

Documentation for Genetics Data Sets

GENETICS DATA ANALYSIS FILES	2
1. General description.....	2
2. Cross reference of dataset names with exact source.....	3
3. Dataset structure and contents	3
4. Condition of data	4
5. Dataset index formulation and key variable mapping	7
6. General strategies for manipulating and merging the data	7
7. Recent updates to dataset.....	7
LISTINGS.....	8
APPENDIX I: Released SNPs and Genotyping Methodology for DNA Dataset	
APPENDIX II: European Ancestry Admixture Data	
APPENDIX III: European Ancestry Illumina Data	
APPENDIX IV: Proc Contents Genetics Datasets	
APPENDIX V: Notes on the Use of Telomere Length Data	

Documentation for Genetics Data Sets

GENETICS DATA ANALYSIS FILES

There are now four datasets posted on the Health ABC website under the Datasets&Documentation link, sublinks Current Datasets/Genetics Datasets:

DNA.sas7bdat
DNARaceGen.sas7bdat
EurAncestry.sas7bdat
EurAncestry2.sas7bdat

DNA (version 2012-11-1) contains genotype data as well as telomere length measurements from several Health ABC ancillary studies. EurAncestry (version 1.15) contains genotyping of 35 ancestry-informative markers, as well as European admixture estimates, for African-American participants in Health ABC. EurAncestry2 (version 1.05) contains Illumina-based genotyping of 1332 different ancestry-informative markers, also for African-American participants. DNARaceGen (version 1.4) lists the Health ABC variable RACE and GENDER for each participant with a DNA sample, identified by barcode only. This file is intended for investigators submitting data to the Coordinating Center to use to calculate Hardy-Weinberg statistics and will not be discussed further.

In addition the following file, not specific to any year but updated each time data are released, can be found at the top of the Current Datasets listing. Note the analyst must download this new format file to decipher the value labels of the DNA variables.

Formats.sas7bdat SAS Format Library

To use the data, please contact the PI at your site.

Listings

The following files are zipped together in self-extracting documents also available on the Health ABC website under the Current Datasets listing

Formats.lst	Text file listing all formats and value descriptions contained in the SAS Format Library
DNAListings	Zipped file containing Proc Contents and Frequencies and Univariates for the DNA dataset
EAListings	Zipped file containing Proc Contents and Frequencies and Univariates for the EurAncestry dataset
EA2Listings	Zipped file containing Proc Contents and Frequencies and Univariates for the EurAncestry2 dataset

1. General description

DNA The DNA file contains genotype and other DNA-based information about the participants enrolled in the study. Key variables included are:

HABCID	Health ABC Enrollment ID# without the 2-letter prefix
VERSION	Data Analysis File Version #

There are 3,074 observations in the DNA file. 51 participants did not have buffy coat collected at baseline, or buffy coat was collected but good quality DNA could not be extracted, so they have no genetic data, although they are included in the dataset. Eight participants were added in version 1.1 as a result of further cleaning of formerly unsuccessfully extracted samples. Thus, these 8 participants have results only for investigators who received DNA after the re-extraction. Additional DNA samples were extracted more recently, and 15 participants have no data for SNPs genotyped before version 1.3. Finally, 6 of 8 participants have been reinstated who were temporarily dropped in version 1.4 until questions about the true identity of the DNA sample have been resolved (see section 4a. Known data errors, below).

In some cases, all participants with available DNA have been genotyped; in others only a subset was genotyped (See Appendix I).

EurAncestry The EurAncestry file contains admixture data and ancestry estimates for 1236 of the African-American participants in Health ABC (see page 4, section 4.a. Known data errors, for why this N has gone down by one in version 1.14). In addition to the raw data, ancestry estimates (estimated percent European ancestry, PcntEA), upper and lower limits of the 95% confidence interval, and the number of SNPs successfully genotyped (SNPs_typed) are also included.

EurAncestry2 The EurAncestry2 file contains admixture data for 1155 of the African-American participants in Health ABC (DNA from 81 did not amplify well enough to be used with this method, and one was excluded for reasons explained on page 4, section 4.a, Known data errors).

2. Cross reference of dataset names with exact source

A complete list of variable names can be found in Contents.lst (search under DNA, EurAncestry, or EurAncestry2, respectively), or in Appendix I (DNA), Appendix II (EurAncestry), and Appendix III (EurAncestry2). Data received with sufficient documentation to date are included. New variables will be added periodically as data and documentation are received.

3. Dataset structure and contents

All genetics datasets contain a single observation per participant. Point mutation genotypes are coded as 0, 1, or 2, for the three possible combinations of alleles (one triallelic SNP is coded as 1, 2, 3, or 4 for the four combinations found in Health ABC participants). These values are formatted when the SAS format library is used. Microsatellite repeats are coded as xxx/yyy, to indicate the number of repeats on one

allele (xxx) and the other allele (yyy). X-linked SNPs are shown as homozygotes for men, unless they were anomalously genotyped as heterozygotes. Only females are used Hardy-Weinberg equilibrium calculations of X-linked SNPs.

Key variables:

HABCID HABC Enrollment ID without the 2-letter prefix

4. Condition of data

a. Known data errors: There has been some question about the genetic race group of a small number of participants. In four pairs of baseline DNA samples (one African-American and one white participant in each pair), it has been confirmed that the baseline DNA does not match DNA extracted from buffy coats stored for the same participant in a later year, and the fact that the DNA was switched within pairs has been confirmed on a very large panel of SNPs (1299 SNPs, data not yet included). **These IDs were corrected as of version 1.12, and it is no longer necessary for investigators to compare analyses with and without these samples, since we are now completely sure of their identity.** Several more suspect-race samples were confirmed to be correctly identified (presumably, these are just cases of mixed-race participants whose self-identified race does not match their admixture results).

In addition, eight samples were identified that appeared to be the wrong gender. Again, later-year buffy coats for many of these were successfully extracted and the identity mix-ups for six were corrected in the panel of 1299 SNPs, along with a seventh that was previously undetected (in this case, two males and one female had their DNA interchanged). **Six of the eight samples removed from version 1.4 have been restored, and all seven ID corrections have been made as of version 1.5. The remaining two samples, most likely not correctly identified, but with no known correct identification, continue to be excluded from the DNA, EurAncestry, and EurAncestry2 datasets (the latter two datasets each had only one suspect sample).**

In the first versions of EurAncestry (version 1.0 and 1.01), the upper limit of the 95% confidence interval for the European ancestry estimate was inadvertently doubled. This error has been corrected as of version 1.2. **Investigators who may have used the confidence interval in earlier analyses are advised to repeat their analysis with version 1.3 or later data.** In addition, the previous versions of this dataset had omitted 187 participants as the result of a misunderstanding. These participants have been restored.

A previously released SNP (ACE_ID) was recoded as 1:D/X (at least one D allele) vs. 2:I/I (no D alleles) because it was found that the deletion allele tended to over-amplify. As of version 1.2, the participants with D/X calls have been regenotyped, and the coding has been restored to D/D, D/I, I/I.

For the EurAncestry data, some SNPs did not pass QC. These SNPs have been omitted from the dataset. After exclusion of these SNPs, two participants had no successfully

typed markers. For these two participants, all 35 markers have been set to the SAS special missing value .T.

b. Strengths and weaknesses of dataset items: Several important new variables have been added as of version 1.3:

- **TELOBATCH AND TELOPLATE** – The first analyses of the telomere length data have shown that assay variability needs to be controlled for either at the plate level or the batch level (all plates done on a certain date are in the same batch). Although it is not clear which level needs to be controlled for, two new variables, **TELOBATCH** and **TELOPLATE** have been included to allow the analyst the flexibility to experiment. Further considerations for the use of the telomere dataset can be found in Appendix IV.
- **REL1REL, REL1ID, REL2REL, and REL2ID** – Because of the recruitment strategies of Health ABC, some participants are genetically related. Although the number is low, some investigators will want to eliminate relatives or otherwise control for this non-independence. Since no Health ABC participant is known to be related to more than 2 other participants, there are now 4 variables that summarize how participants are related. **REL1ID** and **REL2ID** report the Health ABC Enrollment IDs of the first and second related individuals, respectively. **REL1REL** and **REL2REL** report their respective relationship (0=full sib, 1=half sib, 2=first cousin). This information was collected in Year 7 from among the surviving (and returning) Health ABC participants. Insofar as was possible, the responses were cross-checked (e.g., if participant X reports participant Y as a full sib, then participant Y is checked to see if they reported participant X). If a participant was deceased, cognitively impaired, or did not have a visit, but were reported as a relative by another participant, this information was included. However, if no one reported being related to a participant who did not respond to this set of questions, then their relationship to other Health ABC participants is unknown. Cases where participants disagreed with each other about their relationships were referred to the clinics to investigate.

Where available, links to methodology documentation have been included in Appendix I in the “Allele” column. The genotype designations provided by the investigators whose ancillary study provided the data have referred to various nomenclature systems. Where possible, the NCBI database identifier (rs number) is included in the table and in the variable label. In order to make these data as useful as possible to the widest group of analysts, the Coordinating Center has made an effort to provide meaningful variable names and labels, within the constraints of SAS v8.

Where information available by looking up the rs number in the NCBI database seemed to conflict with information provided by the investigator, an effort was made to reconcile the information with the investigator. Since methods of numbering position within a gene vary, position was only used if that was the only way to differentiate two SNPs from the same gene. In early cases, when the polymorphism results in a change in amino acid sequence, an effort was made to include that change in the SNP designation. The more recent increase in genetics data being generated precludes continuing that level of detail. Questions regarding the identity of the genotype should be directed to the investigator.

To assist in this process, the investigator is listed in Appendix I, with a link to their email address.

In addition, most investigators included duplicate samples (provided by the Coordinating Center in a blinded fashion), and the resulting duplicate results have been analyzed. Appendix I includes a weighted Kappa result for each genotype for which blind duplicates were included. The weighted Kappa statistic takes into account degree of difference between unlike pairs, so that pairs having both alleles read differently are given more weight than those where only one allele differed. In some cases, it is not possible to calculate a Kappa because the 3x3 table is incomplete. In many of these cases, there is perfect agreement, and the Kappa is reported as 1. In others, the Kappa is listed simply as undefined. In these cases, however, the degree of disparity was scrutinized and, if necessary, discussed with the investigator, before accepting the data for release.

In a few cases, two investigators genotyped the same SNP. Where one investigator was missing a value and the other had a value, the non-missing value was used. When the genotypes found for the same individual by the two investigators did not agree, the value from the investigator with the better weighted Kappa was used. When this occurred, the blind duplicate analysis was run both with and without these additional, unintentional duplicates (before substitution), and both weighted Kappa values are included in Appendix I.

Similar conventions are used in Appendices II and III for the European ancestry datasets, although these data are all from one Ancillary study, so methodology links and investigator information are included as single entries above the table.

Formats are continually being added to the Health ABC formats library. It is important that investigators using these data update their format library in order to be able to decode the genotype designations, which would otherwise be meaningless numbers (0,1,2). The heterozygous genotype was always designated 1, but no effort was made to make the homozygous ancestral genotype consistently either 0 or 2. This information was not readily available in most cases. Rather, to minimize the number of formats, the homozygous genotype with the lowest letter in the alphabet is designated 0, and that with the higher letter is designated 2 (e.g. A/A is always 1; T/T is always 2; C/C and G/G may be either, depending on the other possible allele). One SNP has now been genotyped that has 3 possible alleles. Only 4 combinations were found among Health ABC participants, so these have been numbered 1 through 4 (the least common allele as never found as a homozygote). A new format has been added to the Format library for that SNP. In addition, when a label indicates an amino acid change from x to y, this is not meant to imply that x is the wild type and y the mutation. Although every effort was made to ensure that the variables are correctly formatted, investigators should make sure the frequencies of alleles match their expectations from what is published in the literature.

When available, the rs# from the NCBI database is included in the label. Starting with version 1.2 new variable names will be standardized to include the gene symbol and rs# if available. Old variable names may be replaced later. If the SNP is known to be

associated with a particular gene (generally true in the DNA dataset, not true in either of the European ancestry datasets), the standard gene abbreviation is also included in the variable name. In some cases, genes overlap or a SNP is found just outside the gene but is still useful for haplotyping the gene. These may still be given a name that includes the gene symbol for the gene the investigator was interested in. Where possible, Appendix I and the Genotype Tracking Log list these as XXXX region.

c. Missing Value Conventions: SAS allows for stratification of missing values. The following missing values have been assigned:

. : Missing

Used when a participant was not a part of the sub-group genotyped

T:Missing Due to Technical Problems

Used when a value is missing from the dataset due to technical difficulties.

A:Not Expected

Used when a value is not expected. For example, participants with no relatives in Health ABC, REL1REL, REL2REL, REL1ID and REL2ID are all .A.

General Strategies for Using Special Missing Values

In SAS, when using special missing values in logical expressions, the missing value is no longer only equal to ‘.’

To express a value equal to missing, the code should be written: `<= .Z` or alternately: `le .Z`

To express a value not equal to missing, the code should be written `>.Z` or alternately: `gt .Z`

.Z is the greatest value of missing available in SAS.

5. Dataset index formulation and key variable mapping

The DNA file is sorted by HABCID, which is a unique identifier for each participant.

6. General strategies for manipulating and merging the data

Because the Health ABC datasets are sorted by Health ABC Enrollment ID, the HABCID variable is most useful for merging with other datasets.

7. Recent updates to dataset

Data from 1 ancillary study have been added to the DNA dataset as of release version 2012-11-1, and includes 14 SNPs from Dr. Wen-Chi Hsueh’s ancillary study #AS03-40AM in 2,977 HABC participants.

Data from 5 ancillary studies have been added to the DNA dataset as of release version 2012-4-1. This includes:

1. 1,295 SNPs from Dr. Brock Beamer's ancillary study #AS07-103 in 2,800 HABC participants
2. 7 SNPs from Dr. Robert Ferrell's ancillary study #AS01-21 single SNP methods in 2,678 HABC participants
3. 1,407 SNPs from Dr. Robert Ferrell's ancillary study #AS01-21 illumina SNP panel in 2,835 HABC participants
4. 65 SNPs from Dr. Wen-Chi Hsueh's ancillary study #AS03-40AM in 2,982 HABC participants
5. 2 telomere length measurements and the mean length from Dr. Abraham Aviv's ancillary study #AS10-130 in 762 HABC participants

Note: For ancillary study #AS01-21 single SNP methodology, genotypes of SNPs were sent to the HABC Coordinating Center in batches. In some cases the new batch to be released in this version had genotype data that was previously released. In most cases the SNPs were concordant with prior genotypes released with the exception of two SNPs **TNFSF6_1800682** and **TNFSF6_763110** which had a very high rate of discordance with previously released SNPs. We have removed these SNPs from the dataset and will not release them in future datasets.

LISTINGS

DNAListings

This is a zipped file containing a Contents listing, and Frequencies and Univariates for the DNA dataset.

EAListings

This is a zipped file containing a Contents listing, Formats listing, and Frequencies and Univariates for the EurAncestry dataset.

EA2Listings

This is a zipped file containing a Contents listing, Formats listing, and Frequencies and Univariates for the EurAncestry2 dataset.

Contents listing

This is an electronic copy of the SAS proc contents printout for the dataset. This list can be used to view the contents of the dataset and to search for variable names, descriptions, and formats. A PDF version of the proc contents listings can be found in Appendix III.

Formats listing

This is an electronic copy of the SAS printout of all formats contained in the SAS Formats Library (formats.sas7bcat) and is useful for viewing descriptions assigned to variable values (e.g.: 1=White, 2=Black).

Frequencies and univariates

This is an electronic copy of the SAS printout of frequencies and univariates for all relevant numeric variables.