Health and Retirement Study:
Candidate Genes for Cognition/Behavior

November, 2014

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

Table of Contents

Description of Files 1

Phenotype Category 1

Data Groups 1

File Types 1

APOE files 3

Summary of Files 4

User Recommendations 4

*Correspondence may be addressed to hrsquestions@umich.edu

Description of Files

Phenotype Category

The data package contains the data for the phenotypic categories of **Cognition and Behavior**. Information about the genes and specific SNPs to include in this data release were compiled with input from an expert panel. This is not a complete list of genes and SNPs potentially associated with these phenotype categories; it is merely a selection of the most biologically promising candidate genes.

Data Groups

There are four groups of data included in this release.

- "Greatest Hits" SNPs: This file includes specific SNPs that have been identified from the literature as being associated with cognition and behavior. The file contains SNPs from multiple genes. This data file is called "SNPs_1000G_Congition.txt"
- Full Genes: These files include 56 genes (one data file for each gene). The "start" and "stop" sites of each gene were identified by position as described above. All SNPs within the gene itself and within 5,000 base pairs (5kb) on either side of the gene were included in the data file. All genes that had a SNP included in the "Greatest Hits" file also have a full gene file. The data files are referred to by their gene name. For example, the data file for the ADRB2 gene is called "ADRB2_SNP.txt".
- Alzheimer SNPs: This file includes the SNPs that were identified as being associated with late-onset Alzheimer's disease in a recent genome-wide association study and meta-analysis conducted of 74,046 individuals (Lambert, et al. 2013. Nature Genetics 45: 1452-1458). The file contains SNPs from multiple genes. The data file is called "SNPs_1000G_Alzheimer.txt".
- APOE variants: This file includes information on the two SNPs that comprise the ε2, ε3, and ε4 isoforms of the ApoE protein, as well as the best guess genotype for the isoform itself. The data file is called "APOE_rs7412_rs429358_SNPs.txt". *Please refer to documentation for this file (below) for more information.*

File types

Within each data group, there are three file types.

- Documentation: A list of candidate genes is provided for the Cognition and Behavior phenotypes. This file is called, "Candidate Genes Cognition_Behavior.docx". This file contains the following sections: 1) SNPs of interest, 2) Genes of interest, and 3) References. The "SNPs of Interest" section gives annotation for SNPs which are of special interest to investigators (see Column Descriptions for Annotation Files below). The "Genes of Interest" section provides additional information about each gene, including its chromosomal position, traits that have been studied for association with that gene, the total number of SNPs that were imputes from 1000 Genomes Project, and the number of imputed SNPs with INFO>0.8. There is also a documentation file for the Alzheimer Disease SNPs, called “Alzheimer Disease SNPs.docx”, which lists the annotation for each of the SNPs (see Column Descriptions for Annotation Files below).

- Data Files: These files are **Tab delimited .txt** files. Rows are ID numbers of HRS participants, and columns are SNPs. The value provided for each SNP is its dosage. The dosage for a person is the number of coded alleles that the person has (ranging from 0 to 2). The coded allele receives a value of “1” and the non-coded allele receives a value of “0”. For example, if “A” is the coded allele for SNP rs99876 and “T” is the non-coded allele for that same SNP, a dosage of 0 would mean that the person had genotype TT for SNP rs99876, a dosage of 1 = AT, and a dosage of 2 = AA. Please note that since these genotypes are imputed using external data (1000 Genomes data), the dosage may not be exactly 0, 1, or 2 for each person for each SNP. The dosage incorporates uncertainty about which genotype a person truly has for a given SNP. That is, if a person has a dosage of 1.5, it means that the imputation algorithm cannot determine whether the person’s true genotype dosage is 1 or 2 (there is an equal probability of the person having either of those dosages). Thus, a dosage of 1.5 for SNP rs99876 would mean that the person is equally likely to truly have genotype AT or AA.
  ***The APOE file is set up a bit differently than other text files. Please refer to APOE documentation for more information.*
- Annotation Files: These files are **Space Delimited .txt** files, and they have the word “info” in their file name. For example, the annotation file for the “Greatest Hits” Cognition and Behavior file is “SNPs_1000G_Cognition_info.txt”, and the annotation file for the ADRB2 gene is called “ADRB2_info.txt”. Each data file has an annotation file. The annotation files include the columns below.

Column Descriptions for Annotation Files

Column Name	Description	Type of Field
SNP	SNP name (usually an rs number)	Text
chr	chromosome that the SNP is on	Numeric
position	base-pair location of the SNP on the chromosome	Numeric
coded_allele	allele that is coded as a “1” in the dosage files	A, C, T, or G
non_coded_allele	allele that is coded as a “0” in the dosage files	A, C, T, or G
exp_freq_coded_allele	frequency of the coded allele in the full HRS sample	Percentage (ranging from 0 to 1)
info	measure of the observed statistical information associated with the allele frequency estimate (a measure of SNP imputation quality)	Numeric (ranging from 0 to 1)
certainty	average certainty (posterior probability) of best-guess genotypes (a measure of SNP imputation quality)	Percentage (ranging from 0 to 1)

APOE files

The APOE protein has three major isoforms (ε2, ε3, and ε4). These isoforms are formed from two SNPs, rs7412 and rs429358. We used 1000 Genomes imputed dosages to first get the "best guess genotypes" for each of these two SNPs for each HRS participant. Then, we used the "best guess genotypes" to infer the APOE isoform according to the following algorithm:

rs7412 best guess genotype	rs429358 best guess genotype	APOE genotype
T/T	T/T	ε2/ε2
C/T	T/T	ε2/ε3
C/T	C/T	ε2/ε4
C/C	T/T	ε3/ε3
C/C	C/T	ε3/ε4
C/C	C/C	ε4/ε4

The data file for APOE is called "APOE_rs7412_rs429358_SNPs.txt". Column descriptions for the APOE data file are below. The annotation file for APOE is called "APOE_rs7412_rs429358_info.txt" and has the same columns as all other annotation files (see above).

Column Name	Description	Type of Field
local.subjectID	subject ID	Numeric
rs7412_dosage	dosage for rs7412	Numeric (ranging from 0 to 2)
rs7412_best_guess_genotype	best guess genotype for rs7412	C/C, C/T, or T/T
rs7412_posterior_probability	posterior probability for rs7412 (a measure of SNP imputation quality)	Percentage (ranging from 0 to 1)
rs429358_dosage	dosage for rs429358	Numeric (ranging from 0 to 2)
rs429358_best_guess_genotype	best guess genotype for rs429358	C/C, T/C, or T/T
rs429358_posterior_probability	posterior probability for rs429358 (a measure of SNP imputation quality)	Percentage (ranging from 0 to 1)
APOE_imputed	best guess isoform of APOE	e2/e2, e2/e3, e2/e4, e3/e3, e3/e4, e4/e4

<u>Summary of Files</u>

Below is a table that summarizes the files included in this release. In this summarization, we have not listed all of the full gene files; we have only included one of the full gene files as an example (ADRB2).

File Name	File Format	File Type	Data Group
Candidate Genes Cognition_Behavior.docx	Word document	Documentation	Information on genes related to Cognition and Behavior
SNPs_1000G_Cognition.txt	Tab delimited text	Data file	“Greatest Hits” SNPs
SNPs_1000G_Cognition_info.txt	Space delimited text	Annotation	“Greatest Hits” SNPs
ADRB2_SNP.txt	Tab delimited text	Data file	Full Gene
ADRB2_SNP_info.txt	Space delimited text	Annotation	Full Gene
Alzheimer Disease SNPs.docx	Word document	Documentation	Alzheimer SNPs from Lambert, et al. (2013)
SNPs_1000G_Alzheimer.txt	Tab delimited text	Data file	Alzheimer SNPs from Lambert, et al. (2013)
SNPs_1000G_Alzheimer_info.txt	Space delimited text	Annotation	Alzheimer SNPs from Lambert, et al. (2013)
APOE_rs7412_rs429358_SNPs.txt	Tab delimited text	Data file	APOE variants
APOE_rs7412_rs429358_info.txt	Space delimited text	Annotation	APOE variants

<u>User recommendations:</u>

Based on the standards in the field, we recommend that users exclude SNPs from their analyses that have low INFO scores. Two commonly used cutoffs are INFO<0.8 (conservative) or INFO<0.3 (liberal).

For the APOE file, we recommend that users exclude subjects with posterior probability <0.8 for either rs7412 or rs429358.

If you wish to convert dosage data into best guess genotype data, we recommend using the following algorithm: genotype = 0 if dosage ≤ 0.5; genotype = 1 if 0.5 < dosage ≤ 1.5; genotype = 2 if dosage > 1.5.