Supporting Information

Extending the theoretical model of giSNPs to diploid populations

To extend our theoretical model to diploid populations, we begin by assuming that N is the number of chromosomes sampled, which corresponds to $N/2$ diploid individuals. Again we consider a variant with minor allele frequency k/N. For the diploid model, equation 1 (the expected fraction of variants with the same minor allele frequency) still holds because the minor allele frequency for a particular variant stays constant whether we consider N chromosomes partitioned into N haploid individuals or N chromosomes partitioned into $N/2$ diploid individuals. Thus we can use this equation to calculate the expected fraction of variants with the same minor allele frequency.

The fraction of variants with minor allele frequency k/N that have an identical allelic configuration is slightly more complicated than for the haploid case. Here, allelic configurations are composed of three classes (homozygous major allele, heterozygous, and homozygous minor allele) rather than two. For a site with minor allele count k, the number of possible phased allelic configurations is

$$\sum_{i=0}^{k} \binom{N/2}{i} \binom{N/2}{k-i} = \binom{N}{k} \tag{1}$$

where the first binomial coefficient represents the configuration of alleles for the first of the two copies of the locus, and the second binomial coefficient the configuration for the second of the two copies of the locus. The right-hand side of the equation follows from the Chu-Vandermonde identity. These allelic configurations are all equally likely under the neutral Wright-Fisher model with infinite-sites mutation.

In the context of giSNPs, we are primarily interested in physically distant pairs of loci (loci that are far apart or on different chromosomes), and phase becomes arbitrary. If we focus on unphased allelic configurations, the number of possible configurations for a site with minor allele count k is

$$\frac{k - k \bmod 2}{2} + 1 = \frac{2k + 3 - (-1)^{k+1}}{4}$$

Unfortunately, these allelic combinations are *not* equally likely. In particular, each heterozygote in the unphased allelic configuration doubles the number of possible phased allelic combinations. Combining this fact with equation 1 yields a formula for the probability of any specific unphased allelic configuration:

$$\Pr(\text{allelic configuration}) = 2^h \binom{N}{k}^{-1}$$

where h is the number of heterozygotes present in the configuration in question.

Overall, it is evident that for a sample of $D = 2N$ diploid individuals, considering a variant with minor allele frequency k/N, for variants with either (1) an unphased allelic configuration with no heterozygotes or (2) a phased allelic configuration, the expected fraction of other variants with which it will be a giSNP is identical to the haploid case with twice as many individuals sampled (sample size $2N$), given in equation 2. For unphased allelic configurations that include one or more heterozygotes, this probability will be inflated by a factor of 2^h, where h is the number of heterozygotes present in the configuration.

Genetically indistinguishable SNPs in humans

To explore the rate of giSNP occurrence in real genome-wide data, we examined giSNPs in 1,093 humans from the 1,000 genomes project (The 1000 Genomes Project Consortium, 2012). We focused on biallelic SNPs, and discarded any sites with over 10% missing data. We obtained SNPs implicated in human genome-wide genotype-phenotype associations from the NIH Catalog of Published Genome-Wide Association Studies (Hindorff *et al.*, 2014). The meiotic recombination rate in humans is roughly 1 cM/Mb (Jensen-Seaman *et al.*, 2004) so in order to focus on effectively randomly assorting loci we ignored giSNPs less than 50 Mb apart on the same chromosome. The large number of SNPs present in this dataset (>38 million) precludes enumeration of LD between all ∼730 trillion pairs of SNPs, so we focused on calculating whether each of 12,607 distinct SNPs with reported significant associations in a GWAS (Hindorff *et al.*, 2014) and any of the other ∼38 million SNPs in the dataset were genetically indistinguishable. Overall, nine GWAS-reported SNPs were members of giSNP pairs with a total of 44,270 other unique SNPs across the genome.

Most of the pairs of giSNPs (44,156 of 44,270 pairs) involved five GWAS SNPs where one individual carried a single copy of the minor allele. In other words, the GWAS SNP minor allele was a singleton private to the individual in question, and thousands of other private variants within that individual formed a cluster of giSNPs. Given that a new exome sequence reveals roughly 200 singletons private to that individual (Tennessen *et al.*, 2012) and that the exome corresponds to approximately 1.5% of the human genome, it is not surprising that there should be thousands of private singleton variants present in each individual. These individuals were not confined to one population but included individuals of African, admixed American, and European descent. The other four GWAS SNPs were genetically indistinguishable from a small number of other variants (9, 10, 32, and 63) at which two individuals carried a single copy of the minor allele.

The nine GWAS SNPs that have giSNPs in the 1,000 genomes dataset were reported in a total of four separate GWAS (Do *et al.*, 2011; Comuzzie *et al.*, 2012; Demirkan *et al.*, 2012; Seppala *et al.*, 2014). In most cases, the SNP did not have a highly significant p-value, and was often reported as suggestive. The sole exception was `rs34637584`, which had a highly significant p-value of 2×10^{-28} (Do *et al.*, 2011) (this association was reported previously to the study in question). However, in all cases, the frequency of the risk allele in the GWAS was either very high or very low (0.998 for two SNPs, ≤ 0.006 for five SNPs; 0.063 for one SNP; unreported for the last SNP). It is well understood that population based association studies are poorly powered to detect associations between traits and causal alleles with low MAF (Long *et al.*, 1997; Ohashi and Tokunaga, 2001; Zondervan and Cardon, 2004). Our results indicate that genotypes with very low MAF are particularly susceptible to "dragging along" many additional non-causal loci that are giSNPs and that, as has been previously noted, any significant GWAS results involving very low MAF variants should be interpreted with extreme caution. This is consistent with the success of collapsing methods in the context of rare-variant association studies, where rare variants are not individually tested for association with phenotypes but, rather, information is aggregated from low-frequency variants across multiple sites to produce a test statistic (Asimit and Zeggini, 2010; Zuk *et al.*, 2014). Thus, giSNPs are not likely to pose significant problems for modern GWAS with large sample sizes that test for association using only SNPs with, e.g., at least a 5% MAF.

In order to examine the dependence of giSNPs in human data on sample size, we added 50 randomly selected humans (100 chromosomes) to our sample of 100 randomly selected individuals from each model organism dataset (*S. cerevisiae*, *D. melanogaster*, and *A. thaliana*). Again, since the absolute number of giSNPs is highly dependent on the number of SNPs in a dataset, we randomly selected 100,000 SNPs from each reduced sample of individuals. After removing differences in sample size, we observed similar overall rates of giSNP prevalence in humans as in the model organisms (Fig. 1A-C). Thus, consistent with the observations made by Lawrence *et al.* (2005), giS-

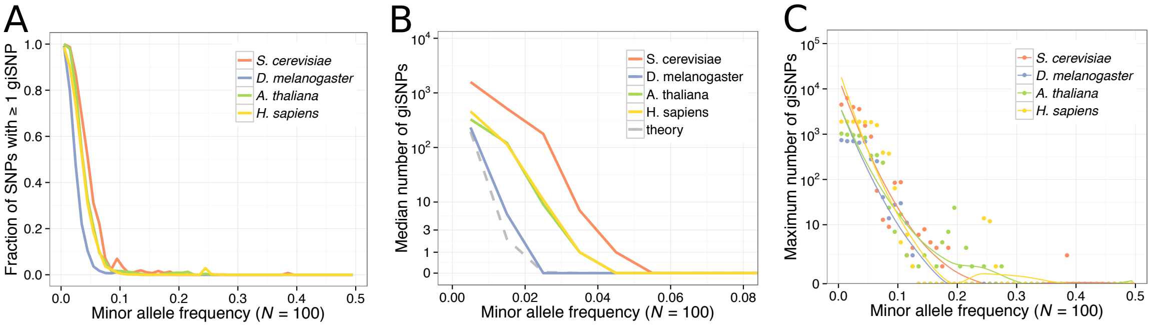

Figure 1: Genetically indistinguishable SNPs in real datasets. This figure is identical to Fig. 4D-F but also includes results calculated from human data. (A-C) Statistics calculated from datasets downsampled randomly to match a sample size of 100 chromosomes and 100,000 SNPs. (A) Fraction of SNPs with at least one giSNP as a function of MAF. (B) Median number of giSNPs as a function of MAF. Small black notches indicate bootstrap 95% confidence intervals on the median. The median number of giSNPs for all MAFs > 0.08 is negligible in all datasets. (C) Maximum number of giSNPs across all allelic configurations, as a function of MAF. Dots indicate the number of giSNPs for the "worst" allelic configuration at each specific MAF. Solid lines provide a local smoothing via the loess method.

NPs would have the potential to be a significant concern in human GWAS in cases where sample sizes are small (e.g. dozens or low hundreds of individuals). Fortunately, this scenario does not apply to most modern human GWAS.

References

Asimit, J. and E. Zeggini, 2010 Rare variant association analysis methods for complex traits. Annu. Rev. Genet. **44**: 293–308.

Comuzzie, A. G., S. A. Cole, S. L. Laston, V. S. Voruganti, K. Haack, *et al.*, 2012 Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS ONE **7**: e51954.

Demirkan, A., C. M. van Duijn, P. Ugocsai, A. Isaacs, P. P. Pramstaller, *et al.*, 2012 Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. PLoS Genet. **8**: e1002490.

Do, C. B., J. Y. Tung, E. Dorfman, A. K. Kiefer, E. M. Drabant, *et al.*, 2011 Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. **7**: e1002141.

Hindorff, L. A., J. MacArthur, J. Morales, H. A. Junkins, P. N. Hall, *et al.*, 2014 A catalog of published genome-wide association studies. [Online; accessed 2-July-2014].

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, *et al.*, 2004 Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. **14**: 528–538.

Lawrence, R. W., D. M. Evans, and L. R. Cardon, 2005 Prospects and pitfalls in whole genome association studies. Philosophical Transactions of the Royal Society of London B: Biological Sciences **360**: 1589–1595.

Long, A. D., M. N. Grote, and C. H. Langley, 1997 Genetic analysis of complex diseases. Science **275**: 1329–1330.

Ohashi, J. and K. Tokunaga, 2001 The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J. Hum. Genet. **46**: 478–482.

Seppala, I., M. E. Kleber, L.-P. Lyytikainen, J. A. Hernesniemi, K.-M. Makela, *et al.*, 2014 Genome-wide association study on dimethylarginines reveals novel AGXT2 variants associated with heart rate variability but not with overall mortality. European Heart Journal **35**: 524–530.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science **337**: 64–69.

The 1000 Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature **491**: 56–65.

Zondervan, K. T. and L. R. Cardon, 2004 The complex interplay among factors that influence allelic association. Nat. Rev. Genet. **5**: 89–100.

Zuk, O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter, *et al.*, 2014 Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U.S.A. **111**: E455–464.