FILE S1

MATERIALS AND METHODS

RELATIVE ABUNDANCE OF INDIVIDUALS' DNA IN POOLS

The estimation of minor-allele frequency assumes knowledge of the relative abundance of each individual's DNA in each pool (implicitly or explicitly). It is therefore important to have a means of estimating these relative abundances. To do this, we took advantage of the genotype data available from the GWAS. In each pool, we selected SNVs that had a genotyping success rate of 100% in the GWAS, which were unambiguously mapped to the genome (using the Varietas portal by (PAANANEN, CISZEK and WONG 2010); the reference genome used was GRCh37) and that were observed in at least 30 reads during sequencing. Furthermore, we required that no indels were found at these sites during the alignment phase. 298,853 such SNVs were available for men, and 298,703 for women. At such SNVs, the proportion of major allele reads out of total reads is expected to correspond to the number of major alleles carried by individuals in the pool, adjusted for the individuals' DNA's relative abundance in the pool. We found the least-squares estimators of the relative abundances in the following manner: in a pool with $h^k/2$ individuals and data for m SNVs, let A be the $m \times h^k/2$ matrix corresponding to the minor allele counts times $1/2$, so that $A_{ij} = 0, 0.5$ or 1 if individual j carries 0, 1 or 2 copies of the minor allele of SNV i, respectively. Let x be an $h^k/2 \times 1$ column vector of relative abundances, so x_i equals the relative abundance of individual i in the pool. Lastly, let b be the $m \times 1$ column vector of the observed minor allele frequencies in the pool, so that b_i equal the proportion of minor alleles read out of total reads of SNV i. The least-squares estimator of the relative abundances vector x is found by solving the following optimization problem:

$$\begin{aligned} \arg\min_{x} \quad & \frac{1}{2}\|Ax - b\|_2^2 \\ \text{subject to} \quad & 0 \leq x_i \leq 1, i = 1, \ldots, h^k/2 \quad \sum_{i=1}^{h^k/2} x_i = 1 \end{aligned} \tag{1}$$

Using the `lsqlin` function in MATLAB, we found the least-squares estimators of relative abundances listed in table 1. In both the men and women pools, relative abundances were generally similar, though not exactly equal.

Individual	Males Pool	Females Pool
1	0.1684	0.1664
2	0.2148	0.2055
3	0.2066	0.2012
4	0.2039	0.2220
5	0.2062	0.2049

Table 1: Relative abundances in pools of individuals' DNA.

ESTIMATION OF SEQUENCING ERROR RATE

To estimate the read error rate of the sequencing platform, we leveraged the GWAS data. We selected a set of 18,163 SNVs in the pool of men (and 16222 in the pool of women) for which the genotype minor allele counts are 0 for all five individuals in the GWAS, and which had at least 50 sequencing reads. We interrogated the proportion of minor alleles out of total alleles read at each such position. For a SNV which was correctly genotyped, this proportion is approximately 0, occasionally with small deviations produced by sequencing error. We discarded 88 SNVs in men (68 in women) which had a proportion > 0.05, as we suspect they might represent genotyping errors. At the remaining SNVs, 2489 of the 1084400 reads in men were minor allele (2203 out of 912938 in women). We thus estimated the sequencing error rate to be 0.229% per base per read in the men's pool and 0.241% in the women's pool, assuming a simplistic error model in which the rate of error is fixed across pools and independent of the position along the read and of the nucleotide being read. It should be noted that in a more realistic error model of high-throughput sequencing platforms, error rates do potentially depend on these factors.

ESTIMATION OF MINOR ALLELE FREQUENCY FROM SEQUENCING DATA WITH ERRORS

The methods presented in this work rely on the estimated minor allele frequencies from sequencing data. To estimate these frequencies, we use a maximum likelihood approach with a simple error model. Note that more sophisticated models are possible, using error patterns specific to the sequencing platform, for example (e.g., (DEPRISTO, BANKS, POPLIN, GARIMELLA, MAGUIRE *et al.* 2011; BANSAL 2010; MCKENNA, HANNA, BANKS, SIVACHENKO, CIBULSKIS *et al.* 2010)). If necessary, such models can be readily substituted for the one presented in this section.

Consider a set of P pools, each containing a mixture of DNA from several individuals (in the case of low coverage sequencing without pooling, the size of each pool is 1). Let h^k denote the number of haplotypes in pool k (thus, pool k contains DNA from $h^k/2$ individuals), and let α_i denote the relative abundance of individual i's haplotypes in the pool, so that $\sum_{i=1}^{h^k/2} \alpha_i = 1$ (the relative abundances are assumed to be known, and a method to estimate them is described above). The pools undergo sequencing, generating observations of the minor and major alleles at each genomic position. Our goal is to estimate p, the minor allele frequency across all pools, for each genomic position.

Let e be the known (or estimated) error rate of the sequencing platform, and for pool k let x^k be the observed counts of the minor allele, y^k the observed counts of the major allele, and $z^k = x^k + y^k$. For individual i in pool k, let t_i^k be the number of i's chromosomes that carry the minor allele, so that $t_i^k \in \{0, 1, 2\}$. Finally, let $\vec{t}^k$ denote the minor allele count vector $(t_1^k, \ldots, t_{h^k/2}^k)$. To estimate p, we observe that $t_i^k \sim B(2, p)$, and that

$$Pr(\vec{t}^k \mid p) = \prod_{i=1}^{h^k/2} Pr(t_i^k \mid p) = \prod_{i=1}^{h^k/2} \binom{2}{t_i^k} p^{t_i^k} (1-p)^{2-t_i^k} \tag{2}$$

Furthermore, when reading a single base from a pool k with the minor allele vector $\vec{t}^k$, the chance

to observe a minor allele, denoted by $f^k(\vec{t^k})$ is

$$f^k(\vec{t^k}) \triangleq (1-e)\sum_{i=1}^{h^k/2}\alpha_i\frac{t_i^k}{2} + e\sum_{i=1}^{h^k/2}\alpha_i\frac{(2-t_i^k)}{2} \tag{3}$$

(to see this, note that we observe a minor allele if we either sample and read a minor allele *without error* or sample and read a major allele *with error*). Therefore x^k, the observed minor allele count in pool k, follows a Binomial distribution:

$$x^k \sim B(z^k, f^k(\vec{t^k})) \tag{4}$$

The likelihood of p for a particular pool k is then:

$$\begin{aligned} L(p; x^k, y^k) &= Pr(x^k, y^k \mid p) \\ &= \sum_{\vec{t^k}\in\{0,1,2\}^{h^k/2}} Pr(x^k \mid z^k, \vec{t^k}) \cdot Pr(\vec{t^k} \mid p) \\ &= \sum_{\vec{t^k}\in\{0,1,2\}^{h^k/2}} \left\{ \binom{z^k}{x^k} (f^k(\vec{t^k}))^{x^k}(1-f^k(\vec{t^k}))^{z^k-x^k} \cdot \prod_{i=1}^{h^k/2}\binom{2}{t_i^k} p^{t_i^k}(1-p)^{2-t_i^k} \right\} \end{aligned} \tag{5}$$

And the full likelihood function is simply the product of the above across all P pools. Note that we can write $a_{\vec{t^k}}^k \triangleq \binom{z^k}{x^k}(f^k(\vec{t^k}))^{x^k}(1-f^k(\vec{t^k}))^{z^k-x^k}$, and then the likelihood function is

$$L(p; \vec{x}, \vec{y}) = \prod_{k=1}^{P}\sum_{\vec{t^k}\in\{0,1,2\}^{h^k/2}} a_{\vec{t^k}}^k \prod_{i=1}^{h^k/2}\binom{2}{t_i^k} p^{t_i^k}(1-p)^{2-t_i^k} \tag{6}$$

in which $a_{\vec{t^k}}^k$ does not depend on p, and can therefore be pre-calculated to speed up calculations. We also denote

$$S_{\vec{t^k}} \triangleq \sum_{i=1}^{h^k/2} t_i^k \quad \text{and} \quad I_{\vec{t^k}} \triangleq \sum_{i=1}^{h^k/2}[t_i^k \in \{2,0\}] \tag{7}$$

(so that $I_{\vec{t^k}}$ is the count of t_i^k's which equal 2 or 0), and note that

$$\prod_{i=1}^{h^k/2}\binom{2}{t_i^k} p^{t_i^k}(1-p)^{2-t_i^k} = 2^{I_{\vec{t^k}}} p^{S_{\vec{t^k}}}(1-p)^{h^k-S_{\vec{t^k}}} \tag{8}$$

To find the value of p which maximizes L, we calculate the natural logarithm of the likelihood function, and take its first derivative:

$$\frac{d}{dp} lnL(p; \vec{x}, \vec{y}) = \sum_{k=1}^{P} \frac{\sum_{\vec{t^k}} a_{\vec{t^k}}^k 2^{I_{\vec{t^k}}} p^{S_{\vec{t^k}}-1}(1-p)^{h^k-S_{\vec{t^k}}-1}(S_{\vec{t^k}} - h^k \cdot p)}{\sum_{\vec{t^k}} a_{\vec{t^k}}^k 2^{I_{\vec{t^k}}} p^{S_{\vec{t^k}}}(1-p)^{h^k-S_{\vec{t^k}}}} \tag{9}$$

It is easy to verify that the likelihood is a concave function of p, and therefore its maximal value can be found using various optimization procedures.

LITERATURE CITED

BANSAL, V., 2010 A statistical method for the detection of variants from next-generation resequencing of dna pools. Bioinformatics **26**: i318–24.

DEPRISTO, M. A., E. BANKS, R. POPLIN, K. V. GARIMELLA, J. R. MAGUIRE, *et al.*, 2011 A framework for variation discovery and genotyping using next-generation dna sequencing data. Nat Genet **43**: 491–8.

MCKENNA, A., M. HANNA, E. BANKS, A. SIVACHENKO, K. CIBULSKIS, *et al.*, 2010 The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res **20**: 1297–303.

PAANANEN, J., R. CISZEK, and G. WONG, 2010 Varietas: a functional variation database portal. Database (Oxford) **2010**: baq016.