MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3028–3035 doi: 10.1111/mec.12105

Population genomics based on low coverage sequencing: how low should we go?

C. ALEX BUERKLE* and ZACHARIAH GOMPERT†

*Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY, USA, †Department of Biology, Texas State University, San Marcos, TX, USA

Abstract

Research in molecular ecology is now often based on large numbers of DNA sequence reads. Given a time and financial budget for DNA sequencing, the question arises as to how to allocate the finite number of sequence reads among three dimensions: (i) sequencing individual nucleotide positions repeatedly and achieving high confidence in the true genotype of individuals, (ii) sampling larger numbers of individuals from a population, and (iii) sampling a larger fraction of the genome. Leaving aside the question of what fraction of the genome to sample, we analyze the trade-off between repeatedly sequencing the same nucleotide position (coverage depth) and the number of individuals in the sample. We review simple Bayesian models for allele frequencies and utilize these in the analysis of how to obtain maximal information about population genetic parameters. The models indicate that sampling larger numbers of individuals, at the expense of coverage depth per nucleotide position, provides more information about population parameters. Dividing the sequencing effort maximally among individuals and obtaining approximately one read per locus and individual (1 × coverage) yields the most information about a population. Some analyses require genetic parameters for individuals, in which case Bayesian population models also support inference from lower coverage sequence data than are required for simple likelihood models. Low coverage sequencing is not only sufficient to support inference, but it is optimal to design studies to utilize low coverage because they will yield highly accurate and precise parameter estimates based on more individuals or sites in the genome.

Keywords: genotyping-by-sequencing, hierarchical Bayesian model, population genomics, resequencing

Received 25 July 2012; revision received 17 September 2012; accepted 26 September 2012

Introduction

For many applications, high-throughput DNA sequencing of whole genomes or targeted fractions of genomes is rapidly supplanting the suite of laboratory methods that were critical for the rapid growth and development of molecular ecology over the last two decades (*resequencing*: e.g., Burke *et al.* 2010; Turner *et al.* 2010; Cao *et al.* 2011; Jones *et al.* 2012; *genotyping-by-sequencing*: van Orsouw *et al.* 2007; Baird *et al.* 2008; Andolfatto *et al.* 2011; Elshire *et al.* 2011; Parchman *et al.* 2012; and *sequence capture*: Gnirke *et al.* 2009; Li *et al.* 2010; Yi *et al.* 2010; Nadeau *et al.* 2011; Rohland & Reich 2012). Whereas much of the enthusiasm about recent advances in sequencing technology has focused on how it is revolutionizing whole-genome shotgun sequencing, the effect on genotyping and resequencing is likely to be equally profound. A key feature in utilizing sequencers for studies of populations is the labeling of DNA templates from individual samples with unique oligonucleotides (often referred to as barcodes) and combining these at some appropriate level into a library for sequencing (e.g., Meyer & Kircher 2010). Thus, the large number of sequences from a single run of an instrument can be divided among many individual

Correspondence: C. Alex Buerkle, Fax: (307) 766-2851;
E-mail: buerkle@uwyo.edu

samples (e.g., 1.6×10^8 sequences in a lane of Illumina HiSeq divided by 768 individuals), each recognizable by its unique barcode, rather than squandering sequencing effort on repeatedly sequencing the same position for one or a few individuals to great and unnecessary depth. All researchers who use sequencing to obtain data for more than one sample face a trade-off among three dimensions: (i) the level of sequence coverage per site and confidence for individual nucleotide positions, (ii) the number of barcoded samples to combine for sequencing, and (iii) the size of the genome fraction that is targeted for sequencing (King *et al.* 2012). In the recent scramble to develop methods for large-scale sequence data, researchers have taken various provisional strategies, including simply sequencing deeply for a small number of individuals (e.g., Nadeau *et al.* 2011), pooling individuals into population libraries (e.g., Burke *et al.* 2010; Gompert *et al.* 2010; Turner *et al.* 2010), or using only the fraction of the data for which a certain threshold of confidence was exceeded and disregarding the rest (e.g., Hohenlohe *et al.* 2010; Manske *et al.* 2012).

Here we present a synthetic review of relevant probability models for population genomics and their application to low coverage sequence data and inference with genotype as an unknown model parameter. Our analysis is focused on inference of population genetic parameters (allele frequencies in this case) from observed sequence reads at loci, rather than on the multiple steps of data cleaning, assembly and variant detection in the same data (e.g., Li *et al.* 2009; McKenna *et al.* 2010). We use a relatively simple, and widely applicable, probability model and simulations to analyze the fundamental trade-off between numbers of individuals in a study and the depth of sequence coverage that can be achieved for a given amount of sequencing effort (we leave aside the consideration of what fraction of the genome to sequence and assume it is dictated by other considerations). Our analysis suggests that, for many population genetic analyses, sequencing to high coverage does not make optimal use of sequencing effort. Instead, researchers will maximize the information they obtain for populations by sampling larger numbers of individuals and accepting the resulting lower sequence coverage at each site. Furthermore, our review of probability models suggests that research will benefit from a clear understanding of the models for focal parameters and fitting the sequencing strategy to the planned analyses.

A family of models for allele frequencies

Classical population models for genotypes need only a minor modification to incorporate the stochastic sampling process that is inherent in the current generation of DNA sequencing instruments. We will first review basic models for population allele frequencies and put them into an explicit Bayesian framework for estimation. We recognize that, with notable exceptions (e.g., Pritchard *et al.* 2000; Hey 2010; Reich *et al.* 2012), a significant fraction of current practice in molecular ecology does not utilize explicit multilocus models for population parameters (Gaggiotti & Foll 2010), nor does it report evolutionary parameters (rather summary statistics are reported; e.g., see distinction between F_{ST} as a parameter and statistic in Balding 2003; Holsinger & Weir 2009). Instead, it is common to estimate parameters like F_{ST} for a genome by taking the numerical average of the single-locus estimates, often without any report of uncertainty in the parameter. One of the possible effects of sampling the genome at thousands of SNPs is to draw attention to proper models for variation across the genome and to question whether current practice is appropriate for a parameter like F_{ST}, which combines information across loci that vary in allele frequency and information content (cf., Jost 2008; Meirmans & Hedrick 2011; Whitlock 2011).

At the most basic level, the Hardy-Weinberg equilibrium (HWE) model is commonly used as the basis for estimation of population allele frequencies. The assumptions of the HWE model lead to the expectation that genotype probabilities are a direct function of allele frequencies in a population. This expectation is true for both likelihood and Bayesian approaches to estimation, and is commonly utilized without explicit reference to the HWE model. Thus, the HWE model gives rise to a Bayesian probability for allele frequencies in a population (p), given the observed genotypes ($\mathbf{g}$; Fig. 1A):

$$P(p|\mathbf{g}) \propto \prod_{i=1}^{N} P(\mathbf{g}|p)P(p) \qquad (1)$$

For biallelic, diploid loci, the $P(\mathbf{g}\,|\,p)$ term (somewhat confusingly, this is commonly referred to as the *likelihood* in Bayesian models) is a binomial probability with parameter p, the frequency of the reference allele (R), with two draws for every diploid individual (i). More generally, for multiallelic, diploid loci, the $P(\mathbf{g}\,|\,\mathbf{p})$ term is a multinomial probability with a parameter vector $\mathbf{p}$, the frequency of each allele in the population, with two draws for every diploid individual.

At the simplest level, the prior on population allele frequencies at a locus ($P(p)$) is a Beta (for biallelic loci) or Dirichlet (for multiallelic loci) distribution, with fixed, minimally informative parameter values, e.g., Beta ($\alpha = 1, \beta = 1$). This specification of a prior on $\mathbf{p}$ has uniform probability density for $p \in (0,1)$ and is equivalent to having a single prior observation of each allele

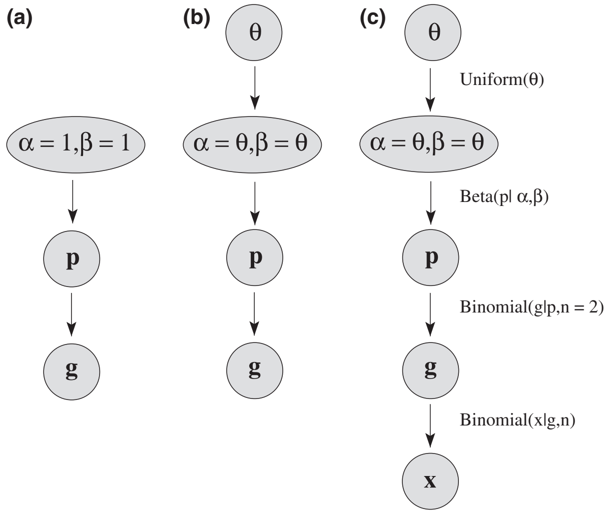

Fig. 1 Directed acyclic graphs of three Bayesian models for population allele frequencies **p**. Probability distributions for each level in the models are listed to the right. A) Bayesian model for observed genotypes (**g**), allele frequencies (**p**), and constant $\alpha = 1$ and $\beta = 1$ parameters of a Beta distribution (see Eqn 1). B) Hierarchical Bayesian model that incorporates a θ parameter as the conditional prior for population allele frequencies (see Eqn 2). C) Incorporation of observed sequence data **x** as a function of the underlying, unobserved genotype (**g**), where **g** is a parameter that is estimated and for which a posterior probability distribution is obtained (see Eqn 4).

copy in a population. The contribution of this prior to the posterior probability will be readily overwhelmed by the data, even with a modest sample of individuals.

Alternatively, we can utilize a conditional prior on allele frequencies in a population, $P(\mathbf{p} \mid \theta)$ (Fig. 1B). The parameter θ gives the expected distribution of population allele frequencies and can be estimated because we typically have data from many loci from the same genome, and these loci share aspects of their evolutionary history and the effects of evolution on their variation. This parameter quantifies genetic diversity and is an estimate of the population mutation rate ($\theta = 4N_e\mu$) under certain conditions (Wright 1931). The shared history among loci can be modeled explicitly and provides better estimates of the **p** parameters. $P(\mathbf{p} \mid \theta)$ is still a Beta (biallelic loci) or Dirichlet (multiallelic) probability distribution, but with parameters $\alpha = \theta$ and $\beta = \theta$, and θ has its own prior probability distribution.

$$P(\mathbf{p}, \theta | \mathbf{g}) \propto P(\mathbf{g}|\mathbf{p})P(\mathbf{p}|\theta)P(\theta) \tag{2}$$

There are several alternatives for specifying the prior probability $P(\theta)$, with a sensible choice being an uninformative, uniform probability over a range such as (0,10 000].

To incorporate sampling of sequence reads into this model, one need only include an additional level in the probability model for the data generating process (Fig. 1C). The sequence reads for each individual (i) and locus (j) can be modeled as independent stochastic samples from the genotype, with possible sequence errors. The appropriate probability distribution for a biallelic locus is the binomial distribution (the multinomial distribution is appropriate for a multiallelic locus).

$$P(\mathbf{x}_{ij}|\mathbf{g}_{ij}, \epsilon) = \frac{n_{ij}!}{\mathbf{x}_{ij}!(n_{ij} - \mathbf{x}_{ij})!} \begin{cases} (1-\epsilon)^{\mathbf{x}_{ij}} \epsilon^{n_{ij}-\mathbf{x}_{ij}} & \text{if } \mathbf{g}_{ij} = rr \\ 0.5^{n_{ij}} & \text{if } \mathbf{g}_{ij} = Rr \\ \epsilon^{\mathbf{x}_{ij}}(1-\epsilon)^{n_{ij}-\mathbf{x}_{ij}} & \text{if } \mathbf{g}_{ij} = RR \end{cases} \tag{3}$$

Here x_{ij} is the count of the reference allele (R) and n_{ij} is the total number of sequence reads, g_{ij} denotes the diploid genotype, and ε denotes the probability of a sequence error (ε might vary among loci, or be constant). The full Bayesian model for genotypes (**g**), population allele frequencies (**p**), and θ is:

$$P(\mathbf{g}, \mathbf{p}, \theta | \mathbf{x}, \epsilon) \propto P(\mathbf{x}|\mathbf{g}, \epsilon)P(\mathbf{g}|\mathbf{p})P(\mathbf{p}|\theta)P(\theta) \tag{4}$$

This basic framework for modeling allele frequencies can be made more elaborate by incorporating other population genetic parameters of interest, including $\mathrm{F_{ST}}$, genomic cline and admixture parameters (Pritchard *et al.* 2000; Falush *et al.* 2003; Foll & Gaggiotti 2008; Gaggiotti & Foll 2010; Gompert *et al.* 2012 a, b). Likewise, any classic population genetic model for genotypic data can be similarly extended to accommodate the stochastic process of obtaining sequence reads from an individual's genotype, simply by adding the likelihood for sequence reads given genotype.

Simulations and Bayesian estimation

We used simulations as a basis for analysis of the trade-off between numbers of sampled individuals and the depth of sequence coverage that can be achieved for a finite sequencing effort. We simulated simple sets of genotypic data for samples of 2–100 individuals (R code for simulations is in Appendix S1, Supporting information). In each case we assumed that the sequencing effort was fixed at 100 sequence reads, which were divided equally among the 2–100 individuals in the sample (2, 5, 10, 20, 25, 50 and 100 individuals), for 50–1 × coverage per individual. The simulations were for biallelic, diploid loci with a true allele frequency of 0.5, 0.1. or 0.01, and the genotypes of the finite sample were stochastic samples from their expected frequencies based on allele frequencies (i.e., according to the HWE model). Sequence reads were sampled stochastically according to a binomial distribution with p = 0,0.5,1, depending on the genotype of an individual, and n = 50–1 (50, 20, 10, 5, 4, 2, and 1) sequence reads. We simulated and analyzed 100 replicate data sets for each combination of number of individuals and read depth, and true allele frequency (100 × 7 × 3). Additionally, we investigated the effect of stochastic variation in coverage among barcoded individuals in a sequencing library by replicating the above simulations, with two modifications: coverage for an individual was a random draw from a Poisson distribution with an expected number of reads of n = 50–0.5 (rather than the previously fixed, equal coverage for all individuals, and the previous minimal coverage of 1×). With a Poisson distribution, at n = 1× coverage 36.8% of the 100 individuals in a simulation should have no reads (at n = 0.5× 60.6% of the 200 individuals will have zero reads). Finally, we performed a set of simulations for p = 0.5 (25 replicates for each coverage) with the modification that all sequence reads were pooled and individual identities not tracked, to generate data of the type that would be utilized if pools of individuals were barcoded rather than individuals.

For analysis of the simulations, we used a basic model that included genotype uncertainty, but that utilized a simple, uninformative Beta ($\alpha = 1, \beta = 1$) prior for allele frequencies: $P(\mathbf{g},\mathbf{p}\mid\mathbf{x}) \propto P(\mathbf{x}\mid\mathbf{g})P(\mathbf{g}\mid\mathbf{p})P(\mathbf{p})$. We specified this probability model in a JAGS model for MCMC estimation of p (Appendix S2, Supporting information) and utilized the rjags interface from R (R Development Core Team 2012) to JAGS (version 3.2.0). For each analysis we report the mean and 95% credible interval (equal tail probability interval from 2.5% to 97.5% quantiles) for the estimated posterior probability distribution for p.

Results

Estimates of population allele frequency were unbiased if the true allele frequency (p) was 0.5, but were biased for small samples of individuals (2–10) and true p of 0.1 or 0.01 (Fig. 2). Bias was evident in that many of the 100 replicate analyses led to point estimates of p that were greater than the true value (for true p = 0.1, at N = 2 all $\hat{p}_{50\times} > 0.1$ and $\bar{p}_{50\times} = 0.23$, and at N = 10 63% of replicates had $\hat{p}_{10\times} > 0.1$ and $\bar{p}_{10\times} = 0.14$; for true p = 0.01, at N = 50 all $\hat{p}_{2\times} > , 0.01$ and $\bar{p}_{2\times} = 0.026$, and at N = 100, 78% of replicates had $\hat{p}_{1\times} > 0.01$ and $\bar{p}_{1\times} = 0.019$). Across replicates, point estimates of p were more widely variable for small samples of individuals. The width of credible intervals for p declined with increasingly large samples of individuals, with diminishing returns beyond 20 individuals (Fig. 2). Overall, the Bayesian credible intervals included the true p values at the expected frequency of 95% of simulation

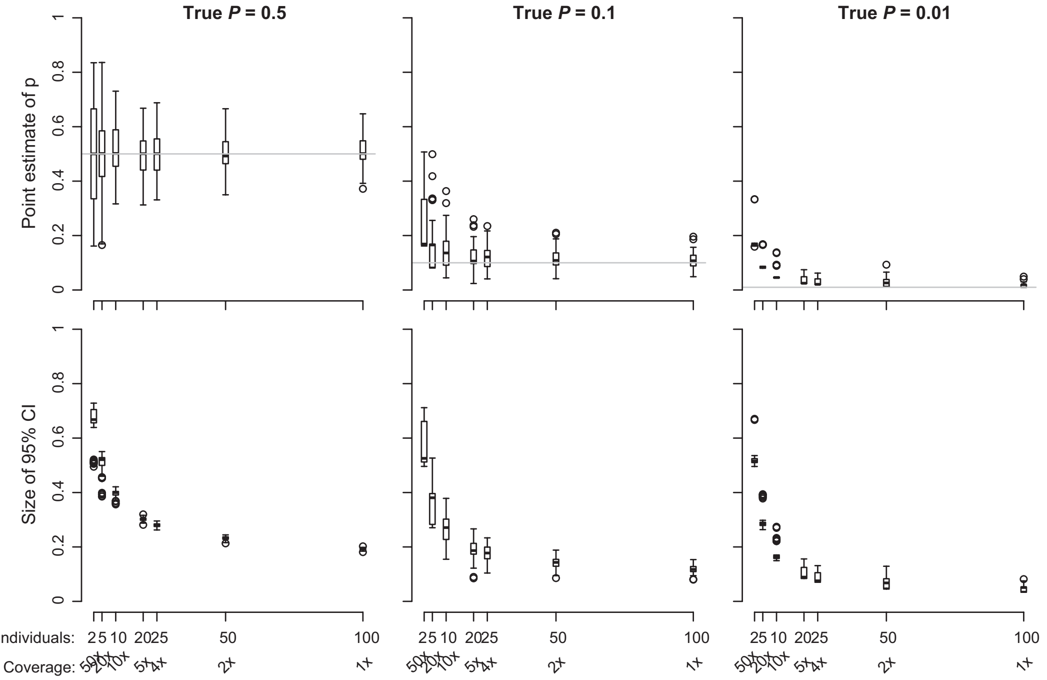

Fig. 2 Distribution of point estimates (mean of the posterior distribution; top row of plots) and size of 95% credible intervals (equal tail probability interval from 2.5% to 97.5% quantiles; bottom row of plots) for population allele frequency in simulations with a finite amount of sequencing (100 reads). Boxplots indicate the median and the interquartile range (IQR), and circles mark any observations beyond 1.5× the IQR. In the simulations, sequencing reads were divided evenly among 2–100 individuals, for 50–1× coverage, with 100 replicate simulations of each combination. Simulations were for a locus with a true allele frequency of 0.5, 0.1 and 0.01.

replicates. For analyses of simulations in which individuals were pooled into a sequencing library, point estimates were similarly more variable for small samples of individuals (Fig. 3). For the pooled samples, the size of credible intervals for p declined with increasing sampling effort, with diminishing returns beyond 20 individuals.

Finally, simulations that included stochastic variation around the expected sequence coverage led to very similar estimates of p as those obtained in the above simulations that utilized fixed, equal coverage among all individuals. At $n = 1\times$ coverage ($N = 100$ individuals), where stochastic sampling of reads leads to 36.8% of individuals with no read data (13.5% at 2×, 0.7% at 5×, and e^{-n} generally), the difference in average point estimates of p across the 100 simulation replicates was 0.0032 and the range of point estimates was very similar for both sets of simulations ($\hat{p}_{1\times}$ for stochastic n: 0.06–0.22, $\hat{p}_{1\times}$ for fixed n: 0.05–0.20). Finally, reducing stochastic coverage further, from 1× ($N = 100$ individuals) to 0.5× coverage ($N = 200$ individuals) resulted in no appreciable change in $\hat{p}$ ($\bar{p}_{1\times} = 0.104$ and $\bar{p}_{0.5\times} = 0.107$).

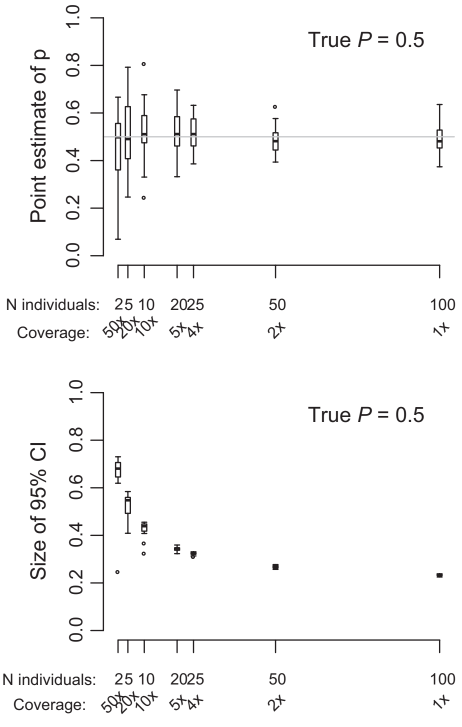

Fig. 3 Distribution of point estimates (mean of the posterior distribution) and size of 95% credible intervals (equal tail probability interval from 2.5% to 97.5% quantiles) for population allele frequency in simulations with a finite amount of sequencing (100 reads) from individuals combined into a pooled library for sequencing (i.e., no individual barcoding). Boxplots indicate the median and the interquartile range (IQR), and circles mark any observations beyond 1.5× the IQR. In the simulations, sequenced pools of individuals contained 2–100 individuals, for an expected 50–1× coverage per individual, with 25 replicate simulations of each combination. Simulations were for a locus with a true allele frequency of 0.5.

Discussion

Many population genetic analyses utilize allele frequencies or other population parameters, including their comparison between samples from different natural populations, experimental treatments or phenotypic categories (e.g., Burke *et al.* 2010; Simonson *et al.* 2010; Turner *et al.* 2010; Yi *et al.* 2010; Jones *et al.* 2012). Given a finite amount of obtainable sequence data, research on populations will be based on the most information if the number of individuals and loci sampled results in approximately one sequence read per locus and individual (i.e., 1× coverage). If twenty individuals from a population were available for study, an optimal sequencing strategy would divide the total number of filtered, assembled sequences across the 20 individuals and whatever number of loci would yield an expectation of 1× coverage per locus and individual. Sequencing to higher coverage would not be beneficial for inference, and sequence reads beyond 1× coverage would be better allocated to other loci or individuals in other populations. This conclusion is supported whether one assumes fixed, equal sequence coverage for all individuals or allows coverage to vary according to a Poisson probability distribution with the same expected number of reads. Beyond this, at lower than 1× coverage and increased numbers of individuals from a population, the potential for increased precision and accuracy of estimates is expected be offset by substantial missing data for individuals. However, the analyses presented here demonstrate that even 0.5× coverage can support inferences with confidence, which is consistent with previous empirical studies (e.g., Gompert *et al.* 2010; Parchman *et al.* 2012; Gompert *et al.* 2012a). This is not a matter of making the best of a bad situation (low coverage), but instead, a larger sample of individuals will lead to greater accuracy and precision for population parameters. This is because each sequence read from a new individual will provide more information about a population parameter (e.g., allele frequency) than any additional reads from individuals that are already represented in the collection of sequences reads.

Researchers might often have the mistaken impression that genotypes must be known with very high confidence, as software for population genetics often requires as input the genotypes of individuals, only to collapse these data into allele counts for populations for certain analyses (based on the HWE model). It is true that most software and models treat genotypes as known and would need a relatively simple modification to account for genotype uncertainty (as in Eqn 3), but there is no inherent requirement that genotypes be known with complete certainty. Likewise, in many cases researchers seek to estimate parameters for individuals (e.g., admixture coefficients, parentage probabilities) and can combine information across loci and thereby could compensate for low coverage at individual loci. Some genetic analyses, such as mapping of continuous phenotypes to individual SNPs, can require genotypic data for each sampled locus in individuals. However, this apparent requirement can be circumvented by methods for directly incorporating uncertainty about genotype (e.g., utilizing full or point estimates of genotype probabilities; Guan & Stephens 2011), or through imputation of unobserved states (Ober *et al.* 2012; Pasaniuc *et al.* 2012). Clearly, if high confidence in genotypes of individuals is required for analysis, higher coverage will be necessary than for estimating parameters for whole individuals or populations. However, given the large-scale differences in precision and accuracy among sampling strategies, it will be beneficial to understand when high coverage data for individual genotypes are required. For example, population analyses of recombination can benefit more from higher marker density and more individuals than from greater sequence coverage per locus (King *et al.* 2012).

The level of desired precision and accuracy of parameter estimates will be determined by the biological application of sequencing. For example, the standards for medical or forensic applications are likely to differ from those for molecular ecology. In any application of sequencing, explicit probability models that incorporate genotype uncertainty can be used to provide guidelines. The models can allow the use of all of the valid data, or can serve as a sound basis for imposing thresholds for data inclusion. More sophisticated models could incorporate information from populations or across loci in the genome to specify prior probabilities for genotypes and would improve accuracy and precision of estimates relative to the simple model used here (e.g., Gompert & Buerkle 2011; Gompert *et al.* 2012a; Parchman *et al.* 2012). For example, allele frequencies in different populations will be correlated due to their common ancestry and can be modeled as draws from a common distribution with a variance that reflects F_{ST} between populations and their shared ancestral population (Falush *et al.* 2003; Foll & Gaggiotti 2008; Gompert *et al.* 2012a). Likewise expectations of recombination in experimental crosses could lead to specifications for prior probabilities for genotype through Hidden Markov models for ancestry blocks along a linkage group (Falush *et al.* 2003). With dense genotyping, it can also be informative to utilize information from linked loci in specifying a prior for a genotype, without reference to a specific model of recombination (e.g., an intrinsic conditional autoregressive prior; Gompert *et al.* 2012b). Given that the simple model analyzed here supports confidence in population parameters at 1× coverage, more elaborate, hierarchical models that share the same likelihood function are expected to also support optimal inference around 1× (e.g., Gompert & Buerkle 2011; Gompert *et al.* 2012a; Parchman *et al.* 2012).

Even the simplest population models, such as those above, are preferable to entirely naive models that treat individuals and loci in isolation and only model single-locus genotypes of the individual, given the sequence reads at a locus. For example, utilizing a binomial probability for an individual's genotype at a locus, one would need six or more reads of the same *R* allele before one could conclude with >95% confidence that an individual is homozygous *RR* (at 5 identical reads, there is still a 6.25% chance that the individual is a heterozygote). Using this model implicitly assumes that all population allele frequencies are equally likely and that no information about allele frequency is available. In contrast, if the population allele frequency were known or inferred to be 0.9, then we can utilize a closed-form Bayesian probability model for the genotype (disregarding error for simplicity): $P(g|x, p = 0.9) = \frac{P(x|g)P(g|p=0.9,n=2)}{P(x)}$, where $P(x \mid g)$ and $P(g \mid p{=}0.9{,}n{=}2)$ are binomials, as above, and the probability of the data ($P(x)$) is the sum of the probabilities in the numerator for each of the three possible genotypes. With this Bayesian model, the prior probability of the common *RR* homozygote is 0.81 and of the heterozygote is 0.18. After observing a single read of the *R* allele, the posterior probability of the *RR* genotype would be 0.9 and of the *Rr* heterozygote 0.1. With two reads of the *R* allele, the posterior probability of the *RR* genotype would be 0.947 and of the *Rr* heterozygote 0.053. If individuals are sampled from populations that can be characterized, as they typically will be, then there are clear benefits of including information about the source population. Given that populations will be known imperfectly, the slightly more complex models above (e.g., Eqn 4) can be used to reflect uncertainty and share information about population allele frequencies. These Bayesian models apply even if the objective is to use a threshold probability to assign genotypes to individuals (e.g., <0.05 probability of assigning an incorrect genotype) before proceeding with further analysis, and are a substantial improvement relative to the binomial probability alone (in this example with p = 0.9, 3 reads rather than 6 reads to reach the <0.05 probability threshold). As noted above, rather than imposing thresholds to obtain point estimates of genotypes, inferences can be based on simple, fully Bayesian models that incorporate genotype uncertainty.

Our analysis of a Bayesian population model suggests that sampling more individuals with even lower coverage makes optimal use of sampling and sequencing effort (an average of a single read per locus and individual in a population model; Fig. 2).

Combining sequencing libraries prepared from individuals into distinguishable pools (and losing individual information) can be a cost-saving measure for some library protocols (e.g., Burke *et al.* 2010; Gompert *et al.* 2010; Turner *et al.* 2010). If inferences are to be made on populations (e.g., allele frequencies and derived statistics or parameters), then little information is lost by labeling all individuals in a pool with the same oligonucleotide barcode (Fig. 3; Futschik & Schlötterer 2010; Gompert & Buerkle 2011). However, in many applications in molecular ecology we are interested in recovering information from allelic states in individuals (e.g., estimating admixture coefficients or locus-specific ancestries for individuals, or linkage disequilibria among loci). Consequently, in these cases pooling will be unnecessary or undesirable, particularly because for many protocols the same laboratory steps would be used to label samples with individually identifying barcodes as would be used to label them with population barcodes. It is simply that a larger number of oligonucleotides would be needed for barcoding individuals. These barcodes are easy to specify and synthesize in large numbers (Meyer & Kircher 2010), they are relatively inexpensive and reusable across projects, and can even be used in resequencing (Meyer & Kircher 2010) and targeted enrichment protocols (Rohland & Reich 2012).

The rapid advances in sequencing technology and the ability to multiplex many samples within a single run of a high-throughput sequencer have initiated a new era in molecular ecology. The increased capacity to sample genome-wide polymorphisms across individuals and populations calls for analyses that are based on explicit models for the dependence and shared information among loci and individuals. Many challenges lie ahead, including how to properly incorporate information from neighboring SNPs, across a genome with heterogeneous recombination rates, while avoiding logical pitfalls that arise from methods like sliding windows (in which a locus's parameter will be contingent on which of several windows presently contains the locus). Even with continued advances in sequencing technology and increased adoption of whole genome resequencing, we can expect for some time to be operating within a realm of finite and variable coverage across the genome, and with trade-offs among sampling individuals, sites in the genome and coverage depth. The continued development and application of probability models for inference will allow us to make good use of the available data.

Acknowledgements

This manuscript was improved by comments from E. Mandeville, C. Nice, T. Parchman, P. Hohenlohe and three anonymous reviewers. This research was facilitated by an NSF grant to CAB (DEB-1050149).

References

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, **63**, 221–230.

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010) Genome wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, **467**, 587–590.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956–963.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

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**, 977–993.

Futschik A, Schlötterer C (2010) Massively parallel sequencing of pooled DNA samples–the next generation of molecular markers. *Genetics*, **186**, 207–218.

Gaggiotti OE, Foll M (2010) Quantifying population structure using the F-model. *Molecular Ecology Resources*, **10**, 821–830.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation population genomics. *Genetics*, **187**, 903–917.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson R, Buerkle CA (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, **19**, 2455–2473.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, A BC (2012a) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Gompert Z, Parchman TL, Buerkle CA (2012b) Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 439–450.

Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, **5**, 1780–1815.

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Holsinger KE, Weir BS (2009) Fundamental concepts in genetics: genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

King EG, Macdonald SJ, Long AD (2012) Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. *Genetics*, **191**, 935–949.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li Y, Vinckenbosch N, Tian G *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency nonsynonymous coding variants. *Nature Genetics*, **42**, 969–972.

Manske M, Miotto O, Campino S *et al.* (2012) Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, **487**, 375–379.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meirmans PG, Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*, **11**, 5–18.

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, pdb.prot5448.

Nadeau NJ, Whibley A, Jones RT *et al.* (2011) Evidence for genomic islands of divergence among hybridizing *Heliconius* butterflies obtained by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 343–353.

Ober U, Ayroles JF, Stone EA *et al.* (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genetics* , **8**, e1002685.

van Orsouw NJ, Hogers RCJ, Janssen A *et al.* (2007) Complexity reduction of polymor phic sequences (CRoPSTM): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE*, **2**, e1172.

Parchman TL, Gompert Z, Mudge J, Schilkey F, Benkman CW, Buerkle CA (2012) Genome wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Pasaniuc B, Rohland N, McLaren PJ *et al.* (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, **44**, 631–635.

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

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reich D, Patterson N, Campbell D *et al.* (2012) Reconstructing Native American population history. *Nature*, **488**, 370-374.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Simonson TS, Yang Y, Huff CD *et al.* (2010) Genetic evidence for high-altitude adaptation in Tibet. *Science*, **329**, 72–75.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260–263.

Whitlock MC (2011) G′ ST and D do not replace F_{ST}. *Molecular Ecology*, **20**, 1083–1091.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 0097–0159.

Yi X, Liang Y, Huerta-Sanchez E *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.

Both authors contributed to the conceptual development, to the simulation and Bayesian modeling and to the writing of this publication..

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 R code for simulating sequence data for individuals and pooled samples. JAGS is used for each data set to obtain samples from the posterior distribution of population allele frequency.

Appendix S2 JAGS models for individual and population pool data.