FILE S1

Haplotype phase uncertainty: Since the *HCN* statistic reflects haplotype patterns, and for many genome-wide SNP datasets consisting of unrelated individuals, haplotype phase would need to be computationally inferred, we wanted to determine how this inference affected the *HCN* statistic. To do this, we simulated 1000 windows with c_{window} = 0.25 cM in a sample size of 100 chromosomes from a bottleneck demographic history (N_{cur}=10,000, N_{mid}/N_{cur}=0.1, N_{anc}/N_{cur}=1.0, t_{cur}=800 generations, t_{mid}=800 generations), where n_{snp}=40. For each window, we then randomly paired the chromosomes into diploid individuals and generated diploid genotypes at each SNP. We next inferred haplotypes from these genotypes using a popular phasing method, fastPHASE (Scheet and Stephens 2006), with the default settings. We chose to use fastPHASE since its performance is comparable to one of the better performing phasing algorithms, PHASE, yet is fast enough to be run on genome-wide datasets. Finally, we compared the *HCN* statistic for the phase-known dataset to the phase-inferred dataset.

Figure S2 shows the *HCN* statistic for a bottleneck model when the correct haplotype phase is known with certainty (left) and when haplotype phase is inferred using fastPHASE (right). The *HCN* from phase-inferred haplotypes has a broader distribution than when haplotype phase is known. In particular the *HCN* constructed using the phase-inferred haplotypes has an excess of windows having many haplotypes (green squares in bins "65-90" and "70-100") as compared to the known phase *HCN*. Although it is a bit more subtle, the *HCN* using the phase-inferred haplotypes also has an excess of windows where the most common haplotype is at a high frequency. This can be seen by the yellow square in the phase inferred haplotypes where there was an orange square in the phase-known *HCN*. Thus, inferring haplotype phase will result in an *HCN* statistic that is slightly different from the true phase-known *HCN*.

Ascertainment bias: To evaluate how the *HCN* statistic is influenced by SNP ascertainment bias, we conducted a variety of coalescent simulations under different demographic models and SNP ascertainment strategies. We then compared the *HCN* from the different ascertainment strategies to the *HCN* with complete SNP ascertainment. We also examined whether another haplotype statistic, H_{pair}, is affected by ascertainment bias.

Since we wanted to address the question of whether discovering SNPs in one population and then typing them in a second population is more biased than selecting the SNPs in the genotyped population, we considered demographic models that consisted of two populations. Briefly, we considered a finite island model (where each population has size N_e=10,000) with a low rate of migration between populations ($4N_em$=9) and high rate ($4N_em$ =99), a population split model where the two populations (each of size N_e =10,000) split 2000 or 5000 generations ago, and a complex model where the two populations split 5000 generations ago and there was a bottleneck in population one (N_{mid}/N_{cur}=0.1, t_{cur}=800, t_{mid}=800). The last model can be thought of as a very crude approximation of the contrast between European (as population 1) and African (as population 2) human populations. For each of these demographic models, we simulated a "genotype" sample of 40 chromosomes from each of the two

populations as well as a SNP discovery sample consisting of an additional four chromosomes from each population. We then examined five different SNP discovery protocols shown in Table S1a. These ascertainment strategies are reasonable ones for many of the human genome-wide SNP datasets like HapMap where many of the SNPs were discovered by comparing two sequencing reads (as in phase I) or from a polymorphism discovery panel with a few chromosomes from multiple populations (phase II SNPs discovered by Perlegen; Hinds *et al.* 2005; International HapMap Consortium. 2005; International HapMap Consortium 2007). For each ascertainment scheme we simulated 1000 windows 500 kb in size with a uniform recombination rate of 1 cM/Mb (c_{window}=0.5 cM) and μ =1 x 10^{-8} per base-pair per generation. To determine whether ascertainment bias becomes a problem for larger datasets containing more than 1000 windows (n_{window}>1000), we also simulated an additional dataset under the complex demographic history consisting of 7000 windows 250 kb in size with uniform recombination rate of 1cM/Mb (c_{window}=0.25 cM) and μ =1 x 10^{-8} per base-pair per generation. Finally, we considered the case where the genotype sample consisted of 120 chromosomes from each population (to mimic the HapMap CEU and YRI samples) and we had data from 7000 windows 250 kb in size with uniform recombination rate of 1cM/Mb (c_{window}=0.25 cM) and μ =1 x 10^{-8} per base-pair per generation. For this set of simulations, the SNP discovery set consisted of 12 chromosomes per population. Here we considered eight ascertainment strategies shown in Table S1b.

For each demographic scenario and ascertainment scheme, we selected a sub-set of 40 SNPs having MAF >10% (n_{snp}=40). If a window had fewer than 40 SNPs, it was dropped from the analysis. We then generated 10 different *HCN* statistics, each with a set of 10 randomly selected SNPs from each window, and then averaged them to generate the final *HCN* statistic. We compared the average *HCN* statistic to the expected statistic under complete ascertainment using a chi-square goodness of fit test. To generate the expected *HCN* statistic under complete ascertainment, we simulated an additional 10^5 windows each consisting of 40 chromosomes and n_{snp}=40 and averaged over 10 different *HCN* statistics, each using a set randomly selected SNPs for each window. From these simulations, we computed the expected *HCN* statistic. Note that, when conducting the chi-square goodness of fit tests, we binned the *HCN* statistic so that we did not have any expected cell counts $\leq$ 5. For the complex demographic history using 7000 windows (for genotype sample sizes of 40 and 120 chromosomes per population) n_{snp}=20 instead of 40.

We find that for all demographic models examined, except for the island model with a high migration rate, ascertainment of SNPs using two discovery chromosomes from one population results in a different *HCN* statistic than that expected under complete ascertainment (Figure S3). This is shown by the low *P*-values for the goodness of fit tests comparing the *HCN* statistic using SNPs polymorphic in two discovery chromosomes to the expected *HCN* under complete ascertainment. The *HCN* statistic constructed from SNPs ascertained in two chromosomes has an excess of windows having a small number of haplotypes and an excess of windows where the most common haplotype is at higher frequency as compared to the complete ascertainment case (Figure S4).

The reason for this pattern is that SNPs polymorphic in the two chromosome discovery sample must occur on branches of the genealogy where one of the two discovery chromosomes carries the mutant allele and the other does not. These branches are a small fraction of the total area of the genealogy. This fact will result in SNPs that are polymorphic in the discovery sample tending to occur on the same branches of the genealogy more often than expected without ascertainment bias. SNPs that co-occur on the same branches of the genealogy will be in LD with each other, resulting in there being fewer haplotypes and the most common haplotype occurring at higher frequency than in the case of less LD among SNPs. When considering SNPs discovered from two chromosomes from the first population, the *HCN*s in both populations differ from the expected *HCN*, suggesting that SNP discovery using two chromosomes does poorly, regardless of whether those two chromosomes are from the population of interest.

SNP discovery using one chromosome from each population is a slight improvement to SNP discovery using two SNPs from population 1. However, we note that for many of the demographic models considered here (Figure S3), the *HCN* constructed from ascertained SNPs differs significantly from the expected *HCN* under complete ascertainment.

However, SNP discovery using four chromosomes from the first population results in a better fit to the expected *HCN* for most of the demographic models considered. In all cases, except for the complex demographic model, the *HCN*s constructed from ascertained SNPs are quite consistent with the expected *HCN* under complete SNP ascertainment. This finding holds true even for the second population which had no SNP discovery, again illustrating that if the two populations have similar demographic histories, ascertainment sample depth may be more important than which population the SNPs were ascertained from in terms of matching the *HCN* statistic. This pattern, however, does not hold for the complex demographic model. Here SNP discovery using four SNPs from the bottlenecked population (population 1) results in a poor fit to the expected *HCN* statistic. The reason for this is that the four SNP discovery chromosomes from the bottlenecked population are less representative of the diversity in the second population that did not undergo a bottleneck (population 2). If, again for the complex demographic scenario, instead of taking four discovery chromosomes from the first population, we take two discovery chromosomes from each population, the *HCN* statistic from the ascertained SNPs more closely matches the expected *HCN* statistic. However, note that if the number of windows of the genome considered is large (n_{window}=7000), the effects of ascertainment bias are still present.

The *HCN* statistic generated using a four chromosome SNP discovery sample from both of the two populations results in an excellent fit to the expected *HCN* for both populations in all demographic scenarios considered. We also found an adequate fit of the expected *HCN* to the observed *HCN* when considering a larger dataset under the complex demographic model. This finding is especially encouraging since the larger number of windows in the dataset ($n_{winidow}$=7000 as compared to 1000 in previous datasets) will have more power to detect subtle departures in the fit of the model. Thus, for the demographic models considered here using n=40 chromosomes, the *HCN* statistic using SNP discovery sample of $\geq$ 4 chromosomes from at least two populations is not significantly different from the true *HCN* statistic.

We also examined whether ascertainment bias is a more severe problem when the genotype sample is >40 chromosomes. To do this, we repeated the above approach for the complex demographic model using n=120 chromosomes and considering larger SNP discovery sample sizes (Figure S5). We find that small SNP discovery sample sizes (here <8 chromosomes) result in significant differences between the *HCN* under SNP ascertainment and the expected *HCN*. However, for larger SNP discovery sample sizes, the effect disappears. This holds true even for the population that has no SNP discovery chromosomes (*e.g.* the solid line at "12 from 2"). To assess the amount of evolutionary variance in the whole process, we performed two completely independent sets of simulations for these demographic and ascertainment models. The results of both replicates are shown in Figure S5. Encouragingly, the variance is reasonably low since the two solid curves (and dotted curves) are similar to each other.

We also evaluated whether the H_{pair} statistic was robust to ascertainment bias. As shown in Figure S1, for all demographic models and ascertainment conditions considered, H_{pair} was severely affected by SNP ascertainment bias. Ascertainment bias results in H_{pair} being higher than expected. This finding is analogous to the effect of ascertainment bias on π, the average number of pairwise differences among DNA sequences (Nielsen *et al.* 2004). Ascertainment bias results in an excess of intermediate-frequency SNPs, which results in there being more pairwise differences between haplotypes than low-frequency SNPs do. Thus, by preferentially selecting intermediate-frequency SNPs, H_{pair} becomes inflated.

Interestingly, we find that for the cases where SNPs were ascertained in population 1 exclusively, the fit of the H_{pair} statistic under ascertainment bias to the expected H_{pair} statistic is actually worse in population 1—the population where the SNPs were discovered in—than in population 2. This pattern is seen for both n_{widnow}=1000 and for n_{widnow} =7000 and for both the "2 from pop 1" and the "four from pop 1" ascertainment strategies. One possible explanation for this counter-intuitive pattern is that the ascertained SNPs from population 1 are more likely to be at intermediate frequency in population 1 (as discussed above), but may have drifted to lower or higher frequency in the second population, resulting in those SNPs being more representative of the true frequency spectrum in that population.

Here are the ms commands to generate the *HCN* statistic in Figure 7:

Growth and Structure:

./ms 40 10000 -t 400 -r 400 250000 -F 4 -es 0.00625 1 0.1 -eM 0.00625 5 -eN 0.00625 0.5 -eN 0.025 0.125 -ej 0.625 2 1 -eM 0.625 0 -eN 0.625 0.25

Growth:

./ms 40 10000 -t 400 -r 400 250000 -F 4 -en 0.01925 1 0.303333

Literature Cited

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072-1079.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851-861.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

Nielsen, R., M. J. Hubisz and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373-2382.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. **78:** 629-644.