Figure e-1. The strategy of data analysis and sample size.

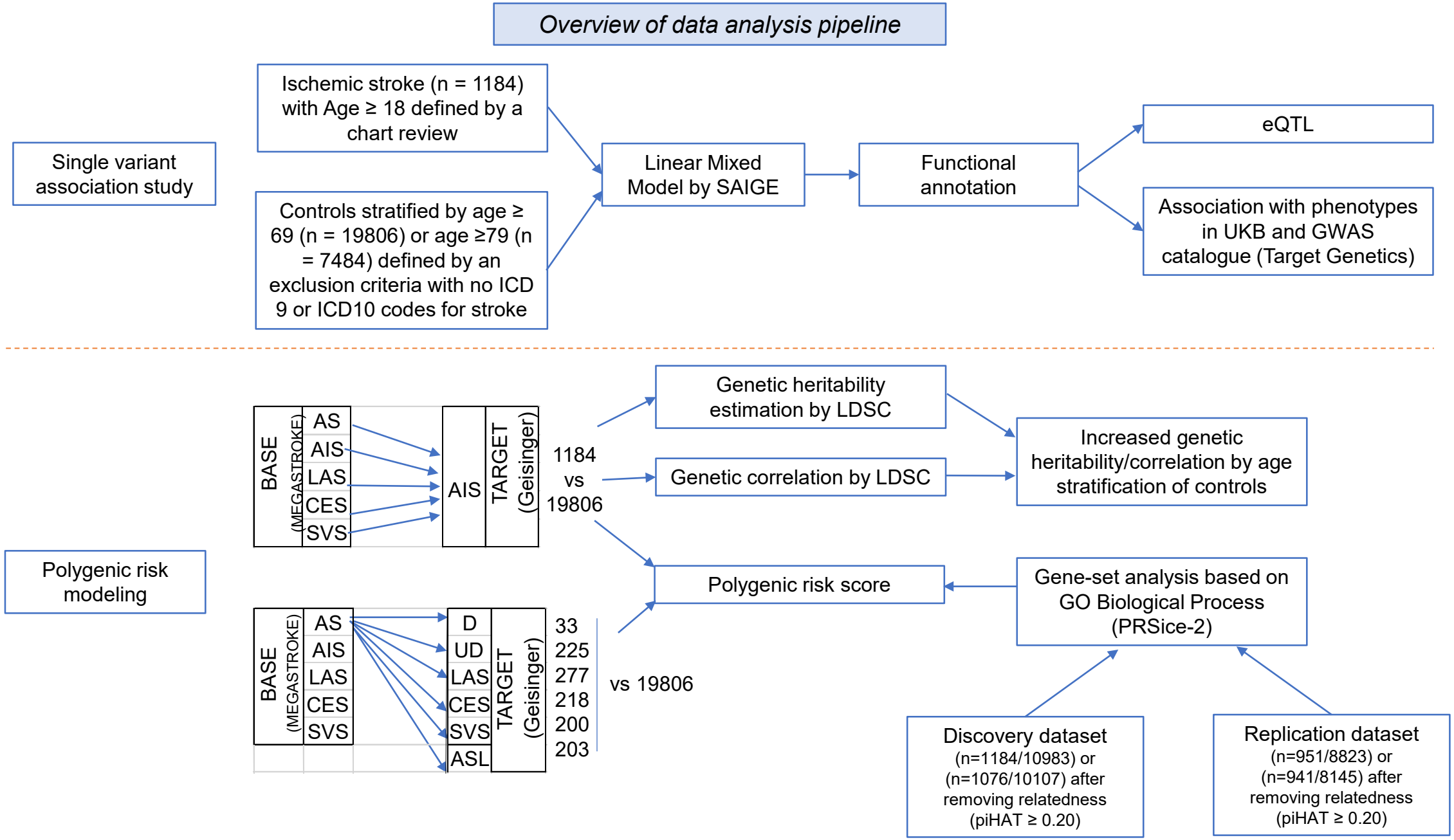


Figure e-2. Principal Component Analysis (PCA) was performed on genotype data of autosomal common SNPs from Geisinger PHASE I (A) and II (B) cohorts. PCA of all Phase I or II subjects against 1092 individuals from 1000 Genome Phase 1 data (AFR, n=246; AMR, n=181; ASN, n=286; EUR, n=379). All Geisinger subjects (purple) selected for this study including cases and controls in Discovery and Replication datasets were closely clustered with 1000 Genome samples with European Ancestry. Unknown subjects were from the Geisinger sample. 60K and 30K represent Geisinger PHASE I (A) and II (B) cohorts.

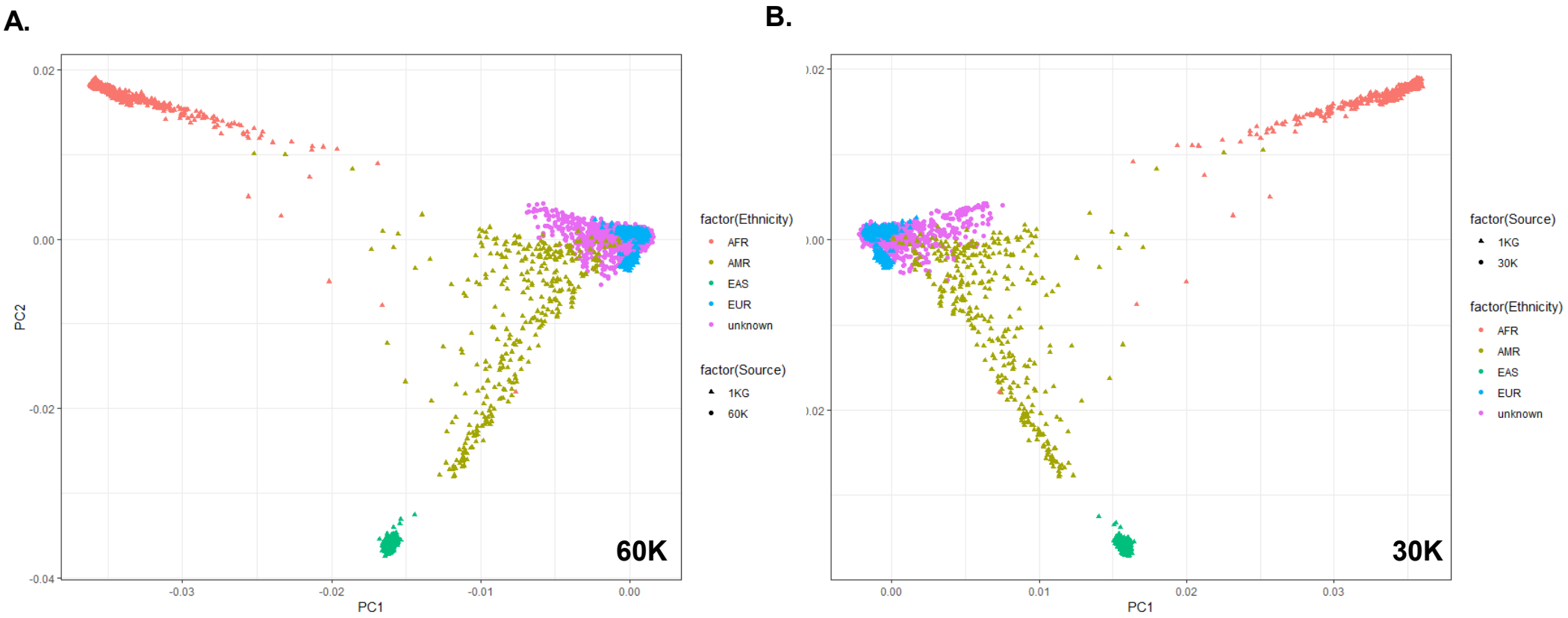
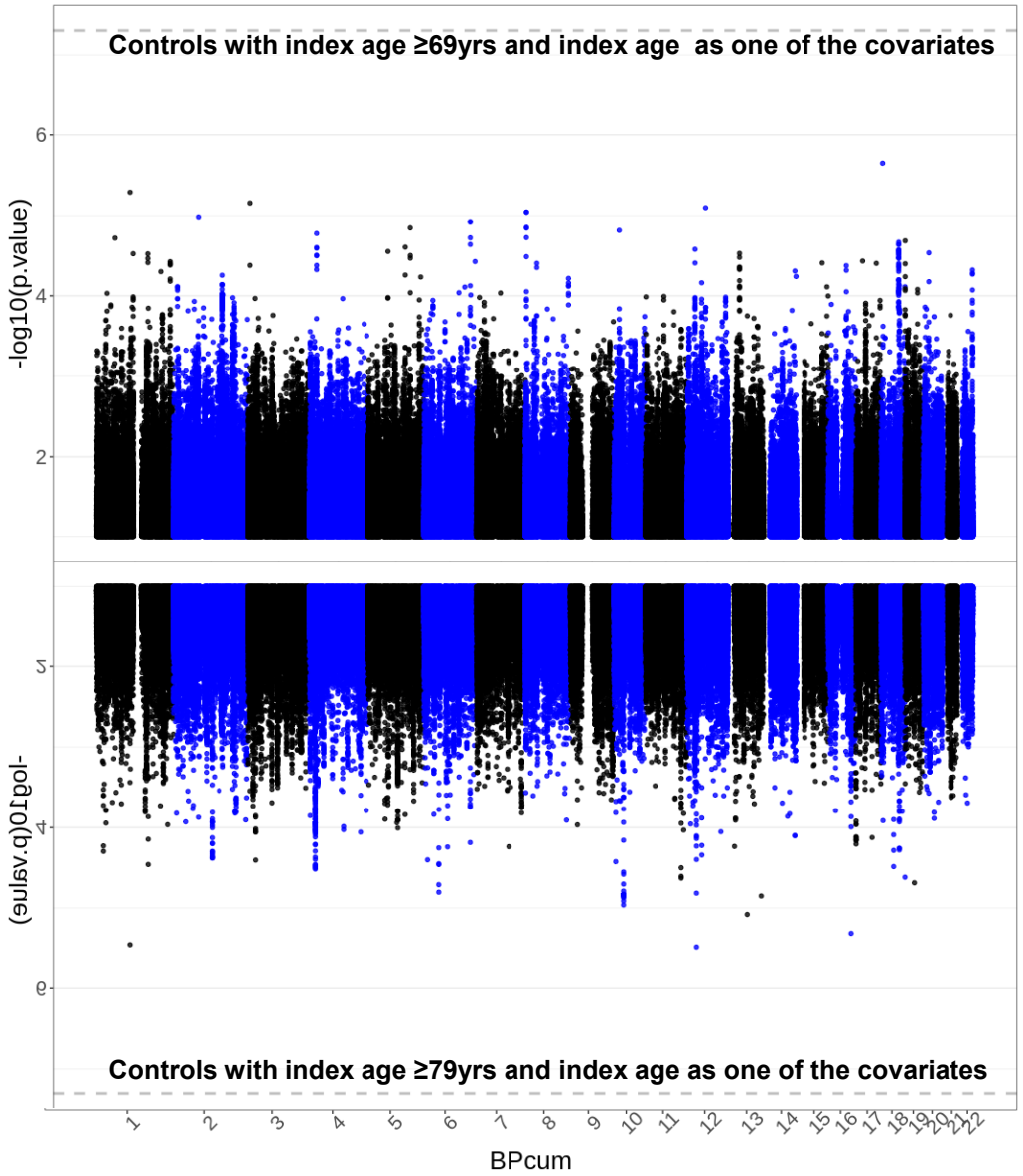
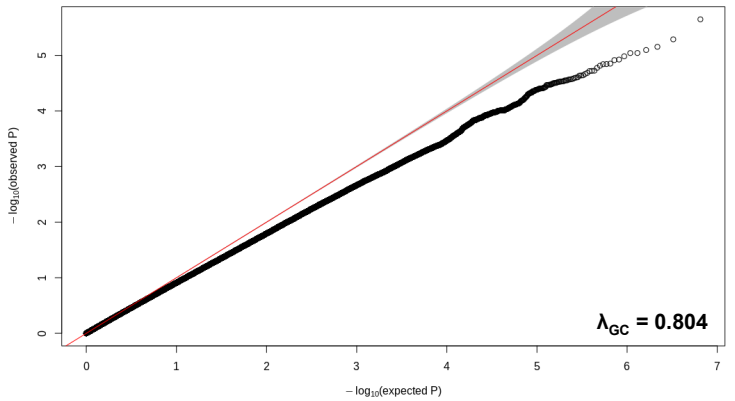


Figure e-3A. Manhattan plots and QQ plots for GWAS result associated with Geisinger ischemic stroke. Case-control design by selecting Geisinger patients with index age ≥69 or ≥79 and without ICD-9/10 codes for stroke as controls. **A.** Sex, Index age, and top five principal components were selected as covariates in the linear mix model using SAIGE with saddle-point approximation to adjust p-value for the association. Index age caused systematic deflation under the linear mix modeling.





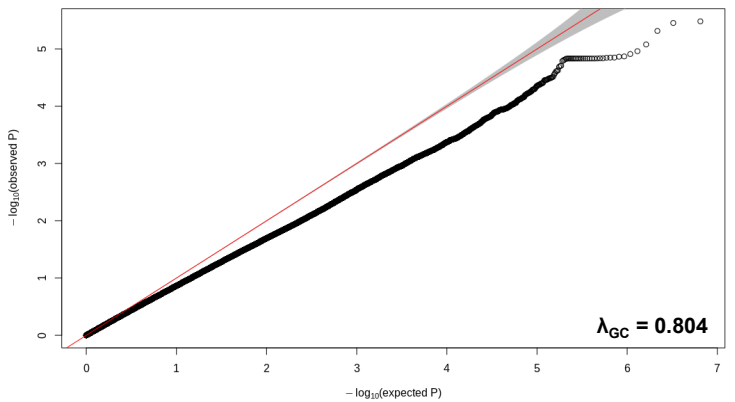
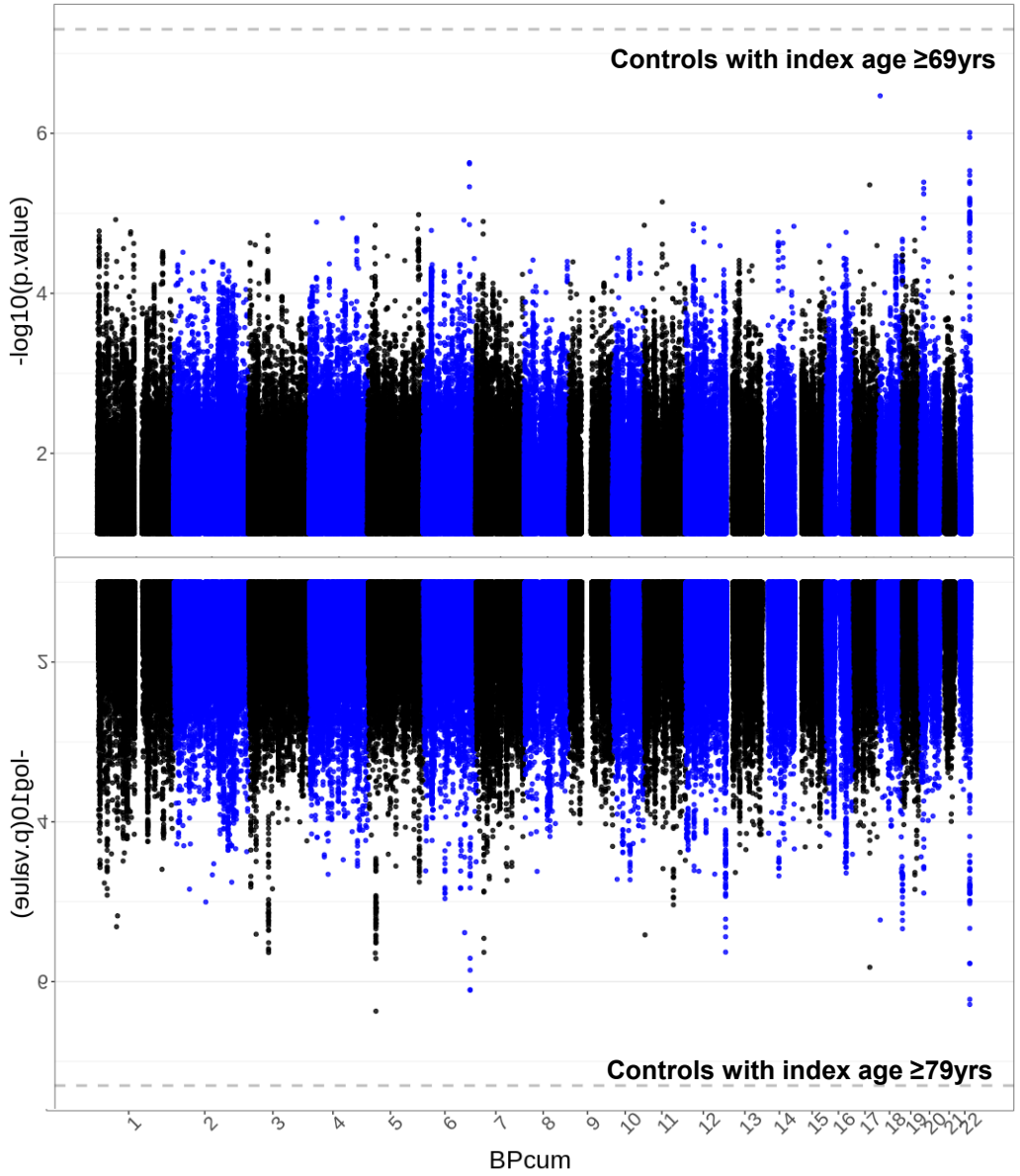
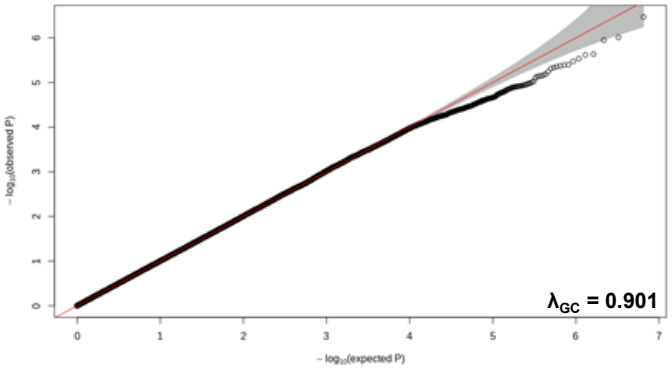


Figure e-3B. Manhattan plots and QQ plots for GWAS result associated with Geisinger ischemic stroke. Case-control design by selecting Geisinger patients with index age ≥69 or ≥79 and without ICD-9/10 codes for stroke as controls. **B.** Sex and the five major principal components were selected as covariates in a linear mix model using SAIGE with saddle-point approximation to adjust p-value for the association. The genomic inflation factor, λ_{GC}, equals to 0.901 and 0.907 for index age ≥69 or ≥79 respectively, suggesting no evidence for systematic inflation of genome-wide test statistics.





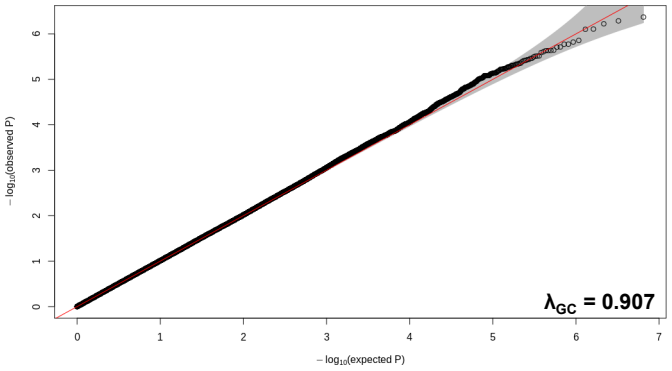


Figure e-4A. Evaluation of predictive power of PRS on Geisinger ischemic stroke. PRSs were derived from MEGASTROKE by PRSice-2 with10000 permutation tests. A. We first applied LD-clumping with an r^2 threshold of 0.1 to all SNPs, followed by p-value thresholding in the testing set. The results were derived from testing over a range of p-value thresholds and picking the thresholding that gave the best predictive performance. Nagelkerke pseudo-R^2 as shown in the y-axis, represents how much variation is explained by the model. the x-axis represents the threshold for a base p-value. P-value on the top of each bar represents the probability of non-zero regression coefficient with the *F* statistic hypothesis testing of the fit of the intercept-only(PRS excluded) model and PRS included model are equal'. The significance with p < 0.001 as the cutoff to show that the PRS included model fits the data better.

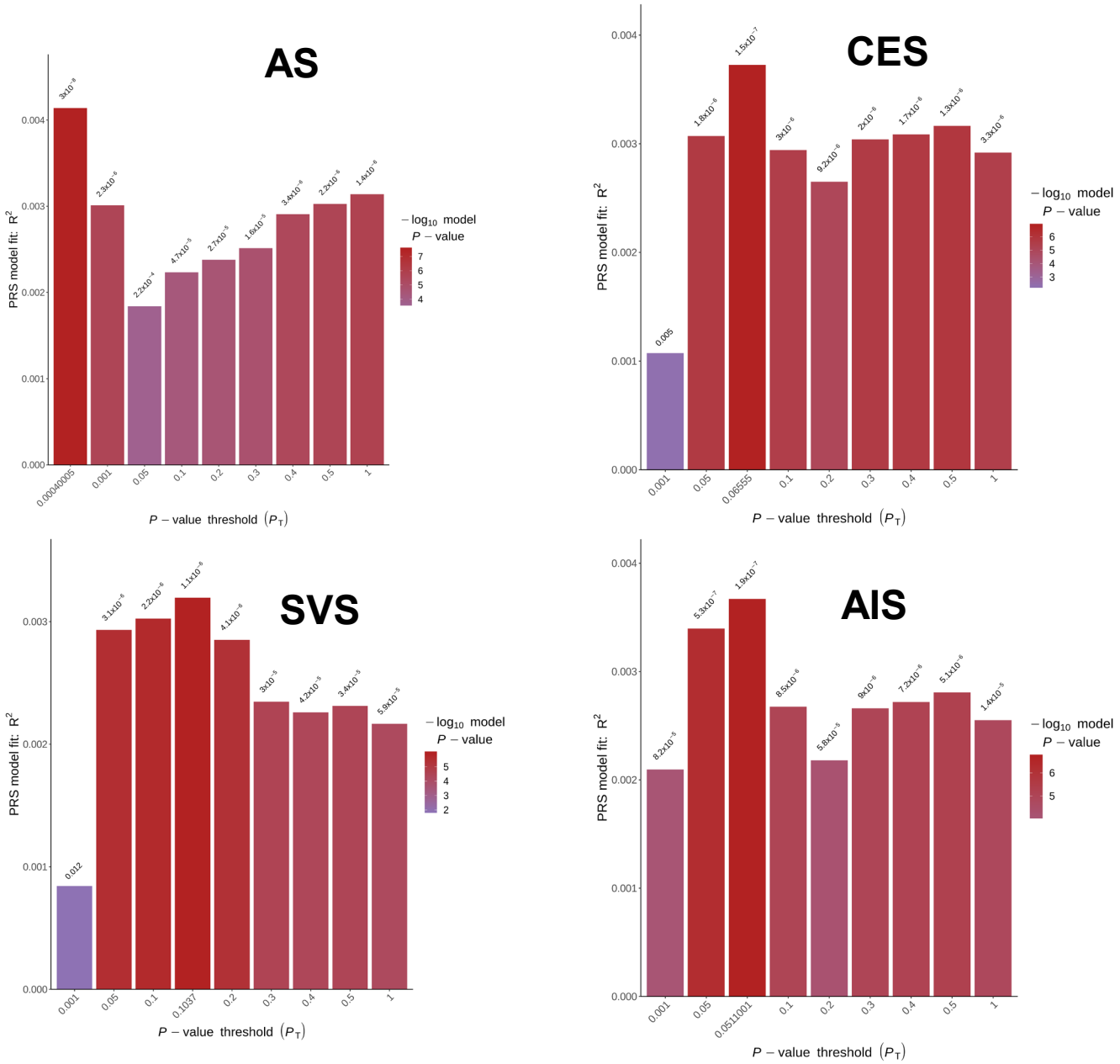
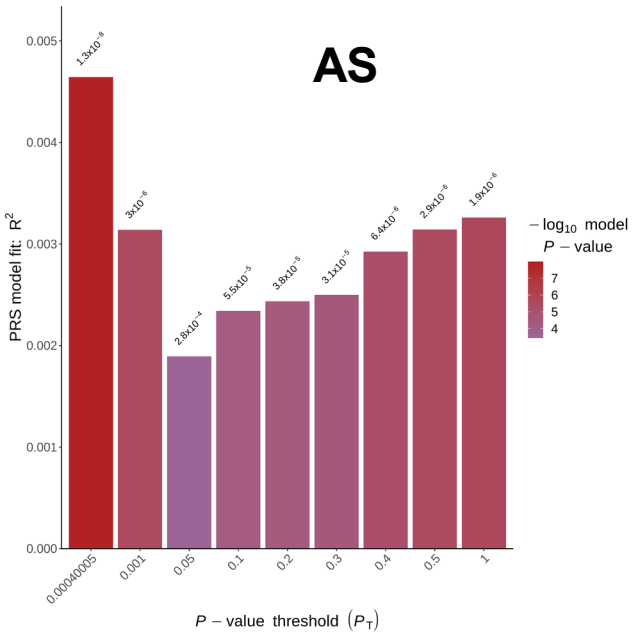
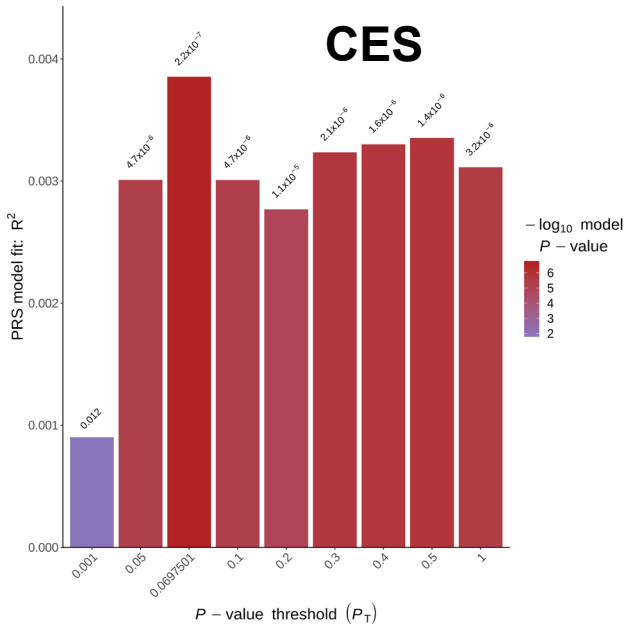
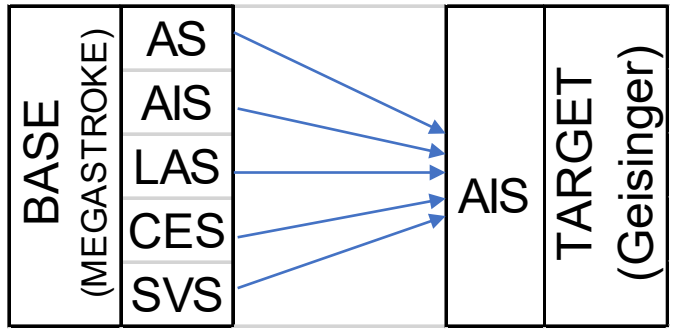
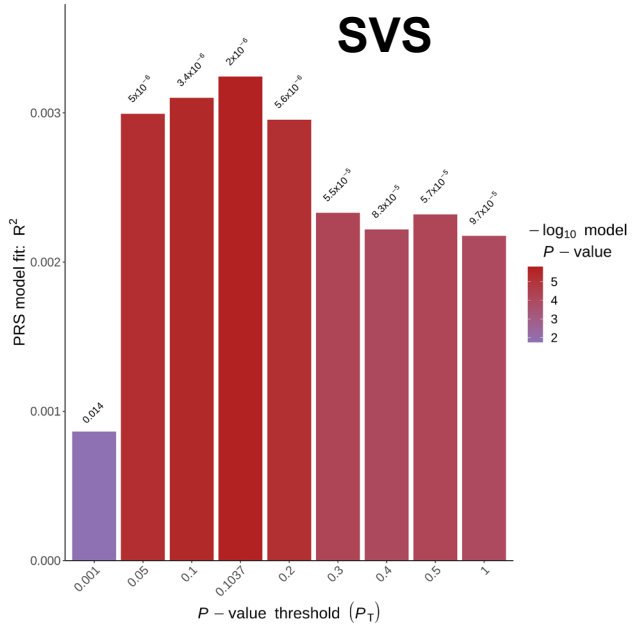


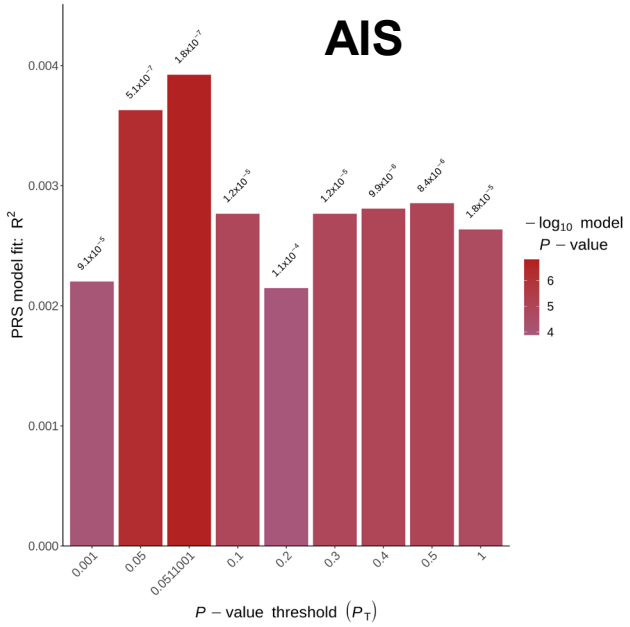
Figure e-4B. Evaluation of predictive power of PRS on Geisinger ischemic stroke. PRSs were derived from MEGASTROKE by PRSice-2 with10000 permutation tests. B. We remove related individuals in the Geisinger sample with paired PI_HAT < 0.2 and maintained the maximum number of cases. We ended up with 1167 cases and 17271 controls. A subgroup analysis was performed by removing individuals from each pair of related individuals (2nd –degree or closer; PI_HAT 0.2)) confirming similar results within these subpopulations.











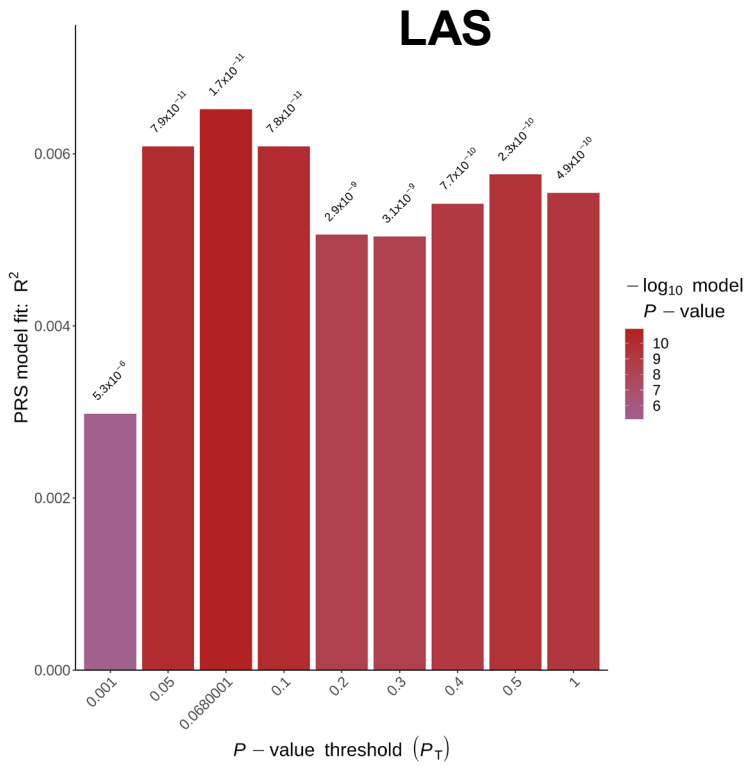


Figure e-5A. PRS derived from common genetic variants with lower minor allele frequency (MAF<0.025) provided the best-fit modeling for the ischemic stroke when PRS was constructed based on the summary statistics of TOAST subtypes such as LAS, SVS, and CES. The high-resolution plots were used to select the consistent cut-off value for MAF threshold for PRS construction and the gene-set analyses. **A.** Discovery Cases (n=1184) vs Controls (n=19806).

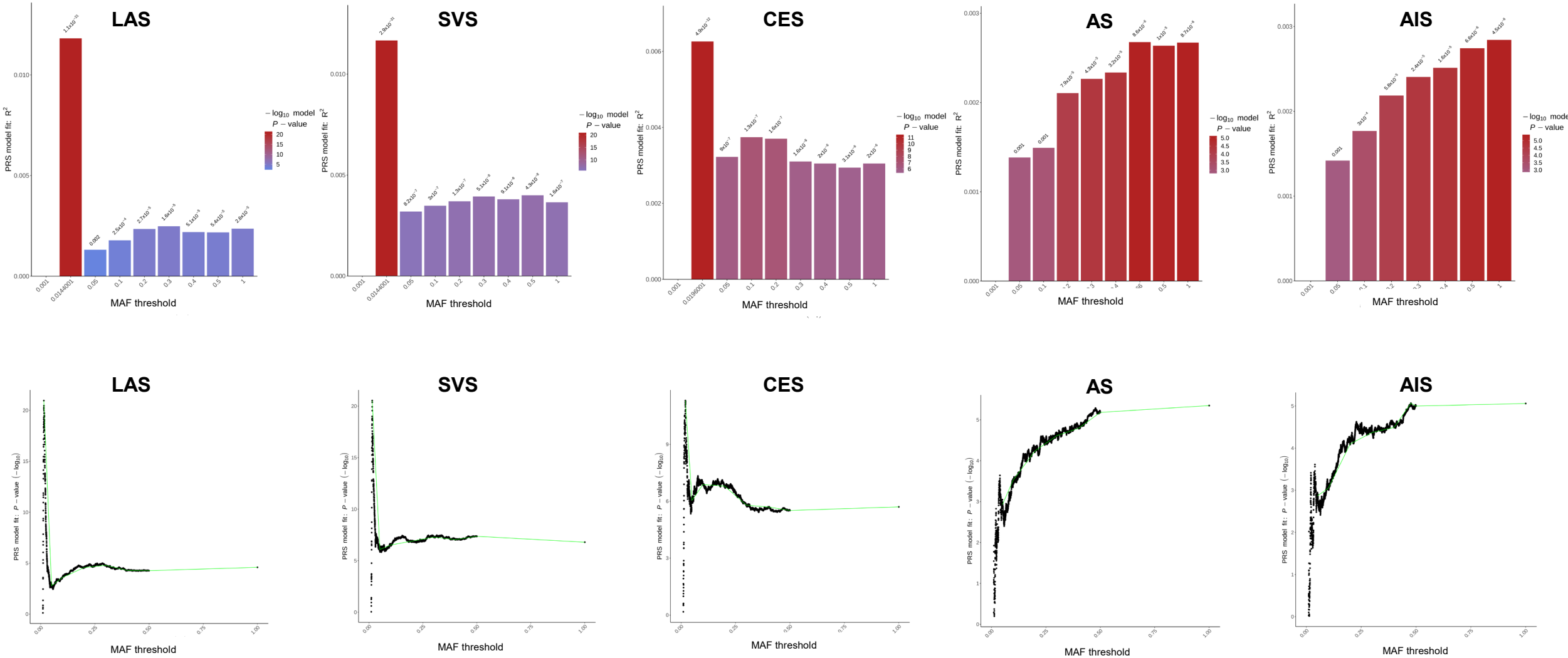


Figure e-5B. PRS derived from common genetic variants with lower minor allele frequency (MAF<0.025) provided the best-fit modeling for the ischemic stroke when PRS was constructed based on the summary statistics of TOAST subtypes such as LAS, SVS, and CES. The high-resolution plots were used to select the consistent cut-off value for MAF threshold for PRS construction and the gene-set analyses. **B.** Replication Cases (n=951) vs Controls (n=19806).

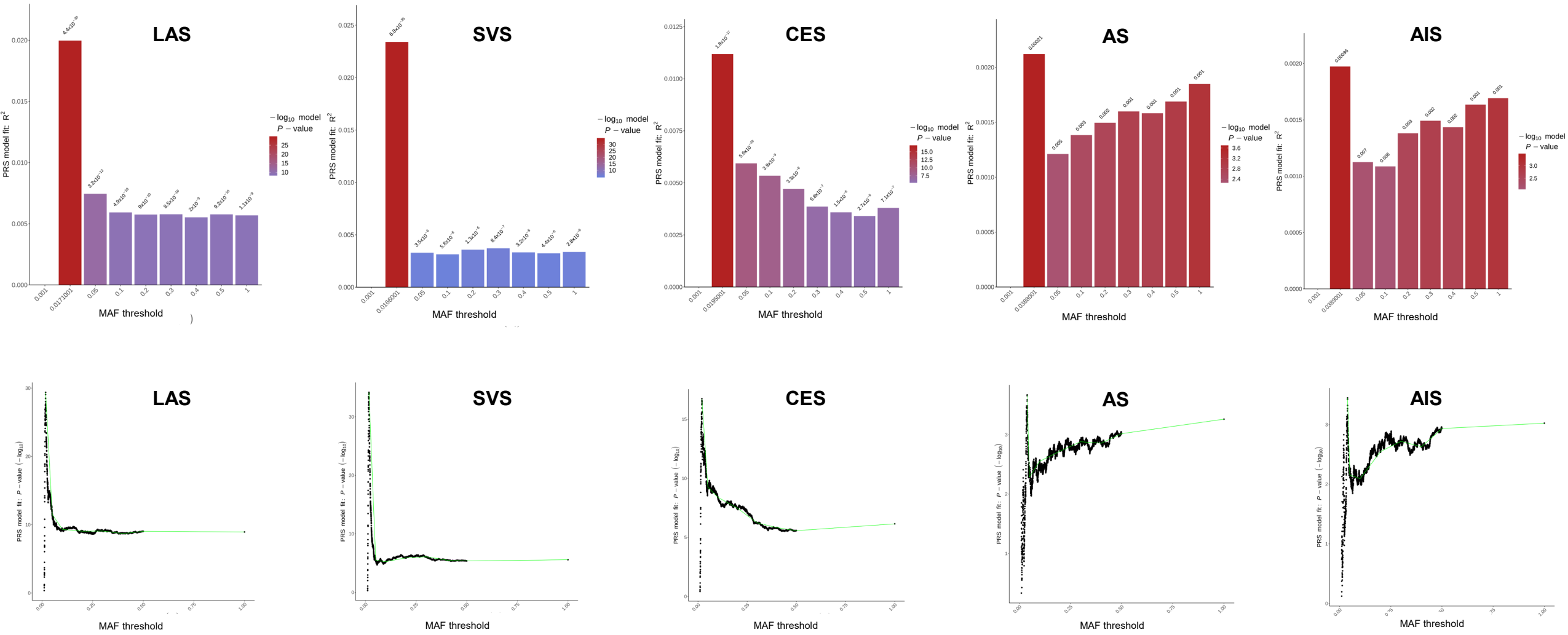


Figure e-6A. Gene-sets analyses illustrated the top five pathways enriched for ischemic stroke (controls with index age ≥ 69yrs) after meta-analysis of Discovery dataset (n=1184/10983) and Replication dataset (n=951/8823) when the PRS was constructed based on each of the five summary statistics of MEGASTROKE.. Control samples with index age ≥ 69 were randomly split into discovery and replication datasets with the same case:control ratio (0.108). The sex and five major PCs were included as covariates in the logistic regression model. The meta-analysis was conducted by Metal with weighted effect size (coefficient) estimates using the inverse of the corresponding standard errors. The global genes were selected as a universal background for gene-sets analyses and the mapping file was "Homo_sapiens.GRCh37.87.gtf". PRSs derived from gene-sets defined by Gene Ontology Biological Process were calculated to test their association with an ischemic stroke under two MAF thresholds (MAF <0.025 or <1) which represents low-frequency common variants or all variants accordingly. 7350 pathways and their related gene sets were defined by Molecular Signatures Database ("msigdb_v7.0_GMTs/c5.bp.v7.0.symbols.gmt"). **A. The dot plot.** PRS derived from genetic variants with relatively lower minor allele frequency (MAF) provided the best-fit modeling for the ischemic stroke (red dots) when PRS was constructed based on the summary statistics of TOAST subtypes such as LAS, SVS, and CES as compared to PRS constructed based on the summary statistics of AS or AIS. Both discovery dataset and replication dataset showed the same profile. The size of the dots represents the R^2, a measure of the proportion of the variance explained by the model. The y-axis represents the significance of the model fit. The total number of variants included in the analysis under two MAF thresholds were also listed on the top. By comparing the result derived from nonrelated individuals with the original datasets, we did not observe any inflated R^2 and p value.

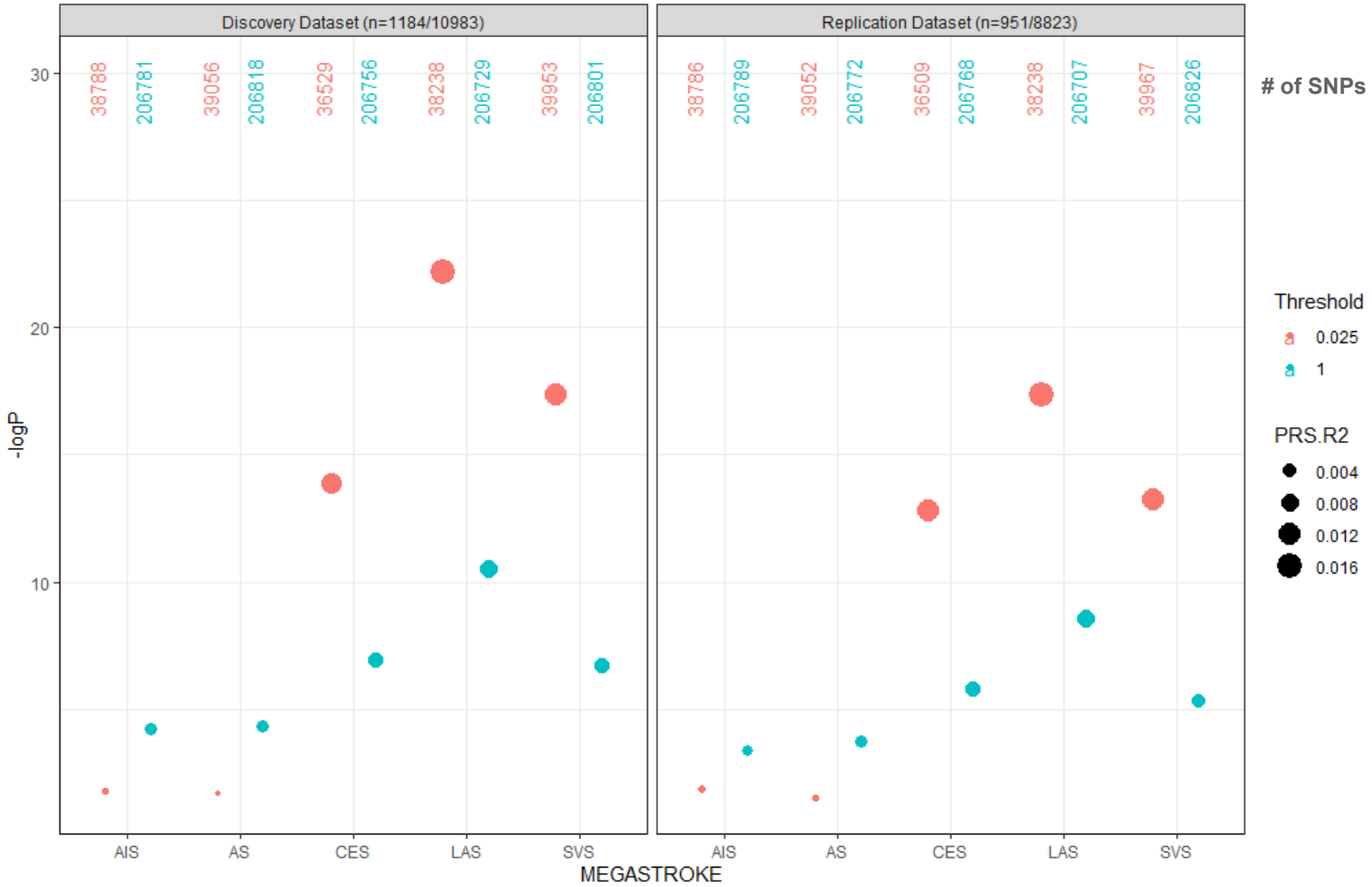


Figure e-6B. Gene-sets analyses illustrated the top five pathways enriched for ischemic stroke (controls with index age ≥ 69yrs) after meta-analysis of Discovery dataset (n=1184/10983) and Replication dataset (n=951/8823) when the PRS was constructed based on each of the five summary statistics of MEGASTROKE..

B. The bar plot. Exploration of the top five pathways enriched from PRS gene-sets analyses after the meta-analysis of discovery and replication datasets using each summary statistics from MEGASTROKE to construct PRS under two levels of MAF thresholding (y-axis). The red or blue bar represents the ischemic stroke has a negative or positive association with the corresponding biological process according to the direction of the coefficient respectively. All the p-values in the x-axis were present as raw. The pathways with adjusted p value not surviving Bonferroni correction ($-\log_{10}(P) < 5.17$) were labeled as '*'.

Figure e-7A. The scatter plots demonstrated the distribution of association p value for gene-sets analyses to determine any inflated p value for the top pathways when including related individuals in the discovery and replication datasets. Two smooth methods, linear model (green line with 95% CI) with formula = y ~ x and generalized additive model (red line with 95% CI) with formula = y ~ s(x, bs = "cs"), were selected to show the overall trend of correlation for the association p value across 7350 pathways between the removerelatedness (x axis) and the original without removing related individuals (y axis) when including all genetic variants (**Figure e-7A**) or only common variants with low allele frequency(**Figure e-7B**) for the PRS constructed from five summary statisics (AIS, AS, LAS, CES, and SVS). Top and bottom rows represent data from the discovery and replication datasets, respectively.

MAF < 1

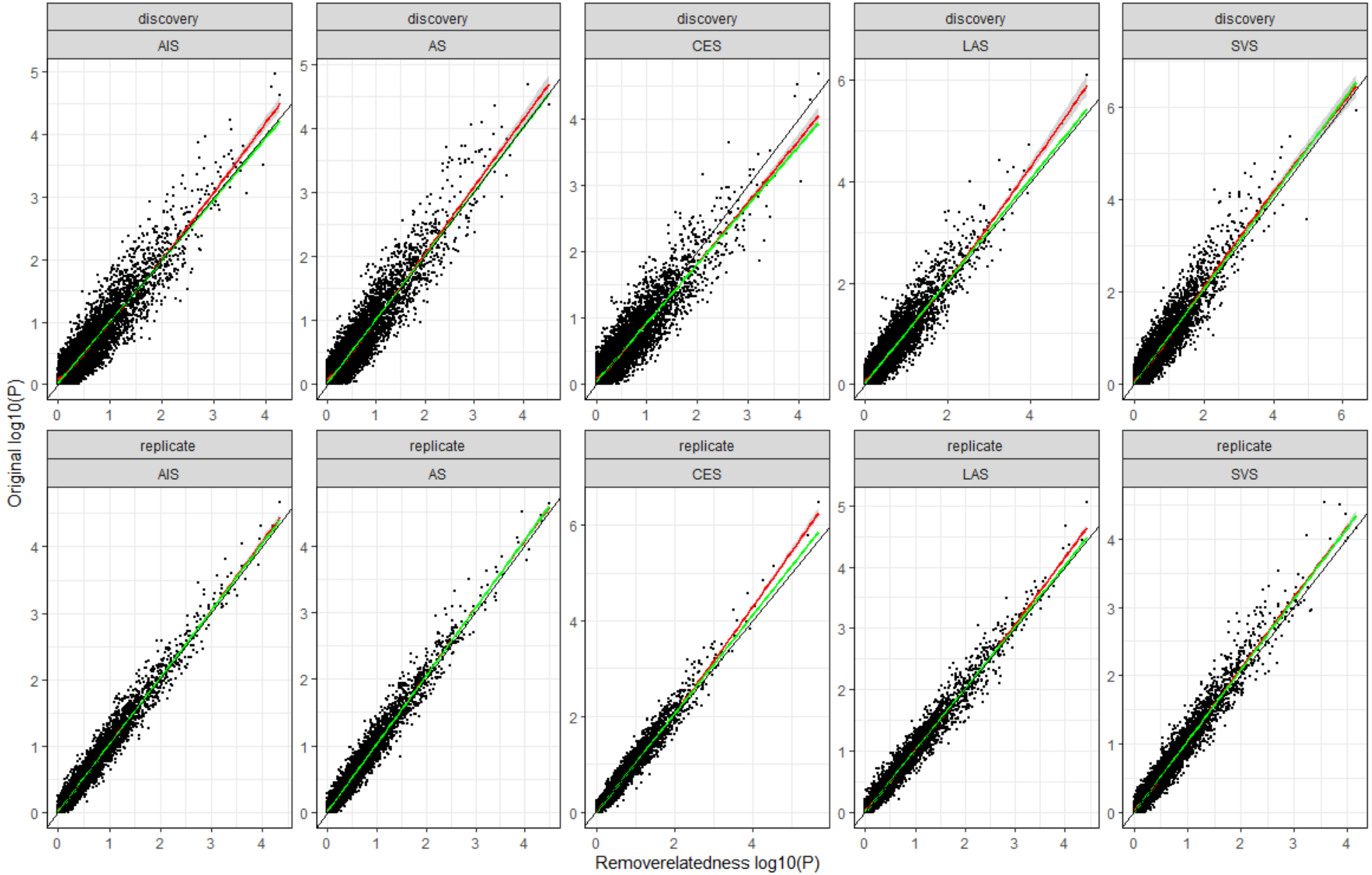


Figure e-7B. The scatter plots demonstrated the distribution of association p value for gene-sets analyses to determine any inflated p value for the top pathways when including related individuals in the discovery and replication datasets. Two smooth methods, linear model (green line with 95% CI) with formula = y ~ x and generalized additive model (red line with 95% CI) with formula = y ~ s(x, bs = "cs"), were selected to show the overall trend of correlation for the association p value across 7350 pathways between the removerelatedness (x axis) and the original without removing related individuals (y axis) when including all genetic variants (**Figure e-7A**) or only common variants with low allele frequency(**Figure e-7B**) for the PRS constructed from five summary statisics (AIS, AS, LAS, CES, and SVS). Top and bottom rows represent data from the discovery and replication datasets, respectively.

MAF < 0.025

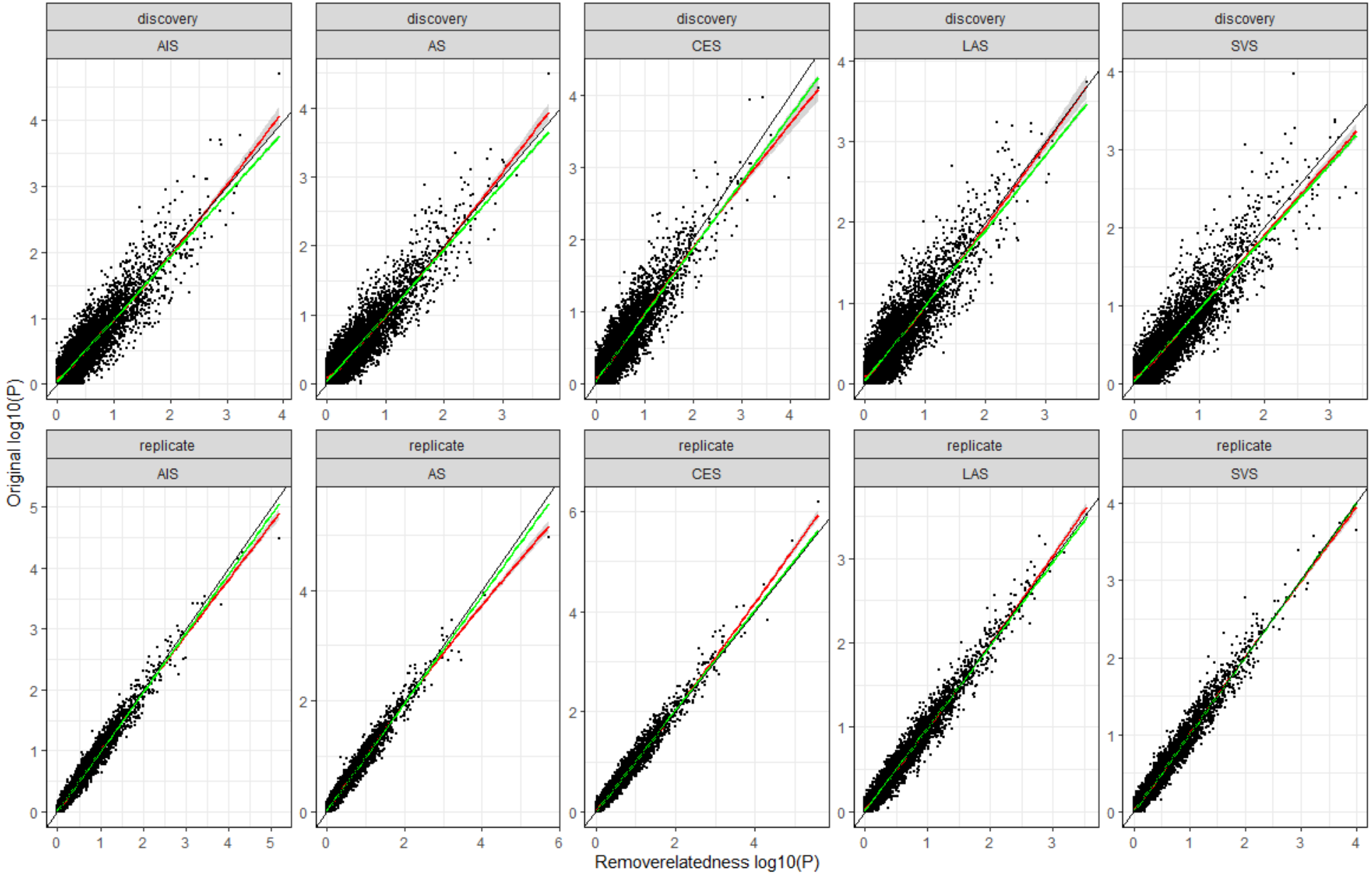


Figure e-8. The performance of PRS derived from five MEGASTROKE summary statistics in prediction of ischemic stroke in the testing dataset.
We showed that the performance of the prediction of ischemic stroke in the testing dataset by the metric, AUC-ROC (Area Under The Curve for Receiver Operating Characteristics), when the two levels of controls (age ≥ 69 or age ≥ 79) and two levels of MAF (<0.025 and <1) were considered. The DeLong's test was conducted to determine the statistical diffencence between AUCs of the base model and the model with additional feature(s), such as normalized PRS. Mod was the base logistic regression model.
Mod <- glm(ischemic_stroke ~ PT_SEX + PC1 + PC2 + PC3 + PC4 + PC5, family = binomial(link = "logit"), data = data). In comparing with PRS derived from AIS or AS of the MEGASTROKE summary statistics with MAF<1, PRS calculated from LAS, SVS, and CES gave the highest prediction power when SNPs with MAF<0.025 used in the discovery dataset. Thus we applied the corresponding 'best fit' model derived from the discovery dataset to the testing dataset to predict IS. We observed a limited but significant improvement of PRS over the base model when using the summary statistics of GWAS derived from the IS subtypes.

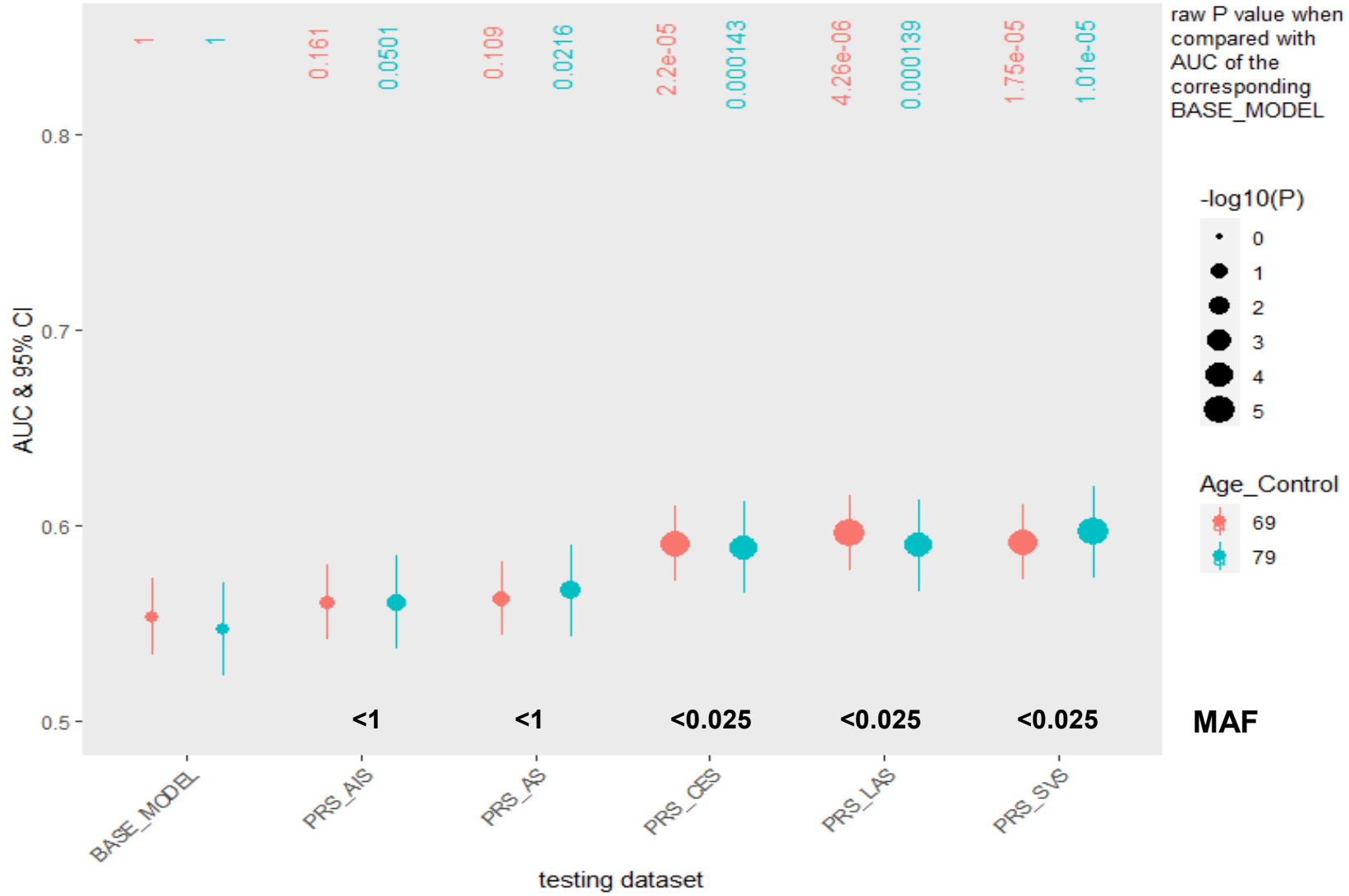


Table e-1. Summary of the setup of the sensitivity analysis by simulation of the same number of controls.
We simulated the same number of controls as to the corresponding controls ≥ 69 or ≥ 79 by a random selection from controls ≥ 59 to determine this augmented predictive power, if any, was largely due to natural selection in aged non-stroke individuals but not due to the change in case:control ratio.

Group	Case		Control						Data type
			>=59yrs		>=69yrs		>=79yrs		
Sex	n	index age	n	index age	n	index age	n	index age	
Male	599	68.8(12.4)	16891	71.3(8.35)	8932	77.8(5.87)	3281	84.5(3.14)	Original
					8825	71.3(8.30)	3363	71.2(8.34)	Random
Female	585	69.6(13.9)	21040	71.2(8.65)	10874	78.7(6.12)	4203	84.8(3.25)	Original
					10981	71.3(8.65)	4121	71.6(8.64	Random
All	1184	69.2(13.2)	37931	71.3(8.51)	19806	78.0(6.01)	7484	84.7(3.21)	Original
						71.3(8.49)		71.3(8.56)	Random

Table e-2. Summary of the top variants associated with ischemic stroke and their pleiotropic effect on stroke-related phenotypes or risk factors. Summary statistics of top variants with p-value < 1×10^{-5} were LD clumped (--clump-p1 1×10^{-5}, -clump-r^2 0.5) and listed here. POS was a genomic coordinate based on the hg19 version. All the p values were raw without correction for multiple testing. PheWAS summary statistics of stroke-related phenotypes obtained from the UK BIOBANK in association with the top loci (https://genetics.opentargets.org/) were listed here. The frequency of risk alleles from the top associated loci was increased in controls with 79yrs or higher, suggesting the protective alleles were enriched in the senior non-stroke population. The risk alleles from the top loci also were associated with increased risk for stroke-related phenotypes or risk factors, suggesting the potential pleiotropy of these variants.

SNP information						≥79yrs					≥69yrs					Case	≥79yrs	≥69yrs	Phewas from UK BIOBANK				
CHR	SNP_ID	POS	A1/A2	Gene	eQTL	AF(A1)	BETA	SE	p.value	p.value.NA	AF(A1)	BETA	SE	p.value	p.value.NA	AF.Cases	AF.Controls		Study	Phenotype	P	BETA	OR
5	rs62349604	24569475	G/A	CDH10	C5orf17	0.141	0.321	0.063	4.272E-07	3.451E-07	0.144	0.262	0.060	1.409E-05	1.258E-05	0.176	0.136	0.142	SAIGE_426	Cardiac conduction disorders	0.0017	0.074	1.08
22	rs41280521	49042116	T/C	SLC7A11		0.053	0.507	0.101	5.180E-07	3.014E-07	0.053	0.475	0.097	9.766E-07	4.826E-07	0.075	0.050	0.051	GCST005065	Cholesterol, total	0.0026	0.240	
6	rs9384568	150946876	T/C	PLEKHG1		0.179	0.284	0.058	7.830E-07	6.805E-07	0.179	0.260	0.055	2.408E-06	2.063E-06	0.216	0.173	0.177	SAIGE_747_12	Valvular heart disease/ heart chambers	0.0083	0.321	1.38
17	rs77711120	52052187	C/A	~KIF2B		0.022	0.760	0.158	1.515E-06	5.566E-07	0.022	0.699	0.152	4.411E-06	1.457E-06	0.035	0.019	0.021	GCST004422	Vascular endothelial growth factor levels	0.0099	0.541	
5	rs77523535	22717800	T/A	CDH12		0.026	0.679	0.144	2.279E-06	1.038E-06	0.027	0.565	0.135	2.686E-05	1.550E-05	0.041	0.024	0.026					
3	rs11130939	63518063	T/G	SYNPR		0.17	0.279	0.059	2.317E-06	2.078E-06	0.173	0.225	0.056	5.781E-05	5.471E-05	0.204	0.164	0.171	NEALE2_20113_2	Stroke | Illnesses of adopted mother	0.0043	0.263	1.3
7	rs10255575	24922835	T/A	OSBPL3	OSBPL3, GSDME	0.316	0.226	0.048	2.331E-06	2.271E-06	0.317	0.199	0.046	1.264E-05	1.227E-05	0.357	0.310	0.314	NEALE2_40001_I251	ICD10: I25.1 Atherosclerotic heart disease | Underlying (primary) cause of death	0.013	0.205	1.23
12	rs12299194	129264838	C/T	~SLC15A4	SLC15A4, GLT1D1	0.092	0.369	0.078	2.341E-06	1.894E-06	0.097	0.240	0.072	7.984E-04	7.769E-04	0.117	0.088	0.096	NEALE2_20002_1077	heart arrhythmia | Non-cancer illness code, self-reported	0.00023	0.182	1.2
11	rs77497298	1521393	A/G	MOB2	KRTAP5-AS1	0.038	0.547	0.118	3.853E-06	2.387E-06	0.038	0.493	0.114	1.412E-05	8.958E-06	0.055	0.035	0.037	NEALE2_20002_1065	hypertension | Non-cancer illness code, self-reported	0.000061	0.047	1.05
3	rs822762	22945204	T/C	~UBE2E2	UBE2E2	0.463	-0.199	0.043	3.928E-06	3.892E-06	0.461	-0.175	0.042	2.481E-05	2.447E-05	0.419	0.470	0.463	NEALE2_2335	Chest pain or discomfort	0.0014	-0.021	0.979
6	rs12189813	132927502	C/G	~VNN1	VNN1	0.366	-0.207	0.045	4.107E-06	4.041E-06	0.366	-0.188	0.043	1.213E-05	1.185E-05	0.323	0.373	0.368	NEALE2_20002_1074	angina | Non-cancer illness code, self-reported	0.001	-0.045	0.956
18	rs11081572	77657319	A/C	KCNG2	FQLC1,HSBP1L1	0.135	0.297	0.065	4.600E-06	4.083E-06	0.136	0.263	0.062	2.105E-05	1.890E-05	0.165	0.131	0.134					
1	rs77731861	60213024	A/T	FGGY		0.024	0.679	0.148	4.871E-06	2.496E-06	0.024	0.625	0.143	1.199E-05	5.608E-06	0.038	0.022	0.023	NEALE2_20003_1140888648	pravastatin | Treatment/medication code	0.00014	0.378	1.46
18	rs56120864	5142233	A/G	~ZBTB14		0.034	0.562	0.124	5.908E-06	3.795E-06	0.031	0.642	0.126	3.397E-07	9.051E-08	0.051	0.032	0.030	SAIGE_411_4	Coronary atherosclerosis	0.00075	0.087	1.09
1	rs10889400	63536077	C/T	~FOXD3		0.289	0.221	0.049	6.664E-06	6.384E-06	0.290	0.193	0.047	3.988E-05	3.846E-05	0.326	0.283	0.288	NEALE2_20002_1372	vasculitis | Non-cancer illness code, self-reported	0.0046	-0.351	0.704
11	rs538800	94239567	T/C	~MRE11	MRE11/GPR83	0.346	-0.206	0.046	9.143E-06	9.009E-06	0.340	-0.160	0.045	3.352E-04	3.322E-04	0.306	0.352	0.342	NEALE2_23127_raw	Trunk fat percentage	0.00007	0.071	
2	rs184559244	101753936	A/G	TBC1D8		0.016	0.818	0.185	9.921E-06	4.485E-06	0.018	0.520	0.161	1.261E-03	1.125E-03	0.027	0.014	0.018					