Supplemental material

Simulation to establish the power and sensitivity of the DNA pooling experiment

To investigate the pooling strategy, we began by creating two pools each of 35 DNA samples of CEPH individuals genotyped as part of the HapMap project [1]. Individuals were assigned to the two pools at random and each pool was genotyped with the 50K Xba array. The average standard deviation (SD) between measurements was 0.02; the correlation between allele frequencies as determined by individual genotyping and from pooling is 0.97; the correlation between differences between groups, obtained from pooling compared to individual genotyping is 0.845. These figures are comparable to the estimates obtained in a previous study using a similar approach [2].

We are testing so many markers that even though there is a high correlation between allele frequency differences estimated from pooled and individual genotyping, there will still be many cases where errors in the pooled estimates yield apparent significant differences between our two groups (false positives) and conversely many cases where there are apparent non-significant differences (false negatives). Therefore we need to determine the sensitivity and specificity of the pooled experiment to know how many markers we should individually genotype. Our first task is to establish that we can accurately simulate the proposed experiment, so that we can obtain the information on how well it is likely to perform.

We used the CEPH data to compare simulation estimates with observed allele frequencies, since we had access to both individual genotypes and to allele frequency estimates from pooled DNA. We calculated the significance, expressed as the logP values, of the allele frequency differences for every individually genotyped marker.

These differences arise purely by chance, since the two groups were chosen at random. We treated any logP greater than 4 as if it were a real effect (this threshold was chosen simply so as to provide a sufficiently large set of markers for the simulated experiment, not because it represents a genome-wide significance threshold). One should also note that since the sample size of the two CEPH pools is small, a logP > 4 translates into large allele frequency differences that are easier to detect by DNA pooling.

We then asked how many of these differences we would be able to detect if we chose from the pooling experiment markers for individual genotyping whose allele frequency differences were significant at a logP of 2 or 3 (SNPs with MAF < 0.05 were excluded from this analysis). There were 1,308 results from the CEPH pools that had a logP of 2 or greater. While only 2.2% of these were true positives (97.8% false positive rate) this included 90.6% of the 32 true positives SNPs (with logP > 4, based on the individual genotypes). When we increased the threshold to logP > 3 for individual genotyping, 6.4% of the results were true positives, but we only detected 65.6% of the true positives.

We then compared a simulation study with these estimates. In the simulation we randomly selected a frequency from the distribution of allele frequencies in the CEPH data, and a SNP specific measurement error from our experiment. We did this ten thousand times. At each iteration we calculated two frequencies whose average equaled the selected frequency and whose difference yielded a logP > 4. In 87% of the cases the effect (of logP > 4) was detected with threshold of logP > 2 and in 67% with a logP above 3. These results are comparable to those obtained from the real data.

1. Altshuler, D. et al. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).

2. Kirov, G. et al. Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics* **7**, 27 (2006).