GAW17 Simulated Mini-Exome Data

The files contained in this CD are comma-delimited. The first line of each file lists the field names.

There two data sets. One consists of a collection of 697 unrelated individuals and their genotypes and phenotypes. These are subjects from the 1000 Genomes Project. They have 7-character IDs of the form NAnnnnn. The second data set is comprised of 697 individuals in 8 extended families and their genotypes and phenotypes. These individuals have 4-digit, numeric IDs. The 202 founders in the family data set were chosen at random from the set of unrelateds.

There is a pedigree file for each data set:

families.ped - ID,FA,MO,SEX,AGE
unrelateds.ped - ID,SEX,AGE,Population

Sex is coded 1=male, 2=female. The sex and population data for the unrelateds are taken from the 1000 Genomes Project. The extended families are loosely based on the families from previous GAWs. The ages in the family data set were assigned randomly to the unrelateds so that the distribution of ages is identical across the two data sets. Note that a subject from the 1000 Genomes Project who was chosen to be a founder in the family data set may have been assigned a different age or sex in the family data set. This is not an issue as only autosomal SNPs were considered.

SNP genotypes were obtained from the sequence alignment files provided by the 1000 Genomes Project for their pilot3 study (go to http://www.1000genomes.org for more information about the 1000 Genomes Project). The UnifiedGenotyper method from the Genome Analysis Toolkit (GATK) package (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit) was used for the detection of SNPs and for the calling of SNP genotypes. A male human genome (human_b36_male.fasta.gz) was used as the reference genome sequence for both male and female alignments.

The UnifiedGenotyper method was run twice on the alignment files. The first time it was allowed to scan freely through the alignments to search for SNPs. Genotypes that were not homozygous for the reference base allele were called for the SNPs detected. A subset of the detected SNPs was then selected by choosing only those SNPs whose genotypes were called from an alignment of 10 or more sequencing reads. During the second run genotypes, including those homozygous for the reference base, were called only for the subset of SNPs selected in the first run.

The 1000 Genomes Project genotypes were not phased, and some genotypes were missing due to incomplete sequence coverage in some individuals. We utilized the program fastPHASE [http://depts.washington.edu/uwc4c/express-licenses/assets/fastphase/] to infer missing genotypes and haplotypic phase. In the family data set, we used the program CHRSIM [Speer et al. 1992] to drop the phased founder genotypes through the rest of the pedigree. Reoombination was taken into account, with a single obligate crossover event occurring on each chromosome.

The genotypes were held fixed for all 200 simulation replicates. The genotypes for the family data set and the unrelateds are stored in files named cN_snps.unr and cN_snps.fam, respectively, where N is the chromosome number (1-22). The SNPs in each of these files are stored in basepair order. The file snp_info contains, for each SNP, the name of the SNP, its chromosome and basepair location, the name of the gene in which it is located, whether the SNP is synonymous or nonsynonymous, and the minor allele and MAF. There is a total of 24487 SNPs, all of which are autosomal. The file gene_info contains an entry for each gene, which gives the gene name, chromosome, start and stop locations,

gene length and orientation. There are 3205 genes.

A total of 200 replicates of the trait simulation were carried out in both data sets. The traits which are being made available are Q1, Q2, Q4, and AFFECTED (coded 0=no 1=yes). For each replicate, files named fam_phen.N and unr_phen.N (N=1,200) were generated for the family data set and the unrelateds, respectively. These files contain the fields ID, SEX, AGE, SMOKE, Q1, Q2, Q4, and AFFECTED. The SEX and AGE fields are identical to those in the pedigree files, but were included in the phenotype files for convenience.The smoking status covariate, SMOKE, varies across the replicates.

For the family data set, fully informative markers were generated at each gene (recombination was not allowed within genes) and used to compute IBDs at each gene location under the rationale that family-based data sets are likely to have previous STR or high density SNP genotyping that could be used to estimate IBDs. The IBDs are stored in comma-delimited files named ibd_GENE (where GENE is one of the 3205 genes). The IBD files do not contain field names on the first line. The fields are ID1, ID2, and IBD. The IBD files are gzipped to save space.