

Genetics Analysis Workshop 16

Problem 2: Framingham Heart Study Data Set

Description of the Framingham Heart Study

In GAW16, we revisit data drawn from the Framingham Heart Study. The Framingham Heart Study -- under the direction of National Heart, Lung, and Blood Institute; NHLBI – began in 1948 with the recruitment of adults from the town of Framingham, Massachusetts. At the time, little was known about the general causes of heart disease and stroke, but the death rates for cardiovascular disease (CVD) had been increasing steadily since the beginning of the 20th century and had become an American epidemic. The Framingham Heart Study is now conducted in collaboration with Boston University.

The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.

Between 1948 and 1953 the researchers recruited 5,209 subjects (2,336 men and 2,873 women) between the ages of 29 and 62 from the town of Framingham, Massachusetts and began the first round of extensive physical examinations and lifestyle interviews that they would later analyze for common patterns related to CVD development. Subjects were recruited from lists of addresses recorded for the town. Two out of every three households were approached for participation in the study. While there was no intention to recruit families for family studies, the plan was to recruit all household members in the ages 30-60 within each house that was selected for study. Hence, many biologically related individuals were recruited, including 1644 spouse pairs. Since 1948, these participants have returned to the study every two years for a detailed medical history, physical examination, and laboratory tests. Now in 2008 at 60 years of follow up, there remain about 500 participants from this cohort.

Between 1971 and 1975 the study enrolled a second-generation group -- 5,124 of the original participants' children and the spouses of these children -- to participate in similar examinations. 2,616 subjects are offspring of the original spouse pairs and 34 are stepchildren. A total of 898 offspring are children of cohort members where only one parent was a study participant and 1,576 are spouses of the offspring. The Offspring Cohort has been followed every four years through 2001 (except between Exams 1 and 2 with an intervening 8 years) using protocols similar to those used for study of the Original Cohort.

Between 2002 and 2005 the study enrolled the third generation (Gen3) of the Framingham Heart Study – 4095 offspring of the second generation. None of their spouses were recruited. An additional 103 parents of this third generation, who were not recruited between 1971 and 1975, were also recruited at this time. The latter group is not included in the GAW16 data. With the recruitment of this third generation, the study has increasingly focused on genetic factors associated with the development of cardiovascular disease and its associated risk factors. To date, there is only one examination of this generation of participants. A description of the recruitment of this third generation and comparison with the earlier generations at their initial recruitment is presented in Splansky et al. [Splansky GL et al., 2007].

Further information on the Study can be found at <http://www.nhlbi.nih.gov/about/framingham/index.html>.

Genome-wide Dense SNP Scan in Framingham Heart Study

Genetic studies did not begin in the FHS until the 1990s. In the late 1980s and through the 1990s DNA was extracted from blood samples of surviving FHS participants. In 2007, the FHS entered a new phase with the conduct of genotyping for the FHS SHARe (SNP Health Association Resource) project, for which dense SNP genotyping was performed using approximately 550,000 SNPs (GeneChip® Human Mapping 500K Array Set and the 50K Human Gene Focused Panel) in 10,775 samples (some duplicates) from the three generations of subjects (including over 900 pedigrees). Affymetrix conducted all genotyping for the FHS SHARe project, using the 250K Sty, 250K Nsp, and the supplemental 50K platforms. Eighty-nine percent of the DNA samples were collected during the 1990s. To maximize the power of the study, we also extracted DNA from 1133 blood samples, drawn from subjects who had no DNA, to include in the SHARe project. These samples had been sitting in our refrigerators for some time, a few as far back as the 1970s. We refer to these DNA samples as the legacy samples. These samples had a higher failure rate in the genotyping process (40%) than the other eighty-nine percent (3%). Affymetrix invoked its own criteria for a sample to succeed in genotyping. All non-legacy samples must succeed on all three platforms, while legacy samples needed to pass on at least one platform. When a sample failed, additional attempts were made. Samples that repeatedly failed 2-4 times were called failures. Other samples failed due to issues of genotyped sex identification not matching our records or low SNP concordance among SNPs common across arrays or contamination. Eighty-nine percent of the legacy samples for which genotyping results are available passed all three platforms. The genotyping data from the 10,043 samples from 9354 subjects that passed the Affymetrix criteria were additionally checked for gender consistency and consistency with family structure, resulting in genotyping data for 9,274 participants in FHS SHARe. Genotype calls were made with the BRLMM algorithm.

The SHARe database is housed at the National Center for Biotechnology Information database of genotypes and phenotypes (NCBI dbGaP) (<http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?id=phs000007>) and contains all ~550,000 SNPs. This genome-wide dense SNP scan and a subset of phenotypes from the Framingham Heart Study are the focus of the Genetic Analysis Workshop 16.

Genetic Analysis Workshop 16 FHS Data

The Framingham datasets for Genetic Analysis Workshop 16 include:

1. Files containing ~550,000 SNP genotype data
2. Cel files for each of the genotyped SNPs
3. A pedigree file that provides the family structure
4. Three files with phenotypic data, one for each cohort (Original, Offspring, Gen3)

All data for GAW16 can be obtained at the NCBI dbGaP website. Information on the procedures to obtain these data is provided on the GAW website, <http://www.qaworkshop.org/qaw2008.htm>.

The phenotypic data are provided for those subjects who have consented to anyone's use, including those at for-profit and not-for-profit institutions. The pedigree file contains all biologically related subjects in the FHS and is not limited to those with full consent. There are a total of 7130 subjects with phenotype data: 373 Original Cohort, 2760 Offspring and 3997 Gen3 participants. Of these 7130 participants, 6879 are members of pedigrees, 251 are not part of the pedigrees. Of the 6879 subjects who are members of pedigrees, 6525 are genotyped; of

251 subjects who are not included in the pedigree file 227 are genotyped. Thus, there are a total of 6752 subjects who are genotyped. There are 765 pedigrees with 2 to 301 genotyped subjects: 134 pedigrees with 2, 123 with 3, 98 with 4, 85 with 5, 177 with 6 to 10, 72 with 11 to 15, 30 with 16 to 20 and 46 with more than 20.

Data from a subset of examinations were selected for Genetic Analysis Workshop 16: Exams 1, 4, 7 and 11 for the Original Cohort, Exams 1, 3, 5, 7 for the Offspring Cohort and Exam 1 for the Gen3 Cohort. These exams were chosen so that data from FHS participants of approximately the same age from the three cohorts were considered. Only one exam has been completed for Gen3 and thus there are data for only one exam for these participants. Age and sex descriptive statistics for these subjects are provided in the following table:

Variable	Original Cohort	Offspring Cohort	Gen3 Cohort
Sample Size	373	2760	3997
Ages (Mean and Std Dev)			
Exam 1	34.9 ± 3.9	33.7 ± 9.3	40.2 ± 8.8
Exam 4 (Original), Exam 3 (Offspring)	40.9 ± 3.9	46.3 ± 9.3	NA
Exam 7 (Original), Exam 5 (Offspring)	47.0 ± 3.9	53.3 ± 9.2	NA
Exam 11 (Original), Exam 7 (Offspring)	54.8 ± 3.8	60.2 ± 9.1	NA
% Female Exam 1	69.2%	54.4%	53.3%

Note that the Original Cohort participants with data are drawn from the select few who survived ~40 years to have DNA collected and to provide consent for the SHARe project.

All files can be linked by a variable called shareid. This is a unique identifier given to each subject in the FHS SHARe dataset.

Genotype Data

Genotype datasets have the shareid and ~550,000 genotypes for each participant. There are two sets of genotypes available for GAW16: one that is consistent with family structure and gender with no additional cleaning and a second that has been cleaned for Mendelian errors. The first set was cleaned for familial relationships. In this step we evaluated whether the genotypes of subjects were consistent with their reported familial relationships. We used PREST and sib-kin from Aspx to perform this analysis within families. Additionally, we checked for unknown (cryptic) relationships between families using PLINK. In some cases familial relationships were altered as a result. Such errors could occur from unknown familial relationships or sample mix up. Cleaning at this stage can result in all the genotyping of some individuals being deleted. In the second set of genotypes additional cleaning was performed to check for random genotyping errors by using Mendelian checks. In this case only the genotypes of the specific SNP were deleted from the members of the nuclear family in which the Mendelian error occurred. Both genotype datasets include the legacy DNA samples, which were of poorer quality. Including these poorer quality samples in checking Mendelian inheritance may have resulted in loss of genotypes that are correct because genotypes of all individuals in nuclear families were set to missing in the presence of Mendelian inheritance errors.

Pedigree (Family) File

The family structure file, defining the pedigree structures is provided. This file has the following variables:

- | | |
|------------|--|
| 1. pedno | Family (Pedigree) id |
| 2. shareid | ID for SHARe project |
| 3. idtype | 0=Original Cohort,
1=Offspring Cohort
2=New Offspring Spouse
3=Gen3 Cohort
. = not a FHS participant |
| 4. fshare | Father's SHARe id (= . for founder) |
| 5. mshare | Mother's SHARe id (= . for founder) |
| 6. sex | 1=male
2=female |
| 7. geno | 1=genotyped for SHARe
0=not genotyped for SHARe
. = not a FHS participant |
| 8. pheno | 1=a FHS participant with phenotype data
. = not a FHS participant |
| 9. itwin | identifies identical twins in a sibship (same number for those in a twin set)
. = not a twin |

There are 8732 subjects in this file who have been genotyped. However, only data for those participants who consented to general use (both for-profit and not-for-profit) are available to GAW16. Participants with phenotype data and who are not in the family file are not members of families and are biologically unrelated to one another.

There are a few additional variables in this dataset located at NCBI dbGaP, which will not be needed for GAW16.

Phenotype Files

Three phenotype files are provided: 1) Original Cohort participants, 2) Offspring participants, 3) Gen3 participants. These files provide information on demographics such as sex and age, the traditional risk factors for coronary heart disease, such as blood pressure, diabetes, smoking and lipid levels, and also data on incident coronary heart disease. Data from multiple exams are provided for the Original and Offspring Cohort subjects, but only one exam is provided for Gen3, since these subjects have undergone only one exam (2002-2005). Follow up for events in these subjects is through 2006. A full description of the phenotypes is provided in **Appendix 1**.

Appendix 1

This appendix details the phenotypic variables that have been provided to Genetic Analysis Workshop 16. First, we provide dates for all examinations so that users of the data are able to line up examinations at approximately the same chronological time, if so desired. Bolded dates indicate the exam data that are available in GAW16.

The start and stop dates of each examination are

09/29/48 - 04/25/53	Original Cohort Exam 1
08/01/50 - 05/02/55	Original Cohort Exam 2
01/18/52 - 11/27/56	Original Cohort Exam 3
01/15/54 - 07/01/58	Original Cohort Exam 4
05/02/56 - 11/26/60	Original Cohort Exam 5
06/10/58 - 02/18/63	Original Cohort Exam 6
03/22/60 - 10/31/64	Original Cohort Exam 7
05/01/62 - 12/14/66	Original Cohort Exam 8
04/23/64 - 10/29/68	Original Cohort Exam 9
05/21/66 - 08/17/70	Original Cohort Exam 10
06/13/68 - 09/02/71	Original Cohort Exam 11
06/09/71 - 05/29/74	Original Cohort Exam 12
01/28/72 - 03/20/76	Original Cohort Exam 13
05/28/75 - 04/24/78	Original Cohort Exam 14
03/22/77 - 11/13/79	Original Cohort Exam 15
05/14/79 - 05/10/82	Original Cohort Exam 16
05/19/81 - 05/21/84	Original Cohort Exam 17
04/12/83 - 11/21/85	Original Cohort Exam 18
04/29/85 - 06/30/88	Original Cohort Exam 19
11/24/86 - 06/04/90	Original Cohort Exam 20
10/18/88 - 05/20/92	Original Cohort Exam 21
12/06/90 - 04/25/94	Original Cohort Exam 22
11/10/92 - 03/19/96	Original Cohort Exam 23
03/10/95 - 01/27/98	Original Cohort Exam 24
06/06/97 - 12/13/99	Original Cohort Exam 25
05/27/99 - 11/27/01	Original Cohort Exam 26
01/17/02 - 11/19/03	Original Cohort Exam 27
02/10/04 - 10/26/05	Original Cohort Exam 28
08/30/71 - 09/03/75	Offspring Exam 1
10/09/79 - 10/27/83	Offspring Exam 2
12/20/83 - 09/30/87	Offspring Exam 3
04/22/87 - 09/11/91	Offspring Exam 4
01/23/91 - 06/29/95	Offspring Exam 5
01/26/95 - 09/02/98	Offspring Exam 6
09/11/98 - 10/26/01	Offspring Exam 7
04/01/02 - 07/08/05	Gen3 Exam 1

Variables available for Genetic Analysis Workshop 16 are indicated in the following table.
Below the table is a description of how variables were measured.

Variable	Original Cohort	Offspring Cohort	Gen3 Cohort
SHARe ID	X	X	X
IDtype (0=Original, 1=Offspring, 3=Gen3)	X	X	X
Sex	X	X	X
Attended Exam 1 (1=Yes, 0=No; yes by definition for Gen3)	X	X	NA
Attended Exam 4 (Original), Exam 3 (Offspring) (1=Yes, 0=No)	X	X	NA
Attended Exam 7 (Original), Exam 5 (Offspring) (1=Yes, 0=No)	X	X	NA
Attended Exam 11 (Original), Exam 7 (Offspring) (1=Yes, 0=No)	X	X	NA
Age in Years Exam 1	X	X	X
Age in Years Exam 4 (Original), Exam 3 (Offspring)	X	X	NA
Age in Years Exam 7 (Original), Exam 5 (Offspring)	X	X	NA
Age in Years Exam 11 (Original), Exam 7 (Offspring)	X	X	NA
Systolic/Diastolic Blood Pressure mm Hg Exam 1	X	X	X
Systolic/Diastolic Blood Pressure mm Hg Exams 4,3	X	X	NA
Systolic/Diastolic Blood Pressure mm Hg Exam 7,5	X	X	NA
Systolic/Diastolic Blood Pressure mm Hg Exam 11,7	X	X	NA
Hypertensive Treatment (1=yes,0=no) Exam 1	X	X	X
Hypertensive Treatment (1=yes,0=no) Exams 4,3	X	X	NA
Hypertensive Treatment (1=yes,0=no) Exam 7,5	X	X	NA
Hypertensive Treatment (1=yes,0=no) Exam 11,7	X	X	NA
Fasting* Total Cholesterol mg/dl Exam 1	X	X	X
Fasting* Total Cholesterol mg/dl Exams 4,3	X	X	NA
Fasting* Total Cholesterol mg/dl Exam 7,5	X	X	NA
Fasting Total Cholesterol mg/dl Exam 11,7	X	X	NA
Fasting HDL Cholesterol mg/dl Exam 1	NA	X	X
Fasting HDL Cholesterol mg/dl Exams 4,3	NA	X	NA
Fasting HDL Cholesterol mg/dl Exam 7,5	NA	X	NA
Fasting HDL Cholesterol mg/dl Exam 11,7	X	X	NA
Fasting Triglycerides mg/dl Exam 1	NA	X	X
Fasting Triglycerides mg/dl Exams 4,3	NA	X	NA
Fasting* Triglycerides mg/dl Exam 7,5	X	X	NA
Fasting Triglycerides mg/dl Exam 11,7	X	X	NA
Cholesterol Treatment (1=yes, 0=no) Exam 1	NA	X	X
Cholesterol Treatment (1=yes, 0=no) Exams 4,3	NA	X	NA
Cholesterol Treatment (1=yes, 0=no) Exam 7,5	X	X	NA
Cholesterol Treatment (1=yes, 0=no) Exam 11,7	X	X	NA
Fasting* Blood Glucose mg/dl Exam 1	X	X	X
Fasting* Blood Glucose mg/dl Exam 4,3	X	X	NA
Fasting* Blood Glucose mg/dl Exam 7,5	NA	X	NA
Fasting* Blood Glucose mg/dl Exam 11,7	NA	X	NA
Height In Inches Exam 1	X	X	X
Height In Inches Exams 4,3 (Exam 5 for Original Cohort)	X	X	NA
Height In Inches Exam 7,5 (Exam 10 for Original Cohort)	X	X	NA
Height In Inches Exam 11,7	X	X	NA
Weight In Pounds Exam 1	X	X	X

Weight In Pounds Exams 4,3	X	X	NA
Weight In Pounds Exam 7,5	X	X	NA
Weight In Pounds Exam 11,7	X	X	NA
Smoking Status (0=never,1=former,2=current) Exam 1	X	X	X
Smoking Status (0=never,1=former,2=current) Exams 4,3	X	X	NA
Smoking Status (0=never,1=former,2=current) Exam 7,5	X	X	NA
Smoking Status (0=never,1=former,2=current) Exam 11,7	X	X	NA
Number of Cigarettes Smoked per Day Exam 1	X	X	X
Number of Cigarettes Smoked per Day Exams 4,3	X	X	NA
Number of Cigarettes Smoked per Day Exam 7,5	X	X	NA
Number of Cigarettes Smoked per Day Exam 11,7	X	X	NA
Number of Equivalent Alcohol ounces per Week Exam 1**	X	X	X
Number of Equivalent Alcohol ounces per Week Exams 4,3	NA	X	NA
Number of Equivalent Alcohol ounces per Week Exam 7,5	X	X	NA
Number of Equivalent Alcohol ounces per Week Exam 11,7**	X	X	NA
Hard CHD (0=Never, 1= incident during study, 2=prevalent at exam 1)	X	X	X
Age at Onset of Hard CHD	X	X	X
Diabetes (0=No, 1= Yes)	X	X	X
Age at First Exam at which Diabetes was noted	X	X	X
Age at Death (No one dead in Gen3)	X	X	NA
Age of Last Follow up (Death or Last Contact) (Age at Exam 1 for Gen3)	X	X	NA

* Non-fasting values for the Original Cohort unless otherwise stated below.

**Exams 2 and 12 for the Original Cohort

1. **Systolic/Diastolic Blood Pressure:** Blood pressure was measured by a physician with the subject in the sitting position for at least five minutes. Systolic and diastolic readings in the left arm of the subject were taken with a mercurial sphygmomanometer and a cuff long enough to fit the most obese arm. Although no specific protocol was followed for accuracy, most measurements are given to the nearest even number.
2. **Hypertensive Treatment:** Subjects who reported using drugs for treating their hypertension were classified as on hypertensive treatment. This variable is coded as 0=No, 1=Yes. No hypertensive treatment was in use for the first examination of the Original Cohort and thus, all values for this exam are assigned a value of 0.
3. **Lipids (total plasma cholesterol, high density lipoprotein cholesterol, and triglycerides):** The Original Cohort was not fasting at most exams. One of the few exceptions is exam 11. Offspring and Gen3 Cohorts come to examinations fasting. For the Offspring and Gen3 Cohorts, twelve-hour fasting venous blood samples were collected in tubes containing 0.1% EDTA. Plasma aliquots were obtained by centrifugation at 2500 rpm for 20 minutes at 4 degrees C. Total cholesterol, HDL-C and triglycerides were measured by automated enzymatic methods. Since the 1970s, lipid analyses were performed at the Framingham Heart Study Laboratory, which participates in the Standardization Program of the Center for Disease Control, utilizing the Lipid Research Clinics Program protocols, when these became available. Earlier in the Original Cohort, the Abell-Kendall method was used. In the 1970's HDL-C was measured for the first time in the Offspring Cohort at its Exam 1 and in the Original Cohort over Exams 10-12. HDL-C and triglycerides have been measured at all Offspring and Gen3 exams. Only one HDL-C measurement was obtained at exams 10-12 for the Original Cohort. Fasting triglyceride was measured at these same

exams, while a non-fasting measure of triglycerides was measured at exam 7 of the Original Cohort. The Original Cohort measurements at exams 10-12 are labeled as Exam 11 for GAW16. LDL cholesterol is calculated in the Framingham Heart Study by the Friedewald method. LDL cholesterol is not calculated for those with triglycerides above 400 mg/dl, as the calculation is not accurate for these subjects.

4. **Cholesterol Treatment:** Subjects who reported using drugs for treating their lipid levels were classified as being on cholesterol treatment. This variable is coded as 0=No, 1=Yes.
5. **Glucose:** Blood glucose measures in the Original Cohort are calculated from non-fasting subjects. These measures are only available at exams 1 and 4 in the Original Cohort. For Offspring and Gen3 Cohorts, plasma glucose was measured with a hexokinase reagent kit (A-gent glucose test, Abbott, South Pasadena, CA) in fasting subjects. Glucose assays were run in duplicate and the intra-assay coefficient of variation ranged from 2% to 3%, depending on the assayed glucose level.
6. **Height and weight:** Height in inches and weight in pounds were measured with the subject in light clothing and shoes off.
7. **Smoking Status:** Obtained by self report, coded as 0=never, 1=former, 2=current
8. **Number of Cigarettes Smoked per day:** Obtained by self-report
9. **Number of ounces of equivalent alcohol per week:** Subjects were asked about the number of beers, glasses of wine and cocktails usually drunk within a specified time period such as a week or a month. These quantities were then converted to alcohol equivalent ounces per week.
10. **Hard CHD:** A hard CHD (coronary heart disease) event is defined as any of the following: recognized myocardial infarction diagnosed through an EKG or enzymes, coronary insufficiency or death attributed to CHD. When a participant is reported to have had a CHD event, all information available (such as medical records from the hospital or doctor and death records) is sought and generally obtained. This information is reviewed by a panel of three experts, who determine whether the event was a coronary event or not. For each event, the precise date of the event is obtained. No dates are provided in this public dataset to protect the privacy of the participants.
11. **Age at Onset of Hard CHD:** The age at onset of the hard CHD event, determined from the date of the event.
12. **Diabetes:** Since the Original Cohort was non-fasting at most exams, the definition of diabetes in the Original Cohort differs from that for the Offspring and Gen3 Cohorts. In the Original Cohort, diabetes is defined as having a casual blood sugar of 200 or greater or on treatment for diabetes. In the Offspring and Gen3 Cohorts, diabetes is defined as having a fasting blood sugar greater than 125 or on treatment. This variable is coded as 0=not diabetic and 1=diabetic. The exam of first occurrence of diabetes is used to calculate the age at onset of diabetes.
13. **Age at First Exam at which Diabetes was noted:** We do not have exact dates of onset of diabetes. Thus, onset of diabetes is interval censored. For the GAW16 dataset, the age of

the first exam at which diabetes was noted is provided for those who have been diagnosed with diabetes.

14. **Age at Death:** age at which death occurred, calculated from the date of death.

15. **Age at Last Follow Up:** age of death for those who have died; age at last contact for those still alive in 2006.

References

Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino RB Sr, Fox CS, Larson MG, Murabito JM, O'Donnell CJ, Vasan RS, Wolf PA, Levy D. 2007. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* 165(11):1328-1335.