Theoretical Population Biology 117 (2017) 51–63

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Variance in estimated pairwise genetic distance under high versus low coverage sequencing: The contribution of linkage disequilibrium

Max Shpak [a,b,c,*], Yang Ni [d], Jie Lu [e], Peter Müller [d,f]

[a] Sarah Cannon Research Institute, Nashville TN 37203, USA
[b] Center for Systems and Synthetic Biology, University of Texas, Austin TX 78712, USA
[c] Fresh Pond Research Institute, Cambridge MA 02140, USA
[d] Department of Statistics and Data Science, University of Texas, Austin TX 78712, USA
[e] Genetics Division, Fisher Scientific, Austin TX 78744, USA
[f] Department of Mathematics, University of Texas, Austin TX 78712, USA

ARTICLE INFO

Article history:
Received 23 February 2017
Available online 24 August 2017

Keywords:
Genetic distance
Linkage disequilibrium
Coverage
Cancer genomics
Pooled sampling
Next-generation sequencing

ABSTRACT

The mean pairwise genetic distance among haplotypes is an estimator of the population mutation rate θ and a standard measure of variation in a population. With the advent of next-generation sequencing (NGS) methods, this and other population parameters can be estimated under different modes of sampling. One approach is to sequence individual genomes with high coverage, and to calculate genetic distance over all sample pairs. The second approach, typically used for microbial samples or for tumor cells, is sequencing a large number of pooled genomes with very low individual coverage. With low coverage, pairwise genetic distances are calculated across independently sampled sites rather than across individual genomes. In this study, we show that the variance in genetic distance estimates is reduced with low coverage sampling if the mean pairwise linkage disequilibrium weighted by allele frequencies is positive. Practically, this means that if on average the most frequent alleles over pairs of loci are in positive linkage disequilibrium, low coverage sequencing results in improved estimates of θ, assuming similar per-site read depths. We show that this result holds under the expected distribution of allele frequencies and linkage disequilibria for an infinite sites model at mutation–drift equilibrium. From simulations, we find that the conditions for reduced variance only fail to hold in cases where variant alleles are few and at very low frequency. These results are applied to haplotype frequencies from a lung cancer tumor to compute the weighted linkage disequilibria and the expected error in estimated genetic distance using high versus low coverage.

1. Introduction

One of the defining empirical problems in evolutionary genetics is the measurement and characterization of genetic heterogeneity in natural and experimental populations. The advent of next-generation sequencing (NGS) provides researchers with a tool set for efficiently generating sequence data from large numbers of genotypes and over extensive regions of the genome, including whole-exome and whole-genome sequencing of multiple individuals. This data has the potential to provide the statistical power necessary to make robust inferences of genotype frequencies and their distributions.

High-throughput NGS technology gives researchers choices between different approaches to sampling genotypes from a population. A standard method, most widely used in studies of multicellular organisms, is to sample individuals and sequence their genomes at high coverage, i.e. generating reads containing most or all of the polymorphic sites of interest for each genome. An alternative approach is to sequence from a pooled set of individuals at a read depth much smaller than the number of genomes in the sample, e.g. Futschik and Schlötterer (2010) and Anand et al. (2016), leading to a very low average coverage per individual genome. Fig. 1 illustrates these two scenarios for a small model population: a sample of n individuals sequenced with full coverage, versus low coverage sequencing at read depth n from a pooled set of individuals.

Sequencing at low coverage is typically used in population genetic studies of microbial assemblages and in cancer genomic studies where genetically heterogeneous assemblages of cancer cells are sampled from a single tumor. However, through single-celled sequencing techniques (Navin, 2015; Gawad et al., 2016), individual sampling with high coverage is also possible for these model systems. Similarly, while individual sampling has been standard in population genetic studies of most multicellular organisms,

* Correspondence to: St. David's Medical Center, 1015 E. 32nd St, Suite 414, Austin TX 78705, USA.
E-mail address: mshpak@austin.utexas.edu (M. Shpak).

http://dx.doi.org/10.1016/j.tpb.2017.08.001
0040-5809/

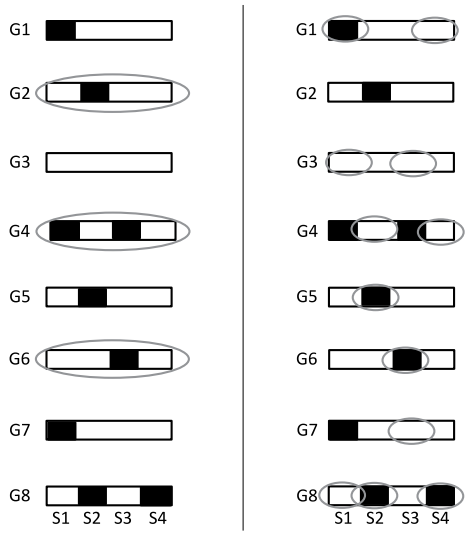

Fig. 1. Illustration of high coverage sequencing (HCS) versus low coverage sequencing (LCS). In this example, the population consists of eight haplotypes G1…G8 characterized by four segregating sites $S1 \ldots S4$. We assume a sampling depth of $n = 3$ and sufficiently many reads to capture all segregating sites. In the left panel, we have a random instance of HCS via the complete sequencing of $G2$, $G4$, $G5$ (gray ovals representing sampling), giving a mean pairwise distance of $\widehat{\pi} = 2$. In the right panel, we have a random instance of LCS, such that G1, G3, G8 are sequenced at S1, G4, G5 and G8 at S2, etc., giving a mean genetic distance $\widehat{\pi} = 8/3$. Note that $E(\widehat{\pi})$ is the same under both modes of sampling, the differences are due to $var(\widehat{\pi})$.

NGS has made pooled sampling with low coverage sequencing inexpensive and practical in studies of animal and plant populations. For example, several recent analyses of genetic variation in *Drosophila* populations (Schlötterer et al., 2014) used low coverage pooled sequencing, drawing reads from a very large pool of macerated flies rather than sequencing fly genomes individually with high coverage.

Sequencing n individuals with full coverage is not statistically equivalent to sequencing at read depth n from a large pool of individuals. High and low coverage results in different estimation errors for population parameters. These include the population mutation rate $\theta = 4Nu$ (where N is the population size and u the genomic mutation rate, with $\theta = 2Nu$ for haploid genomes), which is estimated either from the number of segregating sites (Watterson, 1975) or from the average heterozygosity across sites (Tajima, 1989). Estimates of θ are the basis for a number of statistical tests that distinguish the effects of natural selection and population dynamics from neutral evolution at constant population size. These include the Tajima's D test (Tajima, 1989), which compares θ estimates from the number of segregating sites to those derived from average heterozygosity. Consequently, getting a handle on the variance in estimates of θ and for neutrality test statistics generally is of broad interest and importance in evolutionary genetics (Nielsen, 2001). Several studies have analyzed the contributions of pooling, read depth, and coverage to bias and variance in θ estimates, e.g. Pluzhnikov and Donnelly (1996) and Lynch (2008). For example, given a constant read depth, pooling improves the accuracy in estimated θ due to effectively larger sample size (Futschik and Schlötterer, 2010; Ferretti et al., 2013), while Korneliussen et al. (2013) have shown that low read depth can lead to estimation bias in the Tajima D test statistics.

Considering the effects of coverage on parameter estimation, if the number of genomes sampled is held constant, lower coverage leads to smaller sample size, and consequently greater error. However, Ferretti et al. (2014) have shown that as long as the reduction in coverage is compensated by the number of genomes represented in a sample, low coverage sequencing reduces the error in estimates of θ and the Tajima D statistic. Specifically, if we estimate allele frequencies and θ from n sequences with complete coverage, as opposed to a much larger number of sequences at very low per-genome coverage (so that on average each site is represented by n samples, often from different individuals per site, as shown in panel 2 of Fig. 1), low coverage sequencing reduces the error in estimated θ. Ferretti et al.'s results are explained by the fact that with low coverage sequencing, variant alleles from different segregating sites tend to be sampled from different individuals, corresponding to an effective increase in the number of independent genealogies from which variant allele samples are drawn for each locus. Consequently, their results imply that the degrees to which estimates of θ are expected to improve with low-coverage, large sample sequencing are expected to increase with the strength and direction of linkage disequlibria among polymorphic sites.

In this study, we will consider limiting cases of high and low coverage sequencing to investigate the contribution of linkage disequilibria to estimates of θ. High coverage sequencing (HCS) is represented by complete coverage of all polymorphic sites from n different genomes, as would typically be the case for individual sampling (including single-cell sequencing). Low coverage sequencing (LCS) is represented by a case where a very large sample of genomes is pooled and sequenced at a read depth n for each site, so that allelic variants at different sites are almost always drawn from different genomes. We will compute variances in the Tajima estimator $E(\widehat{\pi}) = 4Nu = \theta_\pi$, which is calculated from the mean pairwise genetic distance in a sample of n genotypes:

$$\widehat{\pi} = \sum_{i,j} \widehat{\pi}_{ij} / \binom{n}{2} \tag{1}$$

(where $\widehat{\pi}_{i,j}$ is the Hamming distance for the haplotype pair i, j summed over all polymorphic sites).

We hypothesize that under most conditions, the variance in $\widehat{\pi}$ estimated using HCS increases with greater linkage disequilibrium across polymorphic sites, i.e. strong linkage disequilibria inflate the estimation error across sites in a haplotype by reducing the number of independent genealogical sample paths. We will investigate this hypothesis analytically, and will additionally validate our results using individual-based simulations. Finally we will also apply these results to NGS data by analyzing allele and haplotype frequencies from cancer cell genomes.

2. The sampling models

Consider a population of N organisms with mutations distributed over S segregating sites. We wish to estimate the mean genetic distance $\widehat{\pi}$ for the population and its sample variance $var(\widehat{\pi})$ under the high and low coverage modes of sequencing. For HCS, we draw $n \ll N$ individual organisms (or cells) from the population and sequence their entire genomes, exomes, or any regions containing the polymorphic sites of interest.

For an idealized model of LCS, we assume a mean coverage depth $n \ll M$, where M is the number of genotypes contributing to the pooled sample (M may be $\ll N$ or of the same order). If reads are short, the majority will contain at most a single polymorphic site. Together, these conditions lead to each polymorphic site being sampled independently of other polymorphic sites with respect to the genome of origin (note that in the second panel of Fig. 1, multiple sites are sampled from the same genome simply because there are very few genomes to draw this random sample from). When computing sample genetic distance, extreme HCS sums over the Hamming distances of all pairs of sampled haplotypes, while extreme LCS results in summing over all pairs for each segregating site sampled from a different genome.