FILE S1

Effects of larger resources for imputation

To determine the effect of using a larger number of resequenced strains on the accuracy of imputation, we assumed that the 62 WTCHG strains were all resequenced, and estimated the imputation accuracy of imputing the WTCHG genotypes from the mouse HapMap SNPs and the gap-filling SNPs, using the leave-one-out cross-validation for each of the 62 strains. Because each strain targeted for imputation now has 61 instead of 16 reference strains, the imputation accuracy is expected to be high. Overall, the errors are reduced from 4.86% to 2.45%, and the errors in the 36 classical inbred strains are reduced from 2.25% to 0.96%. In contrast, the accuracy in the high-confidence genotypes of the classical inbred strains is reduced from 0.35% to 0.16%. More importantly, high-confidence call rate was increased from 88.9% to 95.6% for the 36 classical inbred strains, and from 71.7% to 84.9% for all 47 strains. Several strains such as MRL/MpJ, C57L/J, C57BR/cdJ, PERA/EiJ and PWK/EiJ showed a substantial improvement in imputation accuracy when a larger set of reference strains was used, while many other wild-derived strains still retained imputation errors of greater than 10%.

Since the resequencing of more strains is expected to increase the imputation coverage significantly, we prioritized the strains that might be targeted for resequencing to improve the coverage, based on our analysis of the shared segments. To do this, we picked the strain that maximized the additional genomic coverage of shared segments with the other strains given the coverage by the resequenced reference strains. This procedure is repeated greedily to select the next target of reference strain given the previous set of reference strains. To increase the coverage including the wild-derived strains, many wild-derived strains are prioritized for resequencing. When considering only classical inbred strains, the strains with relatively higher imputation errors tend to be prioritized (Supplementary Table S4).

Next, we estimated the effectiveness of imputation when different numbers of mouse HapMap SNPs are collected. To do this, we selected a range of sparse subsets (10,000 to 1,000,000 markers) of the NIEHS/Perlegen SNPs with complete data in the resequenced strains and estimated the imputation errors for each of 12 resequenced classical inbred strains using leave-one-out cross-validation. As expected, accuracy increased proportionally to subset size. Selecting a 100,000 SNP subset gave an overall imputation error of 1.36% (high-confidence genotype error 0.36% with 93.8% call rate). This is comparable to the imputation accuracy using the current mouse HapMap SNPs. We note that the current size of the HapMap SNP is well powered to capture the majority of variation at low error rates and high confidence. (Supplementary Figure S3). A several-fold increase in SNP map density to 1,000,000 markers further optimizes these rates, and as current genotyping platforms can accommodate this number of assays this would be a viable design for the next generation mouse HapMap.