Estimation of Pairwise Genetic Distances Under Independent Sampling of Segregating Sites vs. Haplotype Sampling

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

[1] Sarah Cannon Research Institute, Austin TX 78705, 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 Science, 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

February 14, 2017

Contents

Running Title : Sample variance of genetic distance

keywords : genetic distance, linkage disequilibrium, Tajima's estimator, cancer genomics, 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

Genetic distance is a standard measure of variation in populations. When sequencing genomes individually, genetic distances are computed over all pairs of multilocus haplotypes in a sample. However, when next-generation sequencing methods obtain reads from heterogeneous assemblages of genomes (e.g. for microbial samples in a biofilm or cells from a tumor), individual reads are often drawn from different genomes. This means that pairwise genetic distances are calculated across independently sampled sites rather than across haplotype pairs. In this paper, we show that while the expected pairwise distance under whole haplotype sampling (WHS) is the same as with independent locus sampling (ILS), the sample variances of pairwise distance differ and depend on the direction and magnitude of linkage disequilibrium (LD) among polymorphic sites. We derive a weighted LD value that, when positive, predicts higher sample variance in estimated genetic distance for WHS. Weighted LD is positive when on average, the most common alleles at two loci are in positive LD. Using individual-based simulations of an infinite sites model under Fisher-Wright genetic drift, variances of estimated genetic distance are found to be almost always higher under WHS than under ILS, suggesting a reduction in estimation error when sites are sampled independently. We apply these results to haplotype frequencies from a lung cancer tumor to compute weighted LD and the variances in estimated genetic distance under ILS vs. WHS, and find that the the relative magnitudes of variances under WHS vs. ILS are sensitive to sampled allele frequencies.

1 Introduction

Genetic variation is the raw material for evolutionary change, consequently, one of the defining empirical questions in evolutionary and population genetics is the measurement of genetic heterogeneity in natural and experimental populations (Lewontin et al., 1974; Ellegren and Galtier, 2016). In addition to its importance to furthering our basic understanding of the evolutionary process (Hansson and Westerberg, 2002), characterization of genetic variation has applied significance in many endeavors relevant to human welfare, including biomedical research. For example, the extent of genetic variation in populations of pathogens can be predictive of their ability to adapt to antibiotic treatment (Martinez and Baquero, 2000; MacLean et al., 2010) and immune response, while genetic heterogeneity among populations of cancer cells is predictive of their potential for metastatic disease and of a tumor's ability to develop resistance to chemotherapies (Dexter and Leith, 1986; Burrell et al., 2013; Sun and Yu, 2015). Estimates of genetic variation are equally relevant to maintaining diversity in crop and livestock strains (National Research Council, 1993; Fu, 2015) and to the maintenance of viable populations in biological conservation (Van Dyke, 2008).

The recent advent of high-throughput technologies for DNA sequencing allows researchers to measure genetic variation within and among populations with very large sample sizes and high statistical power. The methods developed for characterizing genetic variation in studies of multicellular, usually sexually reproducing model organisms can now be applied to genomic studies of typically clonal unicellular organisms such as microbes growing on biofilms, to populations of genetically heterogeneous cancer cells in a tumor, or to viruses in serums. In many cases, both the underlying population genetic models and the descriptive statistics used to measure genetic variation

in microbial or tumor samples must be adjusted to take into consideration the biological characteristics of the populations under study (such as the absence of meiotic recombination), as well as for the statistical properties of what are often different methods of sampling.

One of the most widely used measures of genetic variation in a population is the mean pairwise genetic distance among genomes, which is an estimate of the total heterozygosity across all polymorphic sites. For a sample of n genotypes, the mean pairwise distance is calculated as:

$$\widehat{\pi}_1 = 2\sum_{i,j} \pi_{ij}/n(n-1), \tag{1}$$

where $\pi_{i,j}$ is the Hamming distance for the haplotype pair z_i,z_j, summed over all polymorphic sites in haplotypes i and j, i.e. $\pi_{ij} = \sum_s f(z_{is}, z_{js})$ for $f(z_{is}, z_{js}) = 1$ if site s has different nucleotides in haplotypes z_i, z_j and 0 otherwise.

The parameter $\widehat{\pi}$ is of importance not only as a summary statistic of genetic variation, but as an estimator of key population genetic parameters. Under neutral evolution in an infinite sites model (Kimura, 1969; Tajima, 1996), $\widehat{\pi}$ estimates the population mutation rate (Tajima, 1989), i.e. for a diploid population with N individuals and a per-generation genomic mutation rate u,

$$E[\widehat{\pi}] = 4Nu = \theta_\pi,$$

where θ_π represents the distance-based Tajima estimator for this parameter (in a population of N haploids, $\theta = 2Nu$). As a result, $\widehat{\pi}$ provides an estimate of neutral effective population size in populations when the mutation rate u is known approximately. Additionally, comparisons of θ_π estimated

from genetic distances to the population mutation rate estimated from the number of segregating sites S_n in a sample of n genotypes

$$S_n/(\sum_{i=1}^{n-1} 1/i) = \theta_S,$$

(Watterson, 1975) is the basis for the Tajima D test for selection. Values of θ_π that are inflated relative to θ_S may be the result of diversifying selection or recent population bottlenecks, while values of $\widehat{\pi}$ that are smaller than the value expected from the number of polymorphic sites indicate a history of selective sweeps or, alternatively, a recent population expansion with a relative large number of recent, rare variants. Understanding the error in estimates of $\widehat{\pi}$ relates directly to the error inherent to estimates of population mutation rate θ and tests for neutral evolution derived from this parameter.

In studies of multicellular organisms, $\widehat{\pi}$ is estimated directly from complete haplotypes sampled from n different individuals, each of which has been sequenced over the region(s) containing the segregating sites of interest, i.e. $\widehat{\pi}$ is computed from the Hamming distances among actual haplotype pairs. This pairwise comparison of genotypes across multiple loci is possible because the co-occurring genotypes across sites in the genome are known for individually sequenced genomes. In contrast, for most samples of microbes or of cancer cells, the application of next generation sequencing (NGS) methods (Goodwin et al., 2016) entail sampling reads from an unknown number of different genomes (in contrast to multicellular tissue samples from individual organisms, which are assumed to be genetically homogeneous). In the limiting case, if the read coverage depth at each segregating site is sufficiently small relative to the number of individual genomes in a sample, every read is likely to be drawn from a different individual cell and geno-

type (assuming non-adjacent segregating sites that occur on separate reads). Consequently, sampling in this way for a read depth of n is not statistically equivalent to sequencing n individuals at the same number of sites. The estimated mean pairwise genetic distance for independent sampling of loci from different genomes is:

$$\widehat{\pi}_2 = 2\sum_s \sum_{i_s,j_s} f(z_{i_s,s}, z_{j_s,s})/n(n-1), \tag{2}$$

where $z_{i_s,s}$ is the identity of the ith allele sampled at locus s, which is assumed to be from a different genome (distinct cell or organism) with respect to the ith sample at some other site (in contrast to z_{is} in Eqn. (1), which represents site s in haplotype i). We include the subindex $_s$ in i_s, j_s to highlight this. As in Eqn. (1), $f(x, y)$ is an indicator function equal to 1 if the nucleotide pair is not identical and 0 otherwise.

Throughout this paper, we will refer to these two modes of genotype sampling as as whole haplotype sampling (WHS) and as independent locus sampling (ILS), respectively. The difference between WHS and ILS is illustrated schematically in Figure 1.

FIGURE 1 HERE

Although WHS is usually used to estimate genetic distances when sequencing multicellular organisms while ILS is standard for assemblages of microbes or tumor cells, one can apply single cell sequencing (corresponding to WHS) to microbe and tumor cells (Navin, 2015; Gawad et al., 2016). It is also possible to sample loci via independent reads from different genomes (ILS) in multicellular organisms if one sequences sufficiently many individual

organisms (i.e. more individuals genotyped than there are reads), although it usually isn't practical to do so. Therefore, it is instructive to compare the sampling distributions of $\hat{\pi}$ obtained for WHS and ILS. Even in cases where only either ILS or WHS are practically feasible, it is important to understand the potential sources of error in estimates of genetic distance given the type of sampling being used. The sample variances of $\hat{\pi}$ under ILS vs. WHS are of particular significance, as they determine the expected error in our point estimates of genetic distance in a population, and by extension, the reliability of test statistics for the consequences of natural selection or population dynamics such as Tajima's D.

Below, we will derive the expectations and sample variances of pairwise genetic distance under the two modes of sampling with the fewest possible a priori assumptions about the number and distribution of mutations. We test these analytical predictions against samples from simulated populations undergoing neutral evolution via random mutation and Fisher-Wright genetic drift under an infinite-sites model. We also apply these results to estimating the variances in genetic distances using single nucleotide variant (SNV) frequency data from lung cancer tumors.

2 The sampling models

Consider a population of N organisms with some distribution of mutations over S segregating sites (in the population, as opposed to $S_n \ll S$ in a sample of n). We wish to estimate the mean genetic distance $\hat{\pi}$ for the population and its sample variance $var(\hat{\pi})$ under the WHS and ILS models of sampling. For WHS, 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 simplification but without

loss of generality, assume that the sample consists of n haploid genotypes or known/phased haplotypes, regardless of how they were sequenced or the number of reads (we note that if we were working with diploid genotypes, phasing would not matter if pairwise distances are computed with respect to the per-site genotype).

For an idealized model of ILS in an aggregate sample of microbes or cells, we assume that the number of individual genomes (i.e. from different tumor cells or microbes) that contribute reads to a sample is much larger than the sequencing read depth (mean coverage depth) n. If this is the case, we can assume (approximately) that the majority of reads are sampled different individual genomes. If we make the further assumption that reads are short, the majority of reads will contain at most a single polymorphic site. Together, these conditions imply that the majority of polymorphic sites will be sampled from different genomes, or, more precisely, each polymorphic site is sampled independently of other polymorphic sites with respect to their genome of origin (in the second panel of Figure 1, several sites are sampled from the same genome simply because there are very few genomes to draw this random sample from). When computing average pairwise genetic distance, WHS sums over the Hamming distances of all haplotype pairs, while ILS is the sum over all pairs for each of the S_n segregating sites sampled from different individuals.

Without loss of generality, we also assume an infinite sites model so that there only two alleles per segregating site. This allows an unambiguous binary classification of alleles, with mutations as ancestral "wildtype" vs. "reference" genotype (in the case of tumors, the reference corresponds to the normal germline genotype, with somatic mutations defining the variant genotypes of the clonal lineages), and to specify the direction of linkage dis-

equilibrium. We note, however, that the results derived below are applicable to multiallelic states provided that some allele (usually the most common, or, in the case of cancer genomics, the germline allele) is designated as a reference and all other alleles are pooled together to create an aggregate biallelic state.

Definitions. In this subsection and throughout the manuscript, we will make use of the following definitions and terminology as a formal way of characterizing and distinguishing between Eqns (1) and (2) in the introduction:

Variables: Let z denote a genotype, at either 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 $i = 1...n$. As above, we use z_{is} to denote site s in haplotype i (for WHS), and $z_{i_s,s}$ to denote sample i at site s when sites are sampled independently (ILS).

Pairs: In both cases, that is, for WHS and ILS, respectively, the estimators $\widehat{\pi}_1$ and $\widehat{\pi}_2$ 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 ILS $x_i = z_{is}$ and $\phi(x_i, x_j) = f_{ijs}$ with $f_{ijs} = I(z_{is} \neq z_{js})$(and an additional sum over s, outside the average). In the case of WHS the random variables are $x_i = \boldsymbol{z}_i$ and $\phi(x_i, x_j) = g_{ij} = \sum_s f_{ijs}$. Importantly, while the r.v.'s x_i are independent, pairs (x_i, x_j) and (x_i, x_k) that share a common element are not.

Moments of ϕ_{ij}: We define $E(\phi_{ij}) = \mu$, $var(\phi_{ij}) = \sigma^2$. We also define an

expectation for an indicator function on pairs of pairs with a shared element as $E(\phi_{ij}, \phi_{jk}) = \kappa$.

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

$$\begin{aligned} N_2 &= n(n-1)/2 \\ N_3 &= n(n-1)(n-2)/2 \end{aligned}$$

The value of N_3 follows from the fact that there are $n(n-1)(n-2)/6$ ways to select a triplet i, j, k, and three ways to select a shared element from this triplet. In Appendix A1, we cover some of 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 ILS and WHS,

$$var(\widehat{\phi}_n) = \frac{\sigma^2}{N_2} + 2\frac{N_3}{N_2^2}(\kappa - \mu^2), \tag{3}$$

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 ILS with $\phi_{ij} = f_{ijs}$, and once for WHS with $\phi_{ij} = g_{ij}$.

2.1 Case 1: Independent Locus Sampling (ILS)

For ILS, 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. $p(z_{i_s,s}) = p_s$ for $z_{i_s,s} \in 0, 1$ such that

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

(note that h_s is the heterozygosity at locus s).

The expectation of the indicator function for ordered pairs on pairs includes a covariance term, namely,

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

The sampling estimator for $\widehat{\pi}$ under ILS is given by

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

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

$$var(\widehat{\pi}_{ILS}) = \sum_s var(\widehat{f}_{ns}) = \sum_s \frac{1}{N_2} h_s \left\{ (1-h_s) + \frac{N_3}{N_2}(1-2h_s) \right\} \qquad (4)$$

We remark that in practice, the assumption of independence 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: Whole Haplotype Sampling (WHS)

Computing pairwise differences for independent samples of $z = z_i$ under WHS involves computing moments of sums rather than sums of moments, i.e.

$$g_{ij} = \sum_s I(z_{is} \neq z_{js}) = \sum_s f_{ij,s}$$

For samples of individual haplotypes $i = 1...n$, consider $z_i \sim p(z)$ with $p(z_{is} = 1) = p_s$ as before, but with correlated z_{is}, z_{ir} due to linkage disequilibrium (LD) between sites, i.e. for (arbitrarily labeled) alleles R, r and S, s at the two sites, and defining $q_s, q_r = 1 - p_s, 1 - p_r$ (Lewontin and Kojima, 1960),

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

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

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

With non-zero LD, 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 $(RS, rs) = (z_{i,sr} = RS, z_{j,sr} = rs)$ etc. Therefore

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

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

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

Assuming independence (linkage equilibrium, $D_{sr} = 0$ for all s, r) gives results equivalent to ILS, 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 can construct the sample estimate and variances for g_{ij}. For the WHS model, $\mathbf{z}_i \sim p(\mathbf{z})$, independently, from which we construct

the sample estimate for g_n as:

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

now averaging over haplotypes $\mathbf{z}_i$ (rather than independent counts for each site).

Note that $\widehat{g}_n$ is again an average across pairs, like $\widehat{f}_n$ in the ILS case. We again apply the result in Eqn. (3) to find

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

$$\frac{1}{N_2} \underbrace{\left(\sum_s h_s(1-h_s) + 2\sum_{r<s}(\gamma_{sr} - h_s h_r) \right)}_{\sigma^2} + \frac{2N_3}{N_2^2} \left[\underbrace{\sum_s h_s/2 + 2\sum_{r<s} \delta_{sr}}_{\kappa} - \left(\sum_s h_s \right)^2 \right] \tag{5}$$

2.3 Difference and independence

Using the results in Eqns. (4) and (5), we derive the difference between the sample variances in pairwise differences under WHS vs. ILS as

$$\Delta = var(\widehat{\pi}_{WHS}) - var(\widehat{\pi}_{ILS}) = \frac{2}{N_2} \sum_{r<s} (\gamma_{sr} - h_s h_r) + \frac{4N_3}{N_2^2} \sum_{r<s} (\delta_{sr} - h_s h_r) \tag{6}$$

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

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

For notational convenience, we define:

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

In the absence of 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 WHS and ILS are equal. Otherwise, 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 provided that the sum of weighted linkage disequilibria A_{sr} is positive, i.e.

$$\sum_{sr} A_{sr} = \sum_{sr} (2p_s - 1)(2p_r - 1) D_{sr} > 0. \tag{7}$$

While $E[A_{sr}] > 0$ is a sufficient condition for $\Delta > 0$, it is not a necessary condition. In fact, the variance in mean pairwise distance under ILS may in some cases still be lower than under WHS even for $E[A_{sr}] < 0$. This follows 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, for large sample sizes, the A_{sr} term dominates because it scales as $\sim 1/n$ while the B_{sr} term scales as $\sim 1/n^2$, which means that for many practical cases the sign of $E[A_{sr}]$ predicts that of Δ.

In order to have $E[A_{sr}] > 0$, it is required that on average A_{sr} is positive, i.e. that for most pairs of loci s, r, the "major" alleles (those with $p_s, p_r > 0.5$) are in positive LD, while major and minor allele pairs ($p_s > 0.5, p_r < 0.5$ 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. Our results predict that when the mean weighted LD is positive, the sample variance (error) in estimated pairwise genetic distance will be lower under ILS than under WHS.

2.4 Implications

To understand the conditions under which $\Delta > 0$ holds, we consider the distribution of allele frequencies and pairwise LD under different evolutionary scenarios. Specifically, we ask whether positive weighted linkage disequilibria (the conditions in Eqn. (7)) are general enough to assume that ILS generally leads to a reduced error in estimated genetic distance relative to WHS.

Consider a population undergoing random mutation under an infinite sites model and Fisher-Wright genetic drift in a finite population. At an equilibrium of new alleles acquired via mutations and those lost by genetic drift, the expected number of sites η_k that have k copies of a mutant allele is

$$E[\eta_k] = \theta/k,$$

(Watterson 1975, see also e.g. Ewens 2004 Ch 9, Ch. 2 in Durrett 2008), so that the expected frequency of alleles occurring as k-tuples is $\theta/(S_N k)$.

Because of this harmonic relationship, the majority of mutant alleles in a population are represented as singletons and as other small k-tuples (e.g. $k = 2, 3$, etc). This is consistent with a majority of alleles in the population being rare and of recent origin, with variant allele frequencies close to $p \sim 1/N$. These rare alleles of recent origin are usually lost from the population, while a much smaller subset of alleles in the sample have frequencies $p > 0.5$ and consequently a high probability p of eventual fixation in the population. As a result, for the majority of variant allele pairs in a sample, we have $p_s, p_r \ll 0.5$.

In the absence of recombination, multilocus haplotypes behave as alleles at a single locus, so that the infinite sites model becomes effectively an infinite alleles model (Tajima, 1996). Therefore, every new mutation is in positive LD with the other variant alleles with which it co-occurs and in negative LD with non-co-occurring mutations on other haplotypes. We consider the following scenarios: A) LD among rare, typically non-co-occurring alleles on different haplotypes, B) LD between rare and common alleles when a recent mutation appears on a common haplotype as a genetic background and C) co-occurrence of common alleles on dominant haplotypes.

A) Following the Watterson distribution of k-tuples at equilibrium, there are a large numbers of rare alleles $p_s, p_r \sim 1/N$. However, most of these rare alleles do not co-occur with one another, consequently $P(sr) \sim 0$ and $D_{sr} \sim -1/N^2$. B) Rare alleles typically appear against a background of common haplotypes or subclones defined by high-frequency variant alleles. If we have a recent mutation with frequency $p_s \sim 1/N$ appearing on a background of common alleles at other loci $p_r \sim 0.1$, then the LD $D_{sr} \sim p_s - p_s p_r > 0 \sim 1/N$, because the frequency of the $P(sr)$ haplotype is p_s. By symmetry, new alleles that happen to co-occur on rare haplotypes

will have $D_{sr} \sim -1/N$ with respect to the sites on their genetic background, but there will be an order of magnitude fewer such associations because the majority of new mutations will appear against a genetic background of common haplotypes. C) Similarly, common alleles that co-occur on dominant subclones have $D_{sr} = p_s - p_s p_r \sim 0.1$ (where dominant haplotypes, and therefore allele frequencies can potentially be $p_s, p_r > 0.5$.

These heuristic considerations of scale suggest that the distribution of D_{sr} in the clonal population will be highly skewed, consisting of large numbers of negative but near-zero LD values for the many $p_s \sim 1/N$ rare alleles, and a smaller number of large positive associations associated with mutations defining the common haplotypes. This conclusion is consistent with the highly skewed sampling distributions of D_{sr} for non-recombining loci computed numerically in Golding (1984), i.e. large numbers of weakly negative associations and a small number of high positive LD.

Because of this skew, we hypothesize that populations where the allele frequency distributions are in approximate equilibrium under mutation and drift will have positive $E[A_{sr}]$ and therefore higher sample variance in pairwise genetic distance under WHS than under ILS. In contrast, among populations where all allelic variation is of recent origin and characterized by low frequencies (such as in newly emergent tumors, or in populations that have experienced recent bottlenecks), the negative associations $D_{sr} < 0$ will dominate the distribution due to the fact that recent mutations will initially occur on different reference haplotypes. Because most mutations will occur on disjoint branches of the genealogy, very few haplotypes with significant numbers of co-occurring mutations will have attained high enough frequencies to offset the small magnitude but negative LD values. Therefore, it is possible to observe $E[A_{sr}] < 0$ (albeit with $|E[A_{sr}]|$ and $\Delta \sim 0$)

in populations where most variant alleles occur at near zero frequencies.

We will assess these heuristic predictions about the sign and magnitude of $E[A_{sr}]$ and Δ under different frequency distributions of p and D_{sr} through simulations of mutation and genetic drift for a range of population parameters.

3 Comparison to individual-based simulations

To simulate Fisher-Wright genetic drift in an infinite sites model, we initialized a population of N haploid genotypes characterized by $K = 10^8$ sites with reference genotypes (all alleles set to 0 value, to distinguish them from variant mutations set to 1). 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 progeny was $m \sim Poiss(Ku)$, with the mutations randomly distributed among the K sites. This process was iterated over T generations; in order to approximate a distribution of mutation frequencies near equilibrium, we chose $T \sim 4N$ (because expected coalescent time for all N haplotypes in a population is $E[T_C] = 2N$). In addition, simulations were run for a range of values $T < N$ for comparison to non-equilibrium distributions of allele frequencies and pairwise LD. For each combination of parameters, the simulation cycle was run over 100 replicates.

In order to simulate WHS sampling, n haplotypes were randomly selected without replacement from the model population. The Hamming distances were calculated for all pairs in a sample, while variant allele frequencies and linkage disequilibria were calculated for all individuals and all pairs in the model population. ILS sampling was simulated by selecting n alleles without replacement at every segregating site, summing pairwise distances over all

sites (this can be thought of as sampling with replacement with respect to genomes, but without replacement with respect to each locus). Δ was estimated as the difference in the sample variances between the WHS and ILS pairwise distances. For each simulation replicate, A_{sr} was calculated from the mutation frequencies p_s, p_r and from D_{sr} using Eqn. (6). All simulations were implemented using Python 2.7.3, the code is available from the corresponding author upon request.

Simulation output for population sizes $N = 200, 500$, a sample size of $n = 20$ and a range of generation times T are summarized 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 WHS and ILS (for $n = 20$), as well as their respective sample variances over 100 replicates.

TABLES 1 and 2 HERE

For small time intervals $T < N$, there are few (~ 100) polymorphic sites, all of which are characterized by low variant allele frequencies. Consequently, the mean and variance of genetic distances are of the order ~ 1, ~ 0.1, respectively. For $T \sim 4N$, allele frequencies and genetic distances tend towards the equilibrium values predicted under the neutral infinite sites model, e.g. the estimated pairwise genetic distance $\hat{\pi}$ converges to the Tajima estimator for haploids $\theta = 2Nu$, which is $\hat{\pi} = 150, 300$ for $N = 200, 500$, respectively. Table 2 shows the population mean LDs $\bar{D}_{sr}$ and the sum of weighted LD values $\sum A_{sr} = \bar{A}_{sr} N_2$. We remark that while the mean values of LD are

effectively 0 even for large values of T and S_N, this is not due to individual LD values being near 0. Rather, $\bar{D}_{sr} \sim 0$ is the result of large numbers of positive and negative LD values with high absolute value, as can be seen from the large magnitudes of the summed weighted LD. Figures 2a and 2b show frequency distributions of pairwise LD and weighted linkage LD for a representative model population.

FIGURE 2a-b HERE

Using $\sum A_{sr}$, we compute the predicted difference between WHS and ILS variances Δ_P from Eqn. (6). This predicted value is compared to the simulation estimate $\Delta_S = var_{WHS} - var_{ILS}$. The close correspondence between 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 time intervals (i.e. as the population distribution of allele frequencies and pairwise LD approach equilibrium), in part because of the much larger number of polymorphic sites and the higher frequency of variant alleles at those sites.

With the exception of populations where there are very few mutations and where weighted LD values are very close to 0, we have $\Delta > 0$ for most of the simulated populations. These results conform to our hypothesis that the error in genetic distance estimates based on WHS will be greater than those for ILS for the majority of natural and model populations. The reduction of error through ILS 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 sufficiently 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 shown in Table 2 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 $T = 10$ at both simulated population sizes, with a negative predicted value Δ_P for $N = 500$ (though not for $N = 200$). In these cases, 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 rather restricted conditions corresponding to very small absolute values of Δ and negligible reduction of error in estimating $\widehat{\pi}$ through either WHS or ILS, while for large numbers of segregating sites and increasing allele frequencies, there can be considerable increases in error when $\widehat{\pi}$ is estimated via WHS rather than ILS.

4 Analysis of cancer sequence data

We apply the results of our derivations and numerical analyses to genomic data by estimating $\sum A_{rs}$ and Δ to haplotype frequencies estimated from a lung adenocarcinoma tumor sequence data. The data was obtained from 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 using 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 coverage (depth) of 60-70 fold across sites. Variant calls were performed using GATK (McKenna et al., 2010). The unpublished data were provided to the authors as summaries of variant frequencies and haplotypes by K. Gulukota and Y. Ji.

Through the matching of read ends, somatic mutations co-occurring within $\sim$ 100 bp in single genomes were identified (Sengupta et al. 2015, unpublished). These mutation pairs define two locus haplotypes that can be tallied without the need of phasing. This allows us to estimate the frequencies of haplotypes defined at two adjacent loci directly from the read counts, along with individual allele frequencies. Following the terminology of this paper, while non-adjacent polymorphic sites are sampled as (effectively) ILS, adjacent sites are effectively sampled as whole haplotypes. 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 came from more distant sites or on different chromosomes. The adenocarcinoma data contain estimated frequencies of 69 haplotypes defined by variant alleles at two sites on a single read, and allele frequencies for a total of 138 sites (comparable to the number of somatic mutations identified in the exomes of lung adenocarcinoma and other cancer types, e.g. TCGA 2014, Hoadley et al. 2014). The provided haplotype data is used to determine how the LD values and allele frequencies would effect the error in estimation of $\widehat{\pi}$ for this data set under WHS vs. ILS sampling.

A naive application of Eqn. (6) to the distribution of mutation frequencies and LD values gives $\Delta \sim 0.1$ for $n = 65$, suggesting lower error in $\widehat{\pi}$ estimates from ILS for this data. However, several aspects of cancer genetics complicate this estimate. First, because cancer cells reproduction is

clonal, 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. Variant allele frequencies must be rescaled to reflect these considerations.

Figure 3 shows the distribution of mutant allele frequencies in Sample 1, note the high frequency of values near $p = 0.5$, and the fact that this distribution is not consistent with an equilibrium neutral distribution of $\sim \theta/k$ k-tuples, due to the scarcity of detected rare variants.

FIGURE 3 HERE

Williams et al. (2016a,b) (see also Ling et al. 2015) address the issue of the fixation of heterozygous genotypes by only considering polymorphic, segregating sites when comparing allele frequencies in tumors to those predicted from the neutral model, to the exclusion of sites that are ≥ 0.5 within a margin of sampling error. This also excludes those sites with frequencies $p > 0.5$ due to loss of heterozygosity. In addition, with a range of allele frequencies $p = [0, 0.5]$, the frequencies are rescaled to reflect the frequency of the heterozygous genotype, 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, the genetic distance for a sample where all genotypes at a variant site are 0/1 is 0.

With the assumption of diploidy at all of the genotyped SNV sites and

defining fixation as $p = 0.5$, we find that for $n = 65$, the binomial probability of observing fewer than $x = 26$ mutant alleles is $Bin(x \leq 25|n = 65, p = 0.5) = 0.041$. Thus, we use $f_c = 0.4$ as as a cutoff defining polymorphic sites. Using 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 under WHS being greater than under ILS as before. However Δ is small ($0.034 \leq \Delta \leq 0.070$), suggesting that in practice the estimation errors for $\widehat{\pi}$ are negligibly different for this data set. The small Δ are partly a consequence of the small number of segregating sites (because $\widehat{\pi}_{max} = S_n/2$). Therefore, the variance in $\widehat{\pi}$ estimation under WHS may be expected to increase for greater numbers of segregating sites, as was the case in the simulation data for larger time intervals and S.

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, as an illustration of how for some samples, the variances in $\widehat{\pi}$ may actually be lower under WHS than under ILS.

5 Discussion

Heuristically, the higher error in estimated genetic distance under WHS when weighted LD are positive on average reflects the loss of information due to non-independence across sites. If for most pairs of sites, the most frequent (major) alleles are in positive LD, then any error in estimating frequency and heterozygosity at one site covaries with the error at the other sites. In contrast, with ILS, each site provides independent information and the error across sites is uncorrelated. If there are S_n segregating sites in a sample of n and the variance in estimated genetic distance per site is σ^2, then with independent sampling the error across sites will approach σ^2/S_n. In contrast, 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. In the case of negative LD (i.e. negative association among common alleles), there is an information gain across sites.

On the other hand, a negative association of 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 WHS (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.

Because Δ will either be positive or close to 0 for most distributions of allele frequencies, our results suggest that ILS should be used to minimize error in genetic distance estimation for most natural and experimental populations. 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 WHS and ILS 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 (WHS) rather than individually via ILS will be negligible.

In contrast, for the genomes of clonal, ameiotic 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 (approximating ILS) 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. Moreover, single-cell sequencing usually entails a much smaller sample size n than the coverage depths of 100-1000 that are standard for NGS. Even in cases where $\Delta < 0$ (such as for some of the simulated data with small numbers of rare mutations, or for some truncations of the lung cancer data), the magnitude of the effect is going to be small and outweighed by the reduction of error

588 through high coverage. Moreover, Δ is defined on the assumption of the
589 same effective sample size n for both WHS and ILS, if ILS allows for much
590 larger n, as is often the case, then this is often sufficient to reverse the sign
591 of $var(\widehat{\pi}_{WHS}) - var(\widehat{\pi}_{ILS})$.

592 In addition to providing a summary statistic of genetic variation in a
593 population, $\widehat{\pi}$ is an estimator of population mutation rate θ (and, with a
594 known mutation rate, effective population size N_e) under a neutral model
595 of sequence evolution. As noted in the introduction, these parameter es-
596 timates can be used to detect the population genetic signatures of natural
597 selection and/or demographic histories when compared to θ estimates from
598 the sample number of segregating sites S_n. Consequently, our derivation
599 of the expectation and sample variance in $\widehat{\pi}$ under WHS and ILS are key
600 to calculating the error in estimates of θ and N_e. Sampling error in the
601 Tajima D statistic can be estimated using our derivation of Δ together with
602 an analogous estimate for the sampling error of S_n.

603 Another future research direction suggested by our results is deriving
604 analytically the conditions under which $E[A_{sr}], \Delta > 0$. Eqn. (7) provides
605 the conditions in terms of allele frequencies and LD under which $\Delta > 0$,
606 but does not specify the population genetic conditions under which these
607 distributions hold. For example, showing that an equilibrium distribution
608 of allele frequencies under Fisher-Wright drift both without recombination
609 and for a range of recombination rates leads to $\Delta > 0$ requires deriving a
610 population distribution (as opposed to the distribution within the sample)
611 of pairwise LD values D_{sr}. Computing $E[A_{sr}]$ over a distribution of al-
612 lele frequencies and pairwise LD would essentially formalizing the heuristic
613 argument presented in subsection 2.4

614 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 whole haplotype versus individual locus sampling appear as options under single cell sequencing versus WGS of multicell samples, respectively. The case study 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 Δ to the choice of cutoff f_c defining fixation, even in the diploid cases, bears further investigation as well.

6 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 the following individuals for their helpful comments: Matthew Cowperthwaite, Habil Zare, Mark Kirkpatrick, Jeffrey Townsend, Vincent Cannataro, and Andrea Sottoriva.

7 Statement of Effort

MS proposed the study, wrote the manuscript, ran the simulations, and analyzed the data. YN and PM derived most of the equations in section 2 and in the Appendix. JL wrote the python code used for the simulations.

8 Figures and Tables

Figure 1. Illustration of whole haplotype sampling (WHS) versus individual locus sampling (ILS). In this example, the population consists of 8 haploid organisms 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 WHS via the sampling 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 ILS such that G1, G3, G8 are sampled at S1, G4,G5 and G8 at S2, etc, giving a mean genetic distance $\widehat{\pi} = 8/3$.

Figures 2a-b. Population distributions of pairwise linkage disequilibria D_{sr} (2a) and weighted linkage disequilibria A_{sr} (2b) for a simulated population with $N = 500$ haploid genotypes after $T = 2500$ generations of mutation and Fisher-Wright genetic drift, corresponding an approximate equilibrium allele frequency distribution.

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.

Table 1. A summary of results for a Fisher-Wright model of genetic drift with infinite sites. The table shows a comparison of Δ_P values predicted from Eqn. (6) with simulation the values Δ_S for $N = 200, 500$ and sample size/coverage depth n for a range of time intervals (the last pair of time

values for each population size is of the order $4N$, corresponding to an approximate equilibrium in allele frequencies). The standard error of Δ_S is also shown, where ΔP lies within less than two SE units from Δ_S even for small time intervals where there are few mutations. Mean population pairwise linkage disequilibrium values 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 2. This table shows the number of segregating sites S_n in a sample of $n = 20$, the mean pairwise genetic distances $\widehat{\pi}_W, \widehat{\pi}_I$ (for WHS and ILS, respectively), and the variances in pairwise genetic distance for WHS and ILS. The latter are used to compute Δ_S in Table 1.

Table 3. Calculation 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 $\widehat{A} > 0$ and $\Delta > 0$ for all 4 samples, indicating that the error in pairwise genetic distance estimates for this data set are greater under WHS than under ILS, albeit weakly given the small number of unique haplotypes. Δ is computed using the actual mean coverage depth $n = 65$ for two different cutoffs used to define polymorphic sites. The upper panel shows the values for a cutoff of $f_c = 0.40$, selected based on a binomial probability. 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. Note that $\bar{p}'$ is based on $p' = p/f_c$, rescaled with respect to the diploid cutoff value.

References

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.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345.

Dexter, D. L. and Leith, J. T. (1986). Tumor heterogeneity and drug resistance. *Journal of Clinical Oncology*, 4(2):244–257.

Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer Science & Business Media.

Ellegren, H. and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*.

Ewens, W. J. (2004). Mathematical population genetics. i. theoretical introduction. interdisciplinary applied mathematics, 27.

Fu, Y.-B. (2015). Understanding crop genetic diversity under modern plant breeding. *Theoretical and Applied Genetics*, 128(11):2131–2142.

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.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

716 Hansson, B. and Westerberg, L. (2002). On the correlation between
717 heterozygosity and fitness in natural populations. *Molecular Ecology*,
718 11(12):2467–2474.

719 Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D.,
720 Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V.,
721 et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular
722 classification within and across tissues of origin. *Cell*, 158(4):929–944.

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

725 Lewontin, R. and Kojima, K.-i. (1960). The evolutionary dynamics of com-
726 plex polymorphisms. *Evolution*, pages 458–472.

727 Lewontin, R. C. et al. (1974). *The genetic basis of evolutionary change*,
728 volume 560. Columbia University Press New York.

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

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

735 MacLean, R. C., Hall, A. R., Perron, G. G., and Buckling, A. (2010). The
736 population genetics of antibiotic resistance: integrating molecular mech-
737 anisms and treatment contexts. *Nature Reviews Genetics*, 11(6):405–414.

738 Martinez, J. and Baquero, F. (2000). Mutation frequencies and antibiotic
739 resistance. *Antimicrobial agents and chemotherapy*, 44(7):1771–1777.

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.

National Research Council (1993). *Livestock*. The National Academies Press, Washington, DC.

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

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.

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.

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.

Sun, X.-x. and Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol Sin*, 36(10):1219–1227.

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

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

766 TCGA (2014). Comprehensive molecular profiling of lung adenocarcinoma.
767 *Nature*, 511(7511):543–550.

768 Van Dyke, F. (2008). *Conservation biology: foundations, concepts, applica-*
769 *tions*. Springer Science & Business Media.

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

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

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

778 9 Appendix A1: Ordered Pairs of Pairs

779 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*

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

780 *Proof.* Unbiasedness is straightforward:

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

781 For the variance, note that

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

782 Then

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

783 □

784 *Proof of Eqn. (4).* Let $\widehat{f}_{ns} = \frac{1}{N_2}\sum_{i<j} f_{ij,s}$. From the statistical indepen-
785 dence among sites s located on different reads under ILS, it follows that for
786 a sample of n,

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

787 with

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

788 where the first equality is due to Eqn. (3).

Table 1

N	T	S_N	$\bar{p}'$	$\hat{\pi}_w$	$\hat{\pi}_I$	var_w	var_I
200	5	112.7	0.012	2.94	2.94	0.249	0.241
200	10	181.5	0.016	5.82	5.81	0.468	0.476
200	20	250.1	0.024	11.47	11.47	0.878	0.819
200	50	351.2	0.043	26.79	26.80	2.27	1.58
200	800	770.6	0.315	115.59	114.58	272.0	3.70
200	1000	847.7	0.361	122.90	122.77	415.7	3.97
500	5	308.4	0.0049	2.96	2.96	0.279	0.271
500	10	455.9	0.0066	5.97	5.97	0.537	0.543
500	20	616.9	0.0096	11.61	11.60	1.04	1.04
500	50	875.5	0.0172	28.61	28.68	2.53	2.27
500	100	1078.3	0.0281	54.67	54.67	6.23	3.71
500	2000	2202.4	0.296	301.16	301.22	1532.1	9.09
500	2500	2395.1	0.153	316.06	315.64	2089.8	10.53

Table 2

N	T	$\bar{D}_{sr}$	$\sum A_{sr}$	ΔP	ΔS	$SE(\Delta_S)$
200	5	-4.57×10^{-7}	4.18×10^{-3}	-9.57×10^{-4}	8.01×10^{-3}	4.58×10^{-3}
200	10	1.16×10^{-6}	0.0997	0.0129	-7.85×10^{-3}	0.012
200	20	-3.16×10^{-7}	0.0529	0.0482	0.0587	0.0226
200	50	-9.30×10^{-8}	1.14	0.766	0.687	0.0927
200	800	-8.89×10^{-6}	660.5	297.96	268.34	44.56
200	1000	1.49×10^{-5}	1009.0	444.94	411.68	51.51
500	5	1.16×10^{-7}	4.58×10^{-3}	2.13×10^{-3}	8.08×10^{-3}	3.00×10^{-3}
500	10	-2.83×10^{-7}	-0.0242	-7.92×10^{-3}	-6.23×10^{-3}	7.53×10^{-3}
500	20	7.70×10^{-8}	0.393	0.00	0.0269	0.0213
500	50	-1.43×10^{-7}	0.269	0.256	0.259	0.0546
500	100	-9.80×10^{-8}	4.35	2.74	2.52	0.182
500	2000	-4.46×10^{-6}	3362.8	1606.1	1523.0	213.90
500	2500	5.31×10^{-6}	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

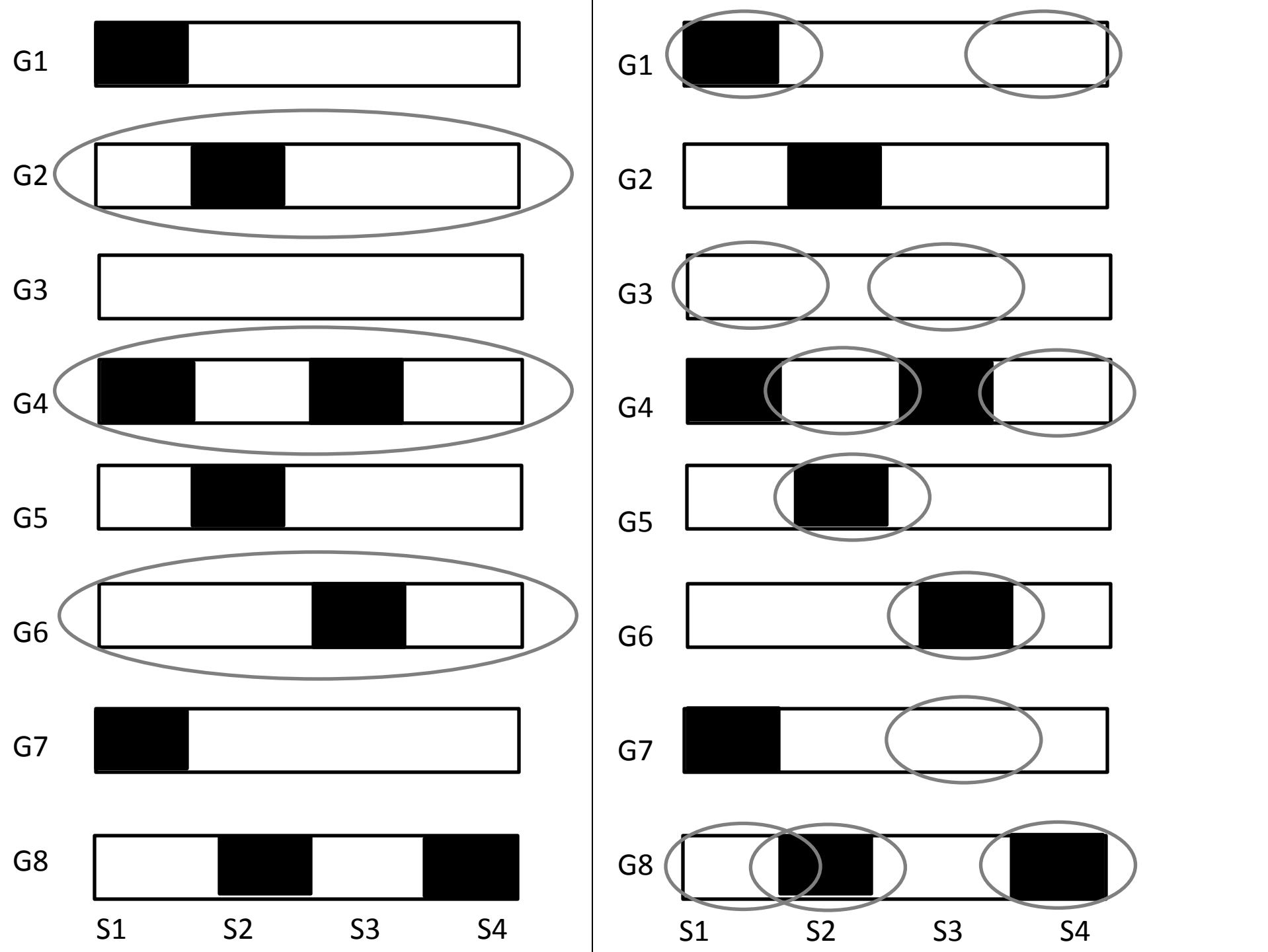
G1
G2
G3
G4
G5
G6
G7
G8
S1
S2
S3
S4
G1
G2
G3
G4
G5
G6
G7
G8
S1
S2
S3
S4

Figure 2a

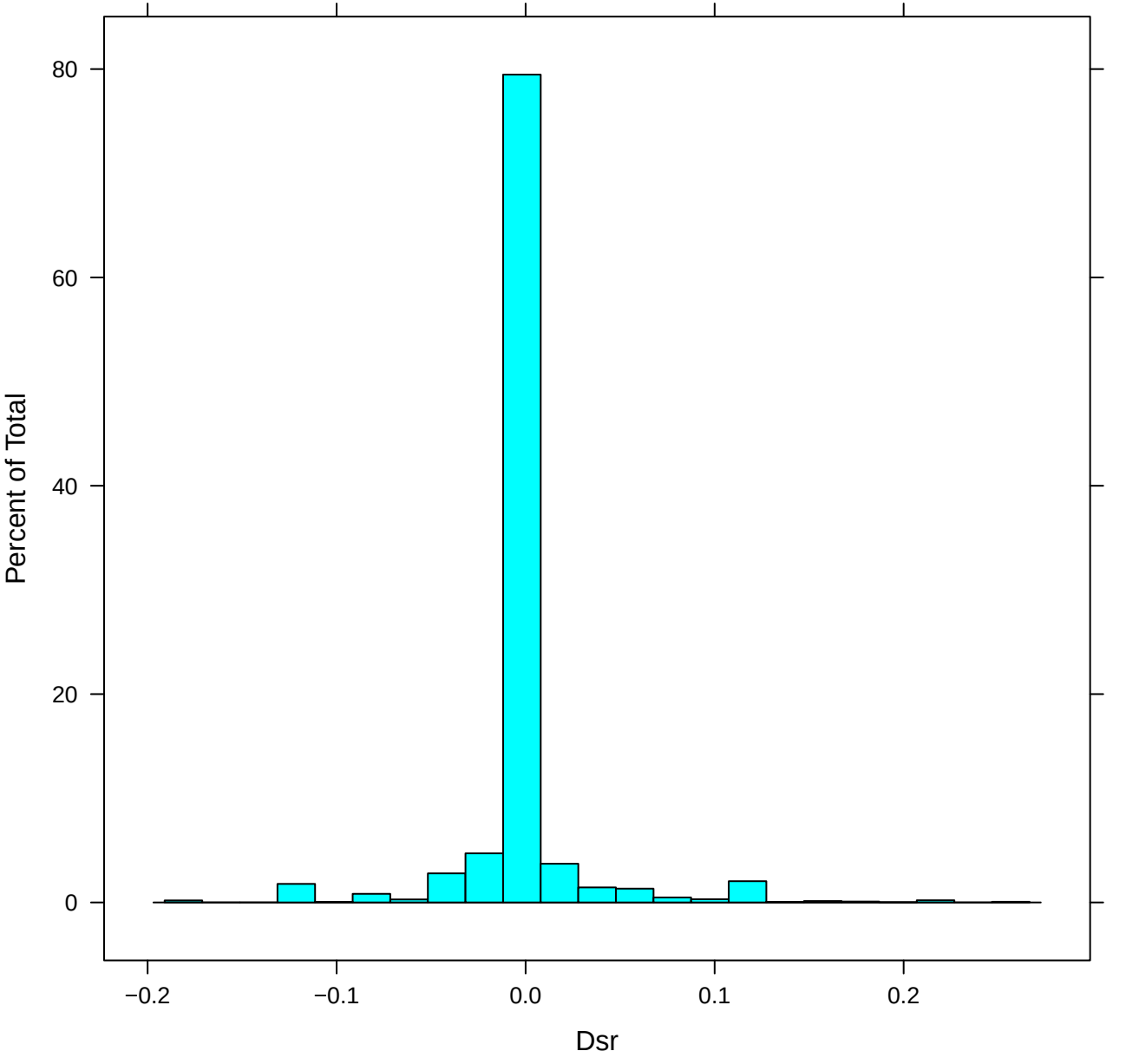

Figure 2b

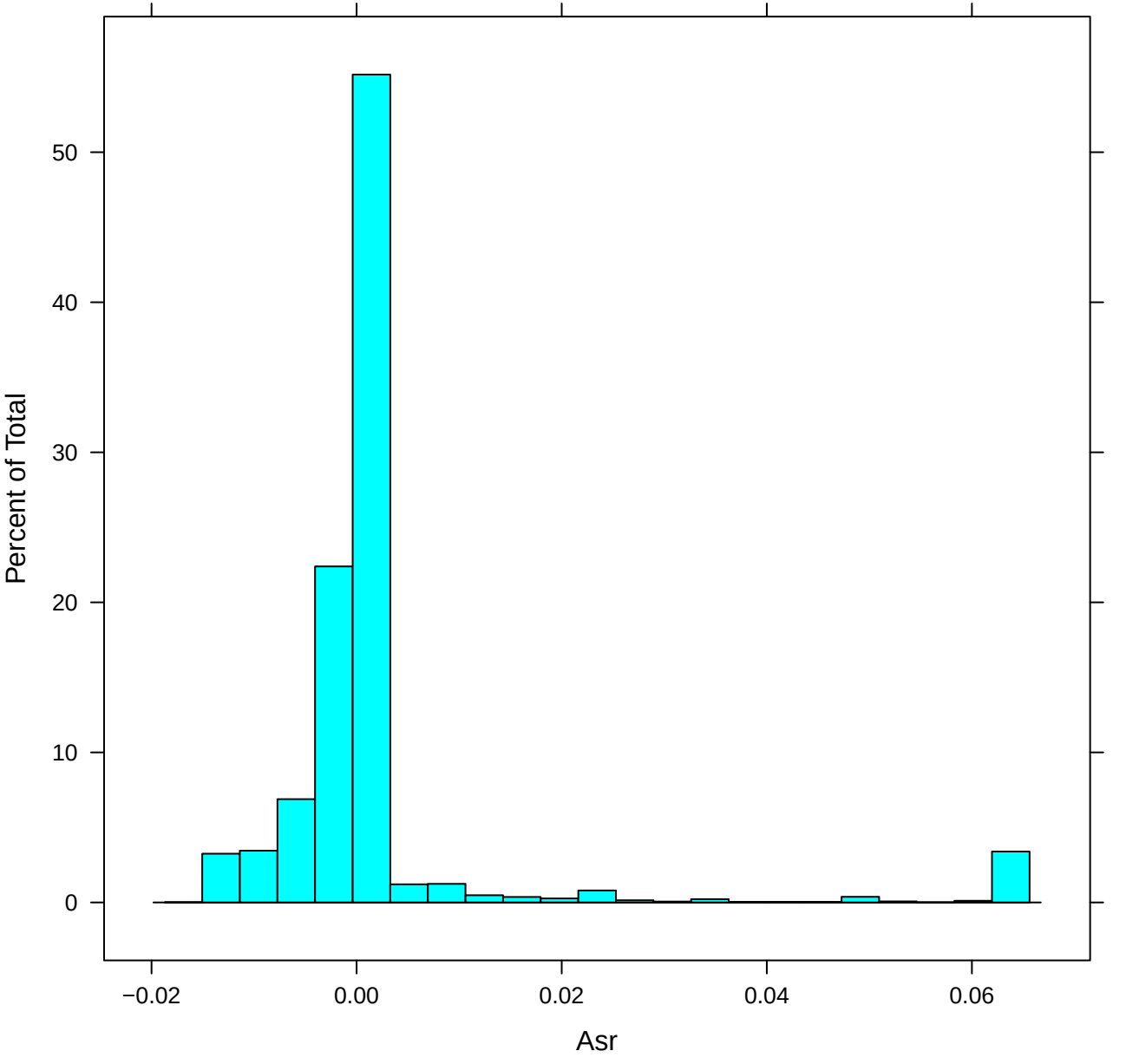

Figure 3