Variance in Estimated Pairwise Genetic Distance Under High versus Low Coverage Sequencing: the Contribution of Linkage Disequilibrium

Max Shpak[1,2,3], Yang Ni[4], Jie Lu[5] and Peter Müller[4,6]

[1] Sarah Cannon Research Institute, Nashville TN 37203, USA

[2] Center for Systems and Synthetic Biology, University of Texas, Austin TX 78712, USA

[3] Fresh Pond Research Institute, Cambridge MA 02140, USA

[4] Department of Statistics and Data Sciences, University of Texas, Austin TX 78712, USA

[5] Genetics Division, Fisher Scientific, Austin TX 78744, USA

[6] Department of Mathematics, University of Texas, Austin TX 78712, USA

May 30, 2017

Contents

Running Title : Sample variance of genetic distance

keywords : genetic distance, linkage disequilibrium, coverage, cancer genomics, pooled sampling, next-generation sequencing

corresponding author : Max Shpak, St. David's Medical Center, 1015 E. 32nd St, Suite 414, Austin TX 78705. Ph: 512-544-8077, Email: shpak.max@gmail.com

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); Anand et al. (2016), leading to a very low average coverage per individual genome. Figure 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.

FIGURE 1 HERE:

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, 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 result in different estimation error 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); 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 degree to which estimates of θ improve with low coverage, large individual sample sequencing are expected to increase with the strength and direction of linkage disequilibria among polymorphic sites.

In this study, we will consider limiting cases of high and low coverage sequencing to investigate the contribution of linkage disequilbria 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:

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

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. 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 Figure 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 haplotype pairs, while extreme LCS results in summing over all pairs for each segregating site sampled from a different genome.

We assume an infinite sites model (Kimura, 1969; Tajima, 1996) so that there are only two alleles per segregating site. This allows an unambiguous binary classification of alleles, with mutations as ancestral "wildtype" vs. "reference" genotype. This also allows us to specify the direction (sign) of linkage disequilibria. For cancer cells, the reference corresponds to the normal germline genotype, with somatic mutations defining the variant genotypes of the clonal lineages. Our methods and results also apply to multiallelic states provided that some allele, usually the most common, is designated as a reference and all other alleles are aggregated to create a biallelic state.

Definitions. Throughout the paper, we use the following definitions and terminology as a formal way of defining Eqn. (1) for high and low coverage sequencing:

Variables: Let z denote a genotype, at either a single locus s or across multiple loci. We define the frequency distribution of z over samples i as

$z_i \sim p(z)$, which are iid among $i = 1...n$. We use z_{is} to denote site s in haplotype i (for HCS). We write $z_{i_s,s}$ to denote sample i at site s when sites are sampled by pooling at low coverage (LCS) when we wish to highlight the fact that z_{is} and z_{ir} are read from distinct haplotypes.

Pairs: For both HCS and LCS, the estimators of $\widehat{\pi}$ include an average $\sum_{i<j} \phi_{ij}/n(n-1)$ of some function $\phi_{ij} = \phi(x_i, x_j)$ of pairs of i.i.d. random variables x_i, $i = 1, \ldots, n$. In the case of LCS $x_i = z_{is}$ and $\phi(x_i, x_j) = f_{ijs}$ with $f_{ijs} = I(z_{is} \neq z_{js})$(and an additional sum appears over s, outside the average). For HCS the random variables are $x_i = \boldsymbol{z}_i$ and $\phi(x_i, x_j) = g_{ij} = \sum_s f_{ijs}$. Importantly, while the x_i are independent, pairs (x_i, x_j) and (x_i, x_k) that share a common element are not, and thus the same for ϕ_{ij}.

Moments of ϕ_{ij}: We define $E(\phi_{ij}) = \mu$, $var(\phi_{ij}) = \sigma^2$. We also define an expectation for the product $E(\phi_{ij}, \phi_{jk}) = \kappa$ for pairs of pairs with a shared element.

Pairs of pairs: Let P denote the set of all ordered pairs of pairs, with $P_3 \subset P$ defining the subset of ordered pairs of pairs with a single shared element,

$$\begin{aligned} P &= \{[(i,j),(k,\ell)] : \ i<j, k<\ell \text{ and } (i,j)<(k,\ell)\} \\ P_3 &= \{[(i,j),(k,\ell)] : \ i<j, k<\ell \text{ and } (i,j)<(k,\ell) \text{ and } |\{i,j,k,\ell\}| = 3\} \end{aligned}$$

Numbers of pairs: The number of ordered pairs, and the number of ordered pairs of pairs with a shared element are, respectively $N_2 = n(n-1)/2$ and $N_3 = n(n-1)(n-2)/2$. The value of N_3 follows from there being $n(n-1)(n-2)/6$ ways to select a triplet i, j, k, and three ways to select a shared element from this triplet. In Appendix A1, we discuss the properties of ordered pairs of pairs, including the derivation of the following relation which we will use below to compute $var(\widehat{\pi})$ under low and high coverage sequencing,

$$var(\widehat{\phi}_n) = \frac{\sigma^2}{N_2} + 2\frac{N_3}{N_2^2}(\kappa - \mu^2). \tag{2}$$

where $\widehat{\phi}_n = \frac{1}{N_2}\sum_{i<j}\phi_{ij}$ is a sample estimate of $E(\phi_{ij}) = \mu$. We will use this result twice, once for LCS with $\phi_{ij} = f_{ijs}$, and once for HCS with $\phi_{ij} = g_{ij}$.

2.1 Case 1: Low Coverage Sequencing (LCS)

For LCS, we use the indicator function at a single site s, $f_{ij,s} = I(z_{i_s,s} \neq z_{j_s,s})$, where $z_{i_s,s} \sim Bern(p_s)$, i.e. $p(z_{i_s,s}) = p_s^{z_{i_s,s}}(1-p_s)^{1-z_{i_s,s}}$ for $z_{i_s,s} \in \{0,1\}$ such that

$$\mu_s = E(f_{ij,s}) = h_s = 2p_s(1-p_s)$$
$$\sigma_s^2 = var(f_{ij,s}) = h_s(1-h_s),$$

where h_s is also known as the heterozygosity at locus s. Under LCS we define the estimator for $\hat{\pi}$ for Eqn. (1) as:

$$\widehat{\pi}_{LCS} = \sum_s \left\{ \frac{1}{N_2}\sum_{i<j} I(z_{i_s,s} \neq z_{j_s,s}) \right\} = \sum_s \left\{ \frac{1}{N_2}\sum_{i<j} f_{ij,s} \right\}.$$

To evaluate the variance of $\hat{\pi}_{LCS}$, first note that the expectation of products $f_{ij,s}, f_{jk,s}$ for ordered pairs is:

$$\begin{aligned}
\kappa_s &= E(f_{ij,s}\, f_{jk,s}) = p(z_{i_s,s} \neq z_{j_s,s}, z_{j_s,s} \neq z_{ks}) = p(z_{i_s,s} = z_{k_s,s} \neq z_{j_s,s}) \\
&= p(z_{i_s,s} = z_{k_s,s} = 1, z_{j_s,s} = 0) + p(z_{i_s,s} = z_{k_s,s} = 0, z_{j_s,s} = 1) \\
&= p_s^2(1-p_s) + (1-p_s)^2 p_s = h_s/2.
\end{aligned}$$

From the assumption of statistical independence among sites s located on different reads with pooling, it follows (Appendix A1) that for a sample of size n,

$$var(\widehat{\pi}_{LCS}) = \sum_s var(\widehat{f}_{ns}) = \sum_s \frac{1}{N_2} h_s \left\{(1-h_s) + \frac{N_3}{N_2}(1-2h_s)\right\} \qquad (3)$$

Approximate statistical independence across sites requires that the number of possible samples of size n is much larger than the number of segregating sites (i.e. $N \gg n$ so that $\binom{N}{n} \gg S_N$).

2.2 Case 2: High Coverage Sequencing (HCS)

Computing pairwise differences among samples of individuals genomes under HCS involves calculating moments of sums rather than sums of moments. For $\mathbf{z}_i \sim p(\mathbf{z})$ sampled independently under HCS, applying $g_{ij} = \sum_s I(z_{is} \neq z_{js}) = \sum_s f_{ij,s}$, we define

$$\widehat{\pi}_{HCS} \equiv \widehat{g}_n = \frac{1}{N_2} \sum_{i<j} g_{ij},$$

and derive its variance below. For a sample of individual haplotypes $i = 1...n$, consider $z_{is} \sim Bern(p_s)$ as before, but with correlated z_{is}, z_{ir} due to linkage disequilibrium (LD) between sites. We define the LD D_{rs} for (arbitrarily labeled) alleles R, r and S, s at the two sites as follows. Letting $q_s, q_r = 1 - p_s, 1 - p_r$ (Lewontin and Kojima, 1960),

$$\begin{array}{lcllllll} p(RS) & = & p(R)p(S) & + & D_{sr} = & p_r p_s & + & D_{sr} \\ p(rs) & = & p(r)p(s) & + & D_{sr} = & q_r q_s & + & D_{sr} \\ p(Rs) & = & p(R)p(s) & - & D_{sr} = & p_r q_s & - & D_{sr} \\ p(rS) & = & p(r)p(S) & - & D_{sr} = & q_r p_s & - & D_{sr}. \end{array}$$

As with LCS, we have, for $h_s = 2p_s q_s$,

$$\mu_f = E(f_{ij,s}) = h_s \text{ and } \sigma_f^2 = var(f_{ij,s}) = h_s(1 - h_s)$$

The probability of different identity among sites s, r in a sample pair i, j is $p(f_{ijs} f_{ijr} = 1) = p(RS, rs) + p(rs, RS) + p(Rs, rS) + p(rS, Rs)$, where $p(RS, rs) = p(z_{i,sr} = RS, z_{j,sr} = rs)$ etc. Therefore

$$\gamma_{sr} = E(f_{ij,s} f_{ij,r}) = 2(p_s p_r + D_{sr})(q_s q_r + D_{sr}) + 2(p_s q_r - D_{sr})(q_s p_r - D_{sr})$$

and similarly, considering triplet samples with shared element j paired with i and k, the probability of different identity between i and j at site r and j and k at site s is $p(f_{ijs} f_{jkr} = 1) = p(R, rS, s) + p(R, rs, S) + p(r, RS, s) + p(r, Rs, S)$, where $p(R, rS, s) = p(z_{ir} = R, z_{jr} = r, z_{js} = S, z_{ks} = s)$ etc. Using these terms, we compute the expectation:

$$\delta_{sr} = E(f_{ij,s} f_{jk,r}) = 2p_s(q_s p_r - D_{sr})(q_s q_r + D_{sr}) + 2(p_s q_r - D_{sr})(q_s p_r - D_{sr})$$

At linkage equilibrium ($D_{sr} = 0$ for all s, r), high coverage sequencing of n individuals and low coverage sequencing of pooled individuals at read depth n are statistically equivalent, i.e. both equations simplify to $\gamma_{sr} = \delta_{sr} =$

$4p_s q_s p_r q_r$.

The mean and sample variance terms for the expected pairwise distances are, respectively,

$$\begin{aligned}
\mu &= E(g_{ij}) = \sum_s h_s, \\
\sigma^2 &= var(g_{ij}) = \sum_s var(f_{ij,s}) + 2\sum_{r<s} cov(f_{ij,r}, f_{ij,s}) \\
&= \sum_s h_s(1-h_s) + 2\sum_{r<s}(\gamma_{sr} - h_s h_r),
\end{aligned}$$

while the covariance κ for the ordered pair of pairs with a shared j element is:

$$\begin{aligned}
\kappa &= E(g_{ij}\, g_{jk}) = E\left\{\sum_s f_{ij,s} \cdot \sum_s f_{jk,s}\right\} \\
&= E\left\{\sum_s I(z_{is} = z_{ks} \neq z_{js}) + 2\sum_{r<s}(I(z_{is} \neq z_{js})I(z_{jr} \neq z_{kr})\right\} \\
&= \sum_s h_s/2 + 2\sum_{r<s} \delta_{sr}
\end{aligned}$$

By incorporating κ, we construct the sample estimate and variances for g_{ij}. Because we are now averaging over haplotypes $\mathbf{z}_i$ (rather than independent counts for each site), $\widehat{g}_n$ is itself an average across pairs, like $\widehat{f}_n$ in the LCS case. Applying Eqn. (2), we find that

$$var(\widehat{\pi}_{HCS}) = \frac{\sigma^2}{N_2} + 2\frac{N_3}{N_2^2}(\kappa - \mu^2) =$$

$$\frac{1}{N_2}\underbrace{\left(\sum_s h_s(1-h_s) + 2\sum_{r<s}(\gamma_{sr} - h_s h_r)\right)}_{\sigma^2} + \frac{2N_3}{N_2^2}\left[\underbrace{\sum_s h_s/2 + 2\sum_{r<s}\delta_{sr}}_{\kappa} - \left(\sum_s h_s\right)^2\right] \tag{4}$$

2.3 Difference and Independence

Using the results in Eqns. (3) and (4), we derive the difference between the sample variances in pairwise differences under HCS vs. LCS as

$$\Delta = var(\widehat{\pi}_{HCS}) - var(\widehat{\pi}_{LCS}) = \frac{2}{N_2}\sum_{r<s}(\gamma_{sr} - h_s h_r) + \frac{4N_3}{N_2^2}\sum_{r<s}(\delta_{sr} - h_s h_r) \quad (5)$$

By collecting terms, we can rewrite the above as

$$\Delta = \frac{2}{N_2}\sum_{r<s} B_{sr} + \frac{4N_3}{N_2^2}\sum_{r<s} A_{sr},$$

where

$$\begin{aligned} A_{sr} &= \delta_{sr} - h_s h_r = (p_s p_r + q_s q_r - p_s q_r - p_r q_s)D_{sr} + 4p_s q_s p_r q_r - 4p_s q_s p_r q_r \\ &= (p_s - q_s)(p_r - q_r)D_{sr} = (2p_s - 1)(2p_r - 1)D_{sr} \\ B_{sr} &= \gamma_{sr} - h_s h_r = 4D_{sr}^2 + 2(p_s p_r + q_s q_r - p_s q_r - p_r q_s)D_{sr} + 4p_s q_s p_r q_r - 4p_s q_s p_r q_r \\ &= 4D_{sr}^2 + 2A_{sr} \end{aligned}$$

For notational convenience, we define:

$$E(A_{sr}) = \frac{1}{N_2}\sum_{r<s} A_{rs}$$

Without linkage disequilibria among pairs ($D_{sr} = 0$ and therefore $A_{sr}, B_{sr} = 0$ for all s, r pairs), $\gamma_{sr} = \delta_{sr} = h_s h_r$ and $\Delta = 0$, i.e. the sample variances under HCS and LCS are equal. Because $B_{sr} \geq A_{sr}$ for $A_{sr} > 0$, $E(A_{sr}) > 0$ is a sufficient condition for $\Delta > 0$. This condition is satisfied when the sum of weighted linkage disequilibria A_{sr} is positive, i.e.

$$\sum_{sr} A_{sr} = \sum_{sr}(2p_s - 1)(2p_r - 1)D_{sr} > 0 \quad (6)$$

We remark that $E(A_{sr}) > 0$ is a sufficient but not necessary condition for $\Delta > 0$. The variance in mean pairwise distance can be reduced with pooled LCS even for $E(A_{sr}) < 0$, because negative A_{sr} may be offset by the positive contributions of D_{sr}^2 to the B_{sr} term when pairwise LD values in the population are sufficiently high. However, with large sample sizes, the A_{sr} term dominates because it scales as $\sim 1/n$ while the B_{sr} term scales as $\sim 1/n^2$; consequently, the sign of $E(A_{sr})$ generally predicts that of Δ.

$E(A_{sr}) > 0$ requires that most "major" alleles (those with $p_s, p_r > 0.5$) at different loci are in positive LD, while major and minor allele pairs at different loci ($p_s > 0.5, p_r < 0.5$ or vice-versa) are in negative LD. The weighted LD A_{sr} provides a measure of the extent to which major alleles are in positive LD, regardless of whether the more common allele is a reference/wildtype or variant/mutant at a particular site. These results predict that when the mean weighted LD is positive, the sample variance (error) in estimated pairwise genetic distance will be reduced by LCS.

2.3.1 Possible Caveats: Random Read Depth and Fractional Coverage

Our results are based on a comparison of two extreme-case scenarios: full coverage sequencing of n sampled genomes versus low coverage sequencing of M sampled genomes at a fixed read depth $n \ll M$, which is an idealization, because in reality the actual read depth under NGS varies across sites.

Random Read Depth. We first consider the effect of having the read depth as Poisson random variable rather than a constant. In Appendix A2, we show that if the read depth $n_s \sim Poiss(n)$ (for mean read depth n), then the variance of the estimated pairwise genetic distance under LCS is:

$$\begin{aligned} var(\hat{\pi}_{LCS}) &= \sum_s \left\{ g(n) + O\left(\frac{1}{n^2}\right)\right\} \\ &= \sum_s \left[\frac{1}{N_2} h_s \left\{(1-h_s) + \frac{N_3}{N_2}(1-2h_s)\right\}\right] + O\left(\frac{1}{n^2}\right) \end{aligned}$$

This result follows because a Poisson random variable with mean n has variance $= n$ (indeed, read depth variance across sites will be typically of $\sim n$ for most other plausible sampling distributions with a mean read depth n), the contributions of the variance in read depth to the variance of the estimate are $\sim O(1/n^2)$, meaning that the $var(\hat{\pi}_{LCS})$ is essentially the same as Eqn. (3) and A_{sr}, so that Δ remain essentially unchanged. Appendix A2 also presents numerical examples confirming these results.

Fractional Coverage. Our derivations for LCS assumed that the coverage in LCS was sufficiently low that effectively only a single locus is sampled per individual, which is an idealization for very short reads and for very large, well mixed pooled sample size M relative to the read depth. With greater coverage, LCS results in the sampling of several polymorphic loci from within one genome, albeit far fewer than complete coverage sequencing of individual genomes. We show that the variances in $\hat{\pi}$ with reduced coverage will be between the limiting LCS case and HCS variances based on a continuity argument. To see the effect of a mode of sequencing where a fraction ρ of each genome is covered, consider the covariance term γ_{sr} used in the derivation of Eqn. (4), $\gamma_{sr} = E(f_{ij,s} f_{ij,r})$. Let $\xi_i = I(z_{is}$ and z_{ir} are phased) denote an indicator for recording phased alleles identified from a single haplotype (such as through the use of long reads).

Different (ξ_i, ξ_j) result in different expressions for γ. If both are on the same haplotype, $\xi_i = \xi_j = 1$, we get the γ_{rs} in the HCS limit (see derivation of Eqn. (4)). If both are independent, $\xi_i = \xi_j = 0$, we get $2p_s(1-p_s) \cdot 2p_r(1-p_r) = h_s h_r$ in the LCS limit. We apply the law of total probability to evaluate

$$\gamma_{rs} = E(f_{ij,s} f_{ij,r}) = \sum_{\xi_i=0}^{1} \sum_{\xi_j=0}^{1} p(\xi_i) p(\xi_j) E(f_{ij,s} f_{ij,r} \mid \xi_i, \xi_j),$$

with $p(\xi_i = 1) = p(\xi_j = 1) = \rho$. The same argument from total probability (applied over cases where i, j, k have s, r sites on various combinations of shared vs. different reads) applies to continuity of δ_{rs}.

Let $\hat{\pi}_\rho$ denote the estimator under a sampling scheme with fractional coverage. It follows from the continuity of δ, γ and from the fact that the marginal expectations $E(f_{ij})$ are the same under HCS and LCS that the resulting $\text{var}(\hat{\pi}_\rho)$ is a continuous function of ρ with $\text{var}(\hat{\pi}_0) = \text{var}(\hat{\pi}_{LCS})$ and $\text{var}(\hat{\pi}_1) = \text{var}(\hat{\pi}_{HCS})$. This result is qualitatively consistent with the relationship between the variance in θ and the fraction of missing sequence data in Ferretti et al. (2014).

3 Dependence on Allele Frequencies and Linkage Disequilibria

Low coverage sequencing reduces the error in estimates of $\hat{\pi}$ when $\Delta > 0$, which holds if the expected weighted linkage disequilibria defined by Eqn. (6) are positive. Ferretti et al. (2014) observed a decrease in variance under low coverage for populations at mutation-drift equilibrium - their result suggests that $E(A_{sr})$ should be positive for equilibrium distributions of allele frequencies in an infinite sites model. Below, we show that this condition holds provided there is a sufficiently high probability density of high frequency mutations, i.e. sufficiently many haplotypes with multiple mutations.

As noted above, $E(A_{sr}) > 0$ requires positive linkage disequilibria among pairs of alleles that are rare ($p < 0.5$) and among pairs of alleles that are common ($p > 0.5$), and by symmetry, negative LD among most pairs of major/minor alleles across loci. For a population of clonal, non-recombinant genomes, we can determine the distribution of LD given a distribution of allele frequencies. In an infinite sites model without recombination, letting p_s, p_r be the frequencies of variant alleles S, R at loci s, r, where $p_s < p_r$, it follows that all occurrences of S must be in S, R haplotypes (conversely, if S co-occurs with r, there can be no S, R haplotypes). Therefore, if $p_s + p_r > 1$ and $p_s < p_r$, then $p(S, R) = p_s$ and the LD between loci s, r $D_{sr} = p_s(1-p_r)$. When $p_s + p_r < 1$, we can compute the expected LD by counting the number of cases that S can co-occur with the R allele versus the r wildtype, since the constraint of S either always or never co-occurring with R when $p_s < p_r$.

3.1 Computing Linkage Disequilibria and $E(A_{sr})$

Let N be the total number of haplotypes and let s, r denote any two loci. We observe variant allele frequencies $p_s = \frac{k}{N}$ and $p_r = \frac{h}{N}$ at loci s and r for $k, h = 1, \ldots, N$. Without loss of generality, we assume $k \leq h$ (corresponding to $p_s < p_r$). Let C_{sr} denote the event that mutations S and R co-occur and S and r don't co-occur, while $\bar{C}_{sr}$ denotes the event that S and R don't co-occur (i.e. S co-occurs with the "wildtype" allele at r). Let $(C_{sr} \cup \bar{C}_{sr})^c$ denote the complement of $C_{sr} \cup \bar{C}_{sr}$ such that S and R co-occur on some haplotype and S and r co-occur on some other haplotype. Because there is no recombination and no multiple mutations per site, $C_{sr} \cup \bar{C}_{sr}$ and $(C_{sr} \cup \bar{C}_{sr})^c = \emptyset$. We compute the probability of co-occurrence $p(C_{sr} | C_{sr} \cup \bar{C}_{sr})$:

(1) $\mathbf{N < h + k}$. Trivially, $p(C_{sr}|C_{sr} \cup \bar{C}_{sr}) = 1$ and $p(\bar{C}_{sr}|C_{sr} \cup \bar{C}_{sr}) = 0$, i.e. $p_r + p_s > 1 \rightarrow p(C_{sr}) = p_s$.

(2) $\mathbf{N \geq h + k}$. We first derive $p(C_{sr})$ and $p(\bar{C}_{sr})$ and then $p(C_{sr}|C_{sr} \cup \bar{C}_{sr}) = p(C_{sr})/(p(C_{sr}) + p(\bar{C}_{sr}))$.

Counting the possible combinations, we find

$$\begin{aligned} p(\bar{C}_{sr}) &= \frac{\binom{N-h}{k}}{\binom{N}{k}} = \frac{(N-h)\cdots(N-h-k+1)}{N\cdots(N-k+1)} \\ p(C_{sr}) &= \frac{\binom{h}{k}}{\binom{N}{k}} = \frac{h\cdots(h-k+1)}{N\cdots(N-k+1)} \end{aligned}$$

The numerators in the first and second equations give the number of ways in which k S alleles can co-occur with r or with R, respectively, given $\binom{N}{k}$ possible positions for the S alleles. Therefore,

$$\begin{aligned} p(C_{sr}|C_{sr} \cup \bar{C}_{sr}) &= \frac{h(h-1)\cdots(h-k+1)}{h(h-1)\cdots(h-k+1) + (N-h)(N-h-1)\cdots(N-h-k+1)} \\ p(\bar{C}_{sr}|C_{sr} \cup \bar{C}_{sr}) &= \frac{(N-h)(N-h-1)\cdots(N-h-k+1)}{h(h-1)\cdots(h-k+1) + (N-h)(N-h-1)\cdots(N-h-k+1)} \end{aligned}$$

Given C_{sr}, $p(SR) = p_s = k/N$ and hence $D_{sr} = p_s - p_s p_r = \frac{k}{N}(1 - \frac{h}{N})$.

Given $\bar{C}_{sr}$, $p(SR) = 0$ and hence $D_{sr} = 0 - p_s p_r = -\frac{kh}{N^2}$.

Therefore, for $C_{sr} \cup \bar{C}_{sr}$,

$$\begin{aligned} E(D_{sr}) &= p(C_{sr}|C_{sr} \cup \bar{C}_{sr}) \frac{k}{N}\left(1 - \frac{h}{N}\right) + p(\bar{C}_{sr}|C_{sr} \cup \bar{C}_{sr})\left(-\frac{kh}{N^2}\right) \\ &= \begin{cases} \frac{k}{N}\left(1 - \frac{h}{N}\right) & \text{for } N < h + k \\ \left(\frac{h(h-1)\cdots(h-k+1)}{h(h-1)\cdots(h-k+1)+(N-h)(N-h-1)\cdots(N-h-k+1)} - \frac{h}{N}\right)\frac{k}{N} & \text{for } N \geq h + k \end{cases} \end{aligned}$$

As expected, in the limiting cases of a new mutation ($k = 1$), $D_{sr} = 0$, while for $k + h > N$, $E(D_{sr}) = p_s(1 - p_r)$.

To compute $E(A_{sr})$, we evaluate the weighted linkage disequilibria over the distribution of η_k,η_h, the number of sites at which there are exactly k, h copies of variant alleles. The expected weighted LD $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$ is

$$
\begin{aligned}
E(A_{sr}|\boldsymbol{\eta}) \;=\; & \Bigg[\sum_{k=1}^{\frac{N}{2}-1} \sum_{h=k+1}^{N-k} \eta_k \eta_h \left(\frac{2h}{N}-1\right)\left(\frac{2k}{N}-1\right) \\
& \times \left(\frac{h(h-1)\cdots(h-k+1)}{h(h-1)\cdots(h-k+1)+(N-h)(N-h-1)\cdots(N-h-k+1)} - \frac{h}{N}\right)\frac{k}{N} \\
& + \sum_{h=\frac{N}{2}+1}^{N} \sum_{k=N-h+1}^{h-1} \eta_k \eta_h \left(\frac{2h}{N}-1\right)\left(\frac{2k}{N}-1\right)\frac{k}{N}\left(1-\frac{h}{N}\right) + \sum_{k=1}^{N/2} \binom{\eta_k}{2}\left(\frac{2k}{N}-1\right)^2 \\
& \times \left(\frac{k!}{k!+(N-k)(N-k-1)\cdots(N-2k+1)} - \frac{k}{N}\right)\frac{k}{N} \\
& + \sum_{k=\frac{N}{2}+1}^{N} \binom{\eta_k}{2}\left(\frac{2k}{N}-1\right)^2 \frac{k}{N}\left(1-\frac{k}{N}\right)\Bigg] \Bigg/ \left(\sum_{k=1}^{N-1}\sum_{h=k+1}^{N} \eta_k\eta_h + \sum_{k=1}^{N}\binom{\eta_k}{2}\right) \\
:= \;& \nu(\boldsymbol{\eta})/\delta(\boldsymbol{\eta}) \qquad (7)
\end{aligned}
$$

with the convention that $\binom{1}{2} = 0$, so that the marginal expectation of A_{sr} is given by

$$
E(A_{sr}) = E\{E(A_{sr}|\boldsymbol{\eta})\} = E(\nu(\boldsymbol{\eta})/\delta(\boldsymbol{\eta})) \approx E(\nu(\boldsymbol{\eta}))/E(\delta(\boldsymbol{\eta})). \qquad (8)
$$

This approximation holds based on a Taylor expansion of the ratio of two random variables, and the fact that $\nu(\boldsymbol{\eta})$ is much smaller than $\delta(\boldsymbol{\eta})$ due to the typically small values of A_{sr} for any pair s, r. The ratio of expectations is:

$$
\begin{aligned}
&\frac{E(\nu(\boldsymbol{\eta}))}{E(\delta(\boldsymbol{\eta}))} \\
=\;& \Bigg[\sum_{k=1}^{\frac{N}{2}-1}\sum_{h=k+1}^{N-k} E(\eta_k\eta_h)\left(\frac{2h}{N}-1\right)\left(\frac{2k}{N}-1\right) \\
&\times\left(\frac{h(h-1)\cdots(h-k+1)}{h(h-1)\cdots(h-k+1)+(N-h)(N-h-1)\cdots(N-h-k+1)}-\frac{h}{N}\right)\frac{k}{N} \\
&+\sum_{h=\frac{N}{2}+1}^{N}\sum_{k=N-h+1}^{h-1} E(\eta_k\eta_h)\left(\frac{2h}{N}-1\right)\left(\frac{2k}{N}-1\right)\frac{k}{N}\left(1-\frac{h}{N}\right) \\
&+\sum_{k=1}^{N/2} E\left\{\binom{\eta_k}{2}\right\}\left(\frac{2k}{N}-1\right)^2\times\left(\frac{k!}{k!+(N-k)(N-k-1)\cdots(N-2k+1)}-\frac{k}{N}\right)\frac{k}{N} \\
&+\sum_{k=\frac{N}{2}+1}^{N} E\left\{\binom{\eta_k}{2}\right\}\left(\frac{2k}{N}-1\right)^2\frac{k}{N}\left(1-\frac{k}{N}\right)\Bigg] \\
&\Bigg/\left(\sum_{k=1}^{N-1}\sum_{h=k+1}^{N} E(\eta_k\eta_h)+\sum_{k=1}^{N} E\left\{\binom{\eta_k}{2}\right\}\right),
\end{aligned}
\tag{9}
$$

The expectations $E(\binom{\eta_k}{2}) = \frac{1}{2}[Var(\eta_k) + E(\eta_k)^2 - E(\eta_k)]$, while $E(\eta_k\eta_h) = Cov(\eta_k, \eta_h) + E(\eta_k)E(\eta_h)$. Therefore, we can approximate the expectation of $E(A_{sr})$ (and at least obtain the correct sign from the numerator) to determine if Δ is positive when the expectations and second moments of an allele frequency distribution are known. The same approach can be used to compute the expectated linkage disequilibria over all η_k, $E[E(D_{sr}|\boldsymbol{\eta})]$, by substitution D_{sr} for A_{sr} (i.e. no factor of $(2p_s-1)(2p_r-1)$) in the equations above.

To calculate $E(A_{sr})$ for an infinite sites model at mutation-drift equilibrium, we use the expectations and variances of η_k (Fu, 1995):

$$\begin{aligned} E(\eta_k) &= \frac{\theta}{k} \\ Var(\eta_k) &= \frac{\theta}{k} + \sigma_{k,k}\theta^2 \\ Cov(\eta_k, \eta_h) &= \sigma_{k,h}\theta^2 \end{aligned}$$

where the coefficients $\sigma_{k,k}$ and $\sigma_{k,h}$ are functions of harmonic series sums a_n and B_n, i.e.

$$\sigma_{k,k} = \begin{cases} B_n(k+1), & \text{if } k+h < n/2 \\ 2\frac{a_n - a_k}{n-k} - \frac{1}{k^2}, & \text{if } k = n/2 \\ B_n(k) - \frac{1}{k^2}, & \text{if } k > n/2 \end{cases}$$

while the covariance coefficients are (for $h > k$)

$$\sigma_{k,h} = \begin{cases} \frac{B_n(h+1) - B_n(h)}{2}, & \text{if } k+h < n \\ \frac{a_n - a_k}{n-k} + \frac{a_n - a_h}{n-h} - \frac{B_n(h) + B_n(k+1)}{2} - \frac{1}{kh}, & \text{if } k+h = n \\ \frac{B_n(k+1) - B_n(k)}{2} - \frac{1}{kh}, & \text{if } k+h > n \end{cases}$$

where

$$\begin{aligned} a_n &= \sum_{i=1}^{n-1} \frac{1}{i} \\ B_n(k) &= \frac{2n}{(n-k+1)(n-k)}(a_{n+1} - a_k) - \frac{2}{n-k} \end{aligned}$$

Ferretti et al. (2014) used the expectations and covariances of η_k to derive sample variances for θ and the Tajima D test statistic, showing that these variances were reduced with LCS for equilibrium allele frequency distributions under an infinite sites model. Using the expectations for (weighted) linkage disequilibria, we show that these results follow as consequences of the distributions of pairwise LD, and compare these predictions to simulation results for representative parameters. We remark that numerical estimates of $E(A_{sr})$ computed from Eqn. (9) will be only approximate for several reasons: first, because the expectations of ratios do not exactly equal ratios of expectations (Eqn. (8)), and second, because the expectation and variance of η_k are derived for sampling distributions where $n \ll N$ (Watterson, 1975; Fu, 1995). However, they usually provide a sufficiently good approximation (Wakeley and Takahashi, 2003) as $n \to N$, so that our estimates of $E(A_{sr})$ should be of the correct sign and magnitude.

4 Comparison to individual-based simulations

To simulate Fisher-Wright genetic drift in an infinite sites model, we initialized a population of N haploid genotypes at $K = 10^8$ sites with reference genotypes. In every generation, N individuals were sampled with replacement from the existing pool, with each individual sampled producing a single progeny. The number of mutations m for each offspring is $m \sim Poiss(Ku)$, with the mutations randomly distributed among the K sites. This process was iterated over T generations; in order to approximate a near-equilibrium distribution of allele frequencies, we ran the simulations for $T \sim 4N$ to assure a coalescent among all lineages in each sample path genealogy. Simulations were also run for a range of values $T < N$ to generate non-equilibrium distributions of allele frequencies and pairwise LD. For each combination of parameters, the simulation cycle was replicated 100 times.

To simulate full coverage sequencing of individuals, n haplotypes were randomly selected without replacement from the model population. The Hamming distances were calculated for all pairs in a sample, variant allele frequencies and linkage disequilibria were calculated over all individuals and all pairs in the model population. Sequencing of pooled samples at very low coverage was simulated by selecting n alleles without replacement at every segregating site and summing pairwise distances over all sites (corresponding to sampling with replacement with respect to genomes, but without replacement with respect to each locus). Δ was estimated from the data as the difference in the sample variances between the HCS and LCS pairwise distances. In every replicate run, A_{sr} was calculated from the mutation frequencies p_s, p_r and D_{sr} using Eqn. (6). All simulations were implemented using Python 2.7.3, the code is available from the corresponding author upon request.

The simulation results for population sizes $N = 200, 500$, a sample size of $n = 20$ and a range of generation times T are shown in Tables 1 and 2. The first table shows the estimated parameter values from which Δ is calculated, including the number of polymorphic sites S_N in the population (as opposed to the sample number of segregating sites S_n), the population mean allele frequency across polymorphic sites, the sample mean pairwise genetic distances under HCS and LCS (for $n = 20$), and their sample variances over 100 replicates.

TABLES 1a-b and 2a-b HERE

Over $T \ll N$ generations, there are very few (~ 100) polymorphic sites, all of which have low variant allele frequencies. The mean and variance of genetic distances are of the order ~ 1, ~ 0.1, respectively. For $T \sim 4N$, allele frequencies and genetic distances are near the equilibrium values, e.g. the estimated pairwise genetic distance $\widehat{\pi}$ converges to the Tajima estimator for haploids $\theta = 2Nu$, which is $\widehat{\pi} = 150, 300$ for $N = 200, 500$, respectively. Table 2 shows the sum of weighted LD values $\sum A_{sr} = \bar{A}_{sr}N_2$. The mean values of D_{sr} are effectively 0 ($\sim 10^{-7} - 10^{-6}$, either positive or negative, not shown in table) even for large values of T and S_N. However, the skewed distribution results in a large postive weighted LD. Figures 2a-c show frequency distributions of allele frequencies, pairwise LD and weighted LD for a representative model population, specifically, for a single sample path where $N = 500, T = 2500$.

FIGURE 2a-c HERE

From $\sum A_{sr}$, we compute the predicted difference between HCS and LCS variances Δ_P using Eqn. (5). The predicted value is compared to the simulation estimate $\Delta_S = var_{HCS} - var_{LCS}$. The consistency of observed and predicted values of Δ is confirmed by the fact that even the largest deviations are within less than two standard error $SE_{\Delta_S} = \sqrt{var(\Delta_S)/n}$ units with respect to the point estimate Δ_S. The fit between analytical predictions and observed values improves for longer generation times as populations approach a mutation-drift equilibrium distribution of allele frequencies.

Except for populations with very few mutations where weighted LD values are ~ 0, we have $\Delta > 0$ for most of the simulations. These results are consistent with LCS reducing the error in sample genetic distance except when variant alleles are rare and at low frequency. This reduction of error through low coverage sampling of many genomes is strongest for near-equilibrium distributions of allele frequencies, for large numbers of segregating sites, and for small sample sizes (corresponding to low coverage depth with NGS). Δ scales approximately as $\sim 1/n$ for large n; consequently, for sample numbers and coverage depths of the order ~ 100, Δ will be smaller by nearly an order of magnitude relative to the values obtained for $n = 20$ (simulations were performed for $n = 10, 50$, the results are not shown due to qualitative similarity to the data in Tables 1-2).

The two observed cases with $\bar{A}_{sr} < 0$ are for $T = 10$, with a negative predicted value Δ_P for $N = 500$ (though not for $N = 200$). Here, the Δ values are effectively zero within a standard error unit, so whether positive or negative values are observed is of purely formal interest (note that for even smaller time intervals $T = 5$ and even fewer polymorphic sites, both $\bar{A}_{sr}$ and $\Delta > 0$, albeit very small). This suggests that at least under neutral evolution, $E(A_{sr}) < 0$ occurs under restrictive conditions corresponding to very small absolute values of Δ and negligible reduction of error in estimating $\widehat{\pi}$ through either HCS or LCS, while for large numbers of segregating sites and increasing allele frequencies, there will be considerable increases in error when $\widehat{\pi}$ is estimated via high coverage sequencing.

For large T, the distribution of allele frequencies in the population approaches mutation-drift equilibrium, so that the mean A_{sr} should approximate $E(A_{sr})$ computed from Eqns (7-9) using Mathematica 11.1. For $N = 200, 500$ and genomic mutation rate $u = 0.3$ ($\theta = 120, 300$), the estimated values of $E(A_{sr}) = 1.11 \times 10^{-3}, 1.06 \times 10^{-3}$, respectively, consistent in sign and magnitude with the mean values of A_{sr} for the simulated populations at $T \sim 4N$. For comparison, evaluating $\bar{A}_{sr} = 2\sum_{s,r} A_{sr}/S(S-1)$ for the number of polymorphic sites S (using the summed values of A_{sr} in Table 2, and the number of segregating sites S in Table 1), we obtain $\bar{A}_{sr} = 1.40 \times 10^{-3}$ for $N = 200, T = 1000$ and $\bar{A}_{sr} = 8.50 \times 10^{-4}$ for $N = 500, T = 2500$.

The positive values of $E(A_{sr})$ result from the fact that even though

$E(\eta_k) \sim \theta/k$ suggests that high frequency alleles that contribute to large positive LD are rare, the high variance in allele frequencies means that any given sample path will usually contain several haplotypes characterized by multiple variant alleles at very high frequency. This means that even though the expectation for a particular η_k will be low for $k \sim N$, most genealogies will be characterized by a high count for some individual large value(s) of k, as is seen in Figure 2a. This generates a positive skewed distribution of pairwise linkage disequilibria (consistent with the distributions of D_{sr} derived numerically for non-recombining loci in (Golding, 1984)), as is seen in Figs. 2b-c. This contributes large positive A_{sr} values for most individual genealogies in the simulations, so consequently, $\bar{A}_{sr} > 0, \Delta > 0$. These results hold not only for populations at mutation-selection equilibria, but also for any populations where variant alleles have had time to accumulate to sufficiently high frequencies, hence the positive (albeit lower) values of Δ observed for all but the smallest time intervals.

5 Analysis of cancer sequence data

In this section, we apply the results of our derivations to genomic data by computing $\sum A_{rs}$ and Δ for haplotype frequencies estimated from a lung adenocarcinoma sequences. Unpublished data on variant frequencies was provided to the authors by K. Gulukota and Y. Ji, who obtained their data via whole-exome sequencing of 4 sections of a primary solid tumor taken from a lung cancer patient. DNA from the samples was extracted using Agilent SureSelect capture probes. The exome library was sequenced with paired-end 100 bp reads on the Illumina HiSeq 2000 platform. Reads were mapped onto the human genome HG19 using BWA (Li and Durbin, 2009), giving a post-mapping mean 60-70 fold coverage across sites. Variant calls were performed with GATK (McKenna et al., 2010). Through the matching of read ends, somatic mutations co-occurring within ~ 100 bp in single genomes were identified (Sengupta et al. 2015). These mutation pairs define two locus haplotypes that can be tallied without the need for phasing, giving estimates of haplotype frequencies (defined by two proximate polymorphic sites) directly from the read counts.

Because reproduction in tumor cells is asexual and ameiotic, estimates of D_{sr} and A_{sr} using a subset of nearly adjacent sites is as representative of other haplotype pairs as if they were located on different chromosomes or on distant loci. The adenocarcinoma data contain estimated frequencies of 69 two-locus haplotypes, and corresponding variant allele frequencies for a total of 138 sites. This provides sufficient data to estimate the LD and weighted LD, and consequently the expected error in estimates of $\widehat{\pi}$ under high versus low coverage sequencing.

A naive application of Eqn. (5) to the distribution of mutation frequencies and LD values gives $\Delta \sim 0.1$ for coverage depth $n = 65$, suggesting much lower error if $\widehat{\pi}$ were estimated for the tumor via low coverage sequencing of pooled tumor cell genomes. However, several aspects of cancer genetics complicate this estimate. First, because cancer cells reproduce clonally, somatic mutations appear in heterozygous genotypes in the absence of mitotic recombination and gene conversion. A SNV frequency of $p = 0.5$ corresponds to fixation of a somatic mutation in a population of asexual diploids. Therefore, if we have heterozygous fixation at a single SNV site, a population consisting of 0/1 (reference and variant) genotypes, a mean genetic distance measure of $\widehat{\pi} = 1/2$ is meaningless because the population is homogeneous with respect to the 0/1 genotype, and variant allele frequencies must be rescaled to reflect this.

Figure 3 shows the distribution of mutant allele frequencies in Sample 1; note the high frequency of values near $p = 0.5$, the skew of the distribution is presumably the result of a low rate of detection of rare variants.

FIGURE 3 HERE

Williams et al. (2016a,b) (see also Ling et al. 2015) address the issue heterozygous genotype fixation by only considering polymorphic, segregating sites when comparing allele frequency spectra to neutral models, to the exclusion of sites that are ≥ 0.5 within a margin of sampling error; which also excludes sites whose frequencies $p > 0.5$ due to loss of heterozygosity. For the truncated range of allele frequencies $p = [0, 0.5]$, the frequencies are rescaled to reflect heterozygosity, which for diploids means mapping $p' = 2p$, or more generally, $p' = p/f_c$ where f_c is the cutoff for the inference of fixation. With this mapping, $\widehat{\pi}$ for a sample where all genotypes at a variant site are 0/1 is 0.

Assuming diploidy at all of the genotyped SNV sites and defining fixation as $p = 0.5$, we find that for sample size $n = 65$, the binomial probability of observing fewer than $x = 26$ mutant alleles is $Bin(x \leq 25|n = 65, p = 0.5) = 0.041$, so we use $f_c = 0.4$ as as a cutoff defining polymorphic sites. By this criterion, and the rescaling $p' = p/f_c$, there are only between 6 (sample 4) and 10 (sample 3) adjacent segregating sites, and consequently between 3 and 5 haplotypes defined by such a pair out of the original 69. The LD and Δ values for this subset of haplotypes are summarized in Table 3. The differences in variances Δ remain positive, consistent with sample variance being lower with pooling. Δ is small ($0.034 \leq \Delta \leq 0.070$), implying that in practice the estimation errors for $\widehat{\pi}$ would be negligibly different between high and low coverage sequencing for this data set. However, the small Δ are partly a result of the small number of segregating sites (i.e. $\widehat{\pi}_{max} = S_n/2$) while $var(\widehat{\pi})$ estimated by individual cell sampling may be expected to increase for more segregating sites, as was the case in the simulation data for larger time intervals.

TABLES 3a-b HERE

The values of Δ are also sensitive to the choice of truncation, as many of the SNVs occur in genotypes that are close to fixation in the tumor. For example, if we use $f_c = 0.49, x = 32$ as a cutoff to define segregating sites rather than $f_c = 0.40$, we obtain $\bar{A}_{sr} < 0$ and $\Delta < 0$ (of the order ~ 0.1). The sign reversal results from some lower frequency SNVs uniquely co-occuring in genomes with other SNVs that are close to fixation. The remaining allele and haplotype distributions contribute negative linkage disequilibria between the high frequency SNVs at one locus and high frequency reference alleles at the other site. The greater absolute value of Δ is a consequence of the fact that with a cutoff of $f_c = 0.49$, there are now 21-28 haplotypes (and 42-56 segregating sites) rather than the 6-10 for the $f_c = 0.40$ cutoff. The negative weighted LDs and Δ with this cutoff are shown in the second panel Table 3b, illustrating that for some samples, the variances in $\widehat{\pi}$ may actually be slightly higher with low coverage sequencing.

6 Discussion

The reduction of error in estimated genetic distance through low coverage sequencing reflects the loss of information due to non-independence across sites through linkage disequilibria. If the most frequent alleles at the majority of sites are in positive LD, any error in the estimated frequency and heterozygosity at one site covaries with the error at the other sites with individual sampling. In contrast, with LCS, each site provides independent information, so that the error across sites is uncorrelated. For S_n segregating sites in a sample of n and a variance in estimated distance per site σ^2, with independent sampling the error across sites will approach σ^2/S_n. In the extreme case where allele frequencies across sites are nearly identical (complete linkage), the sample variance is σ^2 independent of the number of sites. Another way to think of this is to consider the information gain that comes from sampling different loci from different subclasses of individuals in a sample under LCS, so that each polymorphic site has its own sample genealogy. This is analogous to the results of Pluzhnikov and Donnelly (1996), who found that in the presence of recombination, the optimal sequence length per genome for estimating allele frequencies and θ was sufficiently high to provide a large sample size while sufficiently short to provide low coverage per genome when recombination rates are low. With greater recombination rates, there is no information gain through low coverage because all but the closest loci have their own coalescent genealogy.

Conversely and by symmetry, a negative association of major allele frequencies across pairs of sites means that an error in estimated distance at one site will on average be compensated by an error in the opposite direction at another site, leading to reduction in variance under high coverage sequencing of individual genomes (analogous to improved estimation of the mean by sampling positive and negative extremes of a distribution). Both heuristic considerations and simulation results suggest that such a scenario is unlikely except for distributions of allele frequencies that give very small error values regardless, at least under neutral evolution. This was further confirmed by numerical calculation of the expected distribution of pairwise linkage disequilibria in an infinite sites model for clonal, ameiotic organisms.

Our results, like those of Ferretti et al. (2014) suggest that low coverage sequencing over pooled samples should be used to estimate the genetic distance (and consequently, population mutation rate parameter θ and the Tajima D statistic) under most conditions if reduction of estimation error is the sole criterion. However, there are several caveats to this conclusion, some theoretical, others practical. For example, we know that when most pairwise LD are approximately 0, the difference Δ between HCS and LCS estimates will be very small. A number of recent studies have shown that LD are generally among sites that are not physically linked in the genomes of sexually reproducing model organisms, including *Drosophila* (Andolfatto and Przeworski, 2000) and humans (Peterson et al., 1995; Reich et al., 2001). This suggests that any error introduced by sampling alleles from genomes individually with high coverage rather than pooled low coverage may be negligible for non-clonal genomes.

In contrast, in clonal organisms, or for regions of genome under very low recombination in sexually reproducing organisms, LD values will be high. Depending on the distribution of allele frequencies, Δ will be large when evaluated over many polymorphic sites. In the cases of cancer and microbial genomics, the standard NGS approach to sequencing reads from large numbers of cells at low coverage suggests an improved estimation of $\widehat{\pi}$ (and consequently, θ and N_e) relative to what would be obtained from more expensive single cell sequencing approaches. Furthermore, single-cell sequencing usually entails a much smaller sample size n than the coverage depths of 100-1000 that are standard for pooled sequencing. Moreover, Δ is defined on the assumption of the same effective sample size n for both LCS and HCS, when in fact LCS is associated with pooling and high read depth n, as is often the case, then this is often sufficient to reverse the sign of $var(\widehat{\pi}_{HCS}) - var(\widehat{\pi}_{LCS})$ even in the rare cases when $\Delta < 0$ for HCS sample size equal to LCS read depth n. This is because LCS combined with pooling increases the sample size per segregating site.

Finally, we remark that this study was to a large part motivated by efforts to apply the methods and theory of population genetics to cancer biology, where individual cell sequencing at high coverage versus pooled sampling with low coverage are often presented as alternative approaches. The case study computing Δ from lung cancer data in the previous section was used as proof of principle. A more accurate and refined analysis would have to take into consideration a number of potentially confounding variables. These include polyploidy and aneuploidy (so that with ploidy X, fixation corresponds to $p = 1/X$), as well as accounting for the loss of heterozygosity through mitotic recombination, reflected in frequencies $p > 0.5$. The sensitivity of $var(\widehat{\pi})$ to the choice of cutoff f_c defining fixation for both the diploid and polyploid cases is of interest as an area for future research.

7 Acknowledgments

MS and JL were supported by the St. David's Foundation impact fund. YN and PM were supported by NIH grant 2R01CA132897-06A1. We thank Kalamakar Gulukota and Yuan Ji at NorthShore University HealthSystem for providing the lung cancer data used in this paper. We also thank two anonymous reviewers for their helpful critique and comments, as well as the following individuals for advice and assistance: Mark Kirkpatrick, Jon Wilkins, Matthew Cowperthwaite, Habil Zare, Mikhail Matz, and Andrea Sottoriva.

8 Figures and Tables

Figure 1. Illustration of high coverage sequencing (HCS) versus low coverage sequencing (LCS). In this example, the population consists of 8 haplotypes G1...G8 characterized by 4 segregating sites $S1...S4$. 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 $var(\widehat{\pi})$.

Figures 2a-c. Figures 2a,b,c show, respectively, a representative distribution of variant allele frequencies p_s, pairwise linkage disequilibria D_{sr}, and weighted linkage disequilibrium A_{sr} for a simulated population with $N = 500$ haplotypes following $T = 2500$ generations of mutation and Fisher-Wright genetic drift. This is a single sample path (genealogy) rather than an averaged over all replicates, hence the outlier of high frequency mutations associated with a high frequency haplotype with a high density of mutations (Fig 2a). These in turn account for the positive skew (and positive mean value) of A_{sr} in a typical sample path (Fig 2c)

Figure 3. Distribution of allele frequencies p in the first lung adenocarcinoma sample, for $S_n = 138$ polymorphic sites. Values of p near 0.5 indicate heterozygous variant genotypes near fixation. Values $p > 0.5$ are a consequence of loss of heterozygosity via gene conversion during mitotic recombination, these are excluded from our analyses.

Tables 1a-b. A summary of population parameters for a simulated Fisher-Wright model of genetic drift with infinite sites. The table shows a comparison of Δ_P values predicted from Eqn. (5) with simulation the values Δ_S for $N = 200$ (Table 1a) and $N = 500$ (Table 1b) and sample size/coverage depth $n = 20$ for a range of time intervals, including values near $T = 4N$ close to equilibrium. The standard error of Δ_S is also shown, estimated ΔP is within $< 2(SE)$ units from Δ_S even for small T and few mutations. Mean population pairwise LD values (not shown) are all essentially zero for all simulations, while the magnitudes of A_{sr} increase with T as predicted. p is the mean variant allele frequency across all segregating sites.

Table 2a-b. These tables show the number of segregating sites S_n in a sample of $n = 20$, the mean pairwise genetic distances $\widehat{\pi}_{HCS}, \widehat{\pi}_L CS$ (for high and low coverage sequencing, respectively), and the variances in pairwise genetic distance for HCS vs. LCS. The latter are used to compute Δ_S in Table 1. Table 2a shows these summary statistics for $N = 200$, Table 2b for $N = 500$.

Table 3. This table summarizes estimates of Δ from haplotype and allele frequencies in the lung adenocarcinoma sequence data, where haplotype frequencies for sites on individual long reads are known. Note that $\bar{A}_{sr} > 0$ and $\Delta > 0$ for all 4 samples, indicating that the error in pairwise genetic distance estimates for this data set are greater under HCS than under LCS, albeit weakly given the small number of unique haplotypes. Δ is computed from the mean read depth $n = 65$ for two cutoff values defining polymorphic sites. The upper panel shows the values for a cutoff of $f_c = 0.40$, selected based on a binomial probability. We use $p' = p/f_c$, rescaled with respect to the diploid cutoff value. The lower panel shows the same for $f_c = 0.49$, selected arbitrarily close to $p = 0.5$ to show the sensitivity of Δ to the cutoff. The $f_c = 0.40$ calculations are based on 6-10 remaining polymorphic sites, the $f_c = 0.49$ on 42-56 sites, depending on the sample.

Table A2. This table summarizes the same parameter estimates as in Table 2a (for $N = 200, T = 1000$), however, the LCS read depth is now a Poisson random variable with mean $n = 20$ rather than a constant. Note that the simulation values for the means and variances of $\widehat{\pi}$ and of A_{sr}, Δ are largely unchanged due to the negligible contribution of variance in read depth to the error in parameter estimation.

References

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., et al. (2016). Next generation sequencing of pooled samples: guideline for variants' filtering. *Nature Scientific Reports*, 6:Article 33735.

Andolfatto, P. and Przeworski, M. (2000). A genome-wide departure from the standard neutral model in natural populations of drosophila. *Genetics*, 156(1):257–268.

Ferretti, L., Raineri, E., and Ramos-Onsins (2014). Neutrality tests for sequences with missing data. *Genetics*, 191:1397–1401.

Ferretti, L., Ramos-Onsins, S. E., and Perez-Encisco, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, 22:5561–5576.

Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*, 48:172–197.

Futschik, A. and Schlötterer, C. (2010). The next generation of moelcular markers from massively parallel sequencing of pooled dna samples. *Genetics*, 186(1):207–218.

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188.

Golding, G. (1984). The sampling distribution of linkage disequilibrium. *Genetics*, 108(1):257–274.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893.

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

Lewontin, R. C. and Kojima, K.-i. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proceedings of the National Academy of Sciences*, 112(47):E6496–E6505.

Lynch, M. (2008). Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Molecular Biology and Evolution*, 25(11):2409–2419.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303.

Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome research*, 25(10):1499–1507.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86:641–647.

Peterson, A. C., Di Rienzo, A., Lehesjokl, A.-E., de la Chapelle, A., Slatkin, M., and Frelmer, N. B. (1995). The distribution of linkage disequilibrium over anonymous genome regions. *Human molecular genetics*, 4(5):887–894.

Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, 144:1247–1262.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals - mining genomewide polymorphism data without big funding. *Nature Reviews Genetics*, 15:749–763.

Sengupta, S., Wang, J., Lee, J., Müller, P., Gulukota, K., Banerjee, A., and Ji, Y. (2015). Bayclone: Bayesian nonparametric inference of tumor subclones using ngs data. In *Proceedings of The Pacific Symposium on Biocomputing (PSB)*, volume 20.

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

Tajima, F. (1996). Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics*, 75(1):27–31.

Wakeley, J. and Takahashi, T. (2003). Gene genealogies when the sample size exceeds the effective population size. *Molecular Biology and Evolution*, 20(2):2008–2013.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016a). Identification of neutral tumor evolution across cancer types. *Nature genetics*.

Williams, M. J., Werner, B., Curtis, C., Barnes, C., Sottoriva, A., and Graham, T. A. (2016b). Quantification of subclonal selection in cancer from bulk sequencing data. *bioRxiv*, page 096305.

9 Appendix A1: Ordered Pairs of Pairs

Recall the definitions $\mu = E(\phi_{ij})$, $\sigma^2 = var(\phi_{ij})$ and $\kappa = E(\phi_{ij}, \phi_{jk})$.

Lemma 1. *Let $\mu = E(\phi_{ij})$ where the expectation is over pairs $x_i \sim p(x)$ and $x_j \sim p(x)$, independently. Let $\widehat{\phi}_n = \frac{1}{N_2}\sum_{i<j}\phi_{ij}$, denote a sample estimate for μ, averaging over all pairs (i,j) of samples. Then $\widehat{\phi}_n$ is unbiased, $E(\widehat{\phi}_n) = \mu$, and*

$$var(\widehat{\phi}_n) = \frac{\sigma^2}{N_2} + 2\frac{N_3}{N_2^2}(\kappa - \mu^2).$$

Proof. Unbiasedness is straightforward:

$$E(\widehat{\phi}_n) = E(\frac{1}{N_2}\sum_{i<j}\phi_{ij}) = \frac{1}{N_2}\sum_{i<j}E(\phi_{ij}) = \mu.$$

For the variance, note that

$$\mathrm{cov}(\phi_{ij}, \phi_{kl}) = E(\phi_{ij}\phi_{kl}) - E(\phi_{ij})E(\phi_{kl}) = \begin{cases} 0 & \text{when } \{i,j\}\cap\{k,\ell\} = \emptyset \\ \kappa - \mu^2 & \text{when } |\{i,j,k,\ell\}| = 3 \end{cases}$$

Then

$$\mathrm{var}(\widehat{\phi}_n) = \frac{\sigma^2}{N_2} + \frac{1}{N_2^2}\sum_P \mathrm{cov}(\phi_{ij}, \phi_{kl}) = \frac{\sigma^2}{N_2} + \frac{2}{N_2^2}N_3(\kappa - \mu^2).$$

□

Proof of Eqn. (3). Let $\widehat{f}_{ns} = \frac{1}{N_2}\sum_{i<j} f_{ij,s}$. From the statistical independence among sites s located on different reads under LCS, it follows that for a sample of n,

$$\mathrm{var}(\widehat{\pi}_1) = \sum_s var(\widehat{f}_{ns})$$

with

$$\begin{aligned} \mathrm{var}(\widehat{f}_{ns}) &= \frac{\sigma_s^2}{N_2} + 2\frac{N_3}{N_2^2}(\kappa_s - \mu_s^2) = \frac{1}{N_2}h_s(1-h_s) + 2\frac{N_3}{N_2^2}(h_s/2 - h_s^2) \\ &= \frac{1}{N_2}h_s\left\{1 - h_s + \frac{N_3}{N_2}(1-2h_s)\right\} \end{aligned}$$

where the first equality is due to Eqn. (2).

10 Appendix A2: Poisson Distribution of Read Depth

In practice, the read depth with LCS is not constant and varies considerably across sites. Assuming the read depth follows a Poisson distribution with mean n equal to the number of haplotypes in HCS. Computing the Taylor expansion of the variance of genetic distance estimator, we find the difference between fixed and Poisson read depth is $O(\frac{1}{n^2})$.

Let n_s denote the read depth at locus s which follows a Poisson distribution $n_s \sim Poi(n)$ with mean n. Let

$$\begin{aligned} N_2^{(s)} &= \frac{n_s(n_s-1)}{2} \\ N_3^{(s)} &= \frac{n_s(n_s-1)(n_s-2)}{2} \end{aligned}$$

Let $\hat{\pi}_{LCS} = \sum_s \frac{1}{N_2^{(s)}} \sum_{i_s<j_s} \phi_{ij}$ and $\boldsymbol{n} = (n_s)_s$. The variance of $\hat{\pi}_{LCS}$ conditioned on the sampling depths $\boldsymbol{n}$ is given by

$$var(\hat{\pi}_{LCS}|\boldsymbol{n}) = \sum_s \frac{1}{N_2^{(s)}} h_s \left\{ (1-h_s) + \frac{N_3^{(s)}}{N_2^{(s)}}(1-2h_s) \right\}.$$

By the law of total variance, we have

$$var(\hat{\pi}_{LCS}) = E\{var(\hat{\pi}_{LCS}|\boldsymbol{n})\} + var\{E(\hat{\pi}_{LCS}|\boldsymbol{n})\}.$$

The second term is zero because the expectation is independent of $\boldsymbol{n}$,

$$E(\hat{\pi}_{LCS}|\boldsymbol{n}) = \sum_s \frac{1}{N_2^{(s)}} \sum_{i_s<j_s} E(\phi_{ij}) = \sum_s E(\phi_{ij}).$$

The first term is non-trivial. Let $g(n_s) = \frac{1}{N_2^{(s)}} h_s \left\{(1-h_s) + \frac{N_3^{(s)}}{N_2^{(s)}}(1-2h_s)\right\}$ and expand $E(g(n_s))$ at n,

$$\begin{aligned} E\{g(n_s)\} &\approx E\{g(n)\} + E\{g'(n)(n_s-n)\} + \frac{1}{2}E\{g''(n)(n_s-n)^2\} \\ &= g(n) + 0 + \frac{1}{2}E\{g''(n)(n_s-n)^2\} \end{aligned}$$

with

$$\begin{aligned} g(n) &= \frac{1}{N_2} h_s \left\{(1-h_s) + \frac{N_3}{N_2}(1-2h_s)\right\} \\ g''(n) &= \frac{2(an^3 + 3bn^2 - 3bn + b)}{(n-1)^3 n^3} \end{aligned}$$

where $a = 2h_s(1-2h_s)$ and $b = -2h_s^2 + 6h_s$. So

$$\frac{1}{2}E\{g''(n)(n_s-n)^2\} = \frac{1}{2}g''(n)var(n_s) = \frac{1}{2}g''(n)n = O\left(\frac{1}{n^2}\right).$$

Therefore,

$$\begin{aligned} var(\hat{\pi}_{LCS}) &= \sum_s \left\{g(n) + O\left(\frac{1}{n^2}\right)\right\} \\ &= \sum_s \left[\frac{1}{N_2} h_s \left\{(1-h_s) + \frac{N_3}{N_2}(1-2h_s)\right\}\right] + O\left(\frac{1}{n^2}\right) \end{aligned}$$

where the first term of the last equality is the variance of $\hat{\pi}_{LCS}$ given the read depth is fixed at n across all loci. This result holds for any sampling distribution of read depths where the variance is of the order $\sim n$.

The very small $\sim O(\frac{1}{n^2})$ contribution of non-constant read depth to the error in estimated $\hat{\pi}$ is confirmed using simulation results. Compare the variances and other parameter estimates observed in Table 2 for $N = 200, n = 20$ where the read depths are constant to those in Table A1, where the read depth is a Poisson random variable with parameter n.

TABLE A2 HERE

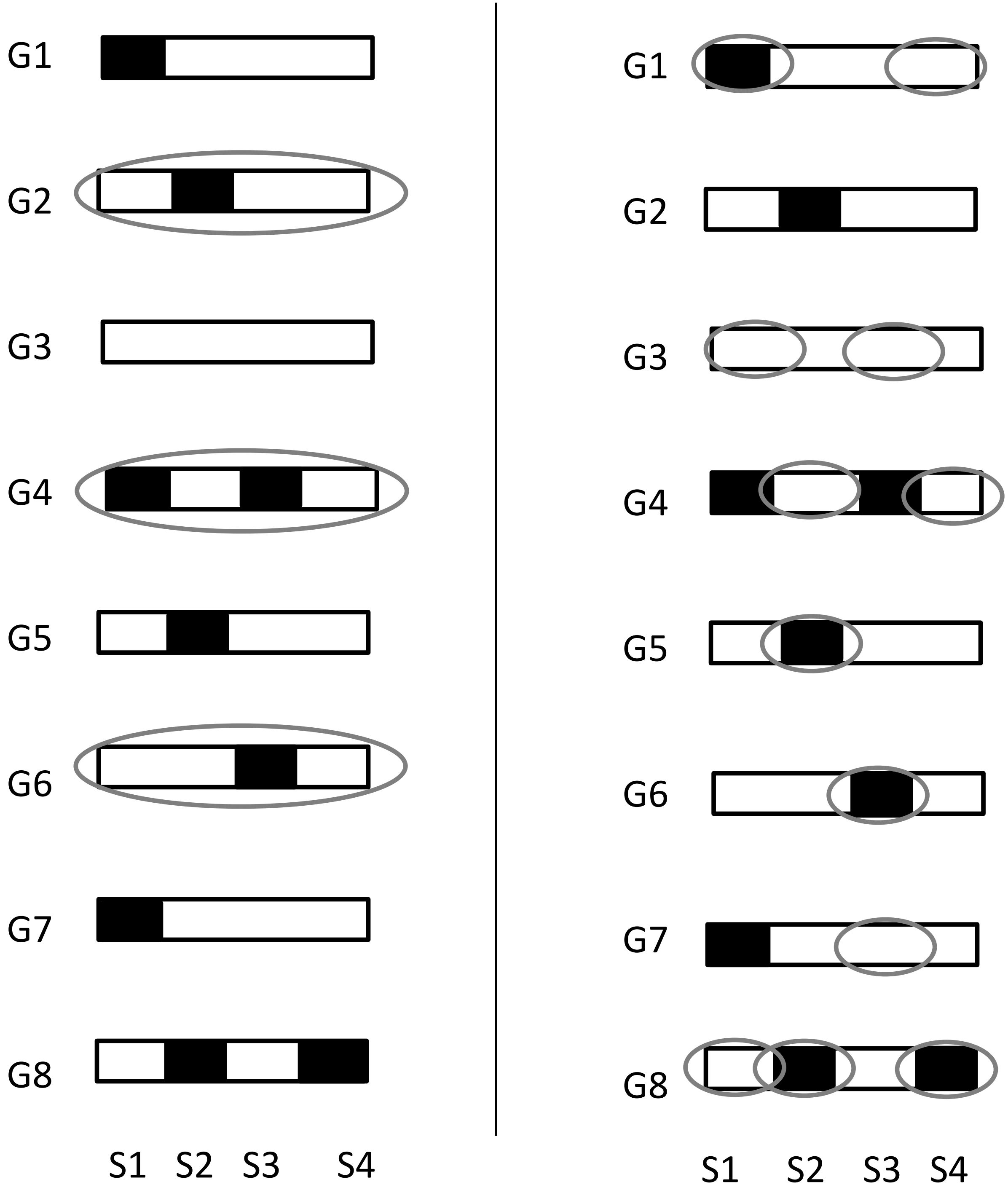

Figure 1

Figure 2a: Variant Allele Freqs

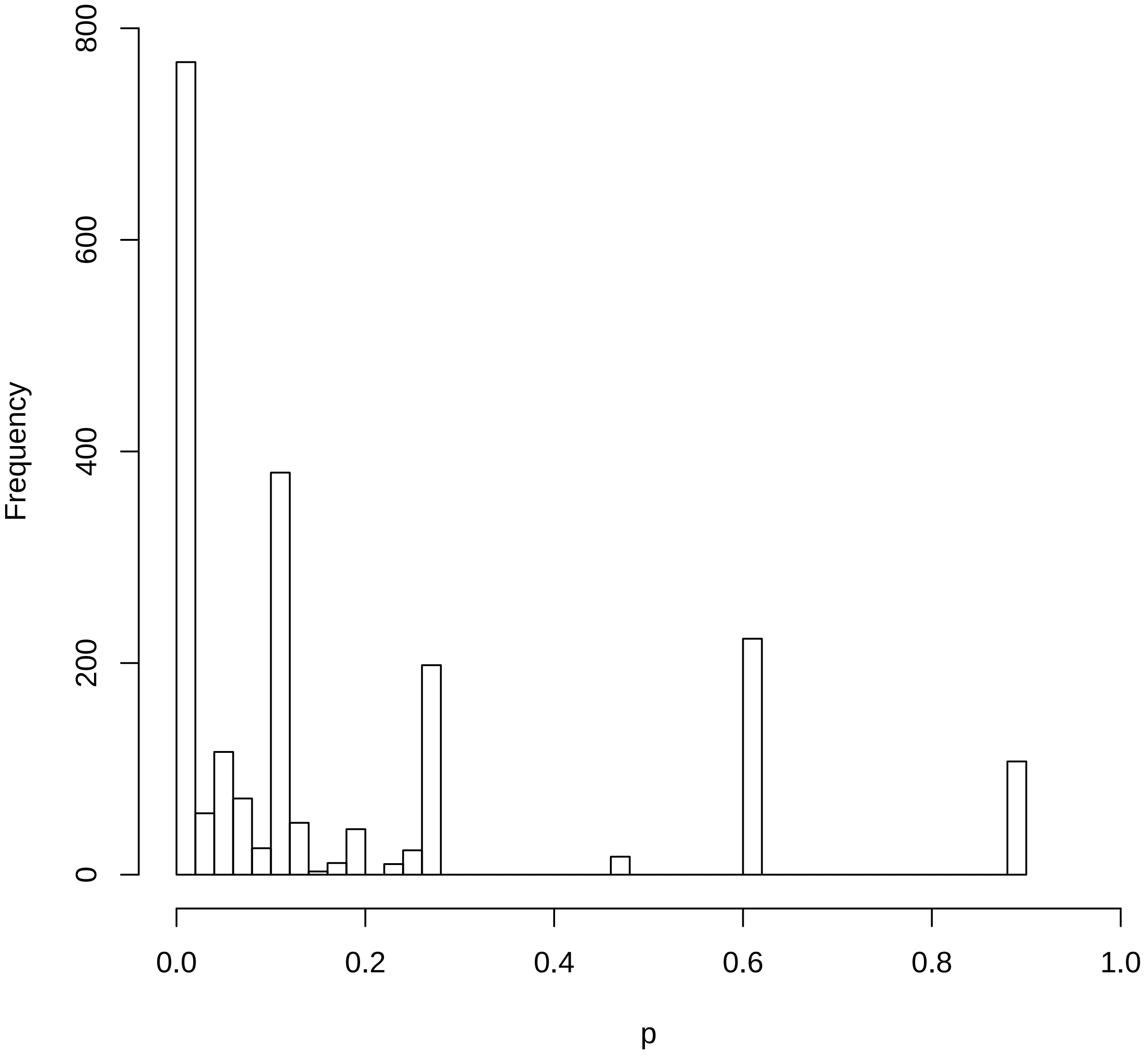

Figure 2b: Dsr

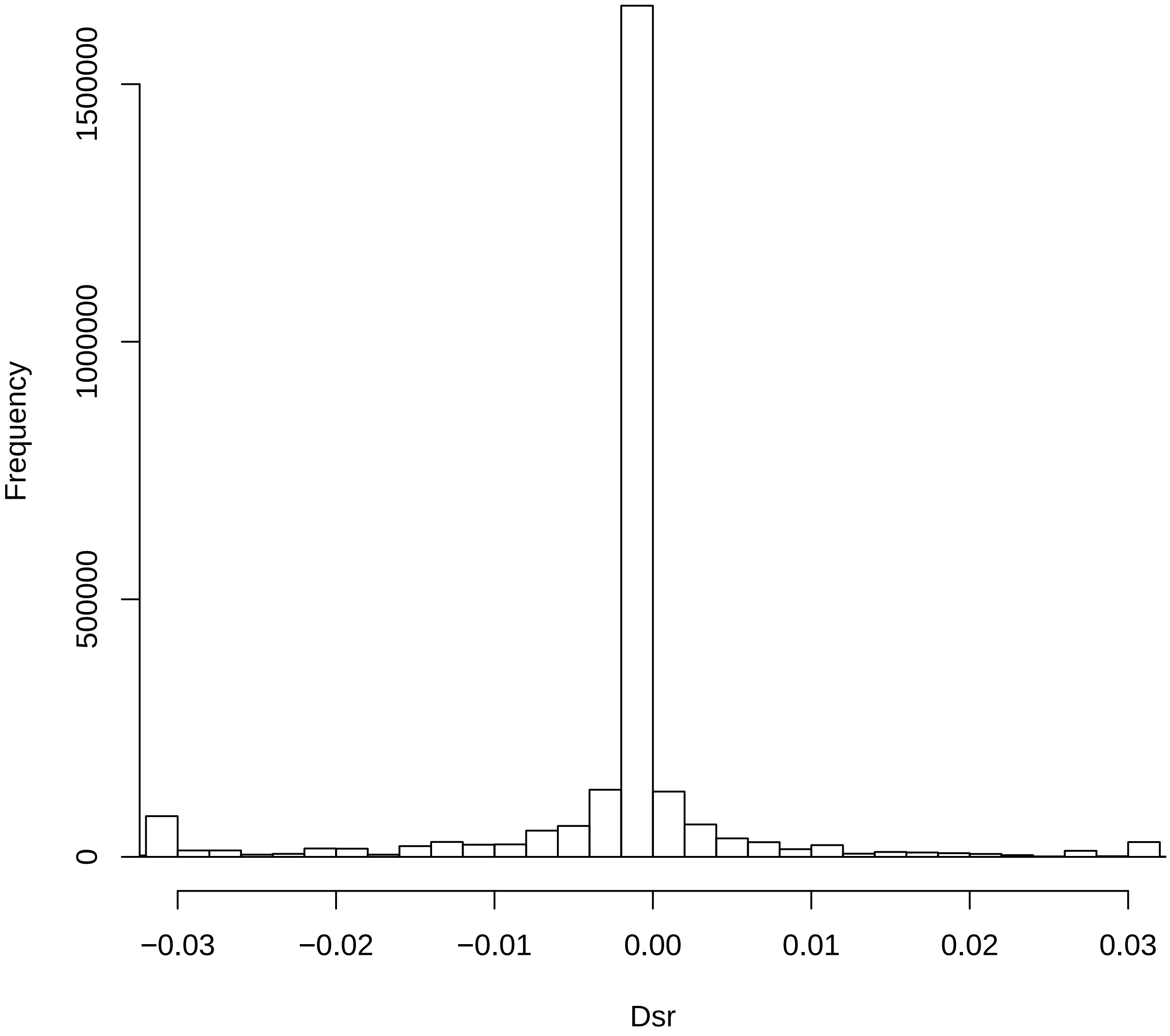

Figure 2c: Asr

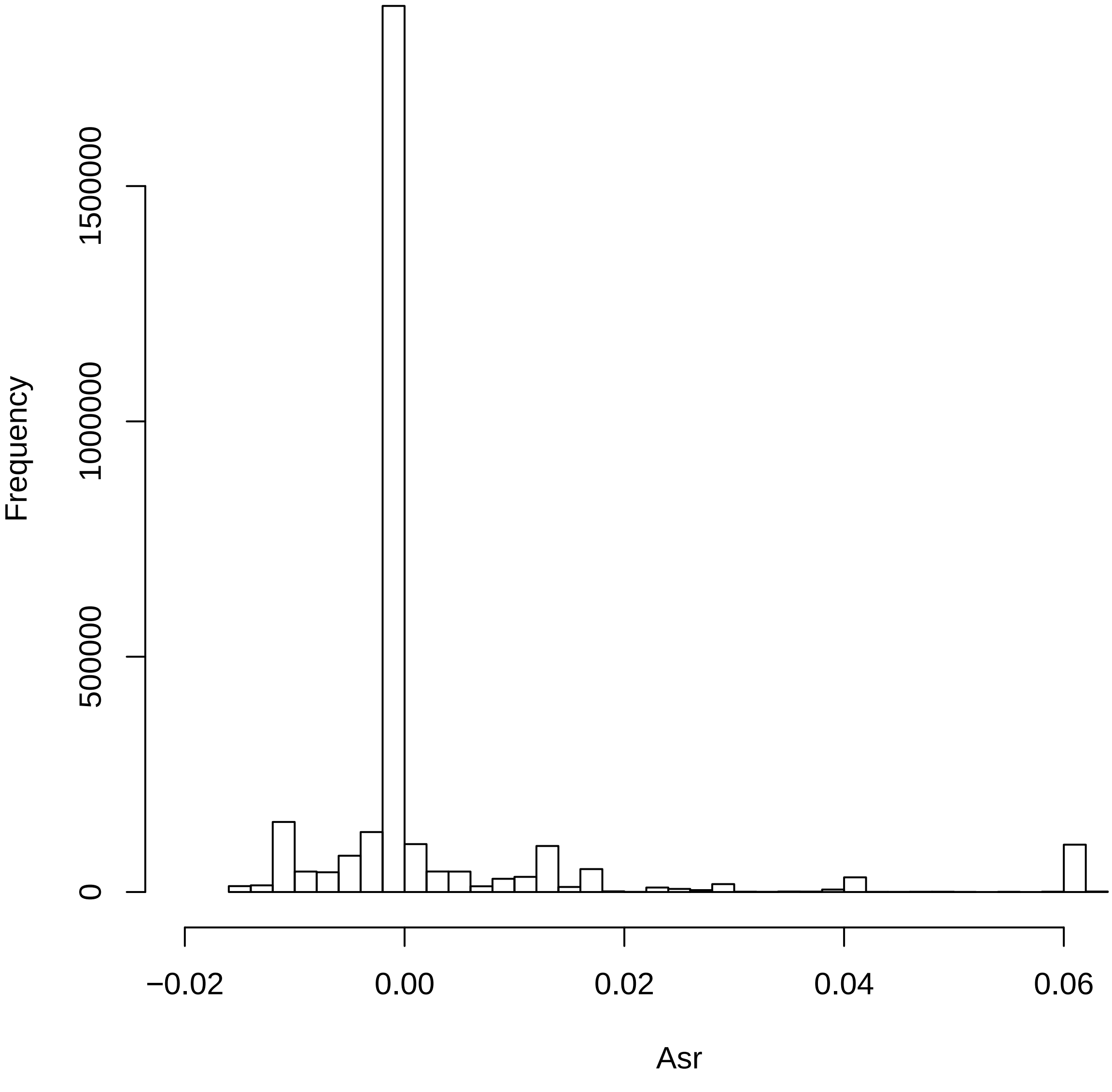

Figure 3

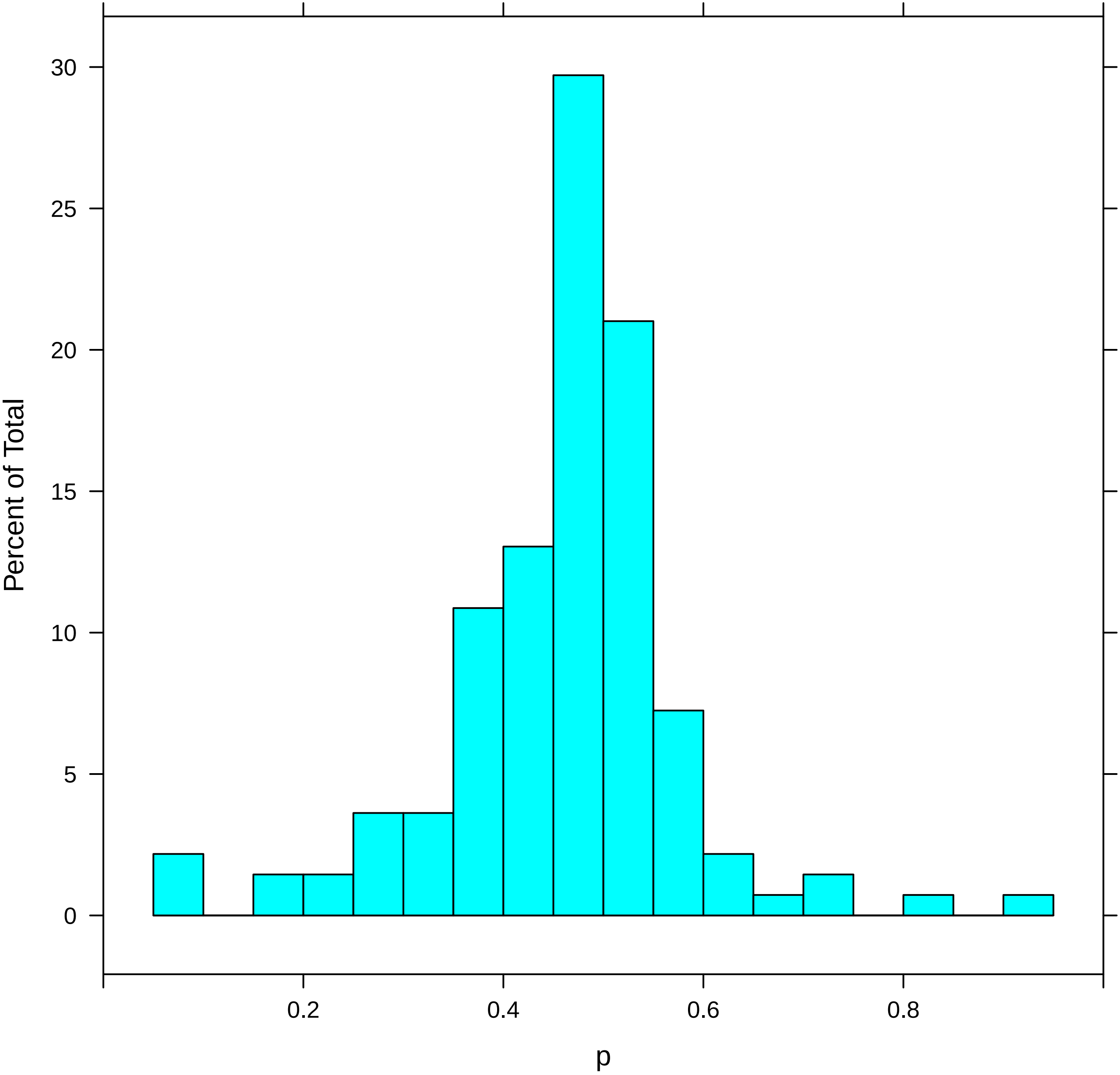

Table 1a

***N*=200**	$\boldsymbol{T}$	$\boldsymbol{S_N}$	$\overline{\boldsymbol{p}}'$	$\widehat{\boldsymbol{\pi}}_{\boldsymbol{HCS}}$	$\widehat{\boldsymbol{\pi}}_{\boldsymbol{LCS}}$	var_{HCS}	var_{LCS}
	5	112.7	0.012	2.94	2.94	0.249	0.241
	10	181.5	0.016	5.82	5.81	0.468	0.476
	20	250.1	0.024	11.47	11.47	0.878	0.819
	50	351.2	0.043	26.79	26.80	2.27	1.58
	800	770.6	0.170	115.59	114.58	272.0	3.70
	1000	847.7	0.178	122.90	122.77	415.7	3.97

Table 1b

N=500	T	S_N	$\bar{p}'$	$\hat{\pi}_{HCS}$	$\hat{\pi}_{LCS}$	var_{HCS}	var_{LCS}
	5	308.4	0.0049	2.96	2.96	0.279	0.271
	10	455.9	0.0066	5.97	5.97	0.537	0.543
	20	616.9	0.0096	11.61	11.60	1.04	1.04
	50	875.5	0.0172	28.61	28.68	2.53	2.27
	100	1078.3	0.0281	54.67	54.67	6.23	3.71
	2000	2202.4	0.144	301.16	301.22	1532.1	9.09
	2500	2395.1	0.153	316.06	315.64	2089.8	10.53

Table 2a

N=200	*T*	$\sum A_{sr}$	Δ_P	Δ_S	$SE(\Delta_S)$
	5	4.18×10^{-3}	-9.57×10^{-4}	8.01×10^{-3}	4.58×10^{-3}
	10	0.0997	0.0129	-7.85×10^{-3}	0.012
	20	0.0529	0.0482	0.0587	0.0226
	50	1.14	0.766	0.687	0.0927
	800	660.5	297.96	268.34	44.56
	1000	1009.0	444.94	411.68	51.51

Table 2b

N = 500	*T*	$\sum A_{sr}$	Δ_P	Δ_S	$SE(\Delta_S)$
	5	$4.58\text{x}10^{-3}$	$2.13\text{x}10^{-3}$	$8.08\text{x}10^{-3}$	$3.00\text{x}10^{-3}$
	10	-0.0242	$-7.92\text{x}10^{-3}$	$-6.23\text{x}10^{-3}$	$7.53\text{x}10^{-3}$
	20	0.393	0.00	0.0269	0.0213
	50	0.269	0.256	0.259	0.0546
	100	4.35	2.74	2.52	0.182
	2000	3362.8	1606.1	1523.0	213.90
	2500	4871.9	2241.3	2079.3	273.37

Table 3a

Pr=0.40	**S**	$\bar{p}$	$\bar{D}_{sr}$	$\sum A_{sr}$	**Δ**
Sample 1	8	0.492	0.223	0.321	0.045
Sample 2	8	0.423	0.555	0.225	0.034
Sample 3	10	0.457	0.408	0.380	0.054
Sample 4	6	0.328	0.500	0.510	0.070

Table 3b

Pr=0.49	**S**	$\bar{p}$	$\bar{D}_{sr}$	$\sum A_{sr}$	**Δ**
Sample 1	42	0.753	-0.713	-1.951	-0.653
Sample 2	56	0.760	0.0352	-3.077	-1.040
Sample 3	46	0.754	-0.0907	-1.998	-0.653
Sample 4	56	0.759	-0474	-2.422	-0.778

Table A2

N=200	*T*	$\sum A_{sr}$	Δ_P	Δ_S	$SE(\Delta_S)$
	5	2.82×10^{-3}	5.76×10^{-4}	3.71×10^{-4}	0.071
	10	0.0541	0.0250	0.0242	0.014
	20	0.0181	0.0696	0.0704	0.034
	50	1.232	0.792	0.620	0.080
	800	754.33	341.14	296.18	42.59
	1000	827.67	373.24	337.70	57.88