Supplementary Material: Epigenomic prediction of cardiovascular risk and interactions with traditional risk metrics

Supplementary Methods

Women's Health Initiative

WHI methylation data came from the BA23 ancillary study, a combined case-control and pseudo case-cohort sampling of 2129 women from the Women's Health Initiative study. WHI is a larger prospective cohort beginning in 1993 that included over 160,000 postmenopausal women from across the United States[1]. Included subjects had no self-reported CVD at baseline, and cases were chosen based on incident centrally adjudicated angina, revascularization, or CHD event during follow-up. Inclusion criteria for methylation measurement resulted in an oversampling of African American and Hispanic participants. Blood samples used for measurement of DNA methylation and clinical biochemistry were taken at baseline. Data are available in the dbGaP public repository (accession: phs000200.v11.p3; downloaded on September 27, 2017).

Framingham Heart Study Offspring Cohort

FHS methylation data came from a substudy of the Framingham Heart Study that measured DNA methylation in 2726 subjects from the Offspring Cohort. The Framingham Offspring Cohort was originally established in 1971 to follow 5209 children of the original Framingham Heart Study participants and their spouses[2]. Fasting blood samples for both methylation and clinical biochemistry were collected from participants at Exam 8, which took place from 2005-8. Blood samples were also provided for clinical biochemistry measurements in previous exams, constituting the "past exposures" examined here. Data are available in the dbGaP public repository (accession: phs000007.v29.p10; downloaded on September 27, 2017). Adjudicated cardiovascular event data was collected through 2015, and events were defined here as any of: myocardial infarction, angina pectoris, stroke (approximately 90% being ischemic), or death from CHD (Framingham event codes 1-29). FHS methylation data were collected in two primary batches in two centers – one in subjects from a nested case-control for CVD measured at Johns Hopkins University (FHS-JHU)[3], and the other in a larger set of remaining Framingham Offspring participants measured at the University of Minnesota (FHS-UM).

Lothian Birth Cohorts

The Lothian Birth Cohorts consist of two birth cohorts (born in 1921 and 1936) established in the Lothian region of Scotland[4,5]. Only the 1936 cohort was analyzed here. Blood samples were collected in three waves starting in 2004, with our primary analyses here focusing on samples from Wave 1 (2004-2007). Cardiovascular outcomes were defined as general CVD or stroke determined at each wave, and event times for survival models were approximated based on the time between Wave 1 and the wave at which the event was reported. LBC data are accessible through the European Genome-phenome Archive (accession: EGAD00010000604).

REGICOR

The REGICOR dataset analyzed here consisted of a nested case-control for myocardial infarction within the larger REGICOR (REgistre GIroní del COR) cohort from the Girona Province in Catalonia (Spain). Whole blood samples were collected from 391 total participants, with those from cases generally collected within 24 hours of the event. Characteristics for this population are available in Supp. Table S3.

DNA methylation data processing

DNA methylation data for all initial cohorts (WHI, FHS, and LBC) were collected using the Illumina HumanMethylation450 microarray platform[6] and downloaded as raw intensity files. Preprocessing was performed using the *minfi* and *wateRmelon* packages for R[7,8]. Sample-wise filters were as follows: robust overall signal in the main cluster based on visual inspection of an intensity plot, less than 10% of probes undetected at a detection threshold of p<1e-16, and a reported sex matching methylation-based sex prediction. Probes were removed using the following criteria: more than 10% of samples undetected at a detection threshold of p<1e-16, location in the X or Y chromosomes, non-CpG probes, cross-hybridizing probes, probes measuring SNPs, and probes with an annotated SNP at the CpG site or in the single-base extension region. Samples were normalized using the Noob method for background correction and dye-bias normalization, followed by the BMIQ method for probe type correction[9,10]. Blood cell fractions for 6 blood cell types (CD4+ T-cells, CD8+ T-cells, B-cells, natural killer cells, monocytes, and granulocytes) were estimated using a common reference-based method[11], and 5 of these (excluding granulocytes) were included in cell count-adjusted statistical models. After quality control and filtering steps, 390597 CpG sites were shared between the 3 datasets, formatted as beta values (ratio of methylated signal to total microarray signal).

DNA methylation data for the REGICOR cohort were collected using the Illumina MethylationEPIC microarray platform[12] and analyzed using the *wateRmelon*[8] and methylumi[13] R packages. Samples were excluded based on detection p-value >0.05 in at least 1% of probes or failure to cluster in the appropriate sex based on X chromosome methylation. Probes were excluded based on detection p-value >0.05 in at least 1% of samples, a bead count <3 in at least 5% of samples, discarding by Illumina based on underperformance (n=1,031) or changes in the manufacturing process (n=977), non-CpG targets, and cross-hybridization (n=43,979). A batch normalization was performed by standardizing beta values to mean zero and unit variance within each bisulfite conversion batch prior to analysis. After quality control and preprocessing, 811,610 CpG sites across 391 individuals were available for analysis. Participants were further excluded from analysis due to unknown smoking habits (n=10) and unavailable information regarding diabetes, hypertension, or hyperlipidemia (n=53). Surrogate variable analysis[14] was used to calculate two surrogate variables, representing potential technical and biological confounders, for adjustment in MRS replication models.

Supplementary Tables and Figures

Table S1: MRS performance in held-out FHS subset without past CVD events

Model	HR per s.d. MRS	p
Unadjusted[1]	1.55	1.9e-08
Basic[2]	1.27	7.3e-03
Plus risk factors[3]	1.30	6.0e-03
FRS only[4]	1.37	7.7e-05

[1] No covariates
[2] Adjusted for age, sex, and estimated cell type fractions
[3] Additionally adjusted for BMI, LDL, HDL, SBP, diabetes status, and current smoking
[4] Adjusted for Framingham Risk Score only

Table S2: MRS stability as evaluated by using multiple within-subject measurements. Generic ICC heuristics for reference: 0-0.5 = poor, 0.5-0.75 = moderate, 0.75 - 0.9 = good, 0.9-1 = excellent.

Cohort	Group type	# of pairs/groups	ICC
FHS	Duplicates	26	0.85
LBC36	Samples over multiple visits	758	0.68
LBC36	Samples over subsequent visits (Wave 1 & 2)	758	0.69
LBC36	Samples over longer time frame (Wave 1 & 3)	758	0.61

Table S3: Description of REGICOR myocardial infarction nested case-control population (continuous variables presented as: mean (standard deviation)

Sample size	391
Prior myocardial infarction	50.1%
Ancestry (% European)	100%
Age	63.2 (6.9)
Sex (% female)	48.6
Smoking	21.5%
Body mass index	28.5 (4.8)
LDL cholesterol	127 (26)
HDL cholesterol	50.0 (10.5)
Systolic blood pressure	135 (18)
Diabetes prevalence	24.7%

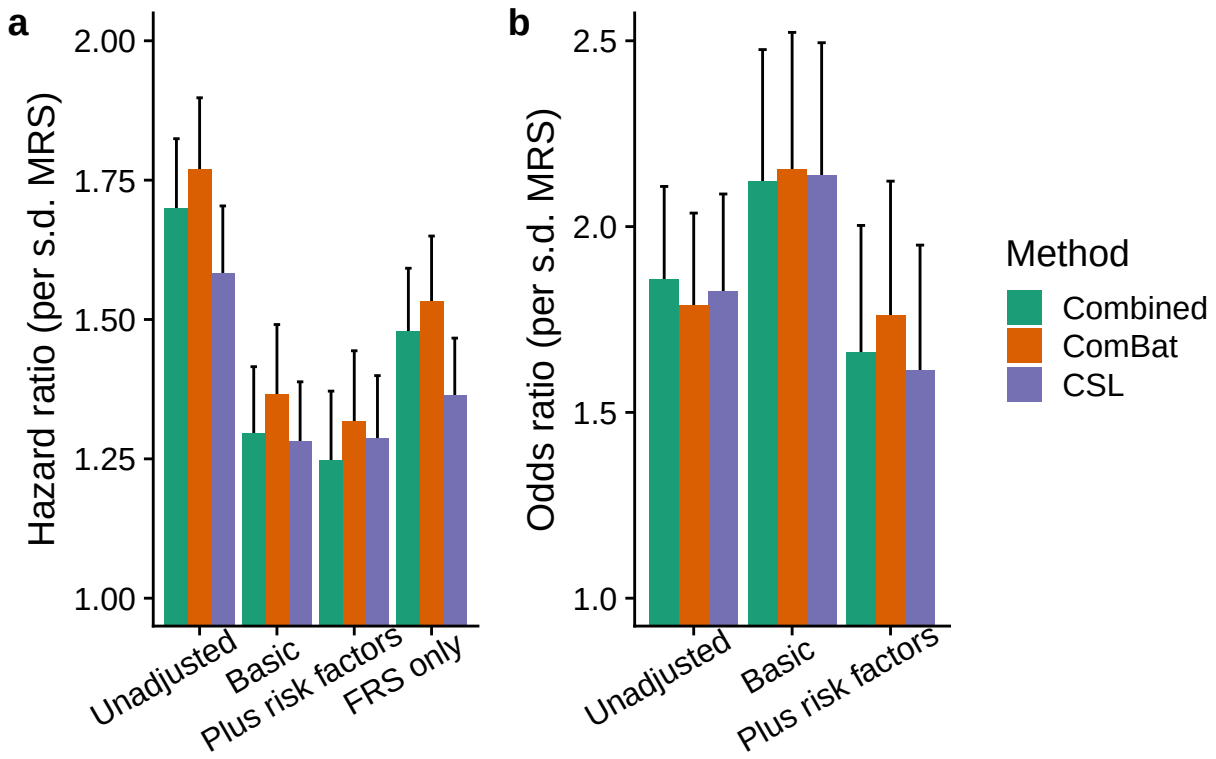

Figure S1: Comparison of modeling approaches. Performance metrics are shown as a function of the test dataset, either FHS-UM (a) or REGICOR (b), and the covariate adjustment. Performance is quantified by either hazard ratio from Cox models (a) or odds ratio from logistic models (b). Covariate sets used for adjustment for models named here are identical to their descriptions for the regression models presented above. Errors bars represent standard errors for the hazard ratio or odds ratio estimates.

Table S4: Validation of Framingham Risk Score

Study	HR per s.d. MRS	p
WHI	1.50	4.7e-61
FHS-JHU	1.40	9.9e-06
FHS-UM	1.63	8.5e-22
LBC	1.01	9.2e-01

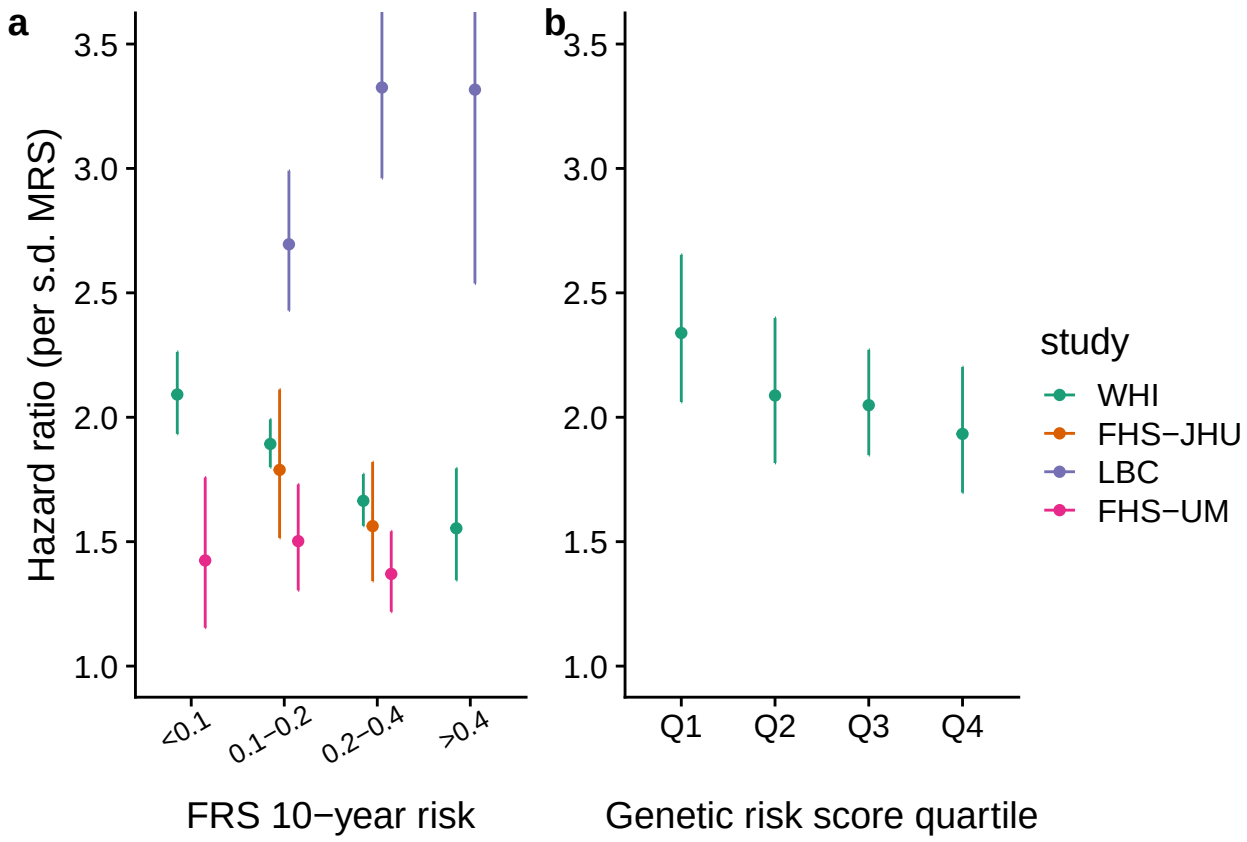

Figure S2: Interactions of MRS with other biomarkers of CVD risk. a) Hazard ratios for the MRS within subsets of 10-year generalized CVD risk according to the Framingham Risk Score. b) Hazard ratios for the MRS within quartiles of a genetic cardiovascular risk score (in white participants only for WHI). Hazard ratios are estimated using the final MRS, which was trained using each of these datasets. Stratum-specific Cox regressions were adjusted for age, sex, and estimated cell subtype fractions. Estimates for strata with less than 25 incident events are not shown. Error bars represent standard errors for the hazard ratio estimates (cut off above in panel (a) for ease of visualization of other points).

Table S5: Risk factor-stratified MRS performance in the REGICOR dataset

Risk factor group	OR per s.d. MRS [95% CI]	N
Q1	4.49 [1.64-12.28]	119
Q2	1.17 [0.67-2.04]	90
Q3	2.58 [1-6.68]	60
Q4	1.2 [0.31-4.59]	55

Supplementary References

1. Anderson GL, Cummings SR, Freedman LS, Furberg C, Henderson MM, Johnson SR, et al. Design of the Women's Health Initiative Clinical Trial and Observational Study. *Controlled Clinical Trials*. 1998;19:61–109.

2. Kannel WB, Feinleib M, Mcnamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families: The framingham offspring study. *American Journal of Epidemiology*. 1979;110:281–290.

3. Joehanes R, Ying S, Huan T, Johnson AD, Raghavachari N, Wang R, et al. Gene Expression Signatures of Coronary Heart Disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2013;33:1418–1426.

4. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*. 2012;41:1576–1584.

5. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*. 2018;47:1042–1042r.

6. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98:288–295.

7. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30:1363–1369.

8. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293.

9. Fortin J-P, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2016;33:btw691.

10. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29:189–196.

11. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86.

12. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*. 2016;17:208.

13. Davis S, Du P, Bilke S, Triche T, Bootwalla O. methylumi: Handle Illumina methylation data. 2019;

14. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*. 2007;3:e161.