Supplementary Methods

Coverage calculations may be confounded by two important factors: First, there is an upward bias in coverage when including the tag SNPs themselves as part of the coverage calcuation. Consider a reference set of SNPs R such as the Phase II HapMap. For a given tag set T, some SNPs are captured either because they are contained in T, or they are in LD with a SNP in T (we call call this set of SNPs L). Thus, the naive estimate of coverage of all SNPs in the genome, G, would be:

$$\frac{L+T}{R} \tag{1}$$

That is, count all SNPs either in the tag set or in LD with a tag and divide this by all the SNPs in the reference set. However, this overestimates the true fraction of captured SNPs which are actually tags, since $G > R$. To correct for this, we define the genome-wide coverage as:

$$\frac{(\frac{L}{R-T})(G-T)+T}{G} \tag{2}$$

The $\frac{L}{R-T}$ part of the formula computes the fraction of the reference set captured because it is in LD with a tag but not a tag itself. Multiplying this fraction by $G-T$ yields a value for the number of SNPs captured by LD genomewide. Adding this number to the number of tags gives an estimate of the total number of SNPs genomewide which are captured by either LD or by being a member of the tag set. Dividing by G yields the coverage estimate. In the case where the reference set is the entire genome (i.e. $R = G$), Equation 2 sensibly reduces to Equation 1.

Coverage estimates will also be biased upwards if all or part of the reference set used for the estimate was used to select tag SNPs. In order to equate coverage of a given reference set to coverage of the genome, the reference set must be considered representative of all common SNPs. If, on the other hand, tag SNPs have been chosen specifically to capture all or part of the reference set then they are "overfitted" to the reference set compared to the set of all common SNPs in the genome. In the case of tag SNPs chosen using the Phase I HapMap data, we attempt to correct for this bias by considering coverage of the Phase II-only subset of R, which is unbiased with respect to the tag set. In that case the correction factor in Equation 2 is split into two parts, representing the overfitted portion of the reference, R_1, and the remainder, R_2:

$$\frac{\frac{L_2}{R_2-T_2}(G-R_1-T_2)+T_2+L_1+T_1}{G} \tag{3}$$

where L_1, T_1, L_2, T_2 are the portion of SNPs captured by LD (L_1, L_2) and the tag SNPs (T_1, T_2) in the Phase I and Phase II data respectively. In this case we estimate the coverage of the markers exclusively in Phase II, $\frac{L_2}{R_2-T_2}$. Multiplying this by the number of commons SNPs in the non-Phase I genome minus T_2 yields

an estimate for the number of SNPs captured genomewide via LD. We then add the number of Phase I and Phase II tags, plus the number of SNPs captured by LD from Phase I and divide this by G to get a corrected estimate of coverage. In the case where there is no overfitting (i.e. $R_1, T_1, L_1 = 0$), Equation 3 reduces to Equation 2. In the case where the two reference sets constitute the entire genome (i.e. $R_1 + R_2 = G$), it reduces to Equation 1.