Investigation and downstream analysis of MRS findings

Kenny Westerman

2017-08-16

Contents

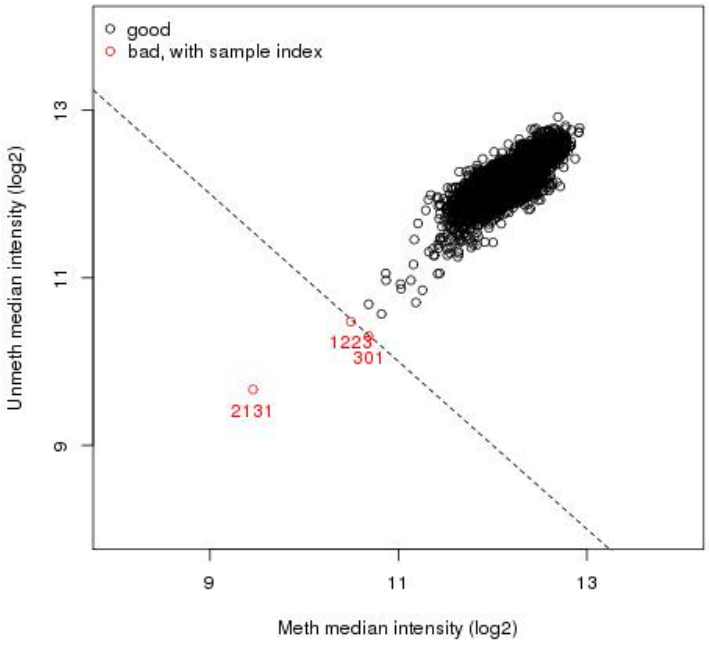
Methylation data

Datasets

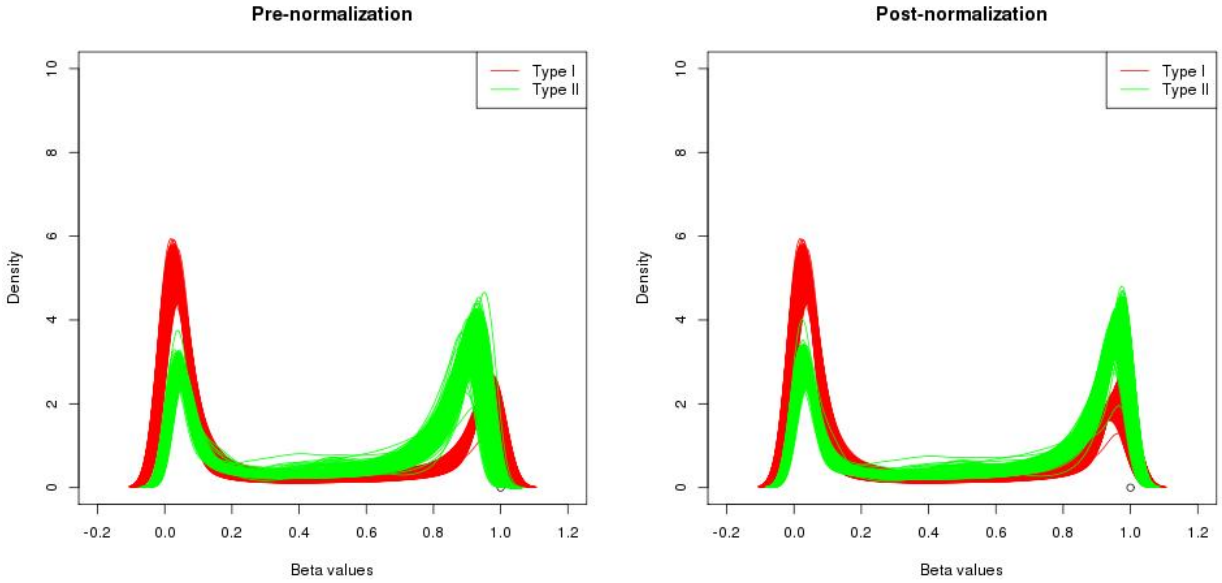
- FHS Offspring in use (dbGaP request likely to be accepted within days)
- Normative Aging Study – in discussion with Joel Schwartz about obtaining the data
- Lothian Birth Cohorts – pending, may be available soon

Preprocessing

- Read in raw .idat intensity files using the minfi package
- Sample QC
  - Mismatch between predicted and reported sex
  - Failed samples with low intensity

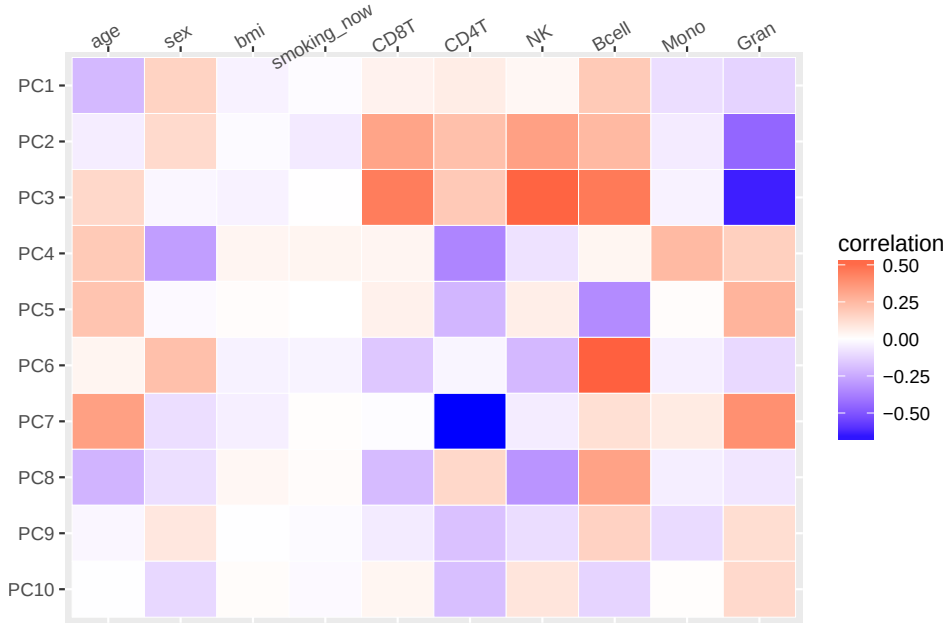


- BMIQ within-array normalization (Teschendorff et al. 2013) for Type I/II bias



- Probe filtering to remove:
  - Probes in sex chromosomes
  - Cross-reactive probes (identified in Chen et al. 2013)
  - Probes measuring SNPs
  - Probes with SNPs at CpG site or single-base extension site
  - Probes measuring CpH (non-CpG) sites
- Methylation-based covariates: Various potential covariates were calculated from the methylation data, including PCA (based on the 50k most variable probes), PCA on control probes (Lehne 2015), SVA, and ReFACTor.

Correlations of the top PCs with basic covariates:



Subject metadata

Population

- Have samples from 2590 unique individuals in the dataset after sample QC
- Most subjects are between 50 and 80 years old
- Approximately equal sex distribution

Cardiovascular events

CVD events were broadly defined as MI, stroke, or death from CHD/CVD. 296 individuals experienced an event between Exam 8 and the available ~3600 days of follow-up. Further characterization of this subgroup:

- 168 males, 128 females
- Very few (8) smoke
- About 100 of these had experienced a previous cardiovascular event.

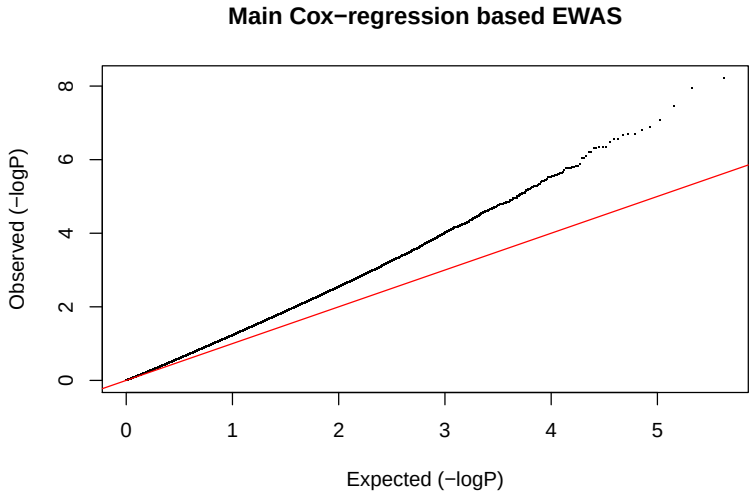
Basic EWAS

A genome-wide scan was performed using methylation M-values ($= log_2\left(\frac{intensity_{methylated}}{intensity_{unmethylated}}\right) = logit(\beta)$) to predict time to CVD event using Cox models. Covariates included age, sex, BMI, blood cell count as estimated by the method of Houseman et al. 2012, and the top 20 PCs.

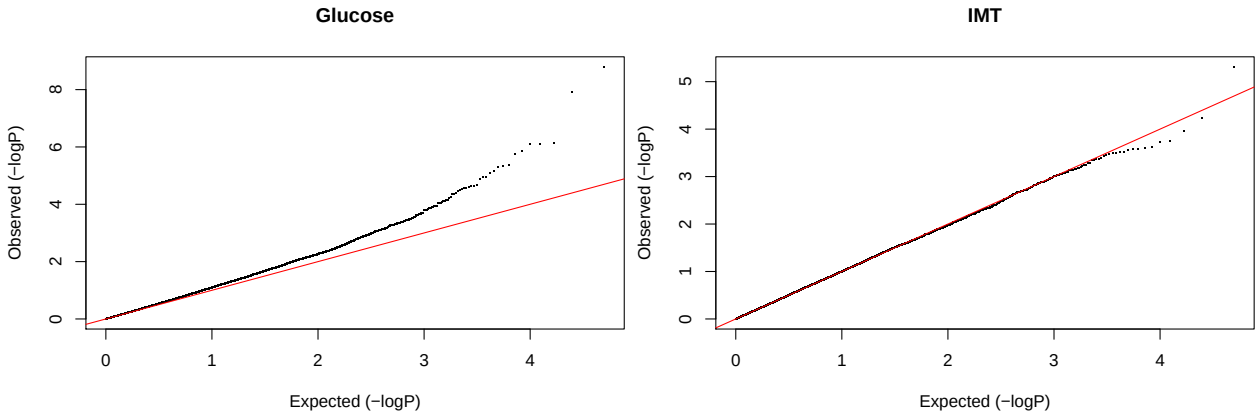
Inflation

Some level of inflation was seen in the overall p-value distribution, using a random subset of 50k probes. Inflation levels were tested across different combinations of the methylation-based covariates mentioned above, and the best performance was seen with a combination of cell counts and principal components, so these covariates were taken forward.

These covariates led to an observed genomic inflation factor λ of 1.31 in the full EWAS.



On Dawn's suggestion, continuous outcomes were also tested using the same set of covariates to understand whether the binary nature of the Cox analysis was contributing to the inflation.



Left: blood glucose as outcome (λ=1.12), right: mean intima-media thickness (from Exam 6) as outcome (λ=1)

The notable inflation reduction with alternate outcomes (despite the same covariates) suggests that either the binary/Cox outcome is somehow the source of the inflated p-values, or there really is a relationship across a significant portion of the genome.

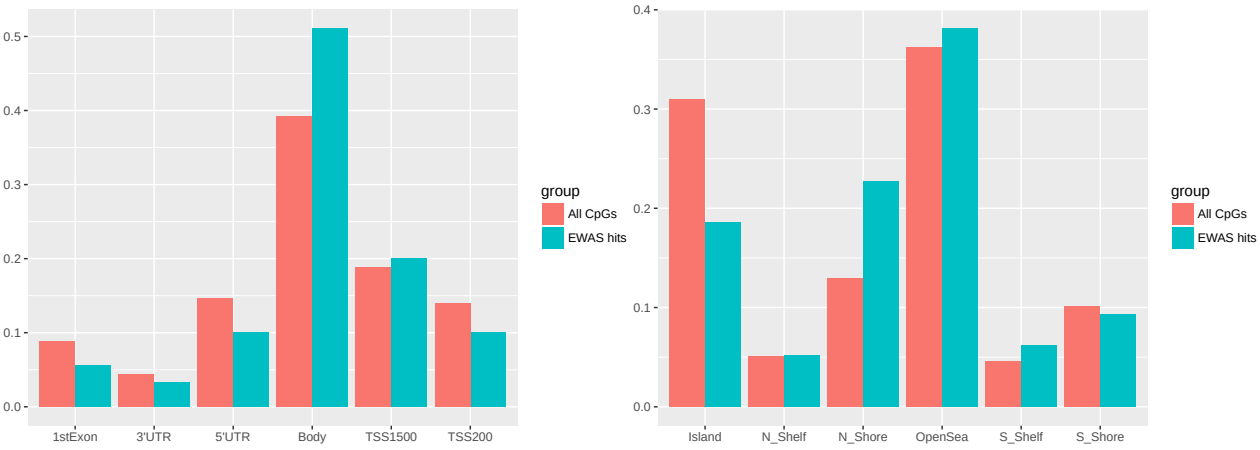
Biological significance of EWAS hits

97 CpG sites are significantly associated with incident CVD at an FDR of 0.05, annotated to 78 unique genes.

Results from a gene set enrichment analysis (simple hypergeometric test) of these genes

Category	p.value	FDR
REACTOME_SIGNALING_BY_GPCR	0.013	0.999
REACTOME_GPCR_DOWNSTREAM_SIGNALING	0.023	0.999
REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	0.109	0.999
REACTOME_CELL_CYCLE	0.143	0.999
REACTOME_GPCR_LIGAND_BINDING	0.152	0.999
KEGG_OLFACTORY_TRANSDUCTION	0.166	0.999
REACTOME_IMMUNE_SYSTEM	0.195	0.999
REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	0.197	0.999
REACTOME_METABOLISM_OF_RNA	0.218	0.999
KEGG_PATHWAYS_IN_CANCER	0.220	0.999
REACTOME_OLFACTORY_SIGNALING_PATHWAY	0.220	0.999
REACTOME_CELL_CYCLE_MITOTIC	0.223	0.999
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	0.245	0.999
REACTOME_METABOLISM_OF_MRNA	0.270	0.999
REACTOME_INNATE_IMMUNE_SYSTEM	0.276	0.999

Genomic location of these CpGs with respect to genes and CpG islands:

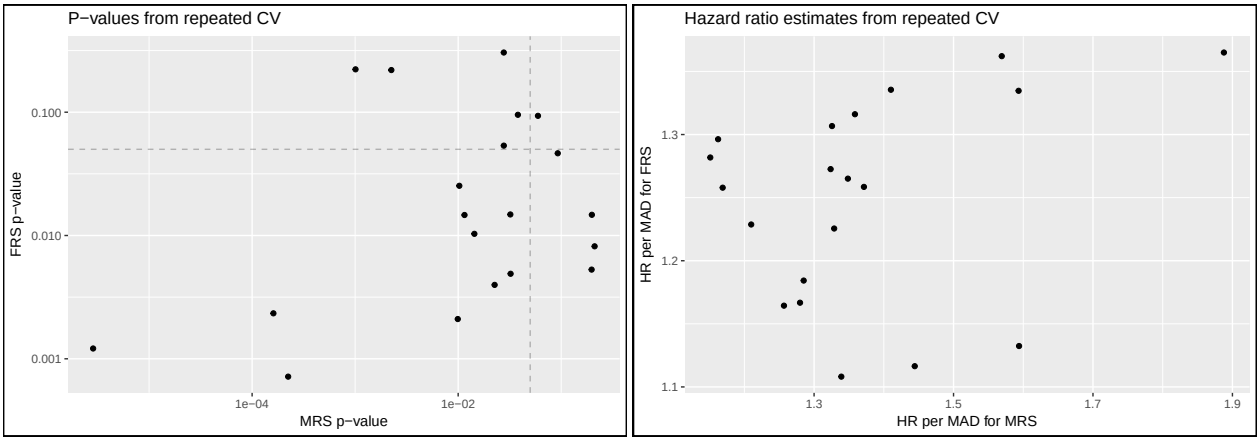


Methylation risk score

Based on the very high-dimensional dataset and existing precedents for similar methylation-based biomarkers (e.g. Horvath DNAm age), a penalized regression approach (elastic net) was applied to the basic Cox model to develop the MRS, using the same covariates as in the EWAS, i.e. *CVD events ~ sex + age + smoking status + WBC counts + 20 PCs + >400k CpGs*. Specifically, all non-zero coefficients from the main regression that corresponded to CpG sites were used in the added sum of methylation M-values that constitutes the MRS.

Cross-validation and stability analysis

Given the current lack of a validation dataset, repeated cross-validation was performed by training the MRS model on a random 50% of the dataset (balanced for total incident CVD events) followed by testing the predictivity of the model on the remaining 50% using a Cox model adjusted for calculated Framingham Risk Score (D'Agostino et al. 2008), for a total of 20 iterations.



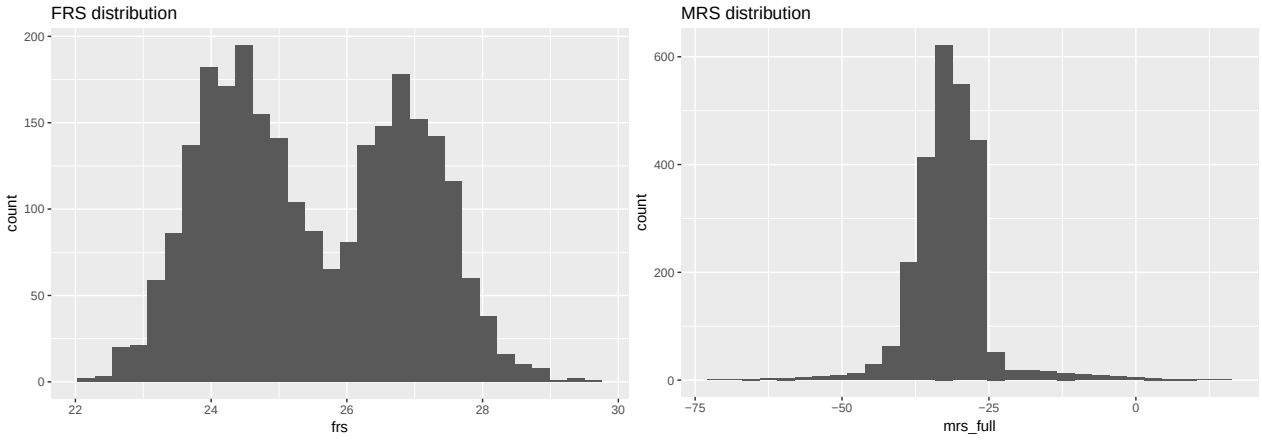
The above plots show result summaries across the 20 repetitions. Left: p-values, right: estimated hazard ratios per median average deviation difference in MRS value. MAD is used in place of SD due to the heavy tails in the MRS distribution.

An average of 989 CpG sites are selected per iteration. Many of the CpG sites are chosen in only one of the iterations (8767 of 19786 total). However, a smaller number appear somewhat consistently, with 38 showing

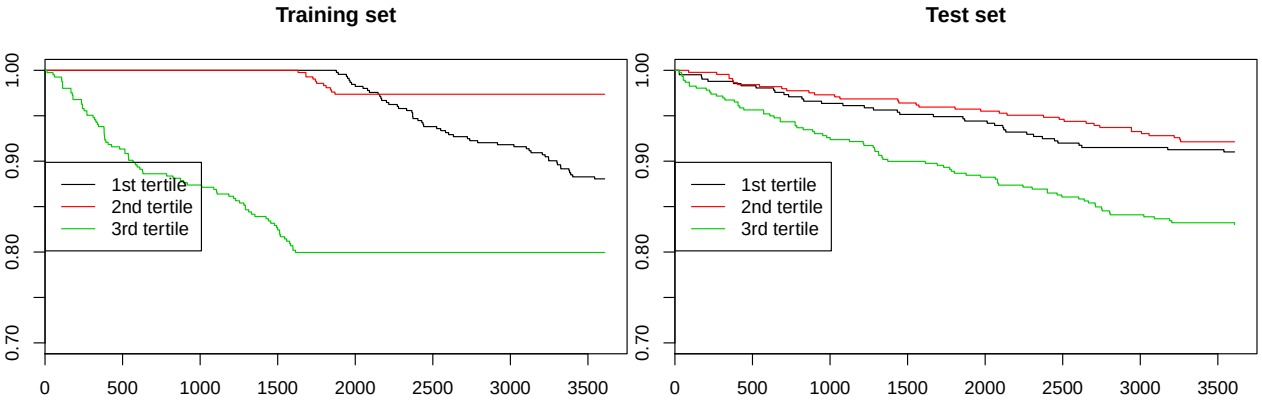
up more than half the time.

Distribution

Though the specific units are arbitrary, the MRS distribution has more extreme values/heavier tails than the set of Framingham Risk Scores in this dataset.

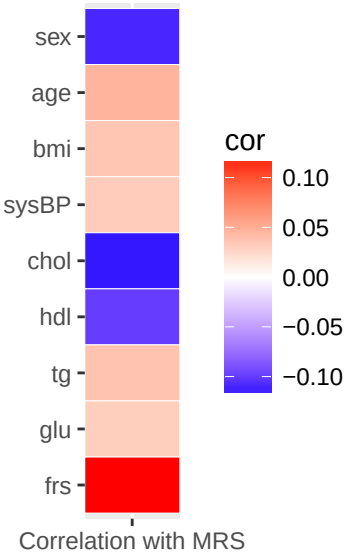


Kaplan-Meier plots



ROC curves

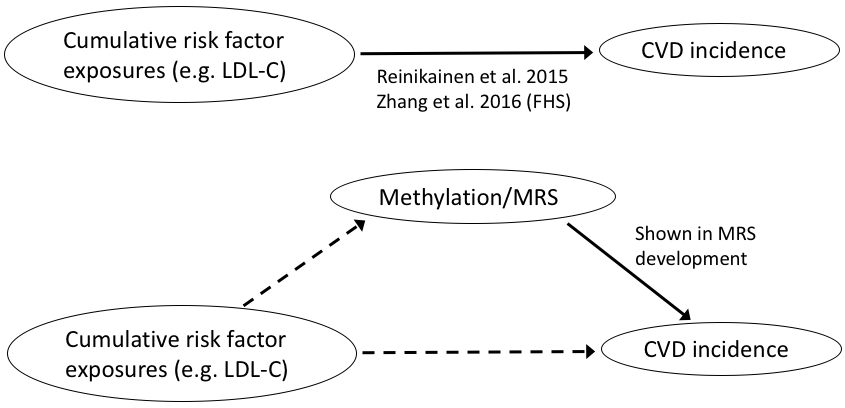
Association with cardiovascular risk factors



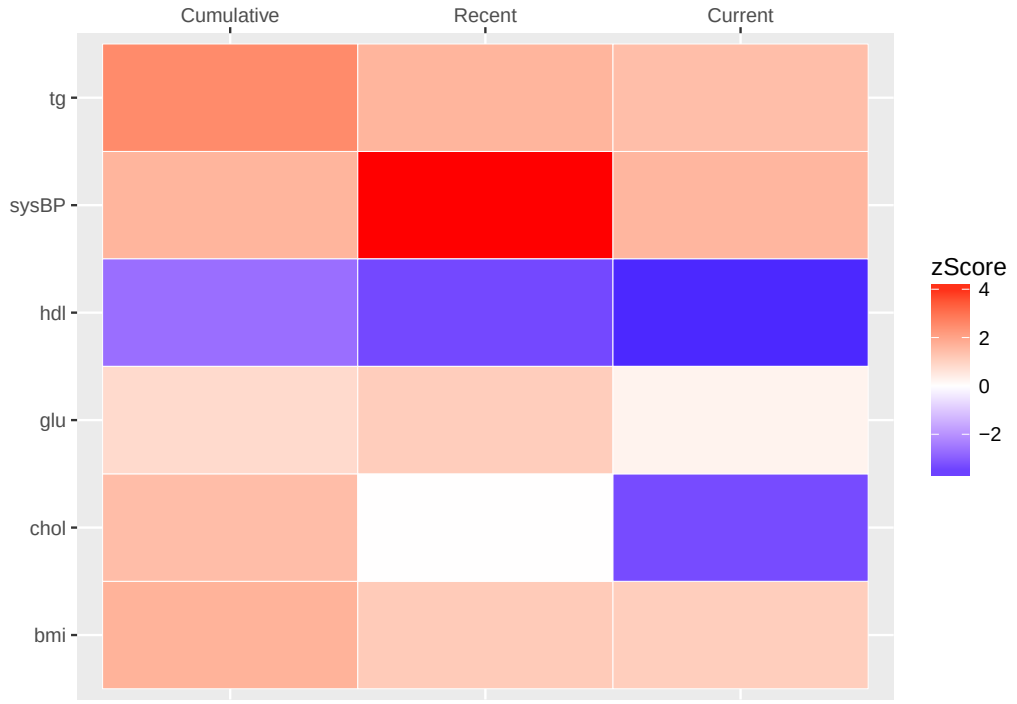
Association with cumulative risk factor exposure

One central question that informed the formulation of my research strategy was whether a CVD MRS would in any sense act as a record of past exposures to risk factors such as blood lipid levels. To investigate this question, laboratory results were retrieved for FHS Offspring participants in exams 1-7 and the following values were established for a series of biomarkers:

- "Cumulative" = mean value over exams 1-7 (~ 35 yrs.)
- "Recent" = mean value over only the more recent exams 5-7 (~ 15 yrs.)
- "Current" = value from Exam 8



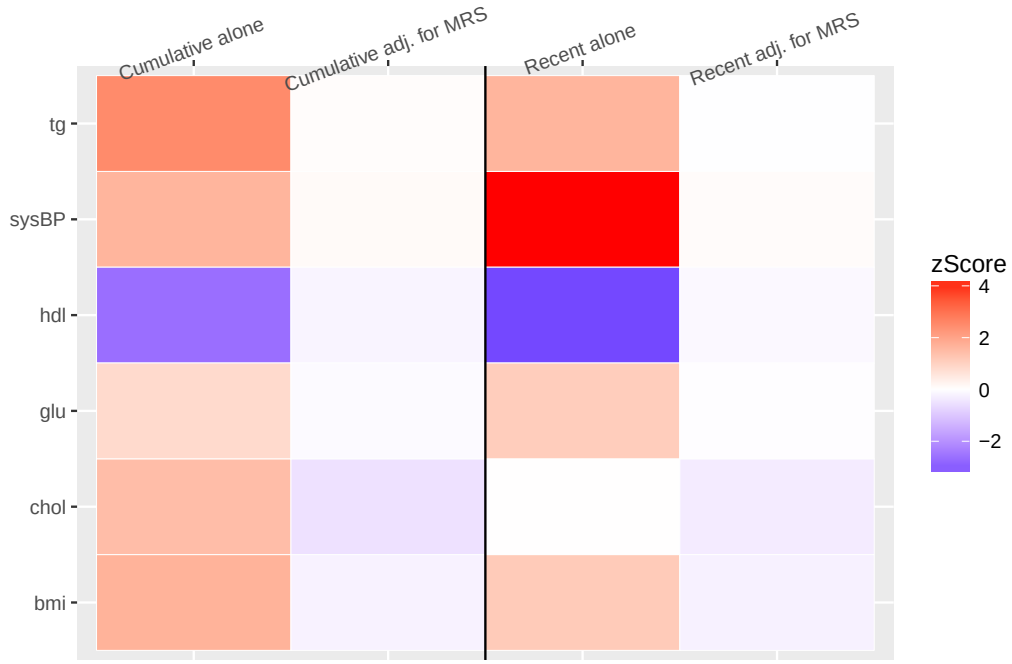
First, do the cumulative exposures have any additive explanatory power beyond current exposure levels (in full dataset) in predicting cardiovascular events? Heatmap represents z-scores from the relevant Cox models.



Next, do the calculated methylation risk scores have an independent association with cumulative biomarker levels (in full dataset)? Heatmap represents t-scores from the relevant linear model predicting MRS from risk factor levels.

Finally, does the MRS potentially mediate these relationships of cumulative biomarker levels with CVD incidence (in "left out" 50% of full dataset)? Heatmap represents z-scores from the relevant Cox models.

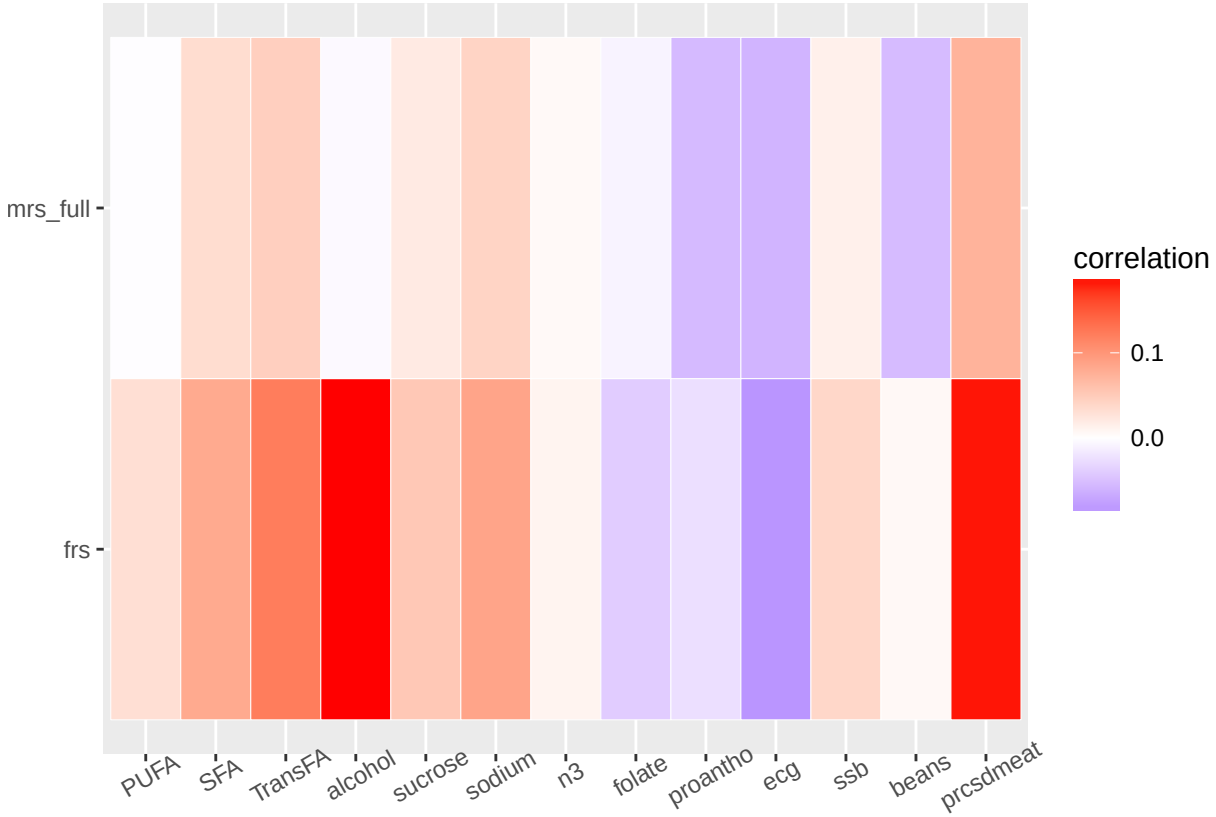
```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 1,2 ; beta may be infinite.
```



Overall, it is interesting that the MRS seems to correlate more strongly with cumulative than current risk factors, but there is not strong evidence that it mediates any cumulative effect of these exposures on CVD risk.

Association with diet

As a preliminary investigation, correlations between specific dietary elements from FFQ data and CVD risk scores were calculated.



Limitations

- Small sample size
- Imperfect data on censoring times due to FHS survival datasets available
- Inclusion of individuals who already have CVD

Future directions

- Any benefit to use of beta-values instead of M-values?
- Acquisition of more datasets for greater statistical power and validation
- Alternative: incorporate a continuous measure (e.g. atherosclerosis progression from imaging) to allow for other datasets to be used or to allow more effective use of the available cohorts
- Incorporation of more 'omics data
  - Genomics
  - Transcriptomics
  - Existing databases, e.g. GTEx
- Further investigation of diet