1 Summary and recommendations for dbGaP users

A total of 9,012 study participants were genotyped on the Illumina HumanOmniExpress array and included in this dbGaP posting. The median call rate is 99.95%, and the error rate estimated from 204 pairs of duplicated study samples is 2.84e-06. Genotypic data are provided for all participants and SNPs. Generally, we recommend selective filtering of genotypic data prior to analysis to remove (1) large ($>$ 5 Mb) chromosomal anomalies showing evidence of genotyping error and (2) whole samples with an overall missing call rate (MCR) $>$ 2%. In this study, all samples had MCR $<$ 2%. There were 35 large chromosome anomalies (31 in study samples) that were filtered, i.e. genotypes in the anomaly regions were set to missing. A composite SNP filter is provided, along with each of the component criterion so that users may vary thresholds (see Table 1). A preliminary association test is described as an example of how to apply the recommended filters. Additional specific recommendations for filtering genotype data are *italicized* in this report.

2 Project overview

The Wisconsin Longitudinal Study (WLS)[1] is a long-term study of a random sample of men and women who graduated from Wisconsin high schools in 1957 and their siblings. The WLS panel started out with a panel of 10,317 members from the class of 1957. Over time a second panel of 8,734 randomly selected siblings of the original graduate panel were recruited for the study. Of these combined panel members 9,027 contributed saliva for genetic analysis. Survey data were collected from the original respondents or their parents in 1957, 1964, 1975, 1992, 2004, and 2011 and from a selected sibling in 1977, 1994, 2005, and 2011. WLS data provide a detailed record of educational, social, psychological, economic and mental and physical health characteristics of a relatively homogeneous population that is almost entirely of Northern and Western European ancestry. Saliva was first collected in 2007-8 by mail. Additional samples were collected in the course of home interviews that began in March 2010. The addition of genetic data to WLS complements the store of extensive WLS phenotypic data and takes advantage of recent developments that have vastly increased opportunities for genetic studies of aging, behavior, cognition, personality, mental health, health, disease, and mortality. Researchers interested in linking the genetic data to the WLS survey data should email wls@ssc.wisc.edu.

3 Genotyping process

A total of 9,472 study samples, including planned duplicates, were successfully genotyped at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. These 9,472 study samples include 223 samples derived from two unique WLS control non-participants, which are not included in the dbGaP posting. There were 198 HapMap samples included as genotyping controls. Except for HapMap controls, all DNA samples were extracted from stored saliva using a modified Oragene extraction protocol, in order to amend the samples that were frozen prior to the extraction. Samples were genotyped in batches corresponding to 96-well plates with one batch per plate. On average, each batch contained two HapMap controls and three duplicate study samples. For 204 pairs of study sample duplicates, the two members of each pair were genotyped on separate plates.

Genotyping was performed at CIDR using the Illumina HumanOmniExpress array (humanomniexpress-24-v1-1, BPM annotation version A, genome build GRCh37/hg19) and using the calling algorithm GenomeStudio version 2011.1, Genotyping Module version 1.9.4, GenTrain Version 1.0. The array consists of a total of 713,014 SNPs. Two updates were made to the initial Illumina manifest. First, prior to genotype calling, CIDR corrected chromosome designations for numerous XY SNPs initially annotated as X or Y. These SNPs occur in pseudo-autosomal (PAR1, PAR2) regions or in the X-translocated region (XTR). Second, prior to genotype data cleaning, genomic positions were adjusted for insertion/deletion (indel) variants to match the convention used in the 1000 Genomes Project imputation

[1] http://www.ssc.wisc.edu/wlsresearch/

reference panels. (See "chrom", "chrom.ilm", "position", and "position.ilm" in "SNP_annotation.csv" for more details on chromosome and position updates.) While the array contains both SNPs and non-SNP variants (i.e., indels), in this report we use the term "SNP" more generically to refer to all genotyped variants.

4 Quality control process and participants

Genotypic data that passed initial quality control at CIDR were released to the Quality Assurance/Quality Control (QA/QC) analysis team at the UW GAC (University of Washington Genetic Analysis Center), the study investigator's team, and to dbGaP. These data were analyzed by the analysis team at UW GAC, and the results were discussed with all groups in periodic conference calls. Key participants in this process and their institutional affiliations are given in Appendix A. The results presented here were generated with the R packages *GWASTools* [1], *GENESIS* [2], and *SNPRelate* [3] unless indicated otherwise. The methods of QA/QC used here are described by Laurie et al. [4].

5 Sample and participant number and composition

In the following description, the term "sample" refers to a DNA sample and, for brevity, "scan" refers to a genotyping instance, including genotyping chemistry, array scanning, genotype calls, etc.

A total of 9,606 samples (including planned duplicates) from study subjects were put into genotyping production, of which 9,472 were successfully genotyped and passed CIDR's QC process (Table 2). The subsequent QA/QC process did not identify any further sample exclusions due to low sample quality; however, 12 scans of questionable identity were identified and are excluded from posting. A further six samples were included in most analyses but later removed from the posted dataset due to withdrawn consent. The set of scans to be posted include 9,231 study participants and 198 HapMap controls. The 9,231 study scans derive from 9,012 unique subjects and include 219 pairs of duplicate scans (Table 3). Note one member of each of 15 monozygotic (MZ) twin pairs is counted as a duplicate scan.

The study subjects occur as 4,601 singletons and 2,263 families of 2–4 members each. Most study families consist of a priori known sibling pairs, though additional first, second, and third degree relationships were discovered by examining genetic relationships, discussed further in Section 8. The 198 HapMap control scans derive from 97 unique subjects, of which all are replicated two or more times. The HapMap controls include 12 singletons, two duos, and 24 trios, and one large CEU family[2].

6 Annotated vs. genetic sex

To compare annotated and genetic sex, we examine both X chromosome heterozygosity and the means of the intensities for X and Y chromosome probes. The expectation is that male and female samples will fall into distinct clusters that differ markedly for both metrics. The plots of X and Y chromosome intensity and heterozygosity in Figure 1 show the expected patterns: two distinct clusters corresponding to male and female samples. There were four discrepancies between annotated (expected) and genetic (observed) sex in this study. Two of the discrepancies could not be resolved or explained. These two samples are excluded from the dbGaP posting due to questionable identity. One discrepancy was explained by a known sex reassignment. The last discrepancy did not result in a sample exclusion because the sample showed the expected genetic relationship with a known full sibling and was thus assumed to have a correct linking between genetic and phenotypic records.

Deviations from expected intensity and heterozygosity on the X and Y chromosomes can also be used to detect potential sex chromosome anomalies. We observed the following potential sex chromosome anomalies: four XXY males, one XYY male, four XY/Y0 or "loss of Y" males, three XXX females, and six XX/X0 females, highlighted in Figure 2. Most of these likely sex chromosome anomalies were also identified by CIDR

[2]CEU: Utah residents with Northern and Western European ancestry from the CEPH collection