Genetic scores for saturated fat response prediction

1 Model intuition

To develop a genetic risk score for "saturated fat responsiveness", the ideal dataset would be a large-scale dietary intervention either increasing of decreasing saturated fat, enabling a model to directly predict changes in a cardiovascular risk factor over the course of the intervention. However, we can create a noisy approximation of this model using cross-sectional data by instead using a product of SFA intake (centered) and some risk factor (centered) as the outcome. The intuition for this approach is based on the fact that we are ultimately trying to model the correlation between SFA and a risk factor, and correlations are defined mathematically as the expected value of a product of two variables.

So, the model being used in each GWAS is the following: $Y = \alpha g + X\beta + \epsilon$, where Y is the product of two standardized variables (SFA and LDL-C), g is a vector of genotype dosages at the SNP in question, and X is a matrix of covariates.

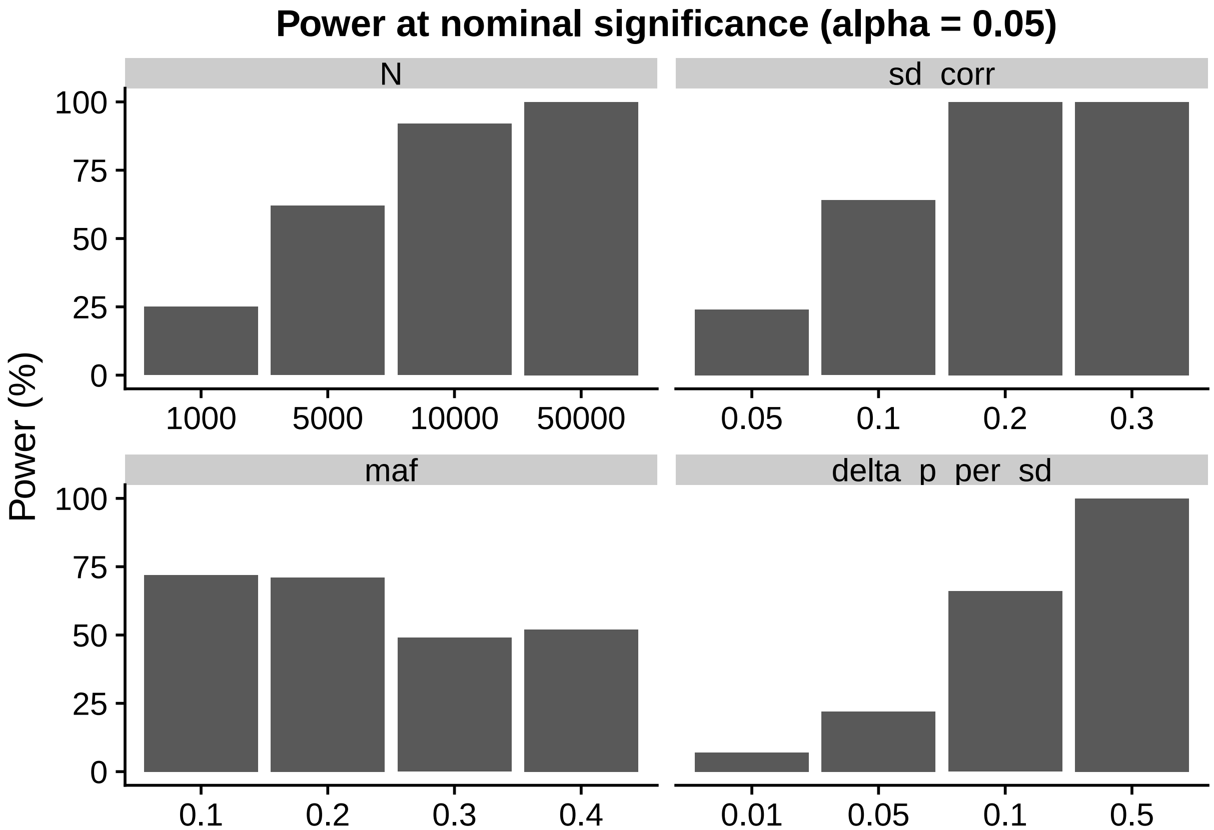
2 Simulations

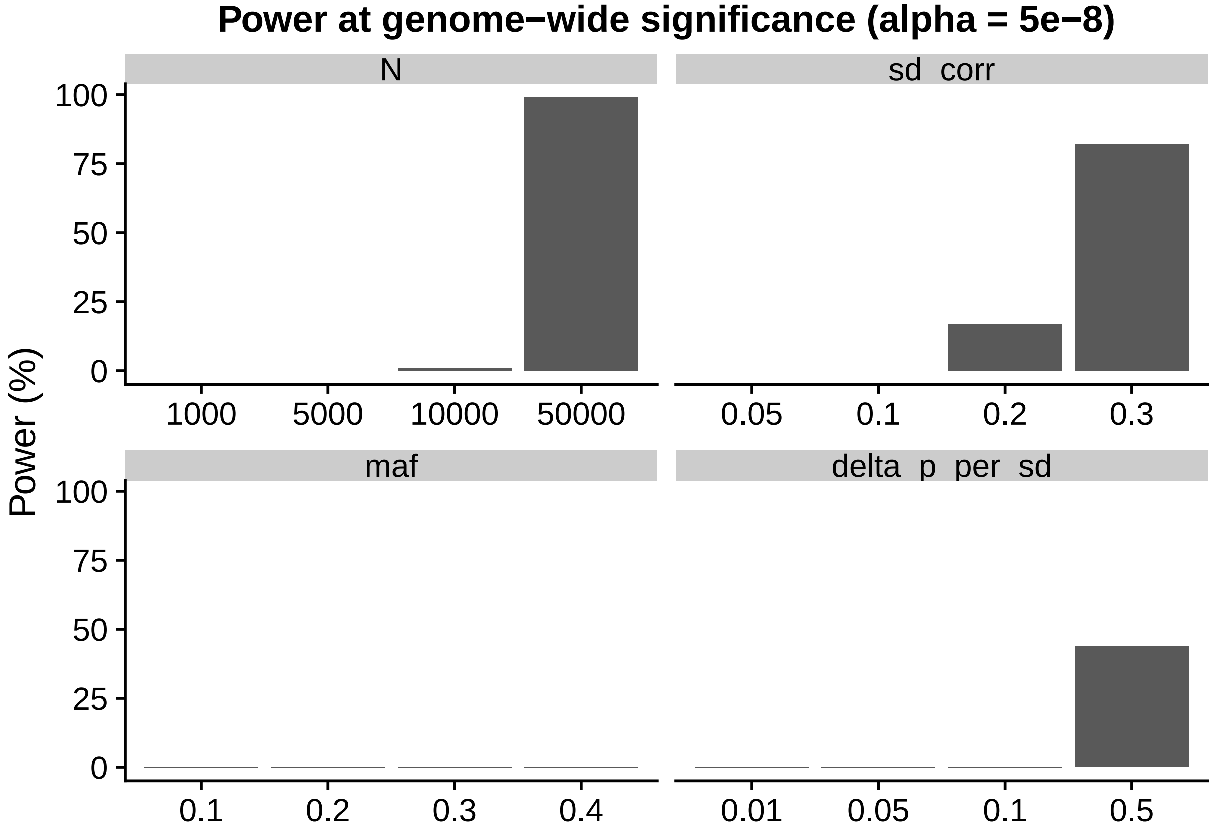
Strategy: for each simulation…

1. Draw N SFA-LDL correlations from a normal distribution ($N(0.1, 0.1)$)
2. Draw a single SFA and LDL value for each correlation from a multivariate normal distribution ($N(\begin{pmatrix} mean_{sfa} \\ mean_{ldl} \end{pmatrix}, \begin{pmatrix} var_{sfa} & cov_{sfa\_ldl} \\ cov_{sfa\_ldl} & var_{ldl} \end{pmatrix})$) with parameters:
    - SFA (mean=20, sd=5)
    - LDL (mean=120, sd=20)
3. Draw a genotype vector (simulating a single variant) as a set of 2 draws per individual from a Bernoulli distribution (0/1 value) with base probability of some minor allele frequency that is a function of the underlying SFA/LDL correlation
4. Use repeated simulations (100) to calculate power for each set of parameters using a linear model predicting $SFA_{scaled} * LDL_{scaled}$ from genotype

Default values for the parameters:

- N = 5000
- Std. dev. of the underlying correlation values = 0.1
- Minor allele frequency = 0.2
- Change in $p_{binomial}$ per std. dev. of the underlying correlation = 0.1


Power at nominal significance (alpha = 0.05)
Power (%)
N
sd corr
maf
delta p per sd
100
75
50
25
0
1000
5000
10000
50000
0.05
0.1
0.2
0.3
0.1
0.2
0.3
0.4
0.01
0.05
0.1
0.5


Power at genome−wide significance (alpha = 5e−8)
Power (%)
N
sd corr
maf
delta p per sd
100
75
50
25
0
1000
5000
10000
50000
0.05
0.1
0.2
0.3
0.1
0.2
0.3
0.4
0.01
0.05
0.1
0.5

3 Individual GWAS results

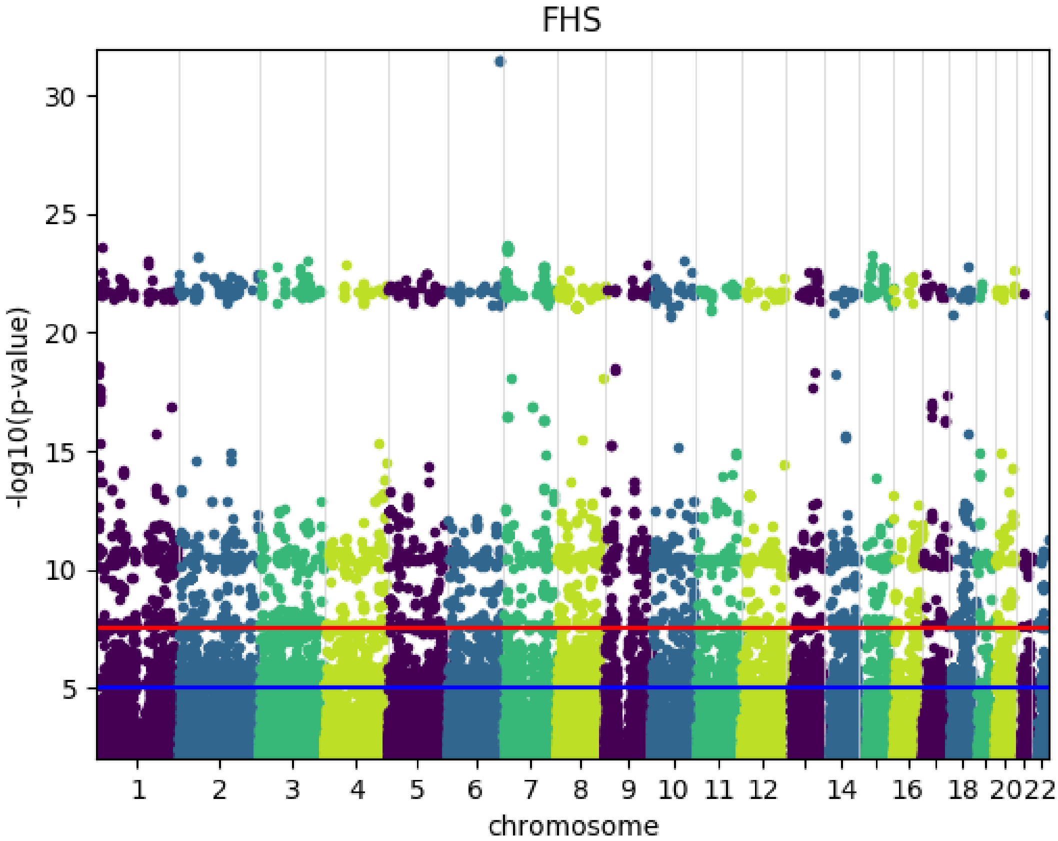
GWAS model (as above): $Y = \alpha g + X\beta + \epsilon$

- Y is the product of two standardized variables (SFA and LDL-C)
  - SFA in units of g/day, normalized by total calories
- g is a vector of genotype dosages at the SNP in question
- X is a matrix of covariates

Inclusion criteria: * White * No lipid medication use

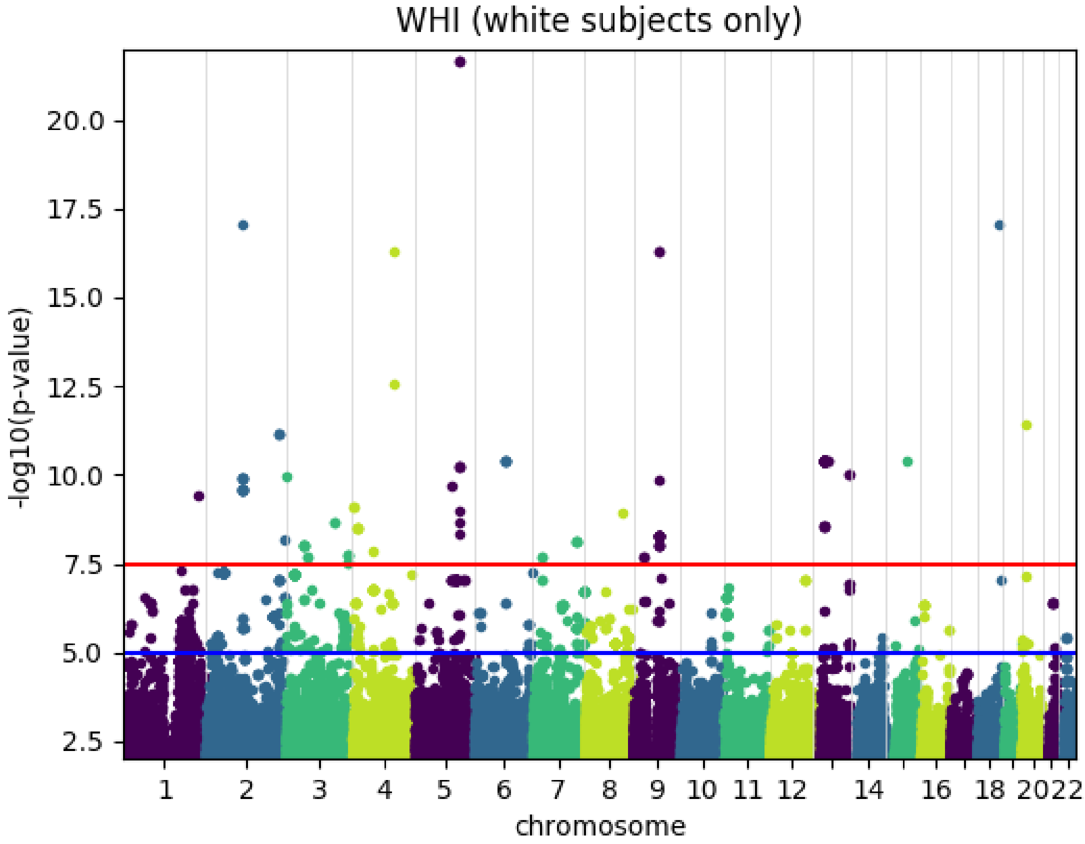
3.1 FHS

- N ~ 750 unrelated individuals
- Covariates: age, sex, BMI, PUFA



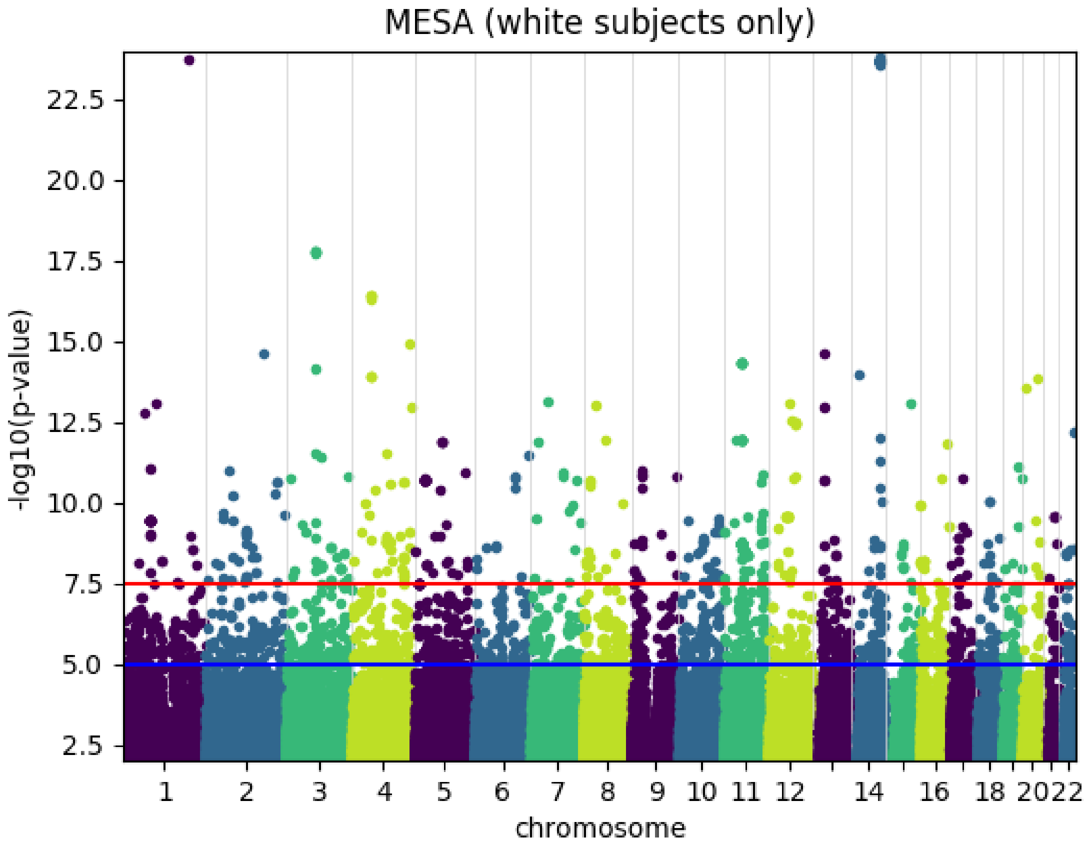
3.2 WHI

- N ~ 5500
- Covariates: age, BMI, PUFA, 5 ancestry principal components



3.3 MESA

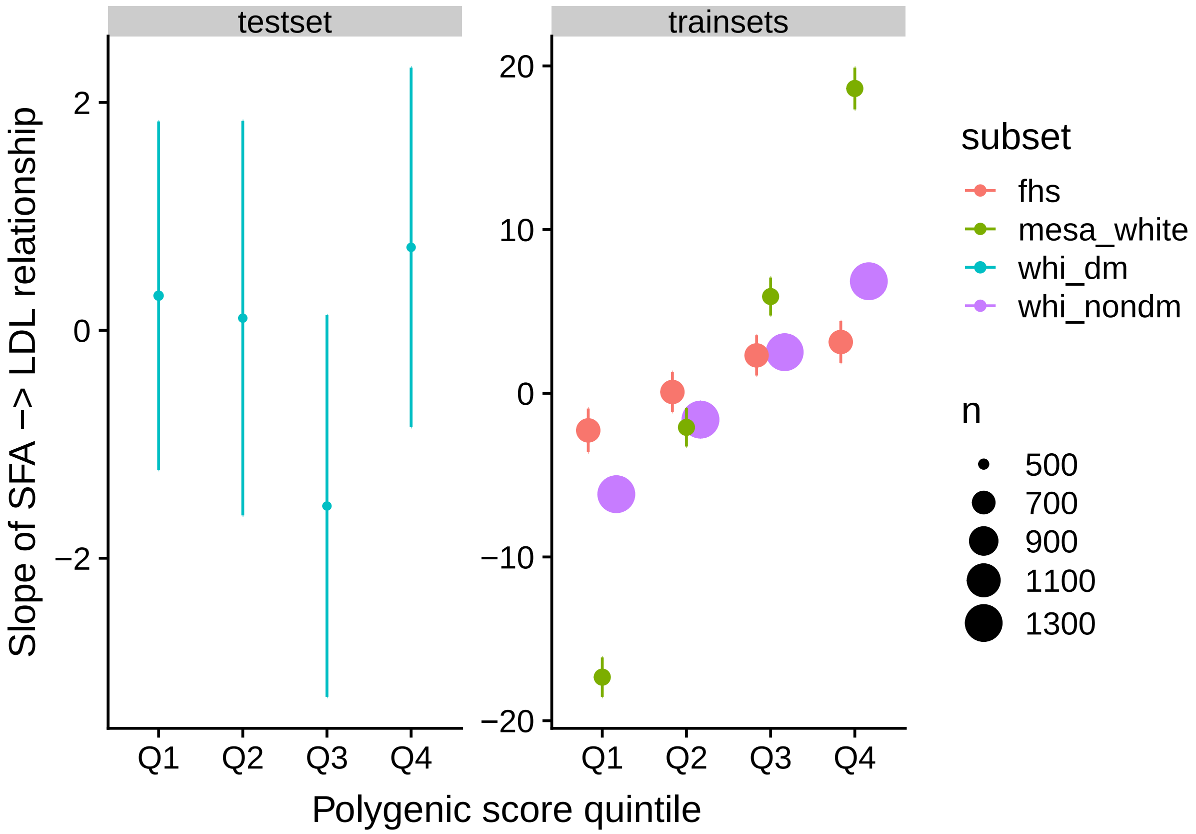
- N ~ 1300
- Covariates: age, BMI, PUFA



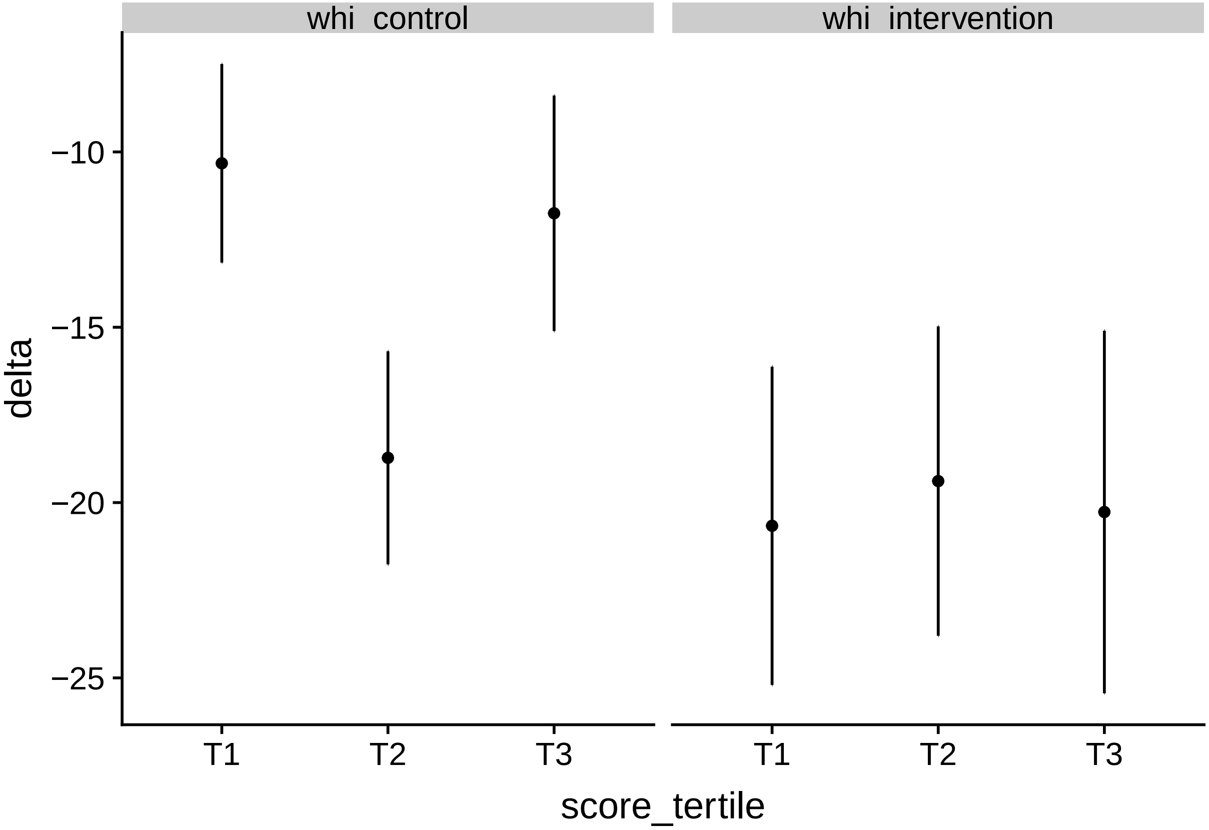
4 Initial results

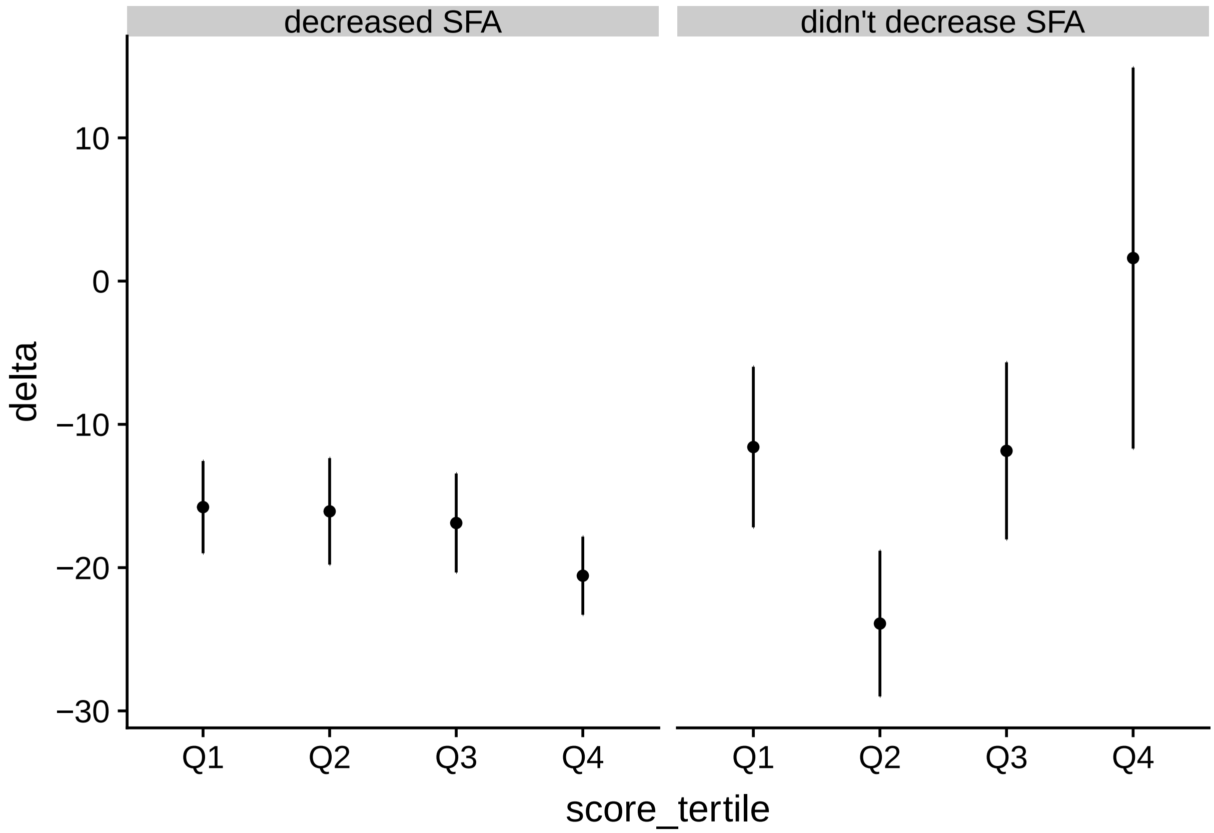
How to combine these cohorts?

- Meta-analysis is an option, but didn't perform it here because of the plans to incorporate other ethnicities and the questionable transference of genetic scores across ancestries.
- Approach based on Patil et al. 2018 – combine individual predictions from each cohort rather than first meta-analyzing the GWAS results
- They are interested in collaborating on this, but here I have implemented a basic version of one of their strategies.


testset
trainsets
Slope of SFA –> LDL relationship
Polygenic score quintile
2
0
−2
20
10
0
−10
−20
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
subset
fhs
mesa_white
whi_dm
whi_nondm
n
500
700
900
1100
1300

5 Longitudinal results in WHI DM subjects





6 Moving forward

- On the question of how to deal with ancestry differences – Parmigiani lab @ Harvard open to collaboration on prediction from models trained in separate cohorts
- Application is in to UK Biobank to try to replicate any findings in the 500k subjects there