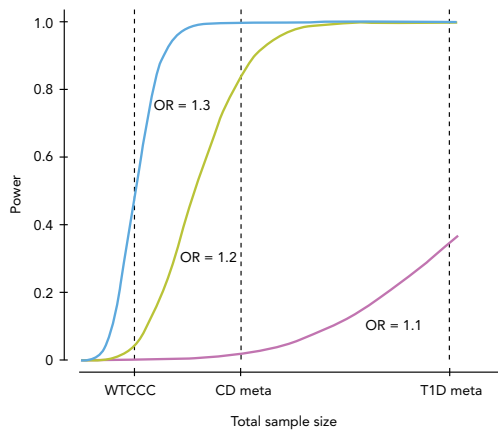




**Figure 2: Power of Meta-Analysis to Detect Disease Association**



Power for a combined analysis to demonstrate genome-wide significant association ( $P < 5 \times 10^{-8}$ ) for various effect sizes (odds ratios 1.1–1.3) for a 20% allele. Total sample sizes (GWAS plus replication) are shown for the WTCCC original design ( $N=5,000$ ), the CD meta-analysis<sup>3</sup> ( $N=14,871$ ), and the T1D meta-analysis<sup>4</sup> ( $N=34,180$ ). Very large sample sizes attained via meta-analysis are necessary to detect weak effects.

these key practicalities in imputation analysis is provided by de Bakker et al<sup>10</sup>. They provide detailed examples about annotating build, strand and allele information for SNPs on commercial platforms, advice on aligning these data to the HapMap, and information about correctly performing association tests on imputed data.

The initial CD meta-analysis used the HapMap2 data set consisting of 2.6 million SNPs in 60 individuals of European ancestry as the reference. Despite the success of this meta-analysis, it is likely that ever larger sample sizes will allow the detection of additional modest risk factors. A new extension to this study will add three additional sets of CD samples genotyped on the Illumina HumanHap550-Duo (comprising 3094 cases and 10,225 controls). This second-generation meta-analysis will use the new HapMap3 data release as a reference, consisting of roughly 1.5 million SNPs from two commercially available platforms (the Illumina Human1M-Duo and Affymetrix Human SNP array 6.0) genotyped in 200 European individuals. Despite having fewer total SNPs than HapMap2, this new resource is recommended for imputation for several reasons. First, data quality affects the accuracy of imputation and HapMap3 is more accurate than HapMap2. While most HapMap2 data were of high quality, the data set contained a small number of poorly performing SNPs, which can have adverse effects when cases and controls have been genotyped on different chips. The HapMap3 has extended the set of populations sampled (which is crucial for GWAS in non-European samples), and increased the number of individuals from each population. The larger reference sample size of HapMap3 offers substantial gains in imputation accuracy, especially for SNPs with <10% frequency. Finally, nearly all of the remaining common SNPs in HapMap2 are highly correlated with one or more SNPs in HapMap3; therefore, the additional SNPs in HapMap2 provide little additional information. The high-quality genotypes and larger sample panel in HapMap3 make it the current state-of-the-art reference set.

A number of different statistical frameworks have been used to tackle

the problem of genotype imputation, each of which has advantages and drawbacks. The initial CD meta-analysis used the popular programs MACH11 and IMPUTE12. These programs yield high accuracy of imputed genotypes via a hidden Markov model that captures certain aspects of population history such as the local recombination rate. The trade off for the complexity of these programs is that they run slowly and require a large amount of memory, making them less suitable for the large HapMap3 reference set. The current extension of the CD project is using another HMM-based program, BEAGLE, which achieves nearly the same imputation accuracy but runs faster and can scale more readily to reference sets with hundreds of samples<sup>13</sup>. BEAGLE's speed and the ease with which it incorporates the HapMap3 data make it a good choice for current imputation analysis, but different tools may be better suited to specific problems, as discussed by Ellinghaus et al<sup>14</sup>. Finally, it is worth noting that the developers of these algorithms are constantly improving their programs (e.g. IMPUTE v215) to enable quicker run-times or to provide new features.

The time required to impute the CD extension data set scaled approximately linearly with the number of GWAS samples. For example, BEAGLE required approximately 2000 CPU-hours to impute the HapMap3 SNPs into the 4686 WTCCC CD samples and 622 hours to impute into the 1452 Belgian/French samples. Imputation can easily be parallelized across sections of chromosomes and subsets of the sample, but each sample subset must contain a consistent mixture of cases and controls<sup>13</sup> to avoid introducing differential bias in the imputed allele frequency estimates. High memory requirements (>8GB) can pose problems, but both IMPUTE v2 and BEAGLE have configurable trade offs between memory usage and processing time. With today's technology, genome-wide imputation cannot be carried out on a standard desktop computer. However, given that it must only be done once for a given experimental data set, and it adds considerable value to expensive GWAS data sets, imputation is a tractable task for groups with even modest computational resources.

## Fast Imputation in a Meta-Analysis of Type 1 Diabetes

In a similar experiment, a GWAS of T1D involving 8000 UK samples was undertaken with the Illumina HumanHap550-Duo BeadChip. We chose the Illumina HumanHap550-Duo because it provided extremely high-quality genotypes and excellent coverage of common variation in European samples. A meta-analysis was then completed using these data and two previous GWAS run with the Affymetrix 500K chips. By using imputation, these data could be confidently integrated across platforms, allowing the selection of the higher coverage Illumina chip<sup>16</sup> for the second GWAS.

Approximately 1500 control samples from the 1958 British Birth Cohort overlapped between the individual GWAS samples and were genotyped on both chips. This set of samples allowed a much simpler imputation method to be employed, where linear regressions of nearby SNPs were used as predictors (implemented in snpMatrix<sup>17</sup>) and with similar accuracy to the HMM methods described above. Imputation accuracy was assessed in the 1958 BBC samples that were typed on both the Affymetrix 500K and Illumina HumanHap550-Duo platforms. The SNPs on the Affymetrix chip were used to impute SNPs on the Illumina chip, and vice versa. The predictions were then compared to the true genotypes on the unused chip in each case.





