2009

# The Genetic Analysis Workshop 16 Problem 3: Simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study

Aldi T. Kraja
*Washington University School of Medicine in St. Louis*

Robert Culverhouse
*Washington University School of Medicine in St. Louis*

E Warwick Daw
*Washington University School of Medicine in St. Louis*

Jun Wu
*Washington University School of Medicine in St. Louis*

Andrew Van Brunt
*Washington University School of Medicine in St. Louis*

BMC Proceedings

Proceedings



# The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study

Aldi T Kraja*[1], Robert Culverhouse*[2], E Warwick Daw[1], Jun Wu[1], Andrew Van Brunt[1], Michael A Province[1] and Ingrid B Borecki[1]

Addresses: [1]Division of Statistical Genomics, Washington University School of Medicine, 4444 Forest Park Boulevard, Campus Box 8506, St. Louis, Missouri 63108, USA and [2]Division of General Medical Sciences, Washington University School of Medicine, 660 South Euclid Avenue, Box 8005, St. Louis, Missouri 63110, USA

E-mail: Aldi T Kraja* - aldi@wustl.edu; Robert Culverhouse* - rculverh@dom.wustl.edu; E Warwick Daw - warwick@dsgmail.wustl.edu; Jun Wu - jun@dsgmail.wustl.edu; Andrew Van Brunt - andrew@dsgmail.wustl.edu; Michael A Province - mprovince@wustl.edu; Ingrid B Borecki - ingrid@dsgmail.wustl.edu
*Corresponding author

*from* Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

*BMC Proceedings* 2009, **3**(Suppl 7):S4 doi: 10.1186/1753-6561-3-S7-S4

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S4



## Abstract

The Genetic Analysis Workshop (GAW) 16 Problem 3 comprises simulated phenotypes emulating the lipid domain and its contribution to cardiovascular disease risk. For each replication there were 6,476 subjects in families from the Framingham Heart Study (FHS), with their actual genotypes for Affymetrix 550 k single-nucleotide polymorphisms (SNPs) and simulated phenotypes. Phenotypes are simulated at three visits, 10 years apart. There are up to 6 "major" genes influencing variation in high- and low-density lipoprotein cholesterol (HDL, LDL), and triglycerides (TG), and 1,000 "polygenes" simulated for each trait. Some polygenes have pleiotropic effects. The locus-specific heritabilities of the major genes range from 0.1 to 1.0%, under additive, dominant, or overdominant modes of inheritance. The locus-specific effects of the polygenes ranged from 0.002 to 0.15%, with effect sizes selected from negative exponential distributions. All polygenes act independently and have additive effects. Individuals in the LDL upper tail were designated medicated. Subjects medicated increased across visits at 2%, 5%, and 15%. Coronary artery calcification (CAC) was simulated using age, lipid levels, and CAC-specific polymorphisms. The risk of myocardial infarction before each visit was determined by CAC and its interactions with smoking and two genetic loci. Smoking was simulated to be commensurate with rates reported by the Centers for Disease Control. Two hundred replications were simulated.

## Background

The Framingham Heart Study (FHS) is a rich platform for the study of cardiovascular disease and the application of novel, imaginative analytic strategies. For Genetic Analysis Workshop (GAW) 16, we use a semi-simulated approach using actual genotypes from the 500 k Affymetrix platform and the 50 k candidate gene chip and building phenotypes on the observed genetic variation. Because blood lipid levels are a major risk factor in the development of cardiovascular disease [1], we modeled disease risk on the lipid pathway, including both genetic and environmental determinants. The FHS has reported that long-term averages of low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglyceride (TG) levels were highly heritable (0.66, 0.69, and 0.58, respectively) [2]. Several familial studies also have reported heritabilities for LDL of 0.50, HDL of 0.54, and TG of 0.39 [3]. Dyslipidemia, as a fundamental component of the atherosclerotic process, is a medically correctable risk factor with established efficacious treatments for reducing risk of coronary heart disease [4]. Thus, we included in our simulation the use and effects of dyslipidemic medications, which have an important role in shaping lipid profiles. This simulation builds in the long tradition of previous simulations for Genetic Analysis Workshops [5,6].

## Methods

The FHS pedigrees, distributed as GAW16 Problem 2, formed the basis of our simulation [7]. In total, there were 6,476 subjects who had genotypes and simulated phenotypes. After the simulations began, additional FHS subjects provided broad consent for data sharing; these additional subjects were not included in the simulations. To ensure comparable data to that which was simulated, we provided a file that defined precisely which subjects were included and their relationships within families. The ~550 k measured single-nucleotide polymorphism (SNP) genotypes, distributed for GAW16 Problem 2 from both the genome-wide scan and the additional candidate gene platform (GeneChip® Human Mapping 500 k Array Set (Nsp and Sty), and the 50 k Human Gene Focused Panel) comprised the genotypes for GAW16 Problem 3. Novel fictitious phenotypes were simulated for subjects.

Although family members of the FHS attended various exams at different times, depending on the generation, we modeled our study as if all subjects were recruited at one time, calculated the family member's relative ages at one particular exam, and then assigned a simulated age for everyone at three time points, with 10-year intervals. The mean age in years (range) for the simulation, by generation and visit, is shown in Table 1.

**Table 1: Mean ages of the simulated data (mean, minimum, and maximum age in years)**

| | Mean age (minimum, maximum) | | |
|---|---|---|---|
| **Generation** | **Visit 1** | **Visit 2** | **Visit 3** |
| 1 | 66 (54, 80) | 76 (64, 90) | 86 (74, 100) |
| 2 | 56 (20, 80) | 66 (30, 90) | 76 (40, 100) |
| 3 | 33 (19, 70) | 43 (29,80) | 53 (39,90) |
| Overall | 43 (19, 80) | 53 (29, 90) | 63 (39, 100) |

The simulation model is depicted in Figure 1. There are up to six "major" genes for the lipid phenotypes HDL, LDL, and TG, and 1,000 polygenes for each trait. Several polygenes have pleiotropic effects (i.e., several of these polygenes affect two or three or trait combinations simultaneously). The identity and effects of the major genes are documented in Table 2. The locus-specific heritabilities of the major genes range from 0.1-1.0% under additive (AA:AB:BB, 0:0.5:1), dominant (AA:AB:BB, 1:1:0), or overdominant (AA:AB:BB, 0:1:0; heterozygotes show higher effect than the two homozygotes) modes of inheritance, with minor allele frequencies at least 5%, with one exception ($\beta 4$), for which the minor allele frequency was 1%. We simulated an overdominant effect ($\gamma 1$) because there appears to be evidence supporting this possibility and this mode of inheritance is rarely, if ever, modeled. The gene $\alpha 4$ is pleiotropic for HDL and TG and interacts with $\beta 5$ in determining LDL (Figure 1). The interaction accounts for 0.7% of the trait variance, and $\beta 5$ has no marginal effect on any phenotype. The locus-specific effects of the polygenes were on average an order of magnitude smaller, ranging from 0.002-0.15%, with effect sizes extracted from negative exponential distributions. All polygenes act independently and have additive effects. HDL, TG, and LDL share 40% of their polygenes in common, and HDL and TG share an additional 20%. The specific identities of the polygenes, their locations, and their generating effect sizes are provided in the Additional Files 1, 2, 3 corresponding to HDL, LDL, and TG. A group of 39 polygenes influencing HDL were clustered within 0.5 Mb on chromosome 11; otherwise, the polygenes for each trait are randomly distributed throughout the genome. The overall effect of each trait-specific polygenic component was scaled to achieve the target total trait heritabilities of 60%, 55%, and 40% for HDL, LDL, and TG, respectively. The remaining variance is uncorrelated among family members, with the exception of a simulated dietary effect (variable: diet) on TG levels that accounts for a correlation of 0.05 among family members, regardless of their coefficient of relationship. The phenotypes generated from this genetic model were scaled to the empirically derived means and variances

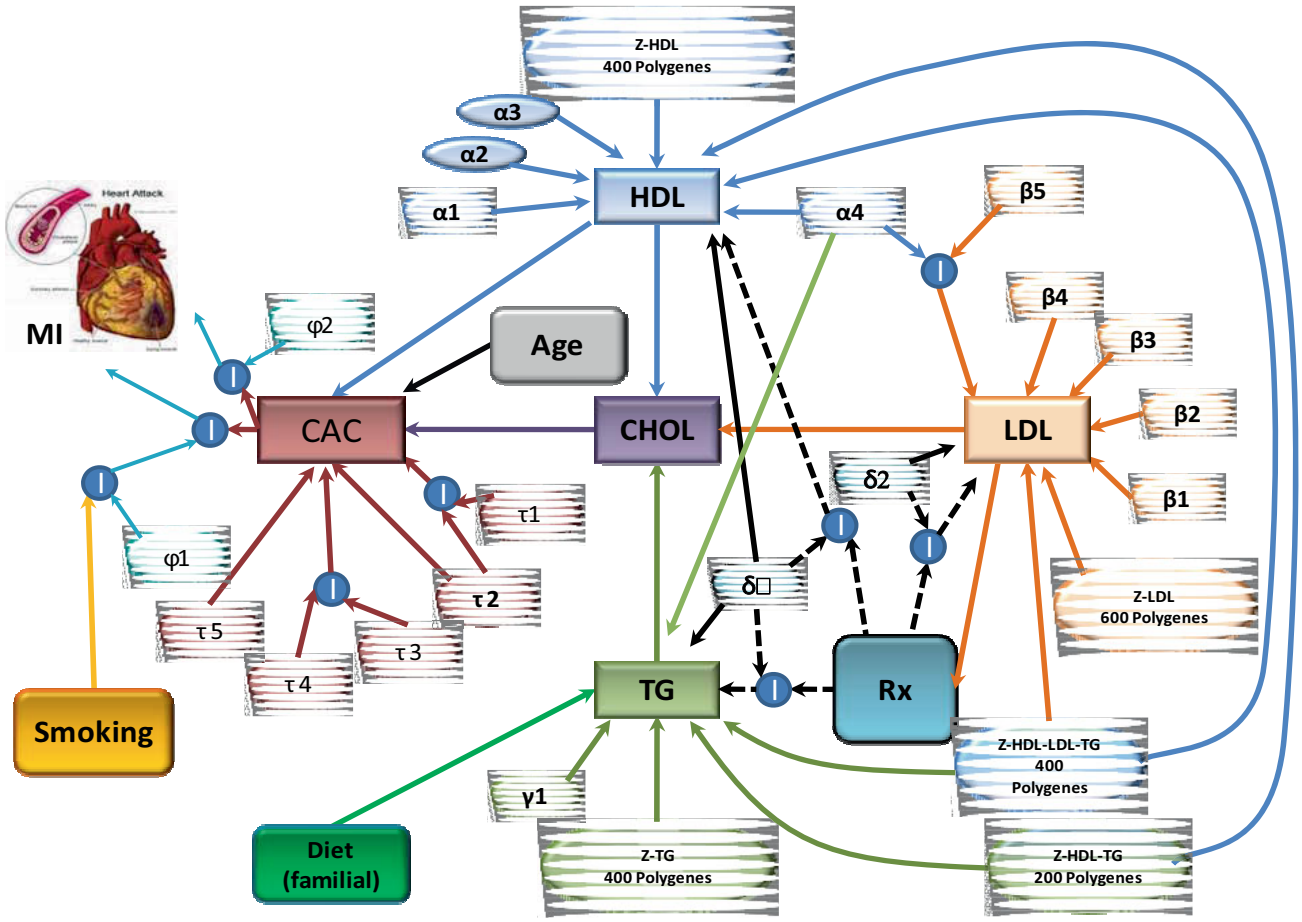


**Figure 1**

**The Genetic Analysis Workshop 16 Problem 3 diagram**. Figure 1 shows simulated phenotypes emulating the lipid domain (HDL, LDL, TG, and CHOL) and its contribution to cardiovascular disease risk (CAC and MI). Simulated major genes are symbolized with Greek letters. There are 1,000 polygenes for each trait HDL, LDL, and TG, several of them with pleiotropic effects. Continued lines and arrows show causality/interaction (I); dashed lines show pharmacogenetic effects only for subjects treated with medication, where response was dependent on the subjects' genotypes. Environmental factors such as diet, smoking, and medication were modeled in the simulation.

for the actual HDL, LDL, and TG traits within 13 age strata (in 5-year intervals) and sex as follows:

$$\begin{aligned}\text{HDL} = {} & \hat{\mu}_{(\text{HDL}|\ \text{age 5 year interval, sex})} + \hat{\sigma}_{(\text{HDL}|\ \text{age 5 year interval, sex})} \\ & \times [h_{\alpha 1} \times a_{\alpha 1} + h_{\alpha 2} \times a_{\alpha 2} + h_{\alpha 3} \times a_{\alpha 3} + h_{\alpha 4} \times a_{\alpha 4} + h_{\partial 1} \times a_{\partial 1} \\ & + \sum_{\kappa=1}^{1,000} (\text{sign} \times \text{apoly}_{\kappa} \times \text{hapoly}_{\kappa}) + \alpha_{\varepsilon} \times h_{\varepsilon}]\end{aligned}$$

$$\begin{aligned}\text{LDL} = {} & \hat{\mu}_{(\text{HDL}|\ \text{age 5 year interval, sex})} + \hat{\sigma}_{(\text{LDL}|\ \text{age 5 year interval, sex})} \\ & \times [h_{\beta 1} \times b_{\beta 1} + h_{\beta 2} \times b_{\beta 2} + h_{\beta 3} \times b_{\beta 3} + h_{\beta 4} \times b_{\beta 4} + h_{\beta 5} \times \text{I}(a_4 \times \beta_5) + h_{\partial 2} \times b_{\partial 2} \\ & + \sum_{\kappa=1}^{1,000} (sign \times bpoly_{\kappa} \times hbpoly_{\kappa}) + b_{\varepsilon} \times h_{\varepsilon}]\end{aligned}$$

$$\begin{aligned}\text{TG} = {} & \hat{\mu}_{(\text{TG}|\ \text{age 5 year interval, sex})} + \hat{\sigma}_{(\text{TG}|\ \text{age 5 year interval, sex})} \times [h_{\gamma 1} \times g_{\gamma 1} + h'_{a_4} \times g_{a_4} + h_{\partial 1} \times g_{\partial 1} \\ & + h_{diet} \times g_{diet} + \sum_{\kappa=1}^{1,000} (sign \times gpoly_{\kappa} \times hgpoly_{\kappa}) + g_{\varepsilon} \times h_{\varepsilon}],\end{aligned}$$

where, for example, $\hat{\mu}_{(\text{HDL}|\text{age 5 year interval, sex})}$ represents the mean of HDL in FHS, given a 5-year age interval and sex; $\hat{\sigma}_{(\text{HDL}|\text{age 5 year interval, sex})}$ is standard deviation of HDL in FHS, given a 5-year age interval and sex; $h_{\alpha 1}$ is the square root of simulated heritability for the $\alpha_1$ SNP (as described in Table 2); $a_{\alpha 1}$ is a simulated effect that reflects in part the penetrance of the $\alpha_1$ SNP; *sign* is a random integer number that takes values (-1) or (+1) with the purpose of randomly changing the contribution direction of polygenes; $apoly_{\kappa}$ represents an instance of each of the 1,000 SNPs effects ($k$ = 1 to 1,000), selected as polygenes for HDL; $hapoly_{\kappa}$ is an instance of the of square roots of heritabilities for 1,000 SNPs selected as polygenes for HDL; $a_{\varepsilon}$ represents the environmental effect that contributes to HDL; and $h_{\varepsilon}$ is

**Table 2: Summary characteristics of the major genes and polygenes for traits HDL, LDL, and TG[a]**

| No. | Trait | Chr | Symbol[b] | Gene | SNP | Role | Total | $h^2$ | Model[c] | MAF | $h^2$ of polygenes |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | HDL | 9q31.1 | $\alpha_1$ | *ABCA1* | rs10820738 | Intron/Exon boundary | | 0.010 | DOM | 6.7 | |
| 2 | HDL | 19q13.2 | $\alpha_2$ | *CYP2B7P1* | rs8103444 | Exon/Intron boundary | | 0.002 | ADD | 24.4 | |
| 3 | HDL | 15q21 | $\alpha_3$ | *LIPC* | rs8035006 | Intron | | 0.003 | ADD | 32.5 | |
| 4 | HDL | 8p22 | $\alpha_4$ | *LPL* | rs3200218 | Exon downstream | | 0.003 | DOM | 21.7 | |
| 5 | HDL | 19q13.2 | $\delta_1$ | *CYP2B6* | rs8192719 | Exon/Intron boundary (Rx) | | 0.003 | ADD (up 10%)[d] | 24.9 | |
| | | | | | | | Total | 0.021 | | | |
| 6 | HDL | 1,000 SNPs | [apoly] | | | | | 0.58 | ADD | ≠ | min = 0.0003; max = 0.0015; avg = 0.00058 |
| | | | | | | | Total | 0.60 | | | |
| 1 | LDL | 4p16.3 | $\beta_1$ | *LRPAP1* | rs7672287 | Intron | | 0.003 | ADD | 22.2 | |
| 2 | LDL | 12q13 | $\beta_2$ | *LRP1* | rs1466535 | Intron | | 0.002 | ADD | 31.6 | |
| 3 | LDL | 11q13.4 | $\beta_3$ | *LRP5* | rs901824 | Intron/Exon boundary | | 0.001 | ADD | 10.3 | |
| 4 | LDL | chrom1 | $\beta_4$ | | rs10910457 | | | 0.005 | ADD | 1.0 | |
| 5 | LDL | 8p22 × 4q24 | $\beta_5$ | *LPL* × *NFKB1* | rs4648068 | Intron | | 0.007 | INT | 31.0 | |
| 6 | LDL | 22q12.2 | $\delta_2$ | *SLC5A4* | rs2294207 | Intron | (Rx) | 0.010 | ADD (down 30%) | 25.7 | |
| | | | | | | | Total | 0.028 | | | |
| 7 | LDL | 1,000 SNPs | [bpoly] | | | | | 0.52 | ADD | ≠ | min = 0.0003; max = 0.00128; avg = 0.00052 |
| | | | | | | | Total | 0.55 | | | |
| 1 | TG | 11q23 | $\gamma_1$ | *APOA5* | rs603446 | Downstream | | 0.003 | OVERD | 43.3 | |
| 2 | TG | 8p22 | $\alpha_4$ | *LPL* | rs3200218 | Exon downstream | | 0.004 | ADD | 21.7 | |
| 3 | TG | 19q13.2 | $\delta_1$ | *CYP2B6* | rs8192719 | Exon/Intron boundary (Rx) | | 0.003 | ADD (down 15%) | 24.9 | |
| 4 | TG | | diet | | | | | 0.01 | Familial | | |
| | | | | | | | Total | 0.020 | | | |
| 5 | TG | 1,000 SNPs | [gpoly] | | | | | 0.38 | ADD | ≠ | min = 0.0002; max = 0.0009; avg = 0.00038 |
| | | | | | | | Total | 0.40 | | | |

[a]Abbreviations: MAF, Minor allele frequency; min, minimum; max, maximum; avg, average; $h^2$, heritability; [apoly], a vector of 1,000 polygenes that contribute to HDL; [bpoly], a vector of 1,000 polygenes that contribute to LDL; [gpoly], a vector of 1,000 polygenes that contribute to TG.
[b]Symbol represents a locus/a vector of polygenes marked by SNPs and corresponding genes, which can be introns/exons, or a diet effect.
[c]Effects of genes were simulated based on ADDitive (0, 0.5, 1), DOMinant (1, 1, 0), OVERDominant (0, 1, 0), and INTeraction genetic models.
[d]Medication treatment lowered LDL on a varying percentage of participants dependent on visits: 2% of subjects in Visit 1, 5% of subjects in Visit 2, and 15% of subjects in Visit 3. Medication treatment increased HDL 10%, lowered LDL 30%, and lowered TG 15% for the treated subjects and in accordance with specific types of genotypes, shown in the "Model" column. For HDL on chromosome 11, 39 SNPs were simulated as a block of polygenes, starting at 110 Mbp and ending at 134 Mbp, with approximately an average equidistance of 0.5 Mbp.

the square root of HDL variance explained by environmental causes.

As individuals progressed to the next visit 10 years later, their phenotypes were scaled by the appropriate age-sex means and variance, but there are no genes governing longitudinal trends *per se*. Instead, we simulated the complicating effects of medication. The simulated value for LDL at each visit for each subject was checked, and individuals in the upper tail of the distribution were simulated as medicated. The proportion of subjects that are medicated increased across visits to comprise 2%, 5%, and 15% of the subjects in Visits 1, 2, and 3, respectively. These proportions were estimated from the FHS data, and reflected the secular increase in the proportion of individuals being treated for elevated cholesterol levels. The response to treatment is governed by two loci ($\delta 1$ and $\delta 2$) as pharmacogenetic processes.

The $\delta 1$ variant has a marginal effect on both HDL and TG levels via additive effects but also, individuals that are homozygous for the minor allele are non-responders to the treatment. Responders (homozygotes and heterozygotes for the major allele) exhibit a 10% increase in their HDL levels and a 15% decrease in TG levels. Similarly, $\delta 2$ is a variant with an additive marginal effect on LDL, and homozygotes for the minor allele are non-responders to treatment. Responders exhibit a 30% decrease in LDL levels. Total cholesterol (CHOL) level is calculated as 0.8*(HDL + LDL + TG/5), and has no independent genetic effects except those influencing the component phenotypes.

Coronary artery calcification (CAC) was simulated as a quantitative phenotype that takes many years to develop. For this reason, CAC was modeled in two stages. First, age-independent CAC ($CAC_{AI}$) was modeled as a function of total CHOL, HDL, and five other genes ($\tau 1$-$\tau 5$) having direct effects on its development. $CAC_{AI}$ was simulated under the model

$$CAC_{AI} = 500 + 20(\text{Total CHOL} - 200) - 25(\text{HDL} - 53) + \mathbf{ME} + \mathbf{PE} + \mathbf{Het} + \varepsilon,$$

where **ME** is a joint genetic effect from an epistatic interaction between $\tau 1$ and $\tau 2$, the effect of $\tau 1$ is purely epistatic (i.e., $\tau 1$ displays only a minimal main effect) while $\tau 2$ displays an additional measurable additive main effect; **PE** is the joint effect from $\tau 3$ and $\tau 4$, a pair of purely epistatic SNPs, each with no main effect; **Het** is an effect from $\tau 5$, a SNP that displays heterosis (overdominance); and $\varepsilon$ is the residual variation not explained by the factors mentioned above. The term $\varepsilon$, 300 times a random draw from a normal distribution with mean 0 and variance 1 ($300 \times N(0,1)$), represents the sum of normal deviations from the mean of each of the modeled genetic effects and "noise" from unmeasured environmental and genetic effects. Because CAC cannot be negative, $CAC_{AI}$ = 0 if the generated value was negative. The models for the effects on $CAC_{AI}$ due to the **ME** and **PE** genotypes are illustrated in Tables 3 and 4. The minor allele frequency (MAF) for each of the four SNPs $\tau 1$-$\tau 4$ is ~0.5. SNP $\tau 5$, which determines the **Het** effect, has a MAF of 0.2. SNP $\tau 5$ genotype 1/1 (common homozygote) increases $CAC_{AI}$ on average by 25, genotype 1/3 decreases $CAC_{AI}$ by 100, and genotype 3/3 increases $CAC_{AI}$ by 400. CAC is derived from $CAC_{AI}$ by using a piecewise linear age adjustment: subjects under 20 years have not developed measurable levels of CAC, CAC buildup is linear from the ages of 20 to 60, and for subjects older than 60, CAC = $CAC_{AI}$. Table 5 lists estimates of the proportion of the variability of CAC attributable to each of the genetic factors averaged over the 200 replicate datasets.

**Table 3: Mean effects of ME ($\tau 1$ and $\tau 2$) on $CAC_{AI}$**

| | | $\tau 2$ | | | |
|---|---|---|---|---|---|
| | | **2/2** | **2/4** | **4/4** | **marginal affects** |
| $\tau 1$ | **2/2** | - 250 | 0 | 250 | 0 |
| | **2/4** | 150 | 0 | - 150 | 0 |
| | **4/4** | - 250 | 0 | 250 | 0 |
| | **marginal affects** | - 100 | 0 | 100 | |

**Table 4: Mean effects of PE ($\tau 3$ and $\tau 4$) on $CAC_{AI}$**

| | | $\tau 4$ | | |
|---|---|---|---|---|
| | | **1/1** | **1/2** | **2/2** |
| $\tau 3$ | **2/2** | 200 | - 200 | 200 |
| | **2/4** | - 200 | 200 | - 200 |
| | **4/4** | 200 | - 200 | 200 |

Whether a subject smoked during the period before a visit influenced the risk of a myocardial infarction (MI). At first visit, men had a 27% chance to be smokers and women had a 23% chance. Each smoker had an 8% chance of permanently quitting smoking before each subsequent visit. The resulting smoking rates are commensurate with rates reported by the Centers for Disease Control for 1998. The risk of an MI before each visit is determined by CAC and its interactions with smoking and two genetic loci, $\phi 1$ and $\phi 2$. No MIs were fatal in our data. Smoking and $\phi 1$ have an interactive effect on risk of MI. The effect of smoking is to constrict blood vessels, thus increasing the risk that CAC will lead to an MI. The risk of MI for a smoker with the most common $\phi 1$ genotype (3/3) is the same as that of an equivalent non-smoker whose CAC is 10% higher. The risk of MI for a smoker with either of the other $\phi 1$ genotypes is the same as that of a non-smoker whose CAC is 40% higher. The $\phi 1$ genotype has no effect on risk of MI in non-smokers. Carrying the most common $\phi 2$ genotype (3/3) has the same effect on risk of MI as reducing CAC by 5%. The effect of any other genotype is the same as increasing CAC by 5%. The final model for MI risk is

$$MI\_risk = [1 + \exp(24 - (0.02 \times CAC \times (1 + \partial_{smoke} + event)]^{-1},$$

where $\partial_{smoke}$ is the joint effect of smoking and $\phi 1$ (0 if a non-smoker, 0.1 if a smoker with genotype 3/3 at $\phi 1$, 0.4 if a smoker with another genotype); and the value of the *event* variable is -0.05 if the $\phi 2$ genotype is 3/3 and 0.05 otherwise. The MAFs for $\phi 1$ and $\phi 2$ are ~0.3. MI_risk was calculated for each visit and a draw from a uniform distribution determined whether the risk resulted in an MI. The SNPs for CAC and MI event were chosen from

**Table 5: Proportion of explained variability for the genetic factors contributing to CAC[a] (by visit)**

| | | Factor | | L1[b] | | L2 | |
|---|---|---|---|---|---|---|---|
| **Factor** | **Visit** | **Mean** | **(Min - Max)** | **Mean** | **(Min - Max)** | **Mean** | **(Min - Max)** |
| ME | 1 | 0.0053 | 0.0037 - 0.0065 | 0.00002 | 0.0[c] - 0.00012 | 0.00030 | 0.00008 - 0.00063 |
| | 2 | 0.0092 | 0.0075 - 0.0112 | 0.00003 | 0.0 - 0.00013 | 0.00055 | 0.00015 - 0.00093 |
| | 3 | 0.0115 | 0.0091 - 0.0137 | 0.00003 | 0.0 - 0.00019 | 0.00066 | 0.00027 - 0.00127 |
| PE | 1 | 0.0091 | 0.0076 - 0.0115 | 0.00004 | 0.0 - 0.00020 | 0.00003 | 0.0 - 0.00014 |
| | 2 | 0.0176 | 0.0152 - 0.0212 | 0.00004 | 0.0 - 0.00017 | 0.00004 | 0.0 - 0.00019 |
| | 3 | 0.0226 | 0.0191 - 0.0266 | 0.00004 | 0.0 - 0.00021 | 0.00004 | 0.0 - 0.00016 |
| Het | 1 | 0.0021 | 0.0012 - 0.0032 | | | | |
| | 2 | 0.0045 | 0.0032 - 0.0060 | | | | |
| | 3 | 0.0062 | 0.0045 - 0.0080 | | | | |

[a]Models included the genetic factors age, sex, CHOL, and HDL.
[b]Columns L1 and L2 indicate the effect of each of the epistatic loci when analyzed without its mate.
[c]0.0 indicates $R^2<10^{-5}$.

the 50 k SNPs in the Gene Focused Panel based on desired MAF, completeness of genotyping, and lack of linkage disequilibrium between the SNPs. The specific identities of the SNPs $\tau$1-$\tau$5, $\phi$1 and $\phi$2, and their chromosomes are listed in Table 6.

## Results and discussion

The phenotypic simulated files are named simphen#.txt, where # stands for a number from 1-200, representing the replication number. The simulated data are archived in the dbGAP of the National Center for Biotechnology Information under the name "GAW16 Framingham and Simulated Data" [8]. The 200 replications of the data include the indexing variable "shareid" that matches exactly with the same shareid of the Framingham Heart Study and can be used to merge the simulated phenotypic data with the FHS genotypic data. The phenotypic variables provided are sex, simage (simulation age), diet, rx (antihyperlipidemic medication use), LDL, HDL, TG, CHOL, CAC, SMOKE, and MIevent, each associated with a number (1, 2, or 3) to identify respectively variables that were simulated for Visit 1, Visit 2, or Visit 3.

**Table 6: SNPs contributing to CAC and MI event**

| Trait | Factor | RS number | MAF | Chr |
|---|---|---|---|---|
| CAC | $\tau$1 | rs6743961 | 0.4997 | 2 |
| | $\tau$2 | rs17714718 | 0.5000 | 19 |
| | $\tau$3 | rs1894638 | 0.4990 | 6 |
| | $\tau$4 | rs1919811 | 0.4994 | 7 |
| | $\tau$5 | rs213952 | 0.2000 | 7 |
| MI | $\phi$1 | rs12565497 | 0.3001 | 1 |
| | $\phi$2 | rs11927551 | 0.2999 | 3 |

We tested all the simulated traits and causative SNP heritabilities as well as the respective association models. Analyzing and interpreting data obtained as part of a genome-wide association study presents numerous challenges, as well as the promise of improved understanding of the genetic factors influencing complex traits. For validation and a detailed analysis of the simulated model see the Online Supplemental Materials for GAW16 [9]. Many genome-wide association studies have been published recently, and many more are being carried out on virtually every conceivable phenotype of biomedical or public health importance. While the rate of development of genetic technologies has propelled us to this point, development and evaluation of statistical and analytic techniques is still underway, with many issues not yet satisfactorily resolved. Nonetheless, important discoveries have been reported. We hope that the simulated GAW16 Problem 3 provides data with which investigators can test the strengths and limitations of their statistical analytic approaches and software.

## List of abbreviations used

CAC: Carotid arterial calcification; CHOL: Cholesterol; FHS: Framingham Heart Study; GAW: Genetic Analysis Workshop; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; MAF: Minor allele frequency; MI: Myocardial infarction; SNP: Single-nucleotide polymorphism; TG: Triglyceride.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

All the authors contributed equally.

## Additional material

### Additional file 1

*Heritability targets and other characteristics for each polygene affecting HDL.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1753-6561-3-S7-S4-S1.xls]

### Additional file 2

*Heritability targets and other characteristics for each polygene affecting LDL.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1753-6561-3-S7-S4-S2.xls]

### Additional file 3

*Heritability targets and other characteristics for each polygene affecting TG.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1753-6561-3-S7-S4-S3.xls]

## Acknowledgements

This work was partially supported by NIH grants HL08770003, HL08768803, 1RR02499203, DK06833603, and HL08821502. The authors are grateful to the continuous interactions with the GAW16 Steering Committee, especially Jean MacCluer and Laura Almasy; with the Framingham Heart Study and especially with collaborators L Adrienne Cupples and Larry Atwood; the NIH/NCBI, especially Cashell Jaquish and Michael Feolo. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

## References

1. Kannel WB, Dawber TR, Kagan A, Revotskie N and Stokes J III: **Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study.** *Ann Intern Med* 1961, **55:**33–50.
2. Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, Gianniny L, Burtt NP, Melander O, Orho-Melander M, Arnett DK, Peloso GM, Ordovas JM and Cupples LA: **A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study.** *BMC Med Genet* 2007, **8(suppl 1):** S17.
3. Kraja AT, Rao DC, Weder AB, Cooper R, Curb JD, Hanis CL, Turner ST, de Andrade M, Hsiung CA, Quertermous T, Zhu X and Province MA: **Two major QTLs and several others relate to factors of metabolic syndrome in the family blood pressure program.** *Hypertension* 2005, **46:**751–757.
4. Kannel WB: **Risk stratification of dyslipidemia: insights from the Framingham Study.** *Curr Met Chem Cardiovasc Hematol Agents* 2005, **3:**187–193.
5. Almasy L, Terwilliger JD, Nielsen D, Dyer TD, Zaykin D and Blangero J: **GAW12: simulated genome scan, sequence, and family data for a common disease.** *Genet Epidemiol* 2001, **21 (suppl 1):**S332–S338.
6. Daw EW, Morrison J, Zhou X and Thomas DC: **Genetic Analysis Workshop 13: simulated longitudinal data on families for a system of oligogenic traits.** *BMC Genet* 2003, **4(suppl 1):**S3.
7. Cupples LA, Heard-Costa N, Lee M, Atwood LD and for the Framingham Heart Study Investigators: **Genetics Analysis Workshop 16 Problem 2: The Framingham Heart Study Data.** *BMC Proc* 2009, **3(suppl 7):**S3.
8. **dbGaP: Genotypes and Phenotypes. GAW16 Framingham and Simulated Data Study Accession: phs000128.v2.p2.** http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000128.v2.p2.
9. **Online Supplemental Materials for Publication - GAW16.** http://dsgweb.wustl.edu/OSMP/GAW16/OSMPGAW16.html.