Health and Retirement Study: Candidate Gene and SNP Data Description

November, 2014

Jessica Faul, Jennifer Smith, and Wei Zhao
Health and Retirement Study, University of Michigan, Ann Arbor, MI

Table of Contents

*Correspondence may be addressed to hrsquestions@umich.edu

Purpose

The purpose of the HRS Candidate Gene and SNP files is to provide data users access to carefully selected subsets of the HRS genotype data available on dbGaP. These are smaller and more manageable files designed for data users who are interested in a specific gene or SNP. Users must still have dbGaP approval before requesting and gaining access to these files from HRS.

Currently, there are two sets of files available – 1) Cognition & Behavior and 2) Longevity. There are separate file description documents that provide details on each specific data package. This document provides an overview and detailed information on the sample collection, genotyping, imputation and annotation of the HRS genetic data common to the source data for both sets of genetic data files.

Overview

Sample

The Health and Retirement Study (HRS) is a longitudinal survey of a representative sample of Americans over the age of 50. The current sample is over 26,000 persons in 17,000 households. The study interviews respondents every two years about income and wealth, health and use of health services, work and retirement, and family connections. DNA was extracted from saliva collected during a face-to-face interview in the respondents' homes. In 2006, saliva was collected using a mouthwash collection method. Starting in 2008, saliva was collected using the Oragene Collection Kit. These data represent respondents who provided DNA samples and signed consent forms in 2006 and 2008. Data from 2010 and 2012 will be added once they become available on dbGaP.

SNP Genotyping and Imputation

SNP genotyping was performed using the Illumina Omni2.5 Beadchip. SHAPEIT2 was used to pre-phase haplotypes, and IMPUTE2 was used for imputation to the 1000 Genomes Project phase I integrated variant set (v3, released March 2012). Relatedness among HRS participants was calculated by estimating a kinship coefficient, and familial relationships were taken into account during the imputation process when possible. Samples with a missing call rate >2% were excluded prior to imputation, and thus imputation was performed on a total sample size of 12,454. Quality control filters that were applied to the genotype data prior to imputation included SNP Hardy-Weinberg Equilibrium (HWE) p-value: $p<10^{-4}$, SNP missing call rate ≥2%, and other criteria as specified in Table 1 of the "Quality Control Report for Genotypic Data". SNPs were not filtered by minor allele frequency prior to imputation, except for the exclusion of monomorphic SNPs. To ensure an unrelated analysis sample for this data release, 87 participants were removed for a final analysis sample size of 12,367. In addition, 8 participants were excluded from X chromosome imputation due to chromosomal abnormalities; thus genes that are on the X chromosome have imputation data available for only 12,359 participants. For additional details on quality control and imputation of the HRS data, please refer to "Quality Control Report for Genotypic Data" and "CIDR Health Retirement Study Imputation Report – 1000 Genomes project reference panel", both available on the HRS website.

Gene and SNP Annotation

Position Information

Position information for genes was obtained using the ENCODE/GENCODE Complete Version 17 genome build from the UCSC Genome Browser (genome.UCSC.edu). As some genes have multiple transcripts, the "start" and "stop" positions of the gene were determined by taking the most upstream "start" and most downstream "stop" position of all known transcripts for the gene. This ensures that all coding sequences of the gene were captured. SNP position information was from NCBI build 37.

SNP Name Mapping File

There are several different naming conventions for SNPs that are relevant to this data. The most common naming convention is the rs number (the letters "rs" followed by numbers). Other types of names that are used include the kgp number (the letters "kgp" followed by numbers) or a name based on chromosomal location (chromosome followed by basepair location of the SNP).

The HRS data consists of SNPs with all three types of names. In order to help users easily obtain rs numbers for SNPs of interest, we have provided a mapping file called "HRS_1000G_rsid_map.csv". This file maps the SNP name provided in the HRS data ("SNP_name_in_HRS_dataset") to its name in the annotation from the 1000 Genomes Project Reference panel data ("1000Genome_name") or the annotation from the UCSC Genome Browser ("UCSC_genome_name"). To construct the mapping file, we first used 1000 Genome Project Reference panel data downloaded from the IMPUTE2 website (the "ALL (macGT1)" file available at: https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html) to map SNPs by chromosome and basepair position. We were able to obtain rs numbers for the majority of SNPs using this method; however, there were a small number of SNPs that did not have rs numbers in the 1000 Genomes Project Reference panel data. To find rs numbers for these SNPs, we used a separate annotation file downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/snp141.txt.gz).

Flipped Strand Issues and Correction Method

Mid 2014, we became aware of an annotation issue with the Illumina HumanOmni2.5-v1_D manifest that caused strand flip errors in the annotation for approximately 20,000 SNPs genotyped in the Health and Retirement Study 2006/2008 samples. This manifest also served as the foundation for the 1000 Genomes imputation for those samples. We systematically investigated the problem and identified two issues with the Illumina HumanOmni2.5-v1_D manifest that led to the strand flip errors. The issues are described in detail in "HRS1-2_dbGaPUserInfo_v3.docx", available on the HRS website.

In order to correct the flipped strand issue, users should take the following action. SNPs that are affected by the strand swap are flagged as *problem.type2.SNP = TRUE* in "HRS1-2_HumanOnmi2.5v1_D_flaggedSNPs.xlsx", available on the HRS website. Users should switch the coded and non-coded allele annotation for the affected SNPs in the annotation files. For example, if coded_allele=A and noncoded_allele=T, then coded_allele should be T and noncoded _allele should be A.

Calculation of Principal Components

Principal Components in All Unrelated Study Subjects

To investigate population structure, principal components analysis (PCA) was performed on a total of 12,419 HRS samples. Principal components were calculated after excluding related individuals but prior to excluding samples with missing call rates >2%. PCA was implemented in the *SNPRelate* package in R. In this data package, we have included the top 10 principal components (PCs) for the 12,367 individuals in the final analysis sample in a file called "HRS12367_10PCs.txt". For additional information on the calculation of PCs in all unrelated study subjects, please refer to "Quality Control Report for Genotypic Data".

Principal Components in Ethnic-Specific Samples

An analysis sample of 8,652 European Americans was defined using the "hwe.eur" filter described in "Quality Control Report for Genotypic Data". Briefly, this sample includes unrelated, self-identified White study subjects with missing call rate <2% who have relatively homogenous ancestry (defined by falling within 1 SD of all self-identified Whites for eigenvectors 1 and 2 in the PCA of all unrelated study subjects described above). If you are conducting an analysis that includes only European Americans, we recommend that your analytic sample includes only these 8,652 individuals. The top 10 PCs and identifiers for these individuals are provided in a tab delimited text file called "EA8652_10PCs.txt".

An analysis sample of 1,519 African Americans was defined using the "hwe.afr" filter described in "Quality Control Report for Genotypic Data". Briefly, this sample includes unrelated, self-identified African Americans with missing call rate <2% who have relatively homogenous ancestry (defined by falling within 2 SD of all self-identified African Americans for eigenvector 1 and 1 SD for eigenvector 2 in the PCA of all unrelated study subjects described above). If you are conducting an analysis that includes only African Americans, we recommend that your analytic sample includes only these 1,519 individuals. The top 10 PCs and identifiers for these individuals are provided in a tab delimited text file called "AA1519_10PCs.txt".

Principal components were calculated for European Americans and African Americans separately using *SNP & Variation Suite 7* software package from Golden Helix (Bozeman, MT). Input SNPs were comprised of autosomal SNPs in the "SNP_qual_filter_extract.txt" file, created using the "SNP_analysis.csv" file described in the "Quality Control Report for Genotypic Data". The "SNP_qual_filter_extract.txt" file is available on dbGaP. Prior to PCA, we excluded SNPs with minor allele frequency <0.05. We also used *PLINK* to do linkage disequlibrium based SNP pruning in order to keep only independent SNPs prior to PCA calculation. For pruning, we selected SNPs with all pairs having $r^2<0.1$ in a sliding window of 50 SNPs.