of d_{xy} and not to make inferences based on its absolute values

3.5.5. Other analyses based on derived statistics: Many other types population-level analysis can be conducted based on the derived statistics that are mentioned above. For example, several commonly used software tools or analytical approaches can use allele frequency matrices to test for deviation from the Hardy-Weinberg equilibrium (e.g. ANGSD), infer population relationships and potential gene flow (e.g. Treemix (Pickrell & Pritchard, 2012), conStruct (Bradburd, Coop, & Ralph, 2018)), perform selection scans (e.g BayPass (Gautier, 2015), Bayescan (Foll & Gaggiotti, 2008), WFABC (Foll, Shim, & Jensen, 2015)), association analyses (e.g. BayPass) 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 code. Similarly, the population-specific or multi-dimensional SFS estimated from ANGSD can be used to infer demographic history (e.g. dadi (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009), 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, 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 phased data, yet require detailed SNP information for reference populations. Overall, one downside of all these analyses, however, is that uncertainties in the derived statistics cannot be taken into account directly.

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

More data usually results in better inference. But with a limited 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 addressed this question with analysis of simulated data (e.g. Buerkle & Gompert, 2013; Fumagalli, 2013; Nevado et al., 2014). In general, these studies have found that sampling many individuals at low read depth provides both more accurate and more precise estimates of population parameters than higher read depth for fewer individuals. Buerkle and Gompert (2013), for example, showed that dividing the sequencing effort maximally among individuals and obtaining approximately one read per locus and individual (1x coverage) yields the most information about a population for allele frequency estimation. Consistent with this, Fumagalli (2013) also found that 1x coverage maximizes power for inference of population structure. Surveying a broader set of population genetic parameters and demographic histories of the sampled populations, Fumagalli did, however, find that under some circumstances, the highest accuracy was achieved at sequencing depths of 2x, where both alleles are more likely to have been sequenced. Other studies (e.g. Nevado et al. 2014) have suggested that the minimal per-sample depth should be even higher.

To shed more light on optimal experimental designs for lcWGS and how thinly we should spread our sequencing effort among individuals, we used simulated data to compare common types of population genomic inference under different sample sizes and sequencing depths, including <1x, which was not explicitly evaluated in the previously mentioned studies.

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

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

4.1. Population genetic inference of an isolated population

We simulated an isolated population that has reached mutation-drift equilibrium, and evaluated the accuracy of lcWGS in inferring key population genetic parameters, including allele frequencies, the SFS, θ, and Tajima's D. As expected, more sequencing data is always better and the accuracy in allele frequency estimation consistently increases with higher sample size and coverage (as measured by the r^2 values in Figure 2). The number of false negative SNPs (i.e. true SNPs in the population that fail to be called) similarly decreases with higher sample size and higher coverage (Figure S1). Importantly, however, distributing the same total sequencing effort (i.e. sample size x coverage) across more samples, with each sample receiving less coverage (e.g. going from bottom left to top right in Figure 2) also consistently improves allele frequency estimation, even when each sample is sequenced at a coverage as low as 0.25x.

Next, we estimated the SFS and derived estimators of θ and Tajima's D from the SFS from each dataset. Similar to what ANGSD's authors have previously shown (Korneliussen et al., 2014), we found that the genotype likelihood model used for this analysis can strongly affect its result. With the Samtools genotype likelihood model, Watterson's θ is systematically underestimated when the average coverage is low (<4x), although Tajima's θ (π) estimates are more robust in face of lower coverage (Figure S2). Consequently, Tajima's D tends to be overestimated (Figure S3), which may lead to an erroneous inference of demographic contraction. In contrast, when the GATK genotype likelihood model is used, Watterson's θ, Tajima's θ, and Tajima's D can all be accurately estimated even at coverage as low as 0.25x (Figure S4, S5). The difference arises because with the Samtools genotype likelihood model, lower-frequency mutations are less likely to be called as SNPs and are more likely to be interpreted as sequencing errors when the coverage is low. This leads to an underestimation of the number of singleton mutations, and therefore Watterson's θ tends to be underestimated. We note that these low-frequency SNPs have minimal impacts on many other population genomic analyses and are often filtered out as a result, so we do not expect strong discrepancies between the two genotype likelihood models in most types of analysis. We also stress that the sequencing errors modeled in our simulations may not accurately represent the sequencing error profile from different sequencing platforms in real life, so our result should not be interpreted as a recommendation of one genotype likelihood model over the other with real data.

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

Thus far, our simulations of different per-sample sequencing depths have assumed that the sequencing effort is equally distributed among all samples. In actual lcWGS studies, this assumption can often be approximated by sequencing in multiple batches and repooling samples based on their output from the first round(s) to add proportionally more sequence to samples that initially generated less data in follow-up sequencing rounds. This has proved to be highly effective for evening out per-sample sequencing yields in our experience (Figure S6). However, repooling based on sequencing output is not feasible for Pool-seq where samples do not have unique barcodes. The common practice to approximate even coverage in Pool-seq, then, is to pool samples in equimolar amounts, but this is often inaccurate due to measurement and pipetting errors, variation in DNA quality, and sequencing biases. To assess the impact of such inaccuracies, we compiled an empirical distribution of relative sequence coverage achieved among samples from three of our lcWGS projects where we pooled individually indexed libraries by molarity, and we sampled from this distribution to simulate a realistic scenario of inadvertent variation in coverage among samples in a pool (Figure S7). We analysed the resulting sequencing data under both a lcWGS design (assuming samples are individually barcoded) and a Pool-seq design (assuming samples are not individually barcoded). We found that with a lcWGS design, the allele frequency estimation is slightly, yet consistently, less accurate in the uneven coverage scenario as compared to the even coverage scenario, since the effective sample size is smaller if some samples contribute more to the pool than others (Figure 3). When each sample is barcoded (as in lcWGS), this uneven contribution can be recognized and accounted for in genotype-likelihood-based inference. Under a Pool-seq design, allele frequencies are simply estimated from allele counts, so the samples that contribute more to the pool tend to more strongly influence allele frequency estimation, leading to much higher errors (Figure 3). As an example, with any sample size between 5 and 160, a lcWGS design with an average of 4x coverage can generate more accurate allele frequency estimation than a Pool-seq design with an average of 8x coverage per sample (as evaluated by RMSE in Figure 3). It is also worth noting that even if samples could be sequenced at perfectly even coverage in a Pool-seq experiment, the allele frequency estimation is still notably less accurate than in lcWGS, because there can be individuals contributing more sequences than others at each given locus due to sampling variance (Figure S8).

4.2. Inference of spatial structure

To evaluate the power of different lcWGS sampling designs to detect population structure, we simulated a metapopulation consisting of nine subpopulations located in two-dimensional space that have reached mutation-drift-migration equilibrium. Each subpopulation has an effective size of 10^4 and is positioned at a node of a three-by-three grid. On this grid, each pair of neighboring subpopulations are connected by bidirectional gene flow (Figure 4). We took samples from each of these populations, simulated the lcWGS process with different combinations of sample sizes per population and sequencing depth per sample, and performed PCA from the simulated data to characterize the genetic relationship among samples and subpopulations, which should mirror spatial relationships.

We first examined a scenario in which gene flow among subpopulations is low (0.25 effective migrants between neighboring subpopulations per generation on average). In this scenario, the spatial structure among subpopulations can be correctly inferred even with extremely low sample size (5 samples per subpopulation) and coverage (0.125x coverage per sample) (Figure 4A). In addition, migrant individuals and hybrids, when included in the sample, can be identified in the PCA (Figure 4A), which would not be possible with a Pool-seq design.

We then increased the level of gene flow by a factor of four (1 effective migrant between subpopulations every generation on average). As expected, the power of PCA to resolve the spatial structure declines, but interestingly, small sample size appears to cause a greater loss of power than low coverage does (Figure 4B). Subpopulations fail to form discrete clusters in the 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, a correct spatial structure can be inferred with a coverage as low as 0.125x (Figure 4B). The reason is that PCA requires reliable covariance estimation between pairs of samples. With larger sample size, more pairs of samples are likely to share informative SNPs between them that have non-zero coverage (note that it is a quadratic relationship), and the overall population structure is more likely to be extrapolated from these pairs of samples. Therefore, to resolve the spatial structure among subpopulations connected by gene flow, it is probably preferable to distribute a given amount of sequencing effort across more samples rather than aiming for higher coverage per individual. Note, however, that our simulations are only informative about qualitative patterns because the power of PCA will depend on the number of polymorphic sites for which data are available. We only simulated a single chromosome for this analysis due to computational limitations. With real data where the genome size is often much larger than the chromosome size that we have simulated, we expect that the spatial structure among subpopulations connected by higher gene flow can be more accurately resolved by lcWGS data with similar sample size and coverage presented here (see Figure S9 for an example of this).

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

A primary advantage of lcWGS compared to reduced-representation sequencing approaches is the increased resolution for genome scans for signatures of selection, for example in the form of outlier SNPs that show elevated levels of differentiation between populations. To evaluate how tradeoffs between sample size and per-sample sequencing depth affect our ability to detect outliers, we simulated two populations connected by gene flow that have reached mutation-drift-migration equilibrium. We then introduced a number of mutations that are strongly beneficial in one population but strongly deleterious in the other, and ran the simulation for another 200 generations. We estimated F_{ST} between the two populations from lcWGS data to identify the loci under selection (details in the supplementary material).

We first examined a scenario where the size of each population is large (N_e = 5x10^4) and gene flow is high (5 effective migrants per generation on average). In this scenario, eight selected loci are segregating in the two populations after 200 generations of selection, and seven out of the eight show highly elevated F_{ST} values compared to the genome-wide background (Figure 5). The one locus with a low F_{ST} value is likely kept at low frequency in both populations due to Hill-Robertson interference. Their neighboring neutral SNPs, driven by linked selection, also exhibit elevated F_{ST}, creating a distinct pattern of narrow genomic islands of divergence caused by divergent selection in the face of gene flow (Figure 5; Turner, Hahn, & Nuzhdin, 2005). This F_{ST} landscape can be recovered from lcWGS data with a total sequencing coverage larger than 10x in each population (e.g. 40 samples per population and 0.25x coverage per sample, Fig. 5). With lower sample size (e.g. 5 samples per population and 2x coverage per sample), however, the background F_{ST} tends to be overestimated, which can lead to more false positive signals in the outlier detection. With higher sample size or coverage per sample, the F_{ST} peaks become higher in magnitude and the background noise diminishes. With the same total sequencing effort, we also see a decline in the background noise when the sample size is larger (along the diagonal from bottom left to top right in Figure 5).

Next, we examined a scenario with smaller N_e (N_e = 10^4) and lower gene flow (an average of 2.5 effective migrants from one population to the other every generation). With these parameter changes, the background level of differentiation becomes larger, the F_{ST} peaks become wider (due to higher LD), the density of SNPs becomes lower (due to lower θ), and there are more peaks with intermediate F_{ST} values (due to stronger Hill-Robertson interference; Figure S10). In this scenario, lcWGS performs similarly well, where many of the F_{ST} peaks can be recovered with a total coverage larger than 10x per population and where larger sample size can further reduce the false positives (Figure S11). Compared to the scenario with larger sample size and higher gene flow, however, there tends to be more false positive signals due to the higher background F_{ST}.

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

Compared to lcWGS, RAD-seq has the advantage of being able to generate high-confidence genotype calls, but it often suffers from a sparser coverage of the genome, which can be particularly problematic for selection scans (Lowry et al., 2017). Here, we simulated the process of RAD-seq under 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 very high marker density (128 RAD tags per Mb, or 128,000 tags in a 1 Gb genome), RAD-seq tends to miss some of these narrow F_{ST} peaks. With a lower yet commonly used marker density (e.g. 8 tags per Mb or 8,000 variable tags in a 1 Gb genome), an overwhelming majority of the signals are missed regardless of sample size (Figure 6). In the scenario where 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, it is still difficult to identify distinct F_{ST} peaks with the RAD-seq data (Figure S12).

5. Application to empirical data

To supplement our simulation-based evaluation of lcWGS inference power 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 show 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) and analysed them in a genotype likelihood framework. For simplicity, we focus on results for 8x, 2x and 0.5x coverage, as results from 4x and 2x are very similar to 8x and 1x, respectively (not shown).

First, we found a positive correlation between the number of variable sites identified during SNP calling in ANGSD and the mean genome-wide sequencing coverage depth (Figure 7a; quadratic function: $R^2 = 0.98$, p=0.00099). Across all 51 individuals used in the final analyses, the number of SNPs identified ranged from 12,266 at 0.5x coverage to 14,851,731 at a mean coverage depth of 8x. For a dataset with a mean per-individual coverage of 0.25x we could not reliably identify any SNPs using ANGSD at the specified SNP p-value threshold of 1e-6 (Figure 7). These results are congruent with the inferences drawn from our simulation study.

Second, we reconstructed the population structure using principal components analysis, 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 7b). One subspecies (*H. erato hydara*) with samples 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 even at per-individual coverages as low as 0.5x (Figure 7c) (Van Belleghem et al., 2017). At 0.5x coverage, we were able to estimate 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 7c). However, even at this low resolution, we detected differentiated genomic windows in the optix region, albeit the estimated F_{ST} was higher 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 1x and moderate sample sizes per population, we can detect population structure and recover distinct signals of differentiation across the genome in empirical data.

6. Using imputation to bolster genotype estimation from lcWGS

As discussed, lcWGS can be a powerful method of estimating parameters across samples or across sites, but confidence in individual genotypes at sites in the genome is limited. So far, we have considered data on a SNP-by-SNP basis. By contrast, imputation, a method whereby stretches of chromosome shared among individuals are identified, leverages information from flanking alleles to inform missing or low confidence genotypes, and can under some circumstances be used to improve genotype likelihoods and boost individual genotype accuracy. Imputation generally works under the assumption that chromosomes which share a series of alleles flanking a site of interest are likely to also share alleles at that site (Li et al. 2011), leveraging LD patterns inferred from sequenced individuals or reference panels of haplotypes (Pasaniuc et al., 2012). Imputation has been most commonly used to boost the power of genome-wide association studies (GWAS), typically by increasing calling rates for rare SNPs, but can also be used to impute non-SNP variation or SNPs not present in a reference SNP panel (Marchini & Howie, 2010). Perhaps most significant for lcWGS is the capacity for imputation to fill in sporadic missing data and improve posterior genotype probabilities (Browning & Yu, 2009; Y. Li, Willer, Ding, Scheet, & Abecasis, 2010).

Most imputation methods designed for use with lcWGS rely on externally generated haplotype reference panels that are typically unavailable for non-model species. Although programs such as Beagle and findhap can be used without them, they perform best with reference panels (Browning & Yu, 2009; VanRaden, Sun, & O’Connell, 2015). One exception is the program STITCH (Davies, Flint, Myers, & Mott, 2016), which imputes directly from sequence read data without reference haplotype panels. With large numbers of samples (>2,000), STITCH has been shown to perform as well as other imputation methods that rely on reference panels (Davies et al., 2016). However, sample sizes of this magnitude are uncommon among studies of non-model species, and although obtaining thousands of samples may be feasible for some species, for others (e.g. elusive, rare or endangered species) it may be difficult or impossible. To explore the performance of imputation with sample sizes typical of studies of non-model species, we simulated population genetic scenarios to identify the conditions under which imputation may bolster genomic analyses of lcWGS, testing combinations of per-sample sequencing depths and sample sizes under each scenario.

6.1. Simulations and genotype estimation

To explore imputation performance under different scenarios, we used the same framework as in section 4.1 in forward simulation of a 30MB chromosome for three neutrally evolving populations that have reached mutation-drift equilibrium. Here, we varied the effective population size (Ne) and recombination rate (r) to create three different scenarios with different levels of genetic diversity and LD because these parameters are known to affect imputation performance (Pasaniuc et al., 2012). In a neutral population, genetic diversity is proportional to the product of effective population size and mutation rate, whereas LD is inversely proportional to the product of effective population size and recombination rate, and accordingly, our three scenarios were characterized by 1) a low diversity, high LD scenario (r = 0.5 cM/Mbp, Ne = 1,000); 2) a medium diversity, medium LD scenario (r = 0.5 cM/Mbp, Ne = 10,000); and 3) a medium diversity, low LD scenario (r = 2.5, Ne = 10,000). For each simulated scenario, we constructed a series of sampling schemes with sample sizes of 25, 100, 250, 500 or 1000 individuals and similar to our approach in Section 4, we used ART-MountRainier (W. Huang et al., 2012) to simulate sequencing reads to average depths of 1x, 2x and 4x per individual for each sample size.

For each scenario, sample size and sequence depth, we compared the estimated genotype accuracy using no imputation (i.e. called from posterior genotype probabilities in ANGSD), and using two imputation programs, Beagle v.3.3.2 and STITCH v.3.6.2, both run without reference panels. We evaluated the performance of each method by the r^2 between true genotypes and allelic dosage (i.e. the sum of posterior probabilities for the alternate allele times 0 for homozygous reference, times 1 for heterozygous, and times 2 for homozygous alternate), by the proportion of correct genotype calls (genotype concordance), and the proportion of called genotypes (see the Supplemental methods for details on simulations and genotype estimation and imputation).

6.2 Imputed genotype accuracy and genotype concordance

For all sample sizes and sequencing depths across scenarios, the accuracy of genotype estimates varied with allele frequency. At the smallest sample size tested (n=25), there was little to no improvement in accuracy using Beagle, and accuracy actually decreased when imputation was performed in STITCH with 25 samples (Figures S13-S15), suggesting that such small sample sizes are inadequate for reliable imputation; thus we focus the rest of our results on n≥100. The correlation (r^2) between imputed allelic dosage and true genotypes was low for sites with minor allele frequency (MAF) < 0.05 to 0.10, but increased and was relatively consistent across higher MAF bins (Figure S13). Genotype concordance (GC), by contrast, had the opposite relationship with MAF; GC was higher for sites with low MAF and decreased with higher MAF (Figure S14). This is because it is easy to achieve high accuracy by calling the homozygous major genotype when the minor allele is rare. In order to summarize overall imputation performance, we averaged r^2, GC and the proportion of called genotypes across sites with MAF>0.05 for each combination of method, scenario and study design (Figure 8).

Compared to genotypes estimated without imputation, the correlation between estimated and true genotypes was most improved by imputation under the low diversity, high LD scenario (Figure 8A). Under this scenario, the greatest improvements were seen when at least 100 samples were sequenced at 1x coverage and imputed in STITCH. Genotype dosages imputed in STITCH using large sample sizes (n≥500) sequenced at 1x coverage had high accuracy (r^2>0.94), whereas imputation in Beagle performed best with coverage ≥2x from sample sizes

≥250. The pattern was similar for a population with medium diversity and medium LD (Figure 8B), with accuracy of imputation somewhat reduced in STITCH but not in Beagle. Imputation performance was markedly worse in the medium diversity, low LD scenario. For sample sizes <250, imputation in STITCH actually decreased genotype accuracy, but there was still an improvement for both Beagle and STITCH when imputation was applied to large sample sizes.

Genotype concordance (GC) was universally high for all methods and sequencing strategies (GC>0.9), except for imputation of 100 samples from the medium diversity, high LD scenario in STITCH (Figure 8D-F). At 1x coverage, fewer than half of genotypes were called by Beagle and without imputation, especially for sites with higher MAF (Figure S15). GC was similar under the medium diversity, medium LD scenario compared to the low diversity, high LD scenario (Figure 8D-E), except GC was somewhat lower for genotypes imputed in STITCH at 1x coverage. The least improvement in GC using imputation was seen under medium diversity, low LD scenario (Figure 8F). For n≤250 samples sequenced at 1x and 2x coverage, GC for genotypes imputed in STITCH were less accurate than those estimated without imputation.

Overall, imputation seemed to be most beneficial for genotype estimation in the populations with small Ne (i.e. low diversity, high LD) and also confer some improvement in populations with larger Ne (medium diversity, medium LD) with sufficient sample sizes and sequencing coverage. However, the benefits were much more modest in a large population with high recombination rate (high diversity, low LD). This was particularly true for STITCH, which estimates distinct haplotype probabilities within a given region across a mosaic of ancestral haplotypes (Davies et al., 2016), a problem that becomes increasingly complex under high recombination. Imputation showed larger improvements with increasing sample size in STITCH than in Beagle, especially at low coverage (1x), whereas Beagle improved more with increasing sequence read depth (Figure 8).

6.3 Allele frequency estimation from imputed genotype probabilities

Because imputation increased the accuracy of posterior genotype probabilities under most scenarios and study designs, we tested whether there was an improvement in allele frequency estimation using imputed genotype probabilities compared to MAF estimation without imputation. To estimate MAF from imputed genotype probabilities, we summed over the posterior genotype probabilities (-domaf 4 in ANGSD), and compared the results to MAF estimated from genotype likelihoods using the EM algorithm implemented in ANGSD (-domaf 1). Under some scenarios and study designs, imputation resulted in small improvements in accuracy of allele frequency estimation (Figure 9). Imputation yielded the largest improvements for large sample sizes (n≥250) sequenced at 1x coverage from the low diversity, high LD population, and from the medium diversity, medium LD population. For small sample sizes from the medium diversity, low LD population, MAF estimated from genotype probabilities imputed in STITCH were less accurate. Beagle showed more consistent, modest improvements, increasing MAF estimation accuracy when coverage was ≥2x for all sample sizes and scenarios.

6.4 Considerations for using imputation in non-model systems

Choosing whether to apply imputation to real-world datasets may depend on the question of interest as well as the details of the study system. For many questions, there is more to be gained by increasing sample size than sequencing depth. This is because for the same sequencing effort (sample size x coverage), the number of genotypes estimated can be greatly increased with modest reduction in genotype accuracy. For example, in the low diversity, high LD population, genotypes imputed in STITCH from 1000 samples 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 genome-wide association tests, where large sample sizes are necessary for adequate power, genotype uncertainty can be incorporated directly into the analysis (e.g. (Skotte et al., 2012). Imputation has been shown, albeit at larger sample sizes and with reference panels, to increase the power of these analyses (Y. Li et al., 2010), in part by reducing genotype uncertainty at sites with limited or zero sequence read depth. The case for increasing sample size over increasing read depth is also true for estimating allele frequencies, as is the case even without imputation (as shown in Section 4). Under the low diversity, high LD scenario, allele frequency estimates based on genotype probabilities imputed in STITCH from 1000 samples at 1x coverage were slightly more accurate (r^2=0.999) than for 500 samples at 2x coverage (r^2=0.998) and 250 samples at 4x coverage (r^2=0.997). However, given that smaller sample sizes are already sufficient for estimating allele frequencies with high accuracy without imputation (r^2=0.990 for MAF estimated from 250 samples sequenced at 1x coverage; Figure 9), imputation is not likely to contribute to analyses of these types of population-level statistics as much as it would for individual-level and genotype-level analyses like GWAS.

Because the performance of imputation varies depending on the diversity and particularly the degree of LD in populations, knowledge of some details of the study system may help researchers anticipate how well imputation will perform. Typically researchers have an idea of levels of diversity, but perhaps less about LD, which can be highly variable across the genome. A set of “true genotypes” (e.g. from high-depth samples) can be used to assess imputation performance, but in the absence of samples for validation, performance can also be assessed based on quality metrics output by the imputation programs (Browning & Yu, 2009; Davies et al., 2016). The optimal imputation method to use will also depend on the study design for a given system. When coverage is higher than 1x, imputation without a reference panel in Beagle resulted in consistent improvement in genotype estimation under all scenarios, but modest to little improvement with 1x coverage. Imputation with STITCH was more accurate at 1x coverage, but only when sample sizes were large and LD was high to moderate, whereas imputation in STITCH performed poorly with small sample sizes from a large population with low LD.

In general, imputation may reliably benefit genotype estimation in non-model species (i.e. species without a reference SNP panel and typically studied with modest sample sizes) under limited circumstances. Populations with small Ne or that have experienced recent bottlenecks, such as threatened or endangered species, will have low diversity and higher levels of LD (Hayes, Visscher, McPartlan, & Goddard, 2003; Waples & Do, 2010), making them potentially good systems for applying imputation, but only as long as relatively large sample sizes can be obtained (e.g. ≥250 for the scenarios simulated here). For larger populations with lower LD levels, even larger sample sizes are needed. Even though large sample sizes may be more readily obtained when populations are large, imputation has more limited potential to improve lcWGS analysis in such scenarios.

7. Limitations, Developments and Conclusion

Throughout this paper, we have demonstrated the utility of lcWGS for population genomics. We and others have shown that for many types of inference (e.g. allele frequency estimation, principal component analysis, and characterization of genetic differentiation), a lcWGS approach actually can provide more accurate results than higher sequencing coverage of fewer individuals. We have also illustrated that a broad selection of software that allows relatively efficient data processing with genotype-likelihood-based approaches is now available (Table 2). Thus, the promise is great, but there are clear limitations to this data type as well.

First of all, it is important to stress that the potential for improved inference accuracy by spreading sequencing effort over many individuals is only realized if the resulting uncertainty about genotypes is accounted for statistically in downstream analysis, with approaches such as those reviewed in this paper. As discussed, 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 absolutely require hard called genotypes. However, as outlined in Section 3, genotype-likelihood-based inference frameworks have now been developed for most major types of population genomic analysis.

One practical limitation is that some methods based on genotype likelihoods carry much greater computational costs than their counterparts based on called genotypes and/or they may have limited accuracy at very low read depth. For example, SFS estimation from genotype likelihoods in ANGSD is generally robust at medium sequencing depths (Nielsen et al., 2012), but is computationally intensive with very large sample sizes, which may be prohibitive for researchers without access to high memory computational resources. Furthermore, SFS estimation at depths lower than ~2x is potentially sensitive to the choice of genotype likelihood model, which warrants further investigation (see Sections 3.2, 4.1, and Box 4 in Fuentes-Pardo & Ruzzante (2017) for more detailed discussions). Some genotype-likelihood-based tools also tend to perform poorly at very low sequencing depth due to inherent limitations of the model and/or the algorithm used. Estimates of LD, for example, tend to have higher error rates and be more biased for sequencing depths <2x (Fox et al. 2019). But with per-sample sequencing depths of 2x or greater and large sample sizes, both LD and SFS estimation should be robust (Fumagalli 2013; Fox et al. 2019).

It is important to remember, however, that most genotype-likelihood-based tools are based on models that carry specific sets of assumptions (e.g. the Hardy-Weinberg assumption for allele frequency estimation in ANGSD), and violation of those assumptions can bias results. Therefore, as with all population genomic inference, it is important that users carefully review the underlying assumptions of analytical tools and interpret results accordingly.

One major limitation, for which no bioinformatic solution is yet available, is that accurate phasing of lcWGS data without a reference panel has not yet been possible, therefore prohibiting haplotype-based analyses in most non-model organisms. 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 even at low coverage, remains too costly for routine re-sequencing in hundreds of individuals as needed to leverage lcWGS approaches. However, the recent development of an affordable linked-read low-coverage approach (haplotagging; Meier et al., 2020) 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. In addition to advances through such novel sample preparation techniques, new analytical approaches for short-read lcWGS data also continue to emerge. Genotype-likelihood-based equivalents to established approaches, such as implementation of the Pairwise Sequentially Markovian Coalescent (PSMC) model, are, for example, currently under active development (ngsPSMC [https://github.com/ANGSD/ngsPSMC]). While these approaches do not address all the analytical gaps yet and potentially have reduced power, they are promising advances for the use of lcWGS in non-model species.

Finally, we clearly recognize that lcWGS is not an optimal solution for all projects. There are systems in which this approach may never be practical. In particular, for species that are rare or difficult to collect (e.g. endangered species and elusive species), it may be impossible to obtain adequate sample sizes for accurately estimating population genomic parameters with lcWGS. In cases where sample size is constrained, it may be better to sequence fewer individuals at higher depth. Some analyses, for example of demographic history, diversity, selective sweeps and inbreeding levels, can be performed just based on deep sequencing of the genome of a single individual (e.g. Li & Durbin, 2011). For species with very large genomes (e.g. many amphibians and pine species), whole genome sequencing may also remain impractical at any sequence depth from a cost perspective, and a reduced representation approach such as RAD-seq or targeted sequence capture may be preferable (Burgon et al., 2020; McCartney-Melstad, Mount, & Shaffer, 2016). De novo RAD-seq locus discovery without a reference requires a relatively high sequencing depth, but for targeted methods like sequence capture, low-coverage sequencing of larger sample sizes and associated genotype-likelihood-based analysis can, similar to WGS, confer distinct advantages over sequencing fewer individuals at higher depth (e.g. Therkildsen et al., 2019; Warmuth & Ellegren, 2019; Wilder et al., 2020).

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 allows population-scale WGS projects with individual-level resolution even on modest budgets. The toolbox for lcWGS analysis based on genotype likelihoods is rapidly expanding, making it an increasingly promising approach for molecular ecology, conservation and evolutionary biology research.

Acknowledgements

We would like to thank the Therkildsen Lab at Cornell University for comments on an earlier version of this manuscript, and Philipp Messer and Robbie Davies for advice on analysis.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. The NCBI SRA accession numbers for the Heliconius data re-analyzed in this project is available in Table S1.

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 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 divergenceLos taxones de Colaptes auratus son diferenciables con datos genómicos pese a sus bajos niveles de divergenciaGenomic data distinguish Northern Flicker taxa. *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*.doi: https://doi.org/10.1101/2020.07.10.197715.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246.

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

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

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.

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

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.

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.

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

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.

Crawford, J. E., & Lazzaro, B. P. (2012) Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. Frontiers in Genetics 3: 66.

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.

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.

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413.

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.

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.

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., 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: 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. *arXiv*. http://arxiv.org/abs/1207.3907

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.

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.

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

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

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

Ilardo, M. A., Moltke, I., Korneliussen, T. S., Cheng, J., Stern, A. J., Racimo, F., … Willerslev, E. (2018). Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, *173*(3), 569–580.e15.

Jørsboe, E., & Albrechtsen, A. (2019). A Genotype Likelihood Framework for GWAS with Low Depth Sequencing Data from Admixed Individuals. *bioRxiv*. doi: https://doi.org/10.1101/786384

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

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

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

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

Kousathanas, A., Leuenberger, C., Link, V., Sell, C., Burger, J., & Wegmann, D. (2017).

Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*, *205*(1), 317–332.

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., 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.

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

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.

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.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834.

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.

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.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511.

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.

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

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. (2020). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *bioRxiv*. doi:

https://doi.org/10.1101/2020.05.25.113688

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

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.

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.

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.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190.

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.

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

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

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.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2012). Association testing for next-generation sequencing data using score statistics. *Genetic Epidemiology*, *36*(5), 430–437.

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

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

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

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

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.

Toews, D. P. L., Taylor, S. A., Vallender, R., Brelsford, A., Butcher, B. G., Messer, P. W., & Lovette, I. J. (2016). Plumage Genes and Little Else Distinguish the Genomes of Hybridizing Warblers. *Current Biology: CB*, *26*(17), 2313–2318.

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.

Wetterstrand, K. A. (2020). DNA sequencing costs: data from the NHGRI genome sequencing program (GSP) www. genome. gov/sequencingcostsdata. *Accessed August*, *5*.

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.

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*, *4(5), 430-443*.

Zeng, K., Fu, Y.-X., Shi, S., & Wu, C.-I. (2006). Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics*, *174*(3), 1431–1439.

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.

Tables and Figures

Table 1. Total cost per sample for both library preparation and sequencing based on November 2020 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
0.6	16(8)	24(16)	Atlantic silverside, Stickleback, Eastern oyster
1	20(12)	32(24)	Zebra finch, Chicken, Purple sea urchin
3	44(36)	79(71)	Human, Atlantic salmon, African clawed frog

*Cost estimates do not include labor and assume that samples are sequenced efficiently on an Illumina HiSeq X Ten system. The assumed costs break down to 8 USD per library (commercial kit reagents) and 1,300 USD per lane generating 110 Gb sequence data. 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 lcWGS data

Analyses		Software						
Analysis	**Method**	**ANGSD**	**Atlas**	**MAPGD**	**GPAT**	**ngsTools**	**PCAngsd**	**Specialised software**
SNP calling		✓	✓					GATK, Reveel, EBG, Freebayes, BaseVar, etc.
Allele frequency estimation		✓	✓	✓				
Site frequency spectrum		✓				✓		ngs2dSFS
Allele frequency differentiation	pFst				✓			
Population differentiation	Fst	✓			✓			
	Dxy					✓		ngsStat
Within population genetic diversity	thetas (Watterson, π)	✓	✓					
Within population neutrality stats	e.g. Tajima's D, Fay & Wu's H	✓						
Individual level genetic diversity	Individual heterozygosity	✓	✓	✓				heterozygosity-em
Inbreeding	Inbreeding coefficient		✓		✓	✓	✓	ngsF, ngsRelate
	IBD tracts							ngsF-HMM
	Runs of homozygosity							bcftools roh
Population structure	PCA	✓				✓	✓	ngsCovar
	Local PCA							lostruct*
	Individual genetic distance	✓	✓			✓		skmer, ngsDist
	Admixture						✓	ngsAdmix, Ohana, Entropy, evalAdmix
Ancestry relationships	D-statistics/ABBA-BABA	✓	✓		✓			
Individual relatedness	Relatedness			✓			✓	ngsRelate
	Parentage							AlphaAssign
	Pedigree analysis							WHODAD
Linkage disequilibrium				✓		✓		ngsLD, GUS-LD, PopLD
Selection scan	PCA-based; ancestry-corrected						✓	Ohana
Association analysis		✓						SNPTEST
Structural variants								svgem
Quality score recalibration		✓	✓					
Genotype imputation								loimpute v0.18, STITCH, LB-Impute, NOISYmputer, LinkImput, etc.

HWE		✓		✓			✓	
Ploidy inference								HMMploidy
Linkage map construction								Lep-MAP3

Note: References for each software can be found in the main text (Section 3) or in the supplementary material.

*lostruct can be used together with custom scripts that perform the PCA e.g. in PCAngsd.

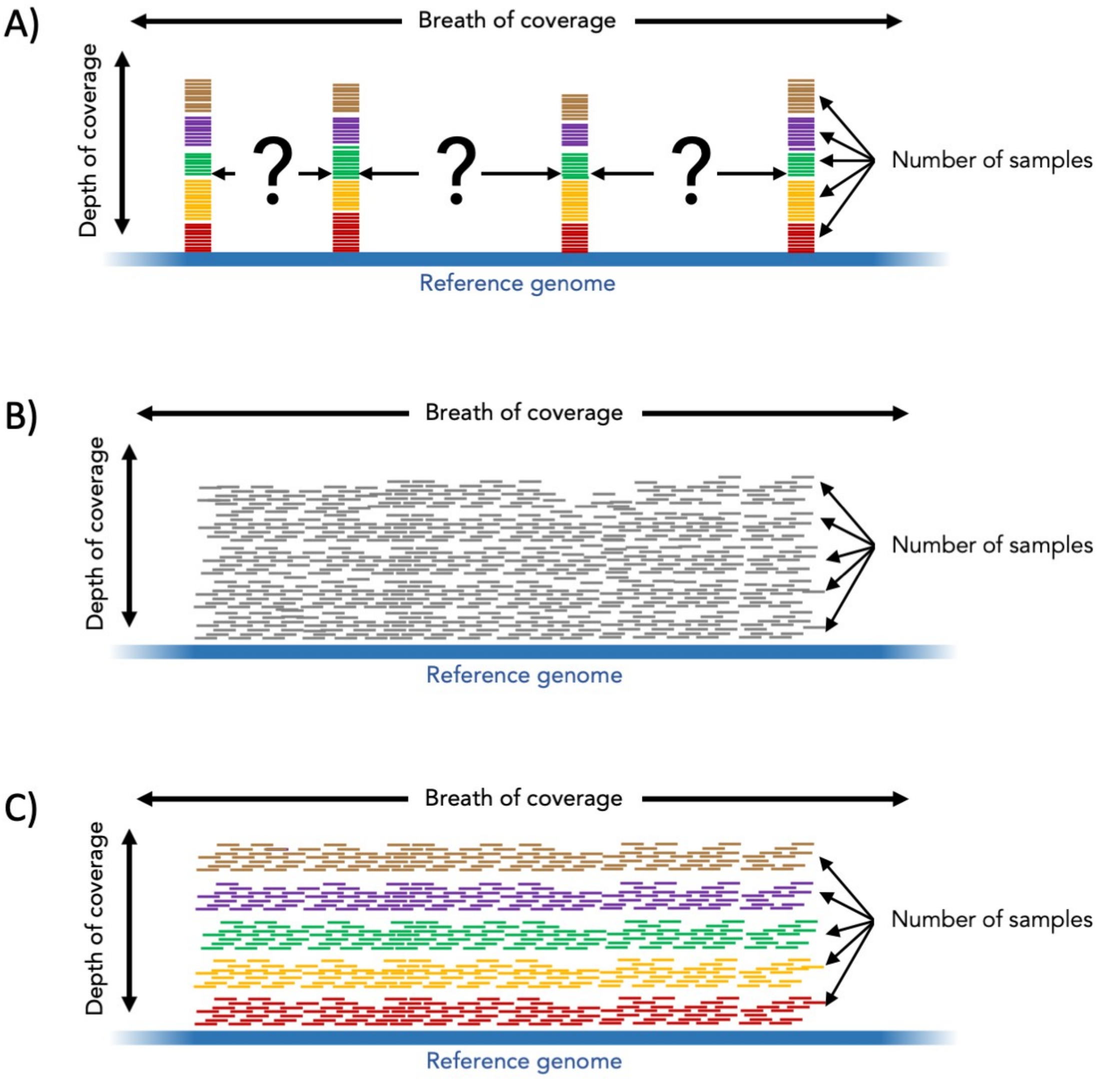

Figure 1. Diagram showing the distribution of sequencing reads mapped to a reference genome under a RAD-seq (A), Pool-seq (B), and lcWGS (C) design.

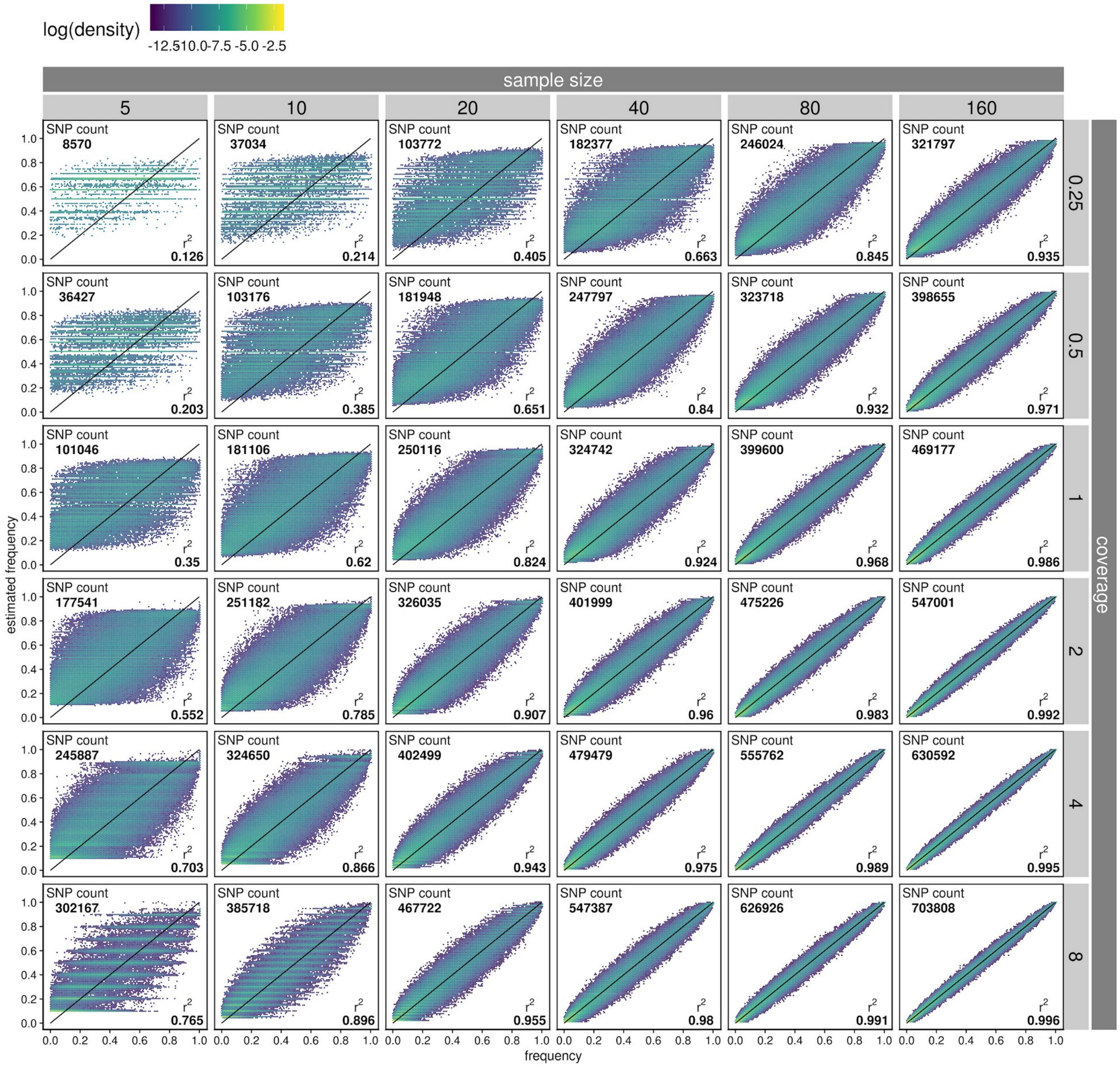

Figure 2. The estimated vs. true allele frequencies at all called SNPs (i.e. true positives + false positives) with lcWGS. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. The color indicates the density of points in the area, 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 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.

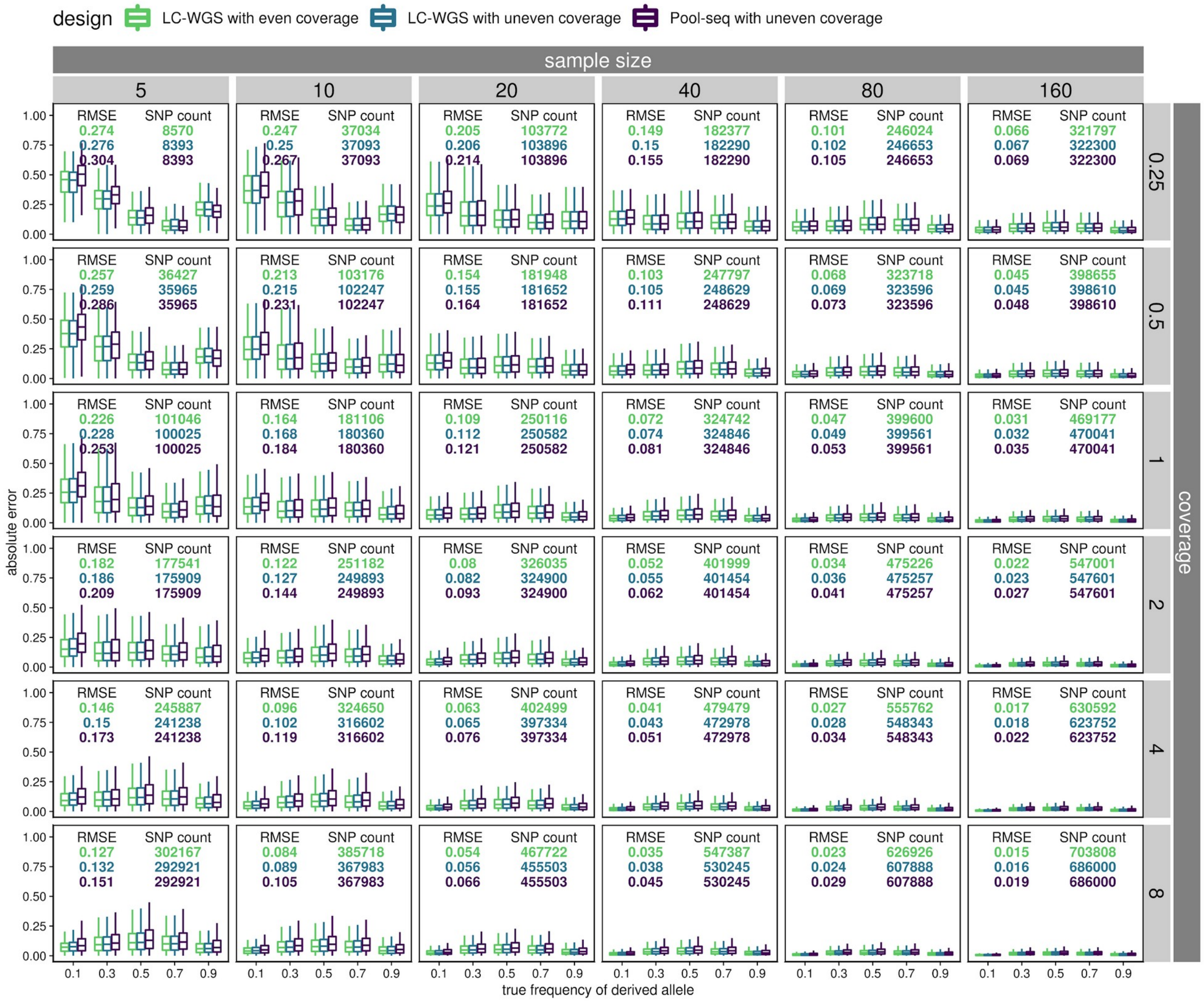

Figure 3. The error in allele frequency estimation with lcWGS and pool-seq data. Derived alleles are binned according to their true frequencies on the x axis, and their absolute errors (|estimated frequency - true frequency|) are shown on the y-axis. Across the different facets, sample size increases from left to right, and coverage increases from top to bottom. The total sequencing effort remains the same along the diagonal from bottom left to top right. Different colors correspond to different sequencing designs, and their root mean squared error (RMSE) and the number of SNPs called (SNP count; this includes the true positives and the false positives) are shown in each facet. False negative SNPs are not included in this figure.

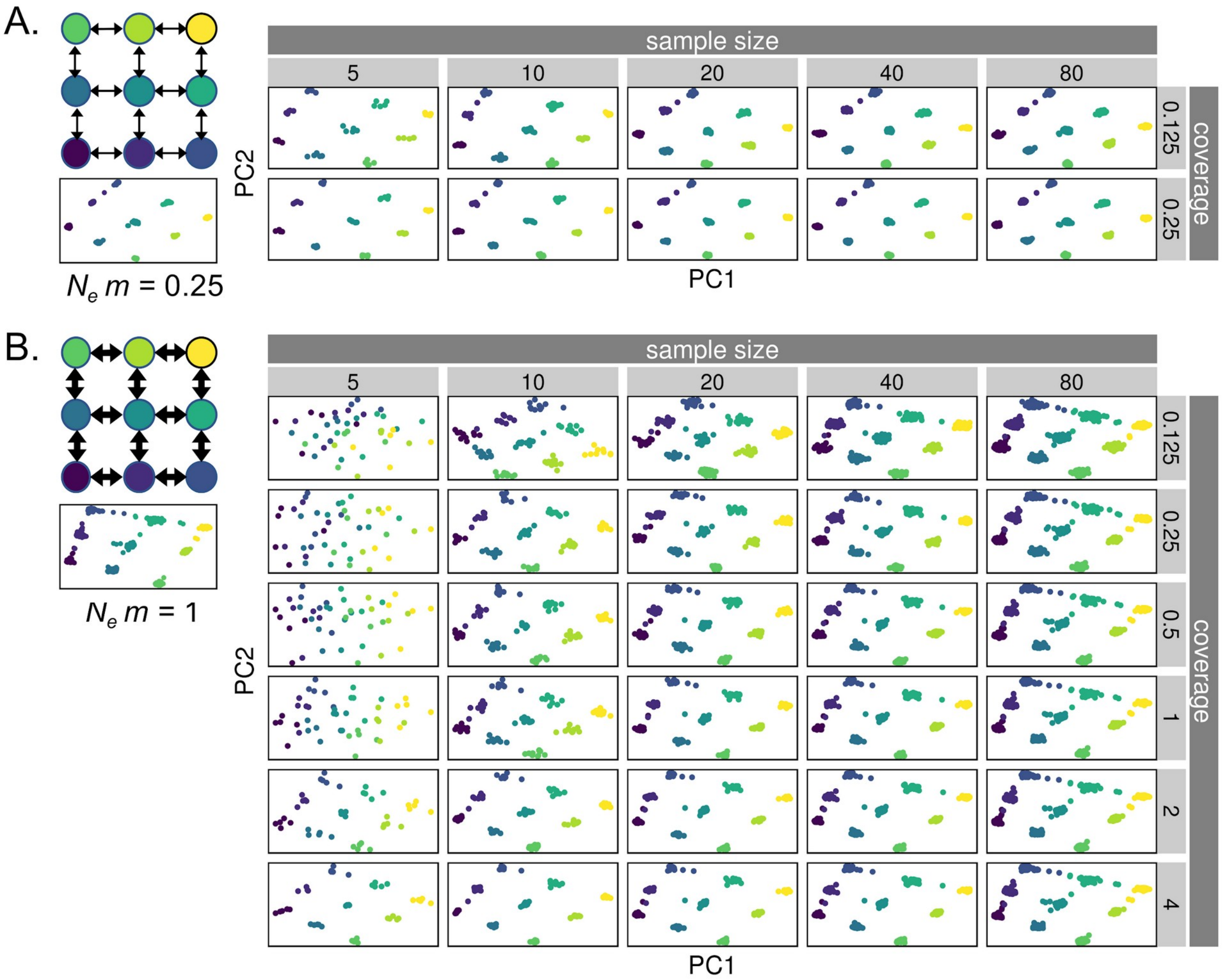

Figure 4. The spatial population structures inferred through principal component analysis (PCA) with lcWGS data. (A) A scenario with lower gene flow (an average of 0.25 effective migrants from one population to a neighboring population per generation). (B) A scenario with higher gene flow (an average of 1 effective migrant per generation). Top left: the true population structures being simulated; each node corresponds to a simulated population, and arrows indicate the direction of gene flow. Bottom left: the first two principal components from PCA performed with the true genotypes of 80 samples per population. 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. Sample size per population increases from left to right, and coverage per sample increases from top to bottom.

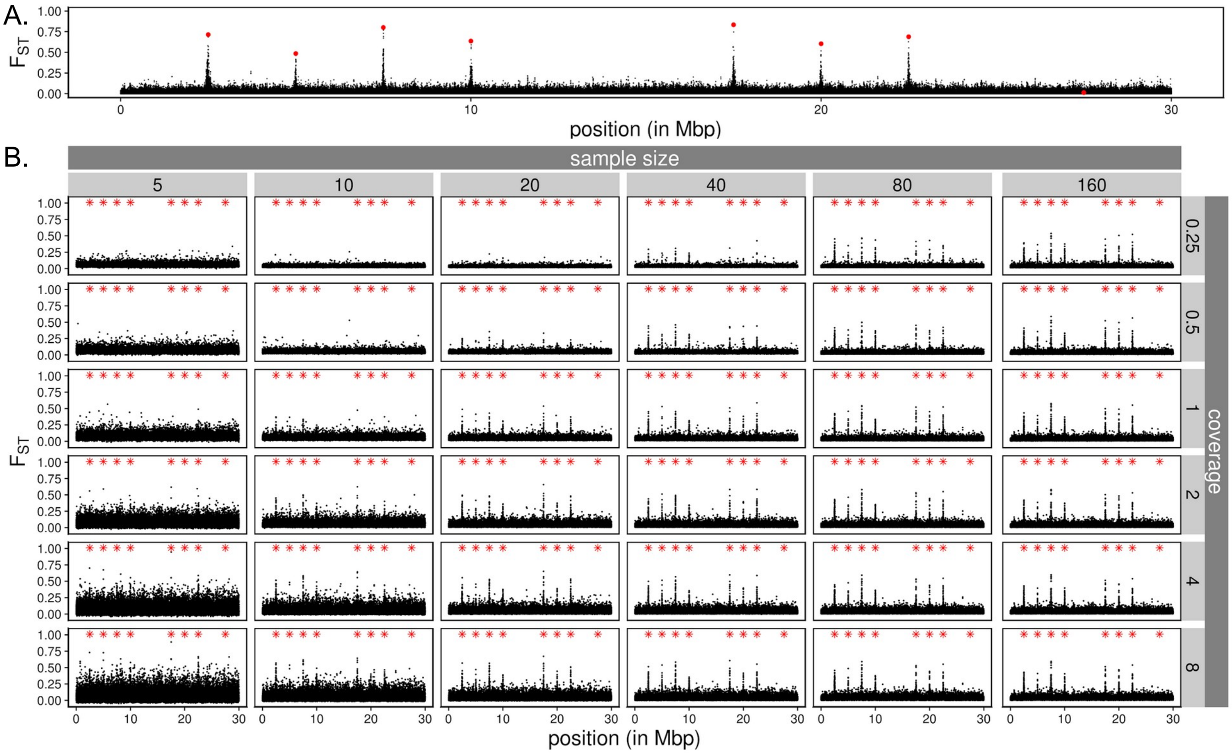

Figure 5. Genome-wide scan for divergent selection with lcWGS data. (A) The true per-SNP F_{ST} values along the chromosome between the two simulated populations. (B) The F_{ST} values inferred from lcWGS data in 1kb windows along the chromosome. Sample size per population increases from left to right, and coverage per sample increases from top to bottom. In (A), the red points mark the position 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).

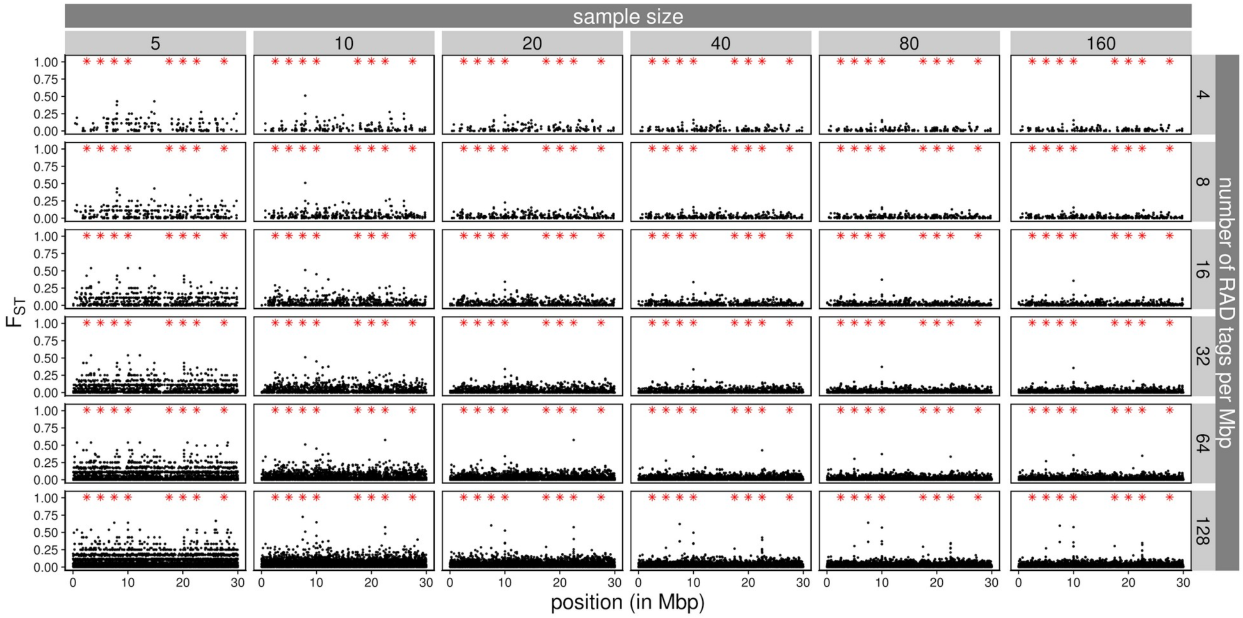

Figure 6. Genome-wide scan for divergent selection with RADseq data. The per-SNP F_{ST} values inferred from RADseq data are shown on the y axis and the SNP positions are shown on the x axis. Sample size per population increases from left to right, and RADtag density increases from top to bottom. The 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).

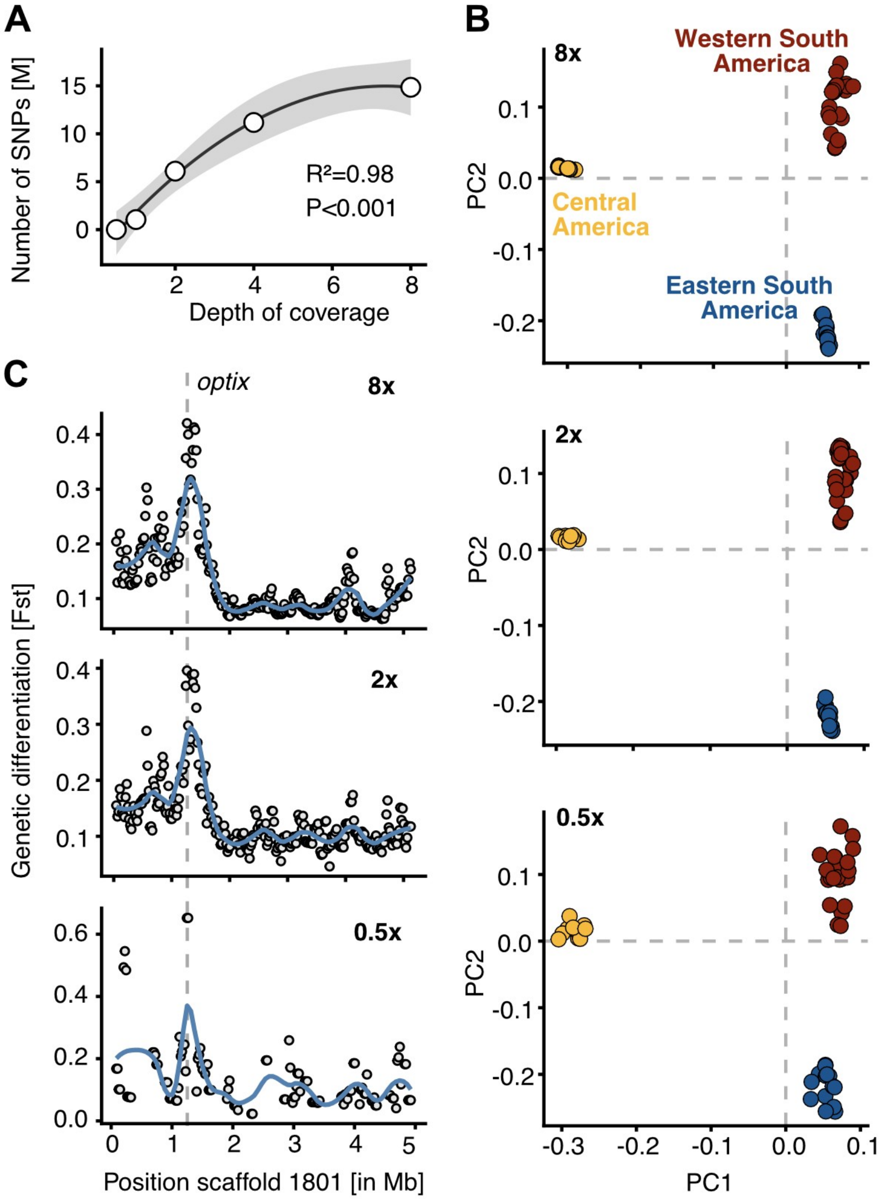

Figure 7. Application of genotype-likelihood-based inference to empirical data. **A)** Correlation between the number of identified SNPs (in millions) with variation in depth of sequencing coverage in the downsampled *Heliconius* dataset. **B)** Principal components analysis for three different coverages (8x, 2x and 0.5x) of 51 samples. Estimates of population structure are highly concordant across coverages. Subspecies are pooled and colored by their broader region of origin. **C)** Estimates of genetic differentiation (F_{ST}) between pooled *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.

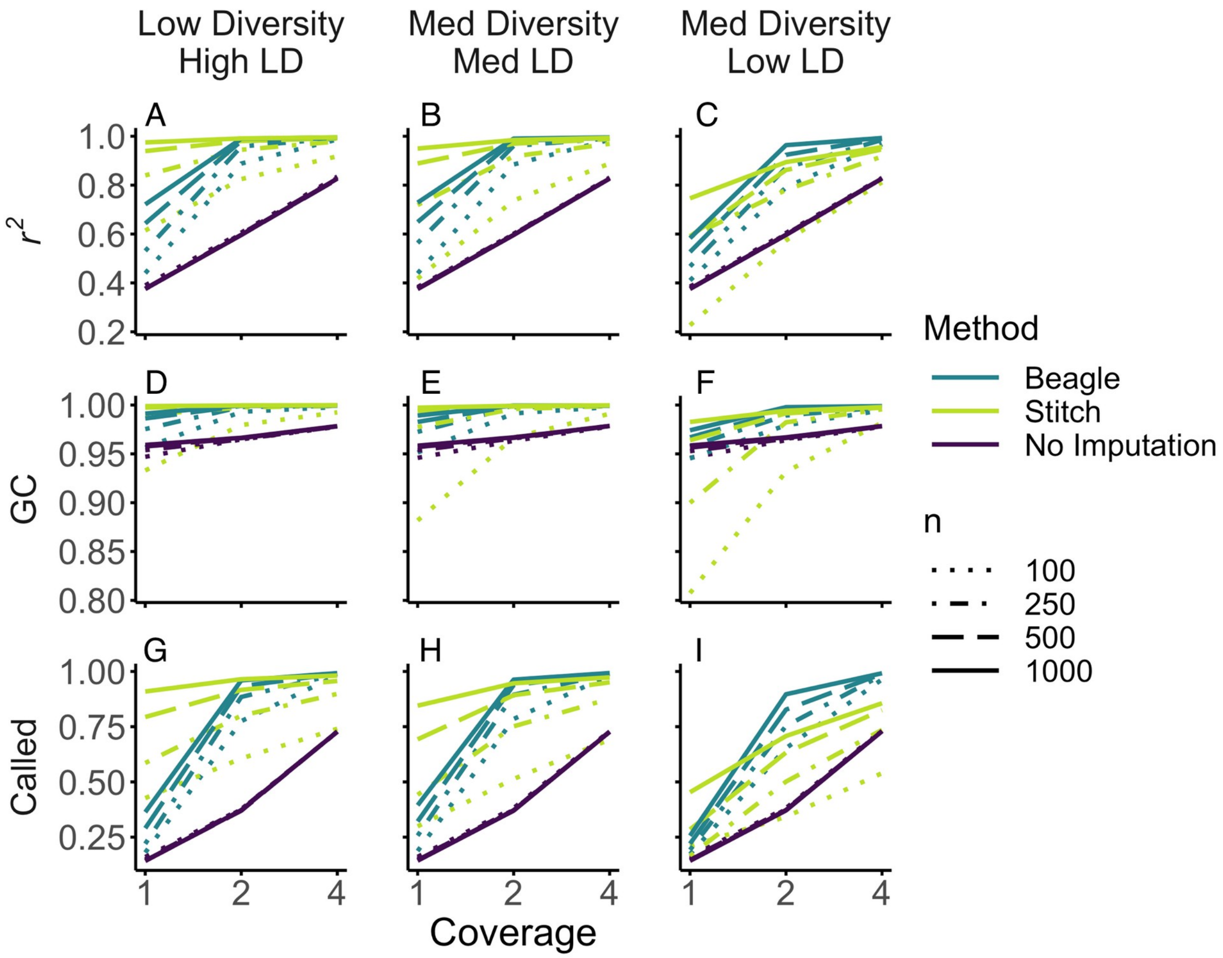

Figure 8. Genotype estimation by imputation in STITCH and Beagle compared to posterior genotypes estimated without imputation for sites with minor allele frequencies (MAF)>0.05. Combinations of sample size (n; with increasing n indicated by more contiguous lines) and sequencing coverage (x-axis) were tested for each method (line colors) under different diversity and linkage disequilibrium scenarios. A-C) Mean r^2 between true genotypes and estimated genotype dosage. D-F) Genotype concordance (GC) between true and called genotypes with posterior genotype probability>0.9. G-I) Proportion of genotypes called with posterior genotype probability>0.9.

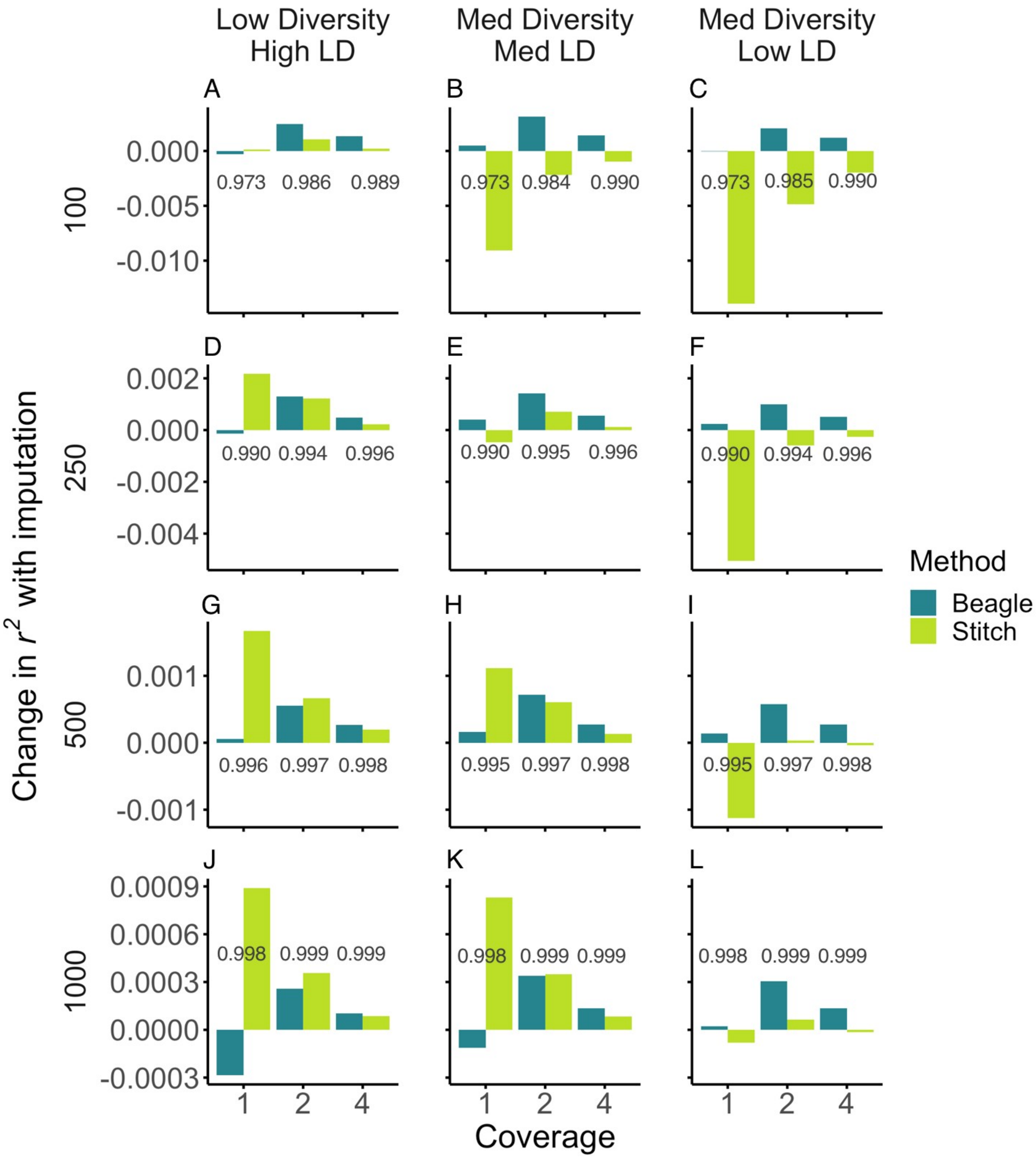

Figure 9. Change in accuracy (r^2) of minor allele frequency (MAF) estimation using imputed genotype probabilities from STITCH and Beagle, relative to non-imputed genotype likelihoods. Values above the x-axis show r^2 for MAF estimated without imputation. The three diversity/LD scenarios are arranged in columns, sample sizes (n=100, 250, 500 and 1000) are arranged in rows, and sequencing depths are shown on the x-axis. Note the different y-axis scales.