

Supplemental Figures for:

Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record

Jacob J. Hughey^{1,2,*}, Seth D. Rhoades¹, Darwin Y. Fu¹, Lisa Bastarache¹, Joshua C. Denny^{1,3}, and Qingxia Chen^{1,4}

¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

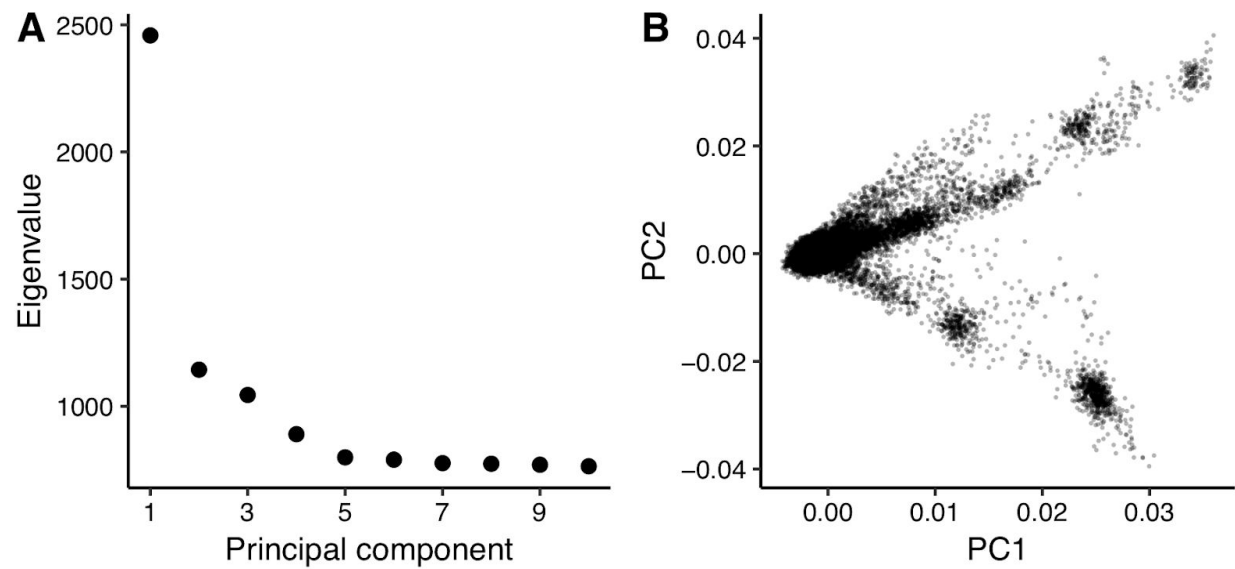
²Department of Biological Sciences, Vanderbilt University, Nashville, TN

³Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

⁴Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN

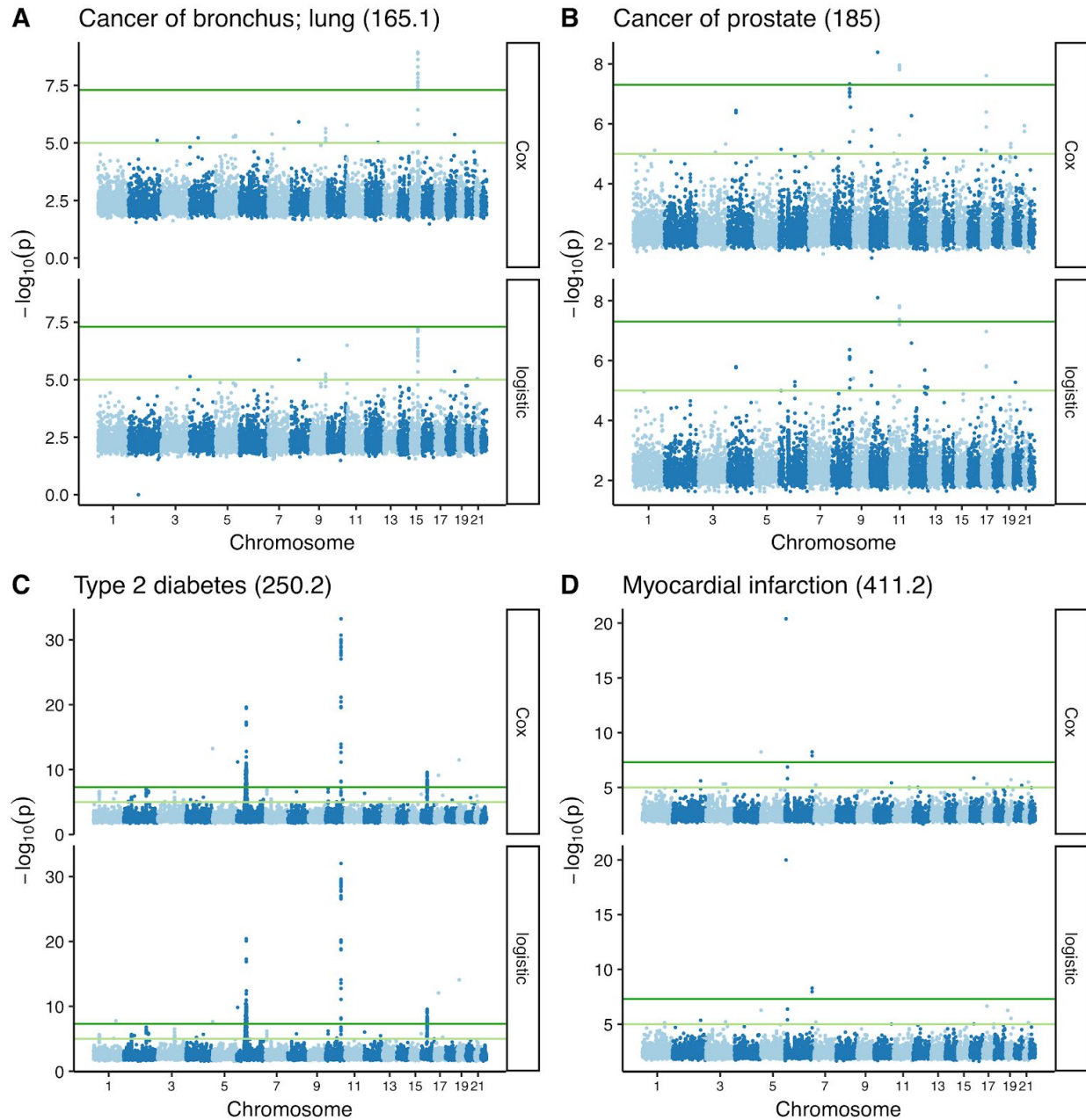
*To whom all correspondence should be addressed: jakejhughey@gmail.com

Figure S1



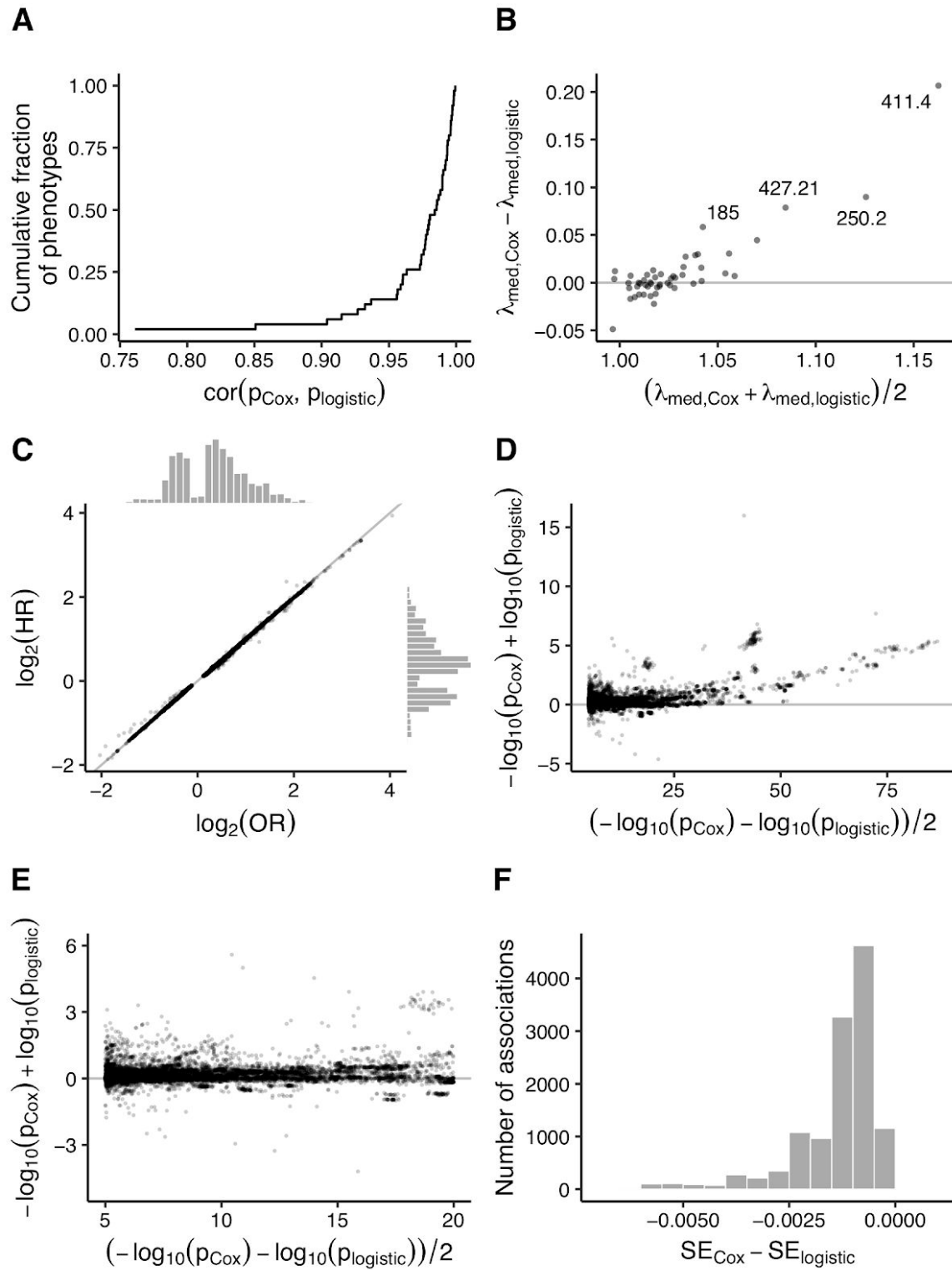
Principal components of genetic ancestry, calculated using SNPRelate. **(A)** Eigenvalues for the top 10 principal components. **(B)** Scatterplot of PC2 vs. PC1, where each point corresponds to a subject in the cohort.

Figure S2



Manhattan plots of GWAS results using Cox and logistic regression for four phenotypes (phecode in parentheses). For each phenotype, only associations having $\text{mean}(-\log_{10}(P)) \geq 2$ are shown. Dark green lines correspond to $P = 5 \cdot 10^{-8}$ and light green lines correspond to $P = 10^{-5}$.

Figure S3



Comparing the results of GWAS based on Cox regression or logistic regression for 50 selected phenotypes. **(A)** Empirical cumulative distribution function of the Pearson correlation between p-values for each phenotype. **(B)** Mean-difference plot of the genomic inflation factor. Each point corresponds to a phenotype. Phenotypes with a difference > 0.05 are labeled with the corresponding phecode (411.4: coronary atherosclerosis, 250.2: type 2 diabetes, 427.21: atrial fibrillation, 185: prostate cancer). **(C)** Scatterplot and marginal histograms of $\log_2(\text{effect size})$ for Cox and logistic regression, calculated as $\log_2(e^\beta)$, where β is the coefficient for genotype. HR refers to hazard ratio from Cox regression, OR refers to odds ratio from logistic regression. In panels C-E, each point corresponds to a SNP-phenotype pair. Panels C-F include only those associations for which $P \leq 10^{-5}$ for either Cox or logistic regression. **(D)** and **(E)** Mean-difference plots of $-\log_{10}(P)$ with different x- and y-axes. **(F)** Histogram of the difference in standard error of the coefficient estimate.