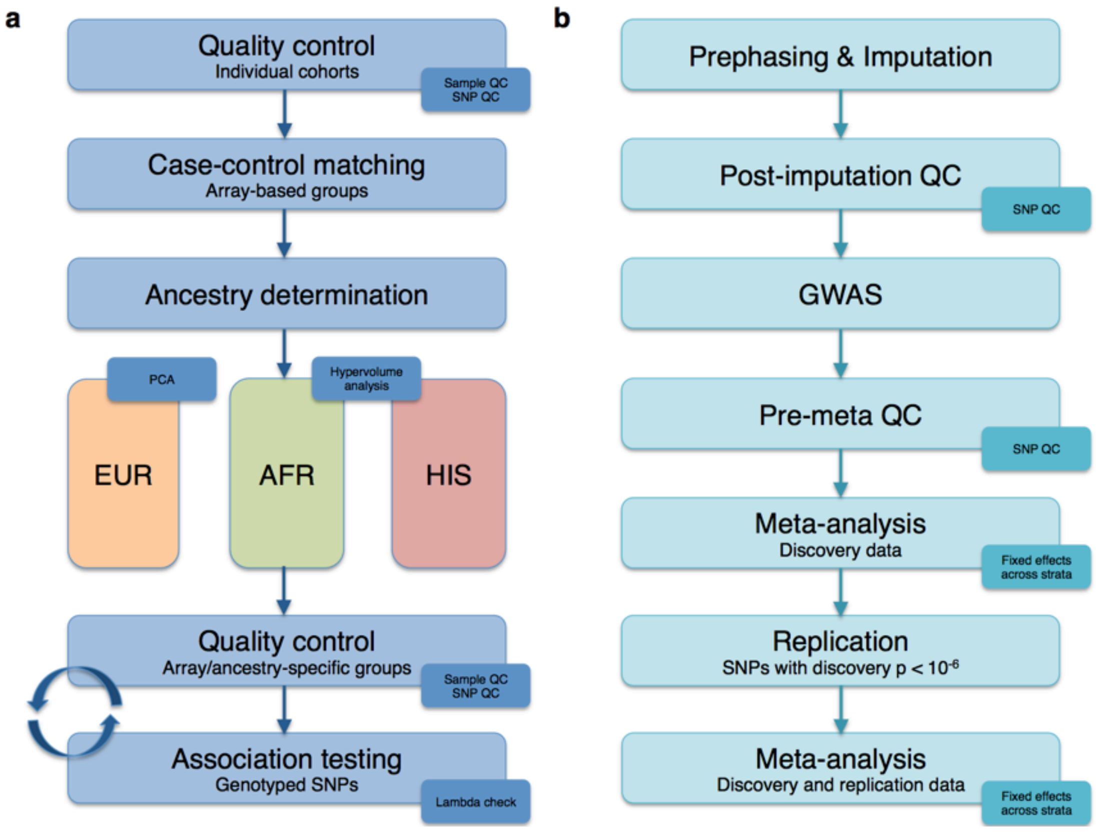
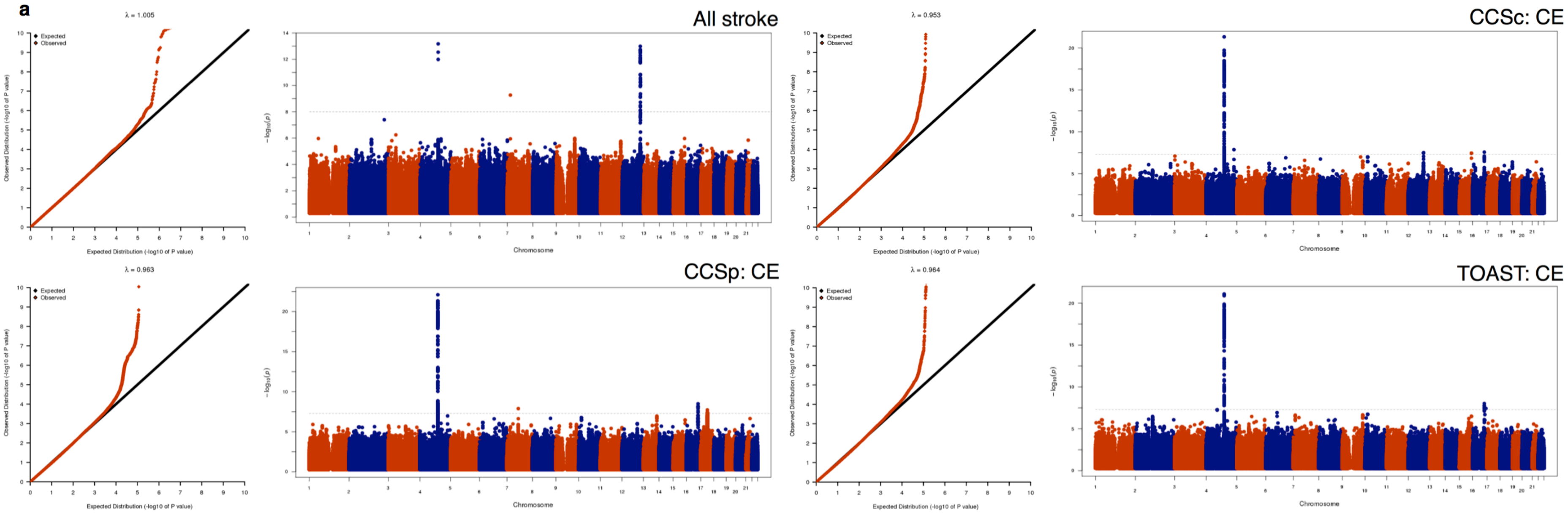
Table of Contents

I. Supplementary Figures



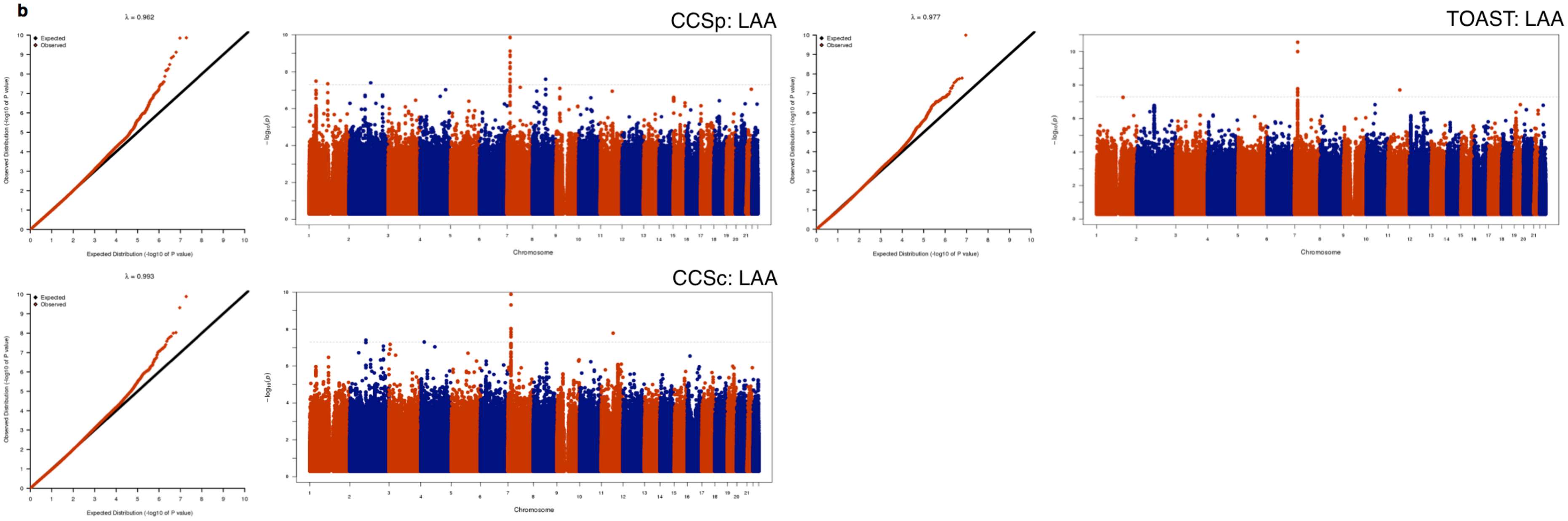
Supplementary Figure 1 | Analysis overview

Extensive quality control (a) was first applied to the data. Individual cohorts were cleaned (**Section 9.1**) and then case and control groups were pooled together based on genotyping platform (**Section 9.2**). Principal component analysis (PCA) and hyperellipsoid analysis were used to define mutually exclusive groups of European-ancestry, African-ancestry, and Hispanic samples (**Sections 9.3.1 – 9.3.4**). Array/ancestry specific groups were then cleaned further (**Section 9.4**), and association testing was used to evaluate effectiveness of QC (**Section 9.5**). QC and association testing were performed iteratively until genomic inflation was well behaved. (b) Post-QC, samples were prephased (**Section 10**) and imputed (**Section 11**); imputed genotypes were then cleaned (**Section 11**). We then performed a stage I GWAS in ischemic stroke and all available subtypes in each array/ancestry-specific stratum (**Section 12**). Summary-statistics were combined using inverse variance-weighted fixed effects meta-analysis and SNPs were selected for stage II replication (**Sections 13** and **14**). Stage I and stage II (replication) data were then combined in a final joint meta-analysis (**Section 15**).



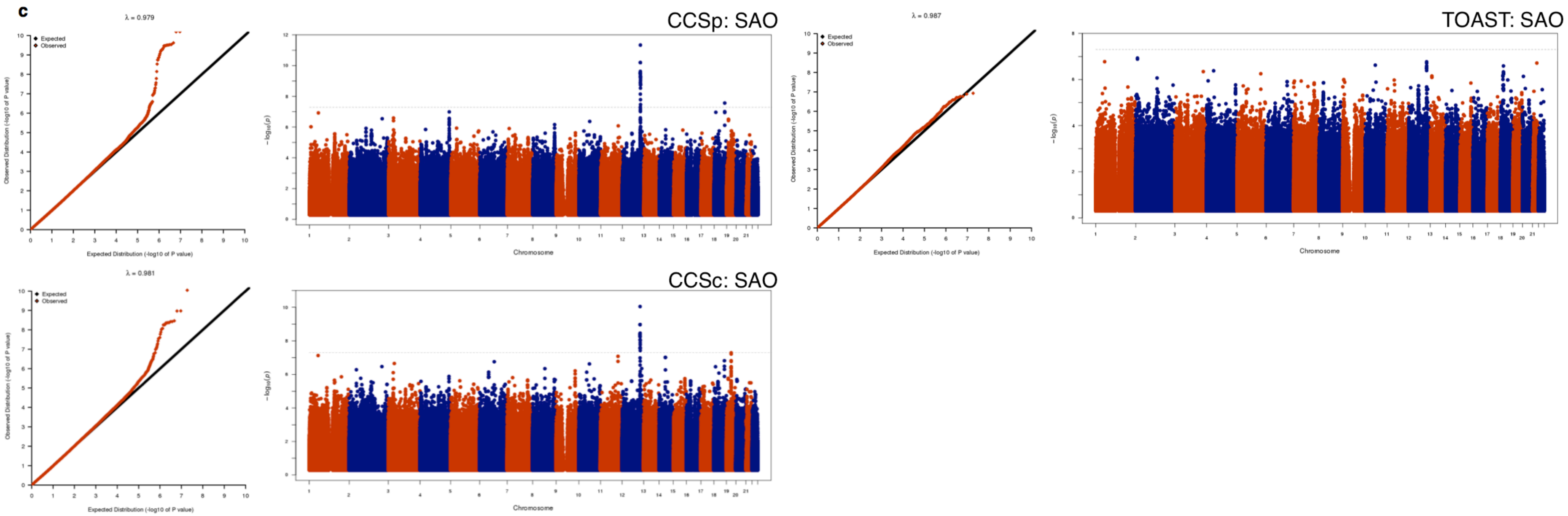
Supplementary Figure 2 | QQ and Manhattan plots of the stage I meta-analyses

Discovery meta-analysis was conducted in all stroke, and each of the CCS Causative (CCSc), CCS Phenotypic (CCSp), and TOAST subtypes (**Section 13.3**). (a) Manhattan and QQ plots for all stroke and cardioembolic (CE) stroke, as determined by CCSc, CCSp, and TOAST.



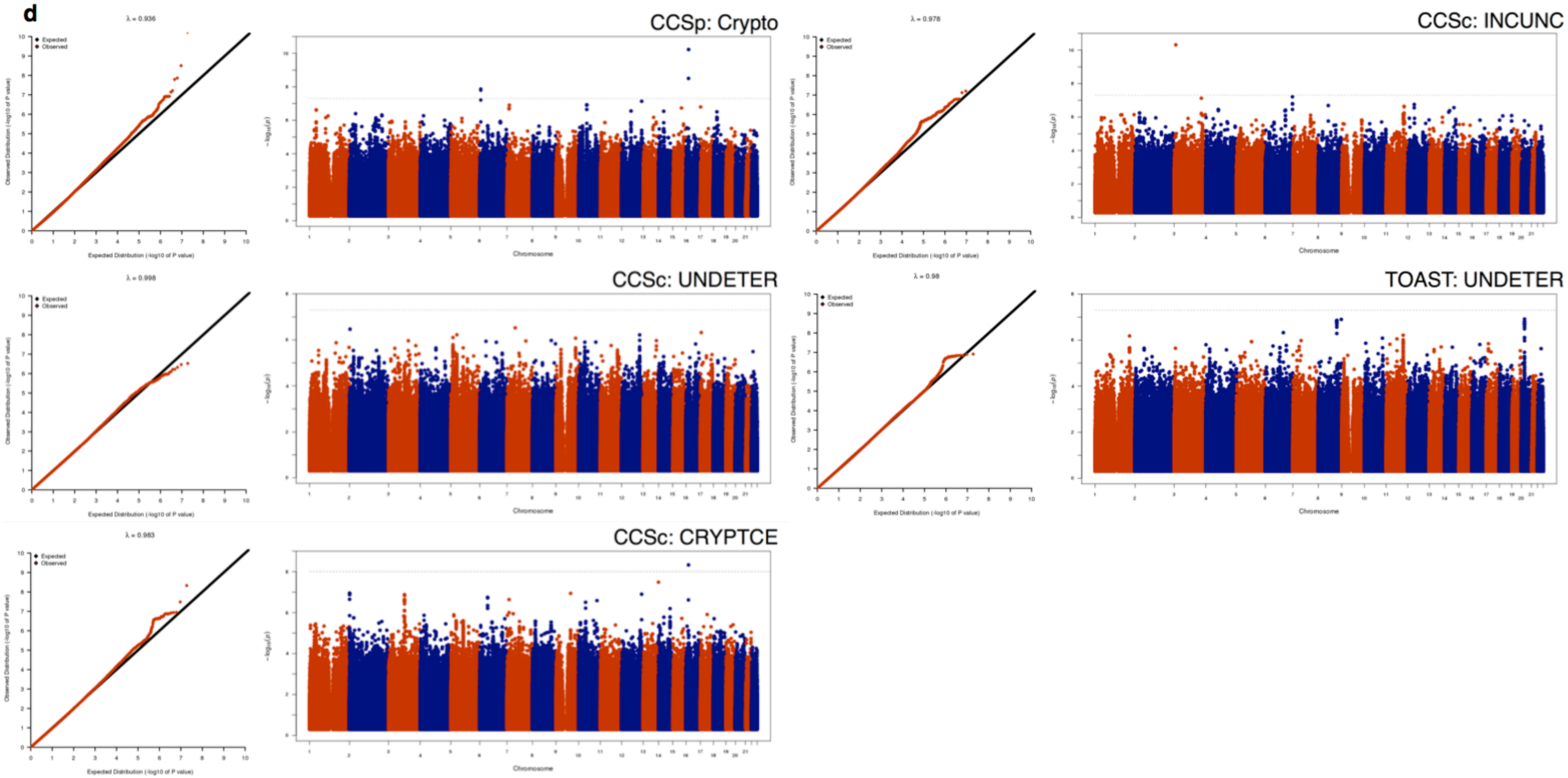
Supplementary Figure 2 (continued)

(b) Manhattan and QQ plots for large artery atherosclerosis (LAA) stroke, as determined by CCSc, CCSp, and TOAST.



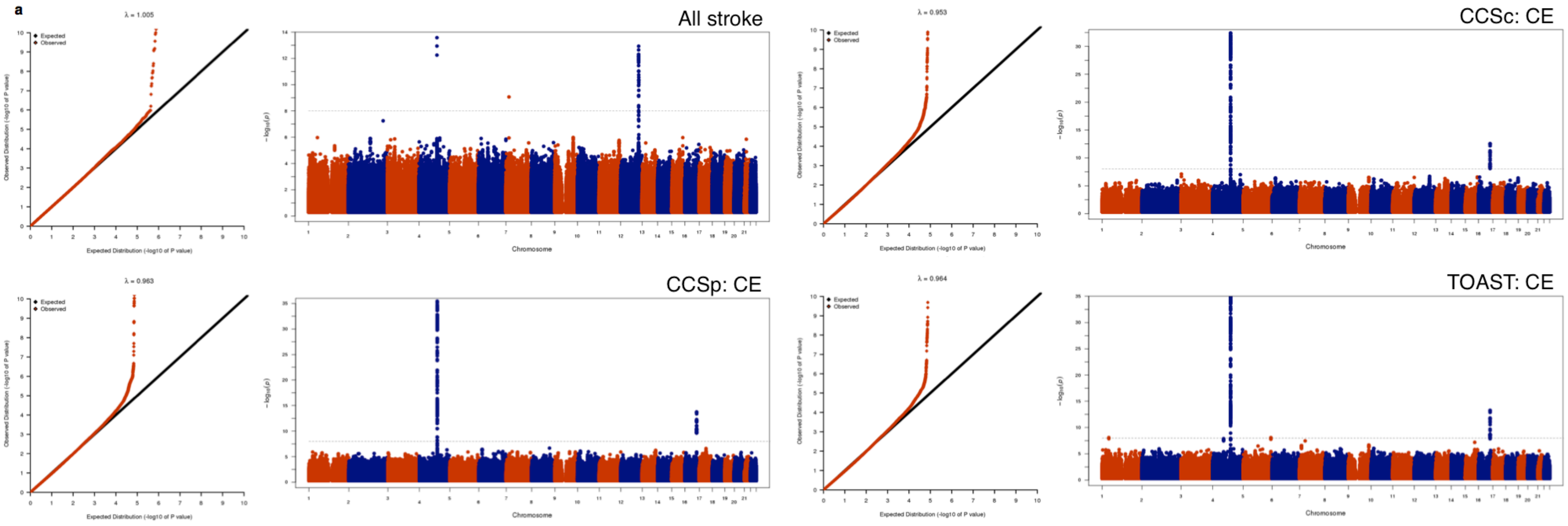
Supplementary Figure 2 (continued)

(c) Manhattan and QQ plots for small artery occlusion (SAO) stroke, as determined by CCSc, CCSp, and TOAST.



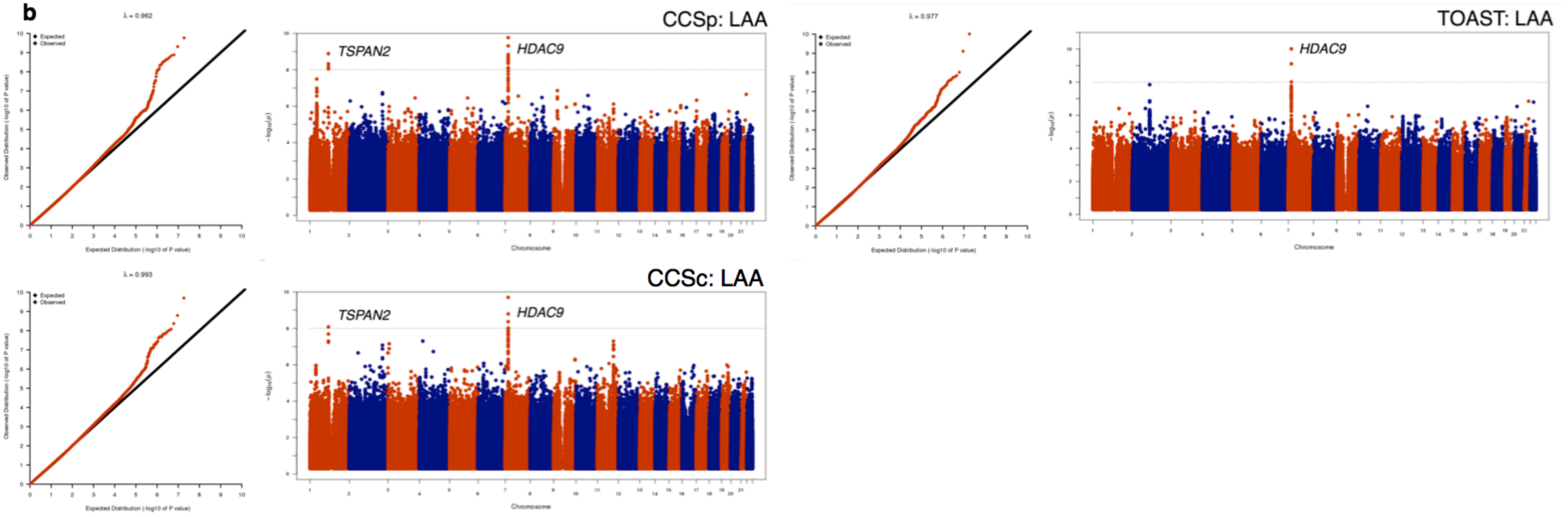
Supplementary Figure 2 (continued)

(d) Manhattan and QQ plots for undetermined stroke, as determined by CCSc, CCSp, and TOAST. CCSc has three undetermined classifications.



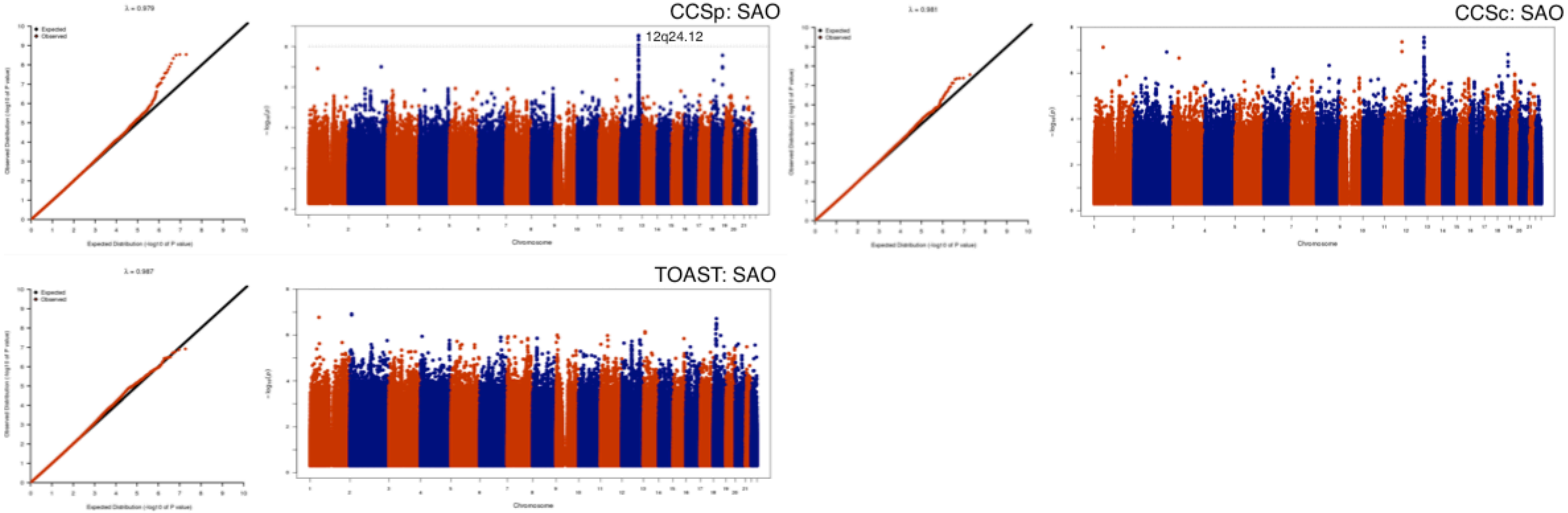
Supplementary Figure 3 | QQ and Manhattan plots of the combined meta-analyses of stage I and stage II data

Genome-wide significant ($p < 1 \times 10^{-8}$) regions are annotated. (a) Manhattan and QQ plots for combined meta-analysis (discovery and replication) of all stroke and cardioembolic (CE) stroke as determined by the CCS Causative (CCSc), CCS Phenotypic (CCSp), and TOAST subtyping systems (**Section 14**).



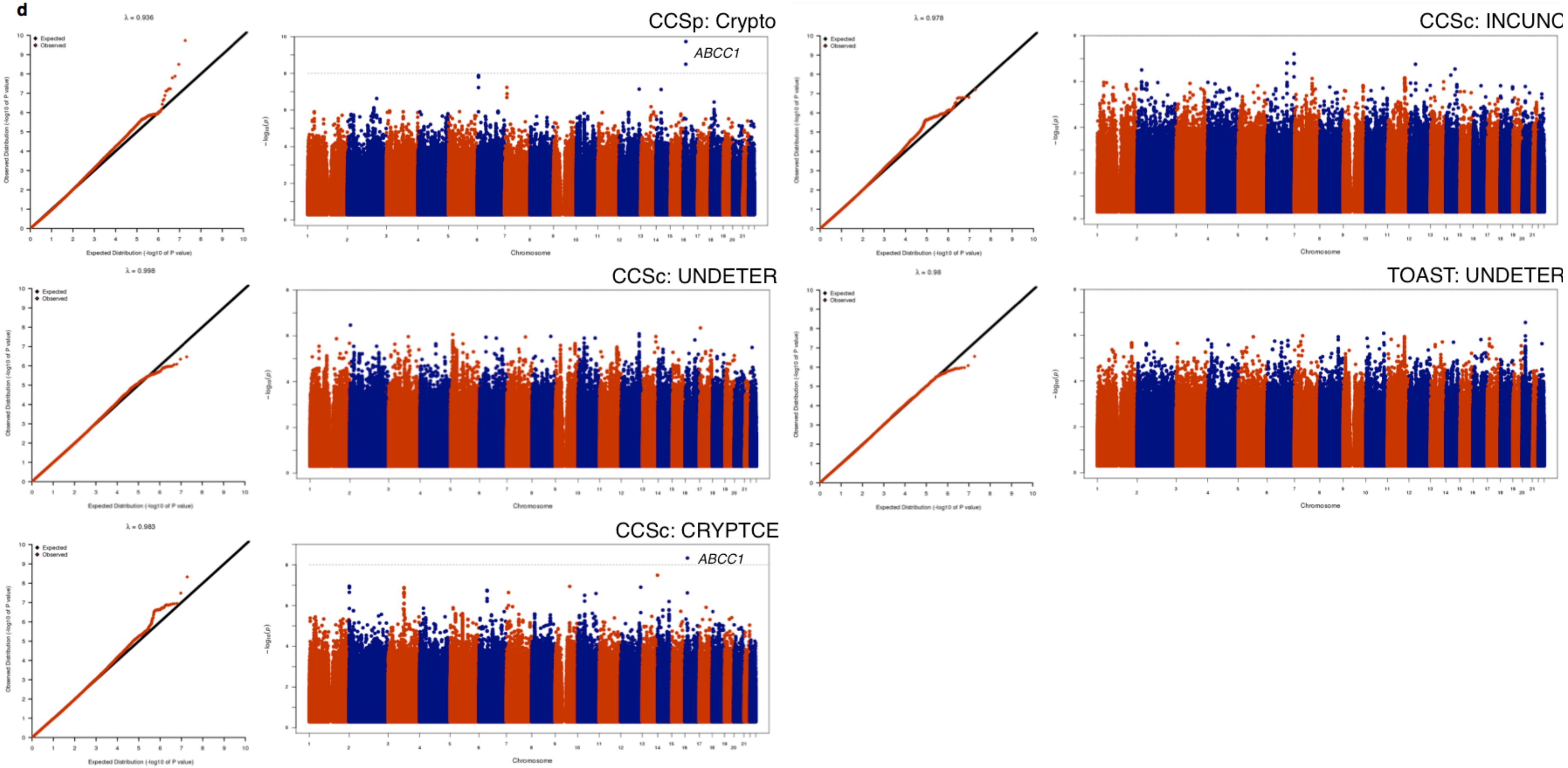
Supplementary Figure 3 (continued)

(b) Manhattan and QQ plots for combined meta-analysis (stage I and stage II) of large artery atherosclerosis (LAA) stroke as determined by the CCS Causative (CCSc), CCS Phenotypic (CCSp), and TOAST subtyping systems (**Section 14**).



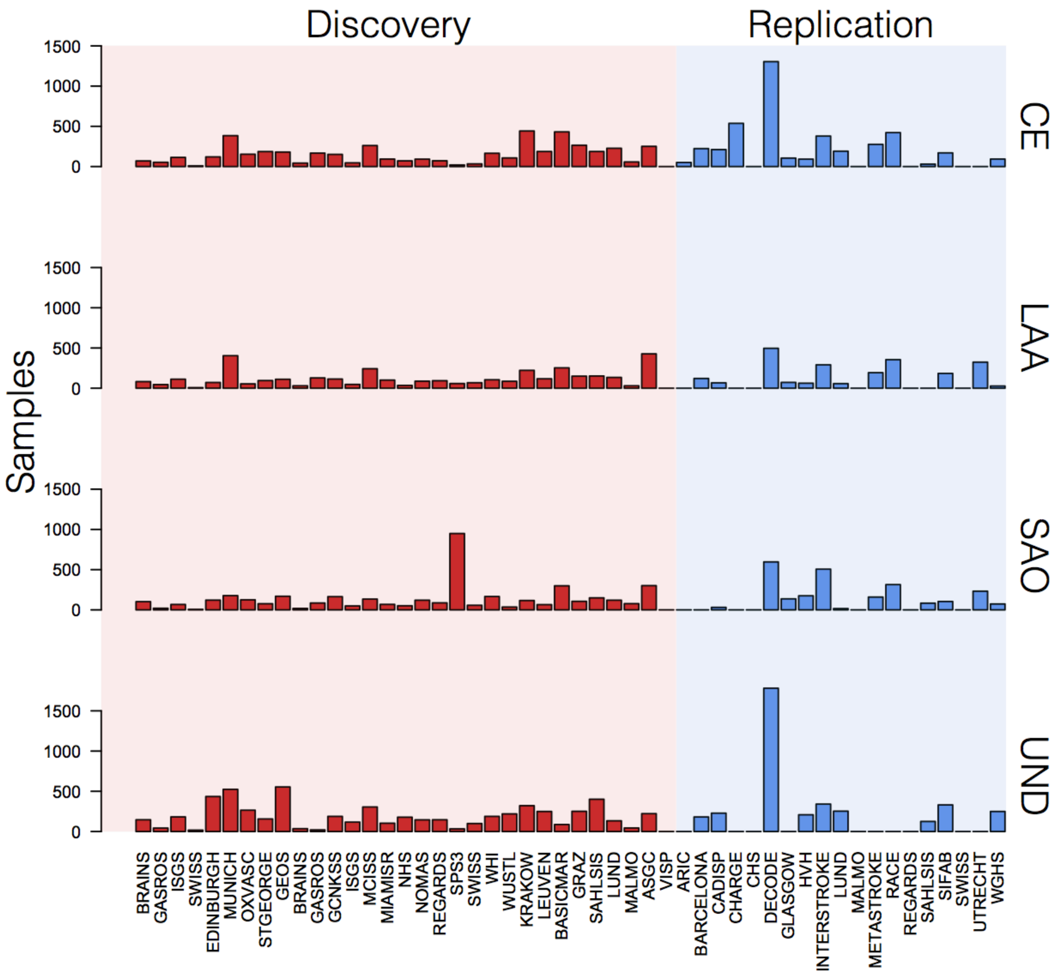
Supplementary Figure 3 (continued)

(c) Manhattan and QQ plots for combined meta-analysis (stage I and stage II) of small artery occlusion (SAO) stroke as determined by the CCS Causative (CCSc), CCS Phenotypic (CCSp), and TOAST subtyping systems (**Section 14**).



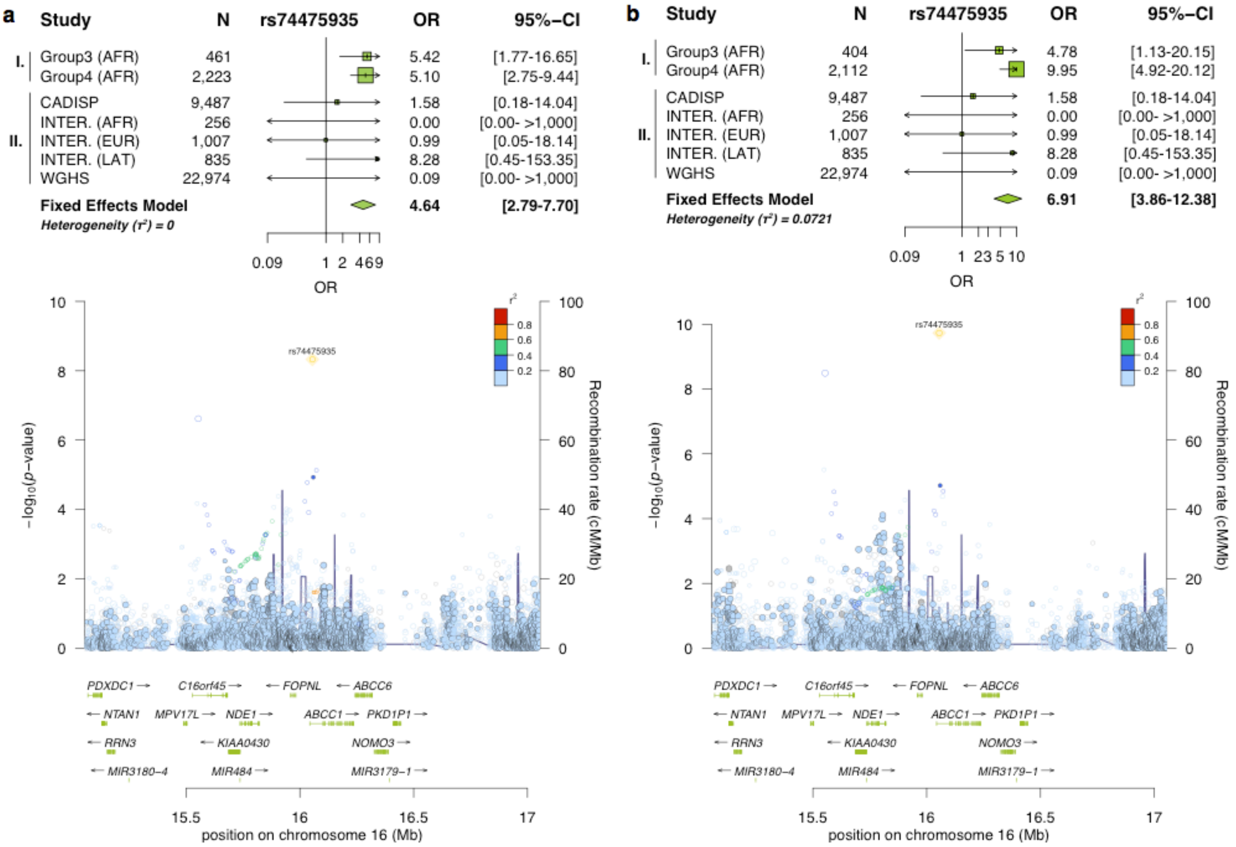
Supplementary Figure 3 (continued)

(d) Manhattan and QQ plots for combined meta-analysis (stage I and stage II) of undetermined stroke according to CCS Causative (CCSc), CCS Phenotypic (CCSp, cryptogenic), and TOAST subtyping systems. CCSc has three undetermined groups (**Section 14**).



Supplementary Figure 4 | Ischemic stroke subtypes in SiGN discovery (stage I) and replication (stage II)

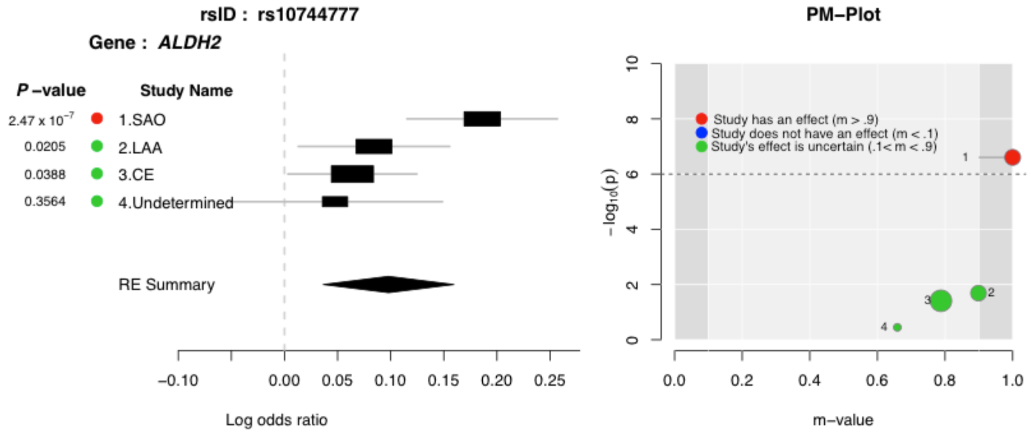
Four primary subtypes were analyzed in the discovery (stage I) and replication (stage II) phases of SiGN: cardioembolic (CE), large artery atherosclerosis (LAA), small artery occlusion (SAO), and undetermined (UND). Counts are based on the union of three subtyping methods: CCS Causative, CCS Phenotypic, and TOAST. Some cases are assigned to multiple subtypes by the different subtyping systems (**Section 8**) and therefore appear in multiple bars in the plot.



Supplementary Figure 5 | Regional association and forest plots for rs74475935

Rs74475935 was associated to undetermined stroke for (a) CCS Causative (cryptogenic and CE minor) stage I and TOAST stage II cases, and (b) CCS Phenotypic (cryptogenic) discovery and TOAST replication cases (**Section 15.3**). This SNP is rare (MAF ~0.1%) in European-ancestry (EUR) samples and low frequency (MAF ~1.5%) in African-ancestry (AFR) samples. It was found in a small number of cases with limited African-ancestry samples available for replication and will require additional data to evaluate its robustness.

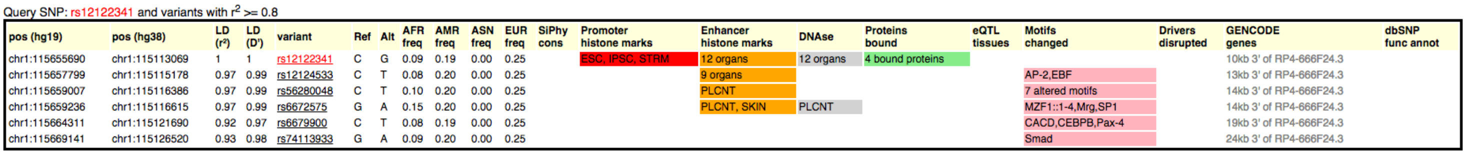
INTER., INTERSTROKE; OR, odds ratio; LAT, Latino/Hispanic.



Supplementary Figure 6 | Testing the specificity of the 12q24.12 locus to the small artery occlusion subtype

Using the discovery results from the CCS Phenotypic (CCSp) subtyping method, we obtained the effect size of each subtype using an inverse-variance weighted fixed-effects model to combine information from multiple strata. The m-value refers to the posterior probability that the effect exists in each subtype. Subtypes are small artery occlusion (SAO), cardioembolic stroke (CE), large artery atherosclerosis (LAA), and undetermined.

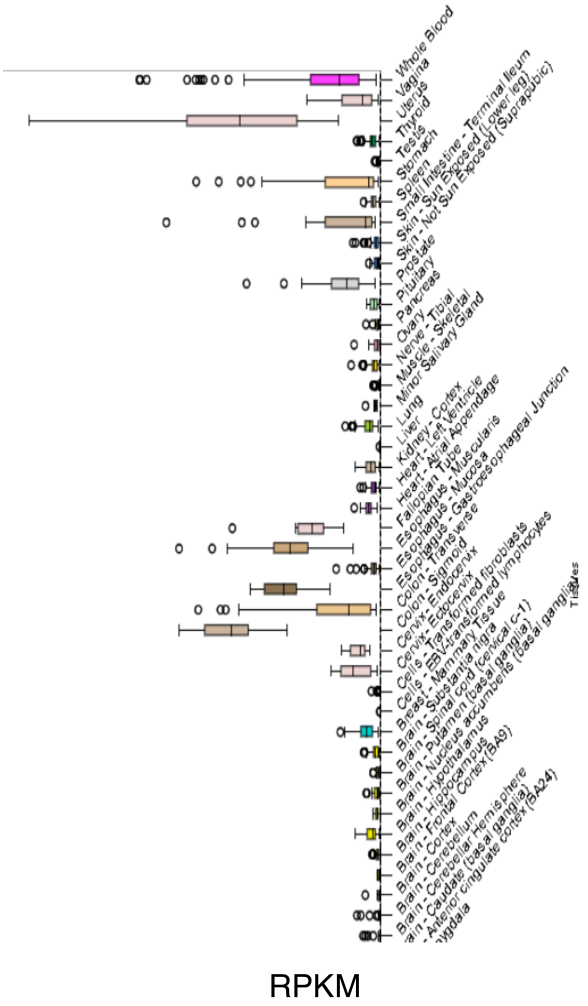
a.



Query SNP: rs12122341 and variants with r^2 >= 0.8

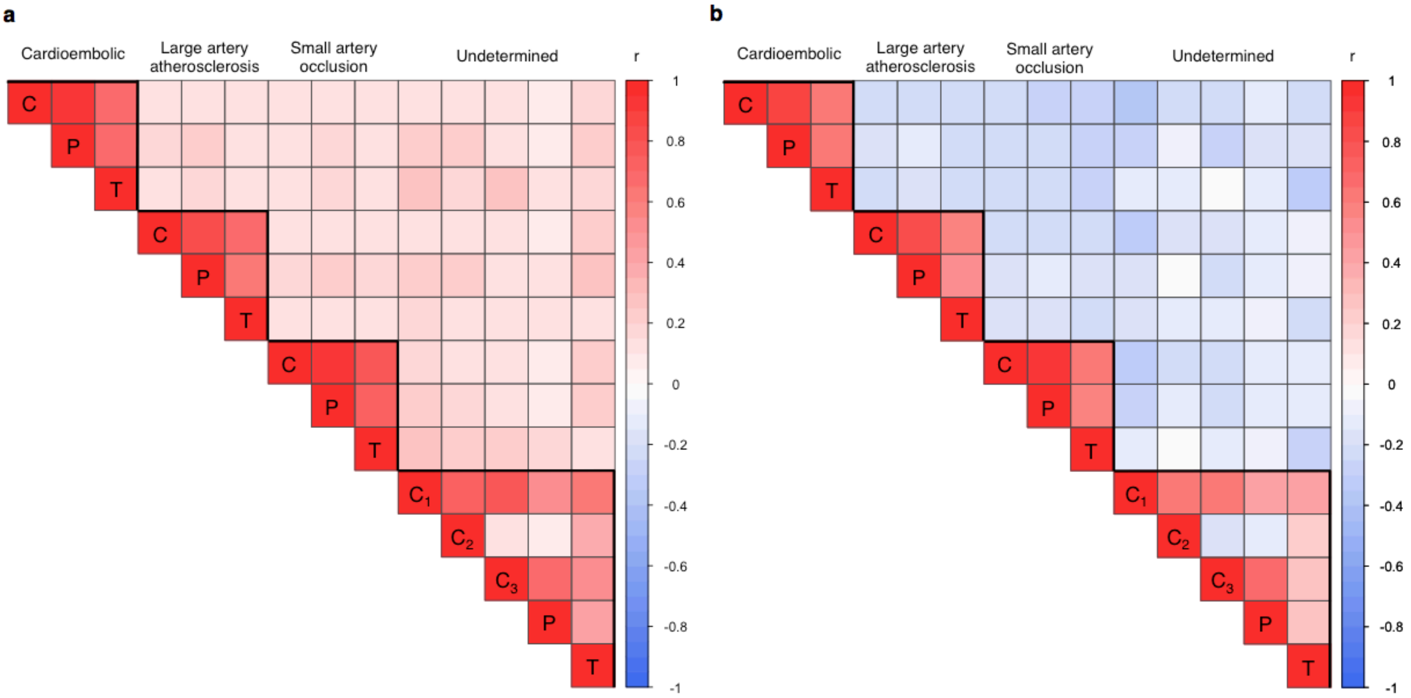
pos (hg19)	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	Drivers disrupted	GENCODE genes	dbSNP func annot
chr1:115655690	chr1:115113069	1	1	rs12122341	C	G	0.09	0.19	0.00	0.25		ESC, IPSC, STRM	12 organs	12 organs	4 bound proteins				10kb 3' of RP4-666F24.3	
chr1:115657799	chr1:115115178	0.97	0.99	rs12124533	C	T	0.08	0.20	0.00	0.25			9 organs				AP-2,EBF		13kb 3' of RP4-666F24.3	
chr1:115659007	chr1:115116386	0.97	0.99	rs56280048	C	T	0.10	0.20	0.00	0.25			PLCNT				7 altered motifs		14kb 3' of RP4-666F24.3	
chr1:115659236	chr1:115116615	0.97	0.99	rs6672575	G	A	0.15	0.20	0.00	0.25			PLCNT, SKIN	PLCNT			MZF1::1-4,Mrg,SP1		14kb 3' of RP4-666F24.3	
chr1:115664311	chr1:115121690	0.92	0.97	rs6679900	C	T	0.08	0.19	0.00	0.25							CACD,CEBPB,Pax-4		19kb 3' of RP4-666F24.3	
chr1:115669141	chr1:115126520	0.93	0.98	rs74113933	G	A	0.09	0.20	0.00	0.25							Smad		24kb 3' of RP4-666F24.3	

b.



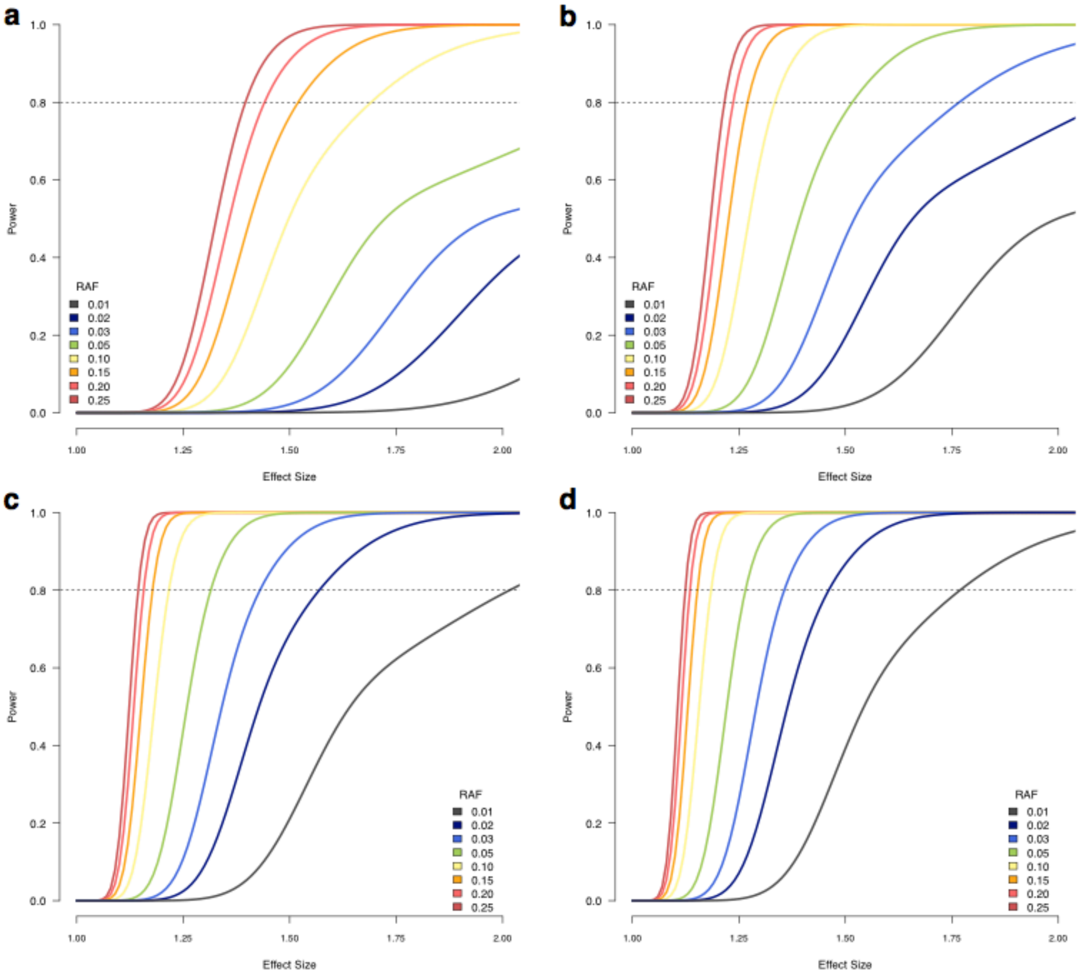
Supplementary Figure 7 | Gene expression pattern of TSPAN2 in various human tissues

A SNP (rs12122341) near *TSPAN2* was implicated in GWAS of large artery atherosclerosis (**Section 15.3**). (a) rs12122341 is located in a DNA sequence immediately adjacent to *TSPAN2* (HaploReg v3). It can be bound by several transcription factor proteins. This sequence is a promotor and enhancer site that is marked by histone modification and DNase hypersensitivity (b) Gene expression patterns of *TSPAN2* (GTEx Project RNAseq data, http://www.gtexportal.org/) are shown. mRNA expression intensity of *TSPAN2* was measured in Reads Per Kilobase per Million mapped reads (RPKM). The vertical axis indicates various human tissues, and the horizontal axis shows the distribution of *TSPAN2* expression intensities in each tissue.



Supplementary Figure 8 | Genetic and phenotypic correlation across the CCS Causative (C), CCS Phenotypic (P), and TOAST (T) subtyping methods

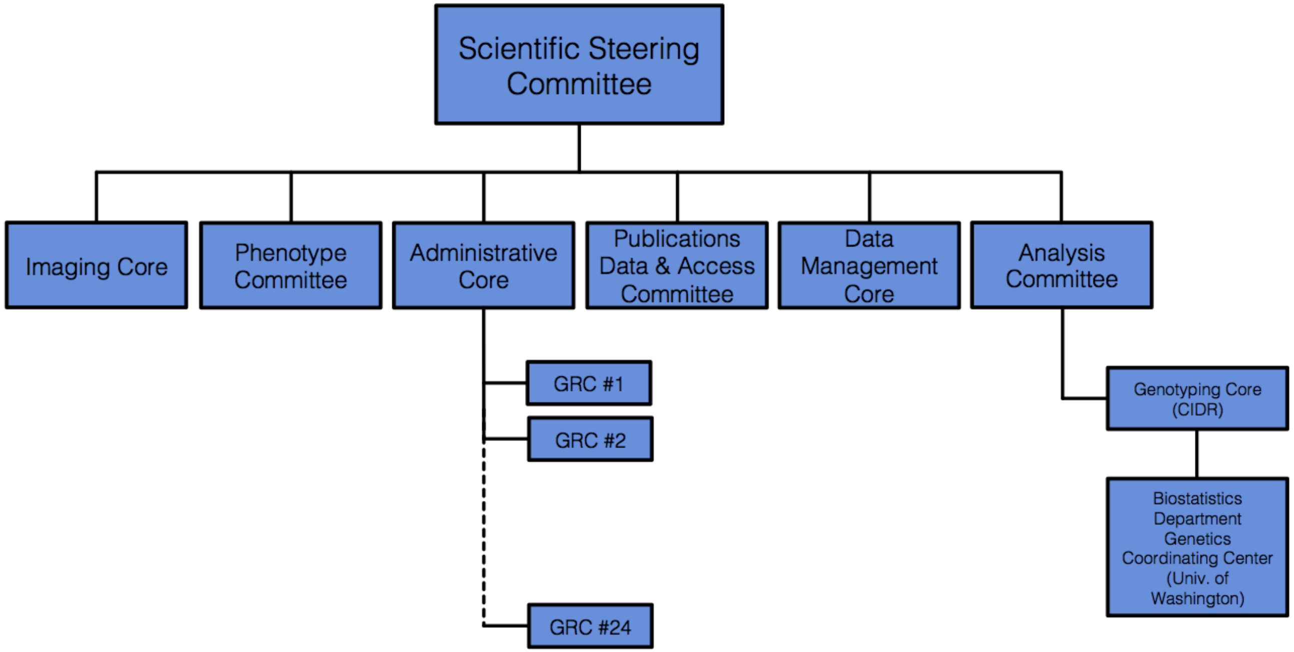
(a) To calculate genetic correlation, genome-wide z-scores were taken from each stage I meta-analysis and used to calculate pairwise correlation (r) in each pair of subytpes (**Section 14.1**, **Supplementary Table 3**). Because all available CCS cases were used in the discovery phase, only TOAST cases were available for replication analyses. There was moderate to strong correlation within each subtype group, indicating that TOAST-subtyped cases could be used in the second phase of SiGN. (b) Pairwise correlation (r) was calculated for all pairs of phenotypes by coding a sample as “1” if it had been subtyped as a case for a particular subtype by a particular subtyping method, and “0” otherwise.



Supplementary Figure 9 | Statistical power (at $p < 5 \times 10^{-8}$) in discovery analysis of the subtypes

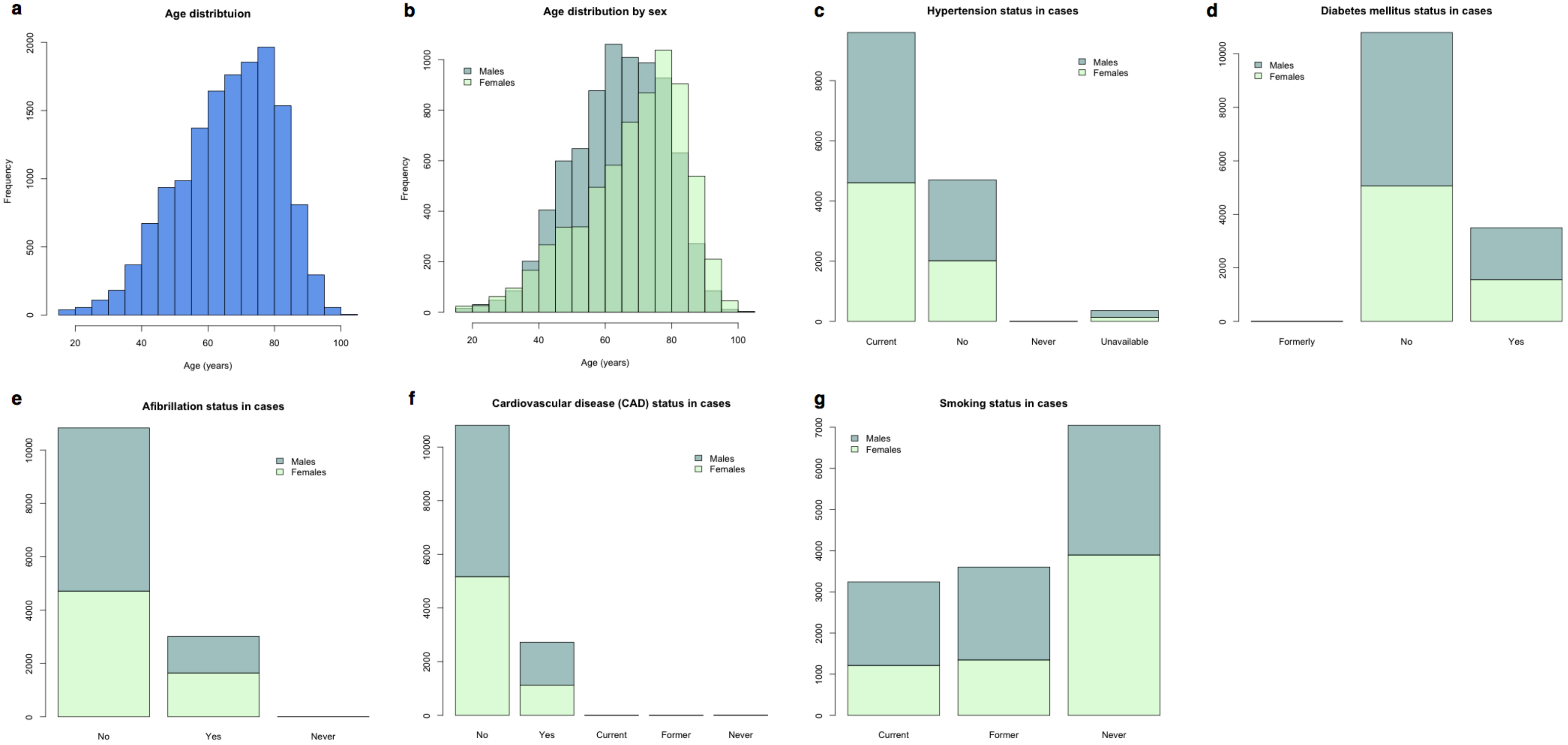
Power curves are shown for SNPs of varying frequency and effect size (odds ratio). Calculations assumed an analysis that included 32,000 controls, a subtype disease prevalence of 1% and (a) 1,000 cases, (b) 3,000 cases, (c) 7,000 cases and (d) 10,000 cases. The “other” subtypes had small numbers of cases (CCSc = 594 cases, CCSp = 718 cases, TOAST = 373 cases) and limited power; it was excluded from all genome-wide analyses (**Section 12**).

RAF, risk allele frequency. Dashed line indicates 80% power.

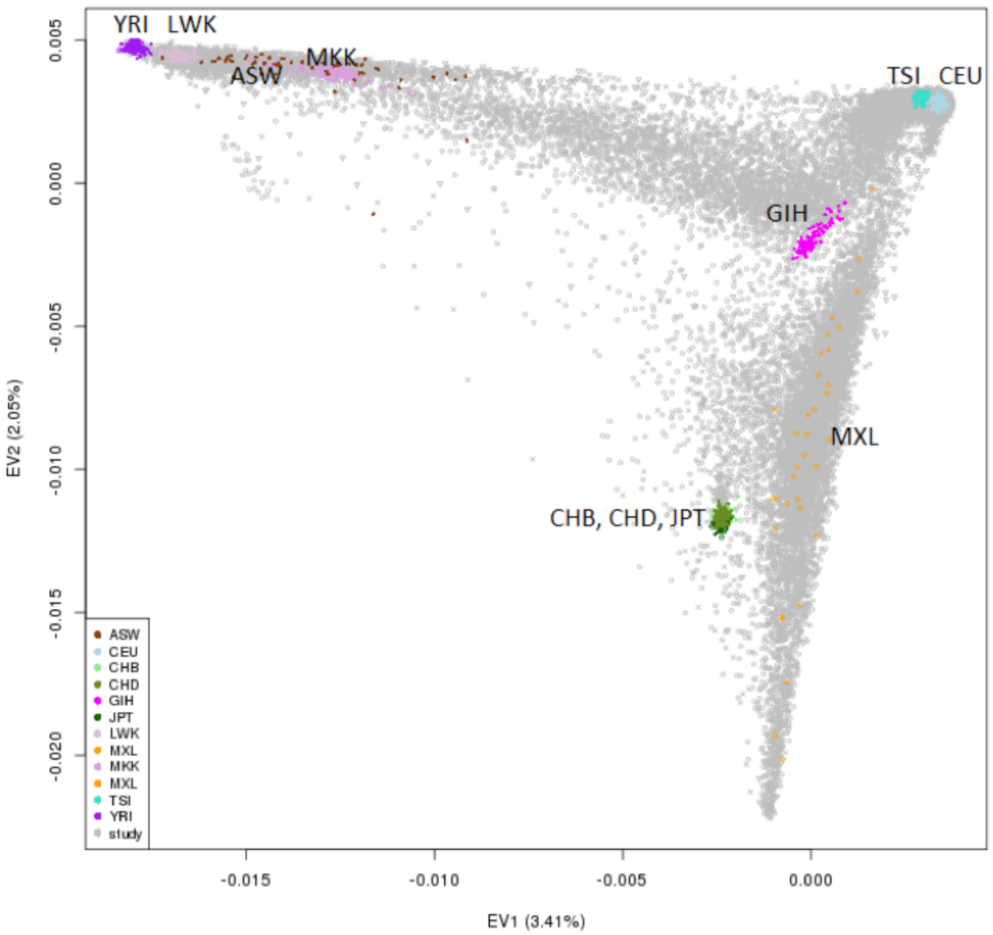

Scientific Steering Committee
Imaging Core
Phenotype Committee
Administrative Core
Publications Data & Access Committee
Data Management Core
Analysis Committee
GRC #1
GRC #2
GRC #24
Genotyping Core (CIDR)
Biostatistics Department Genetics Coordinating Center (Univ. of Washington)

Supplementary Figure 10 | Administrative structure of SiGN

The Scientific Steering Committee (SSC) leads SiGN (**Section 2**). The SSC is responsible for scientific direction and policy decisions. It also oversees the Publications and Data Access Committee. The study's four cores are Administrative, Data Management, Imaging, and Genotyping. The Analysis Committee, composed of genetic epidemiologists and statistical geneticists, advises the Scientific Steering Committee on design issues and is responsible for the analyses of phenotypic and genetic data. CIDR, the Biostatistics Department Genetics (University of Washington), and the Analysis Committee jointly decided on the design of the study. GRC, general research center.



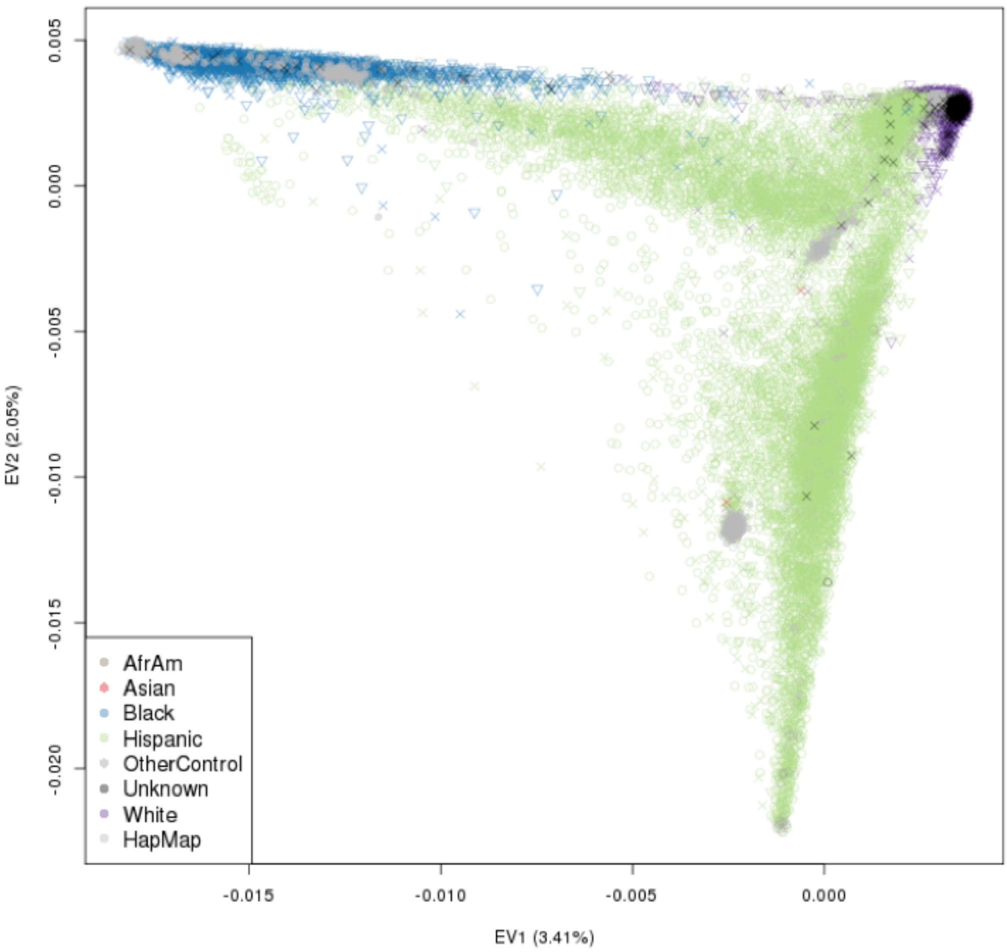
Supplementary Figure 11 | Distributions of stroke risk factors in the SiGN phase I cases (Section 8)



Supplementary Figure 12 | Principal component analysis of unrelated study participants and HapMap controls

PCA of 1,431 HapMap samples and 33.843 study samples (**Section 9.3.2**). Grey symbols denote sample groups (SiGN: ×, HCHS/SOL: O, HRS: Δ, HapMap: ❒). Axis labels indicate the percentage of variance explained by each eigevector. Details on the HapMap 3 populations can be found in **Supplementary Table 10**. Briefly, ASW, MKK, LWK, and YRI are African ancestry. TSI and CEU are European ancestry. CHB, CHD, and JPT are East Asian ancestry. GIH and MXL are admixture populations.

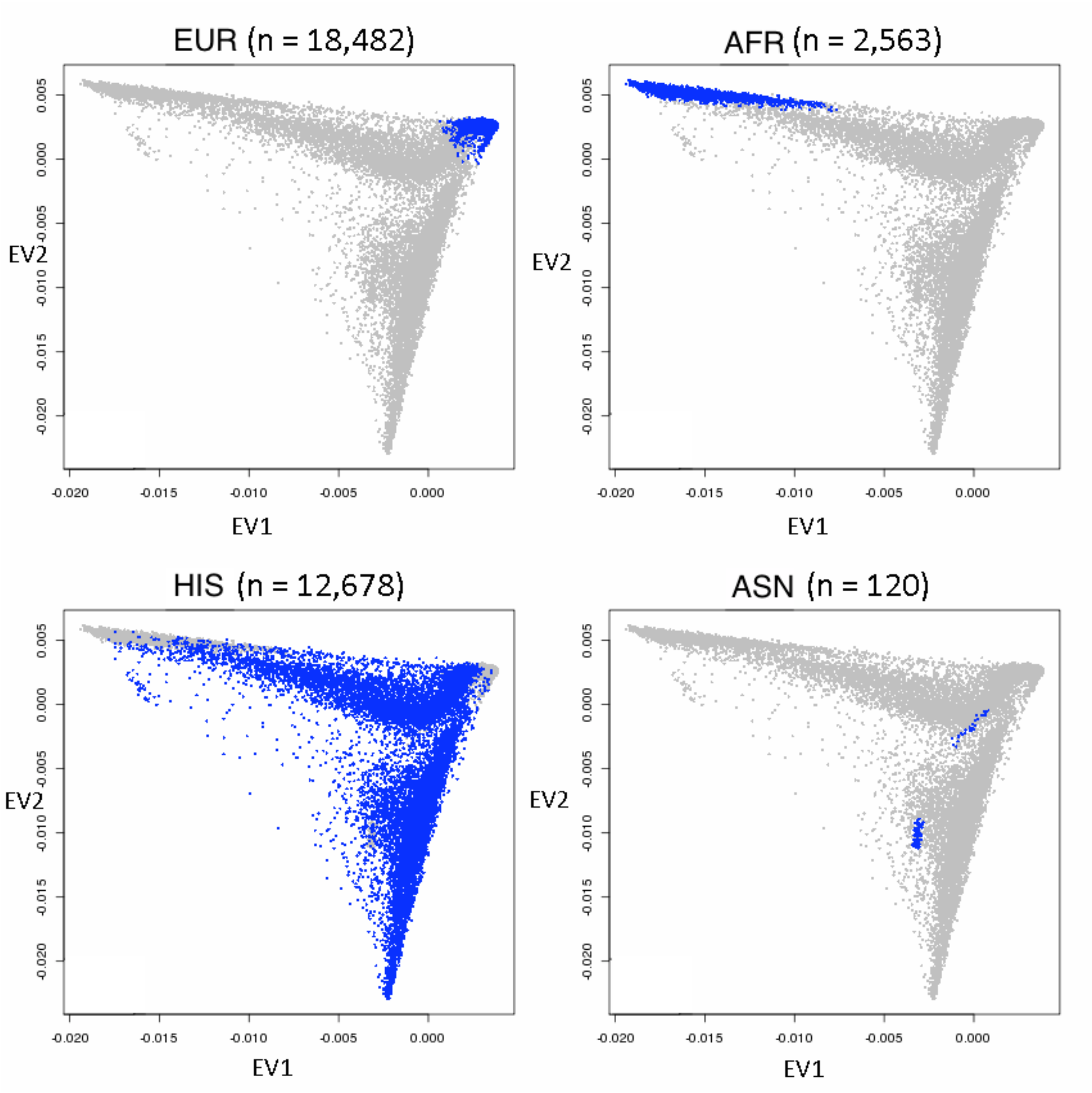
EV, eigenvector.



Supplementary Figure 13 | Principal component analysis of unrelated study participants and HapMap controls

The PCA plot shows 33,843 study samples (color-coded by self-identified race or ethnicity) and 1,431 HapMap controls (grey) (**Section 9.3.2**). HapMap participants are in grey while symbols denote studies (SiGN: ×, HCHS/SOL: O, HRS: Δ, HapMap: ❑). Axis labels indicate the percentage of variance explained by each eigenvector.

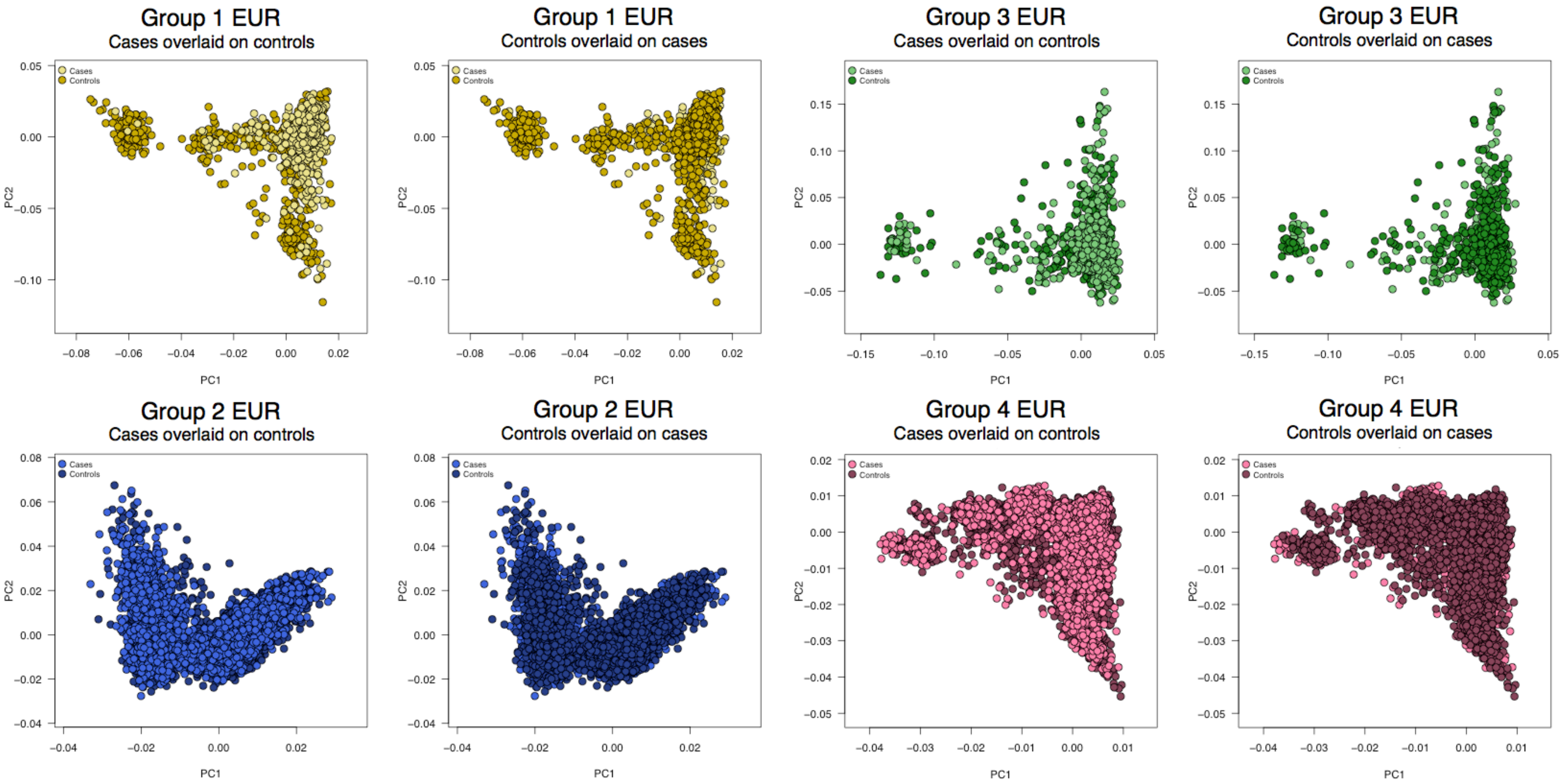
AfrAm, African American; EV, eigenvector.



Supplementary Figure 14 | Principal component analysis of unrelated study cases and controls.

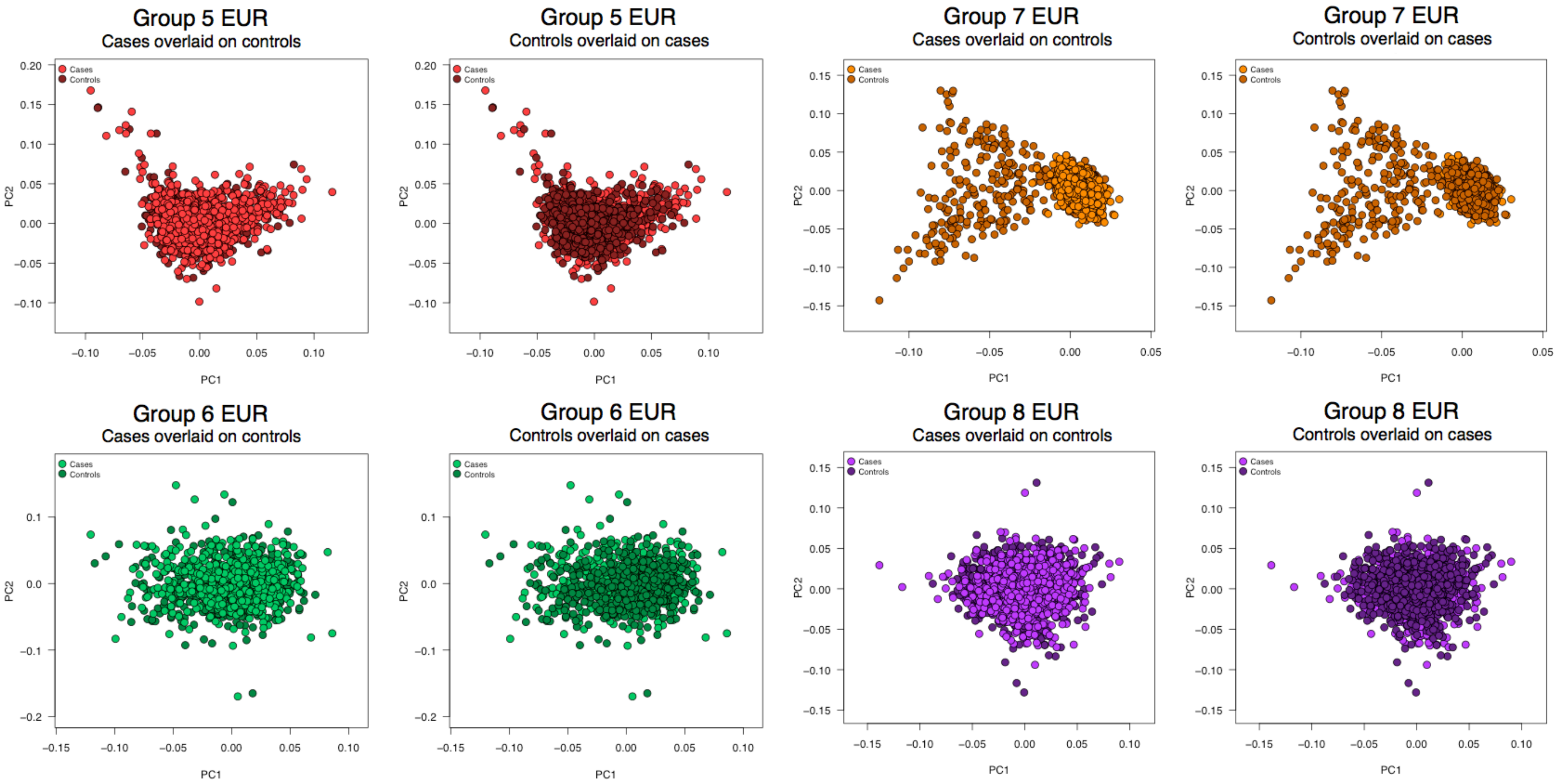
Color-coding shows the composite ancestry group of all 33,843 study participants as determined by the hyperellipsoid clustering technique (**Section 9.3.3**) using at least four PCs (only the first two are plotted).

EUR, white/Caucasian, European-ancestry; AFR, black or African American; HIS, Hispanic/Latino; ASN, East or South Asian; EV, eigenvector.

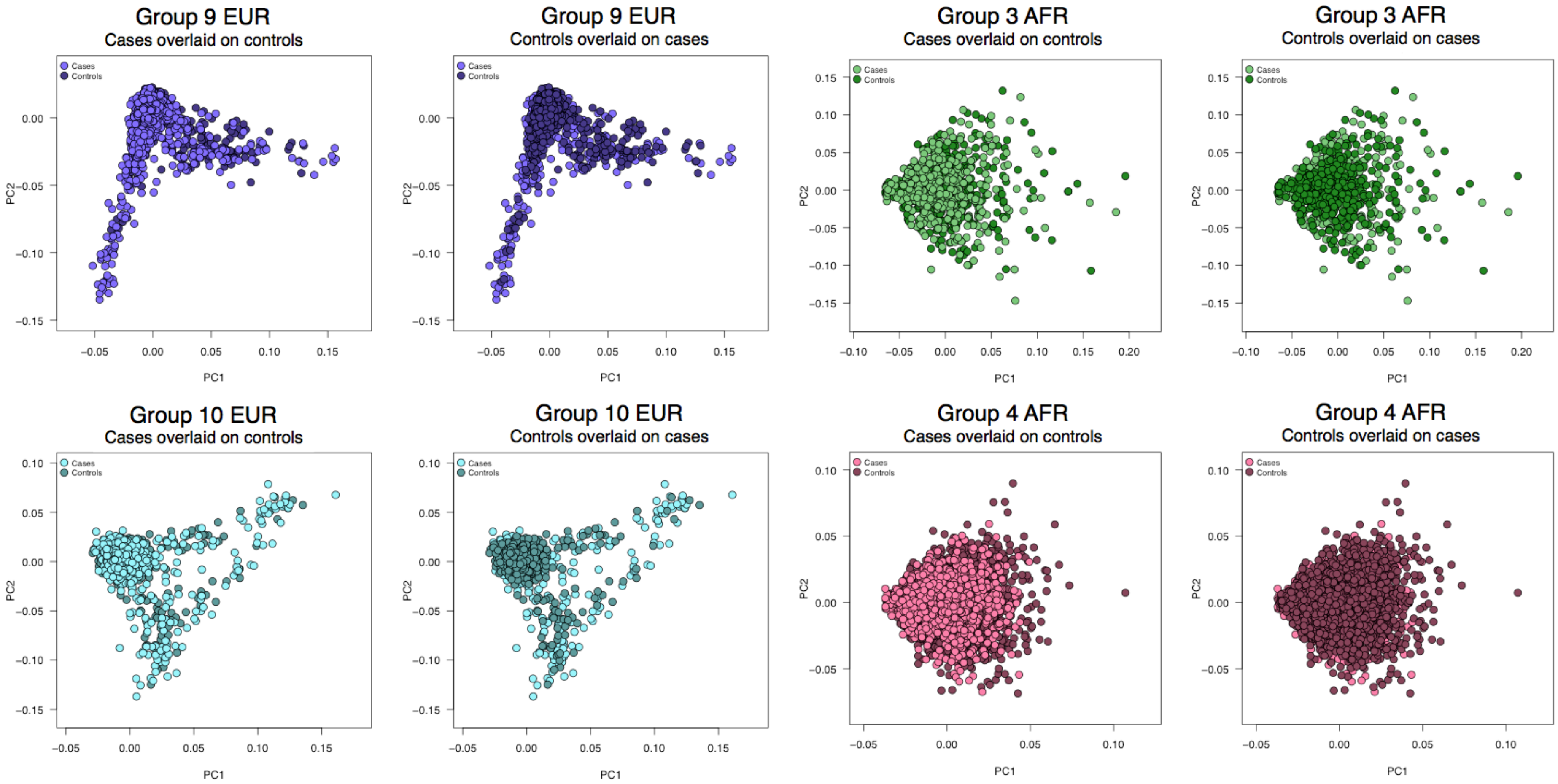


Supplementary Figure 15 | Principal component analysis in European- and African-ancestry case/control stage I strata

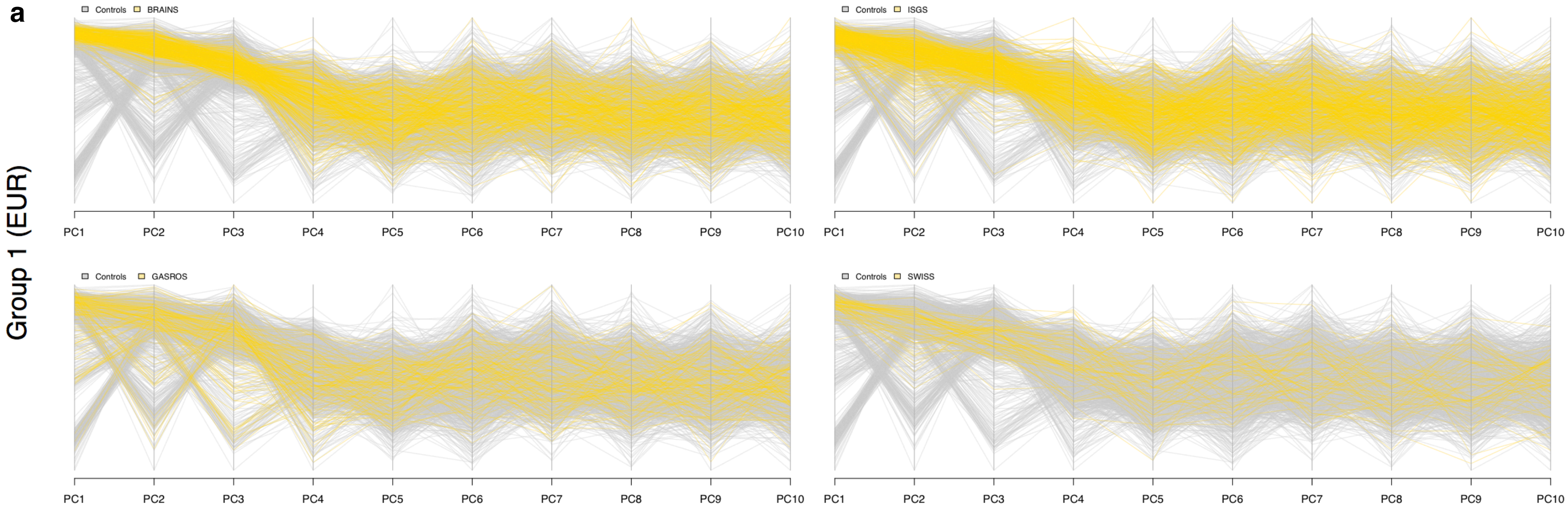
Principal component analysis (PCA) was used to determine ancestral homogeneity in each European-ancestry (EUR) and African-ancestry (AFR) stratum (**Section 9.3.4**). Cohorts within each group are provided in **Section 9.2.**



Supplementary Figure 15 (continued)



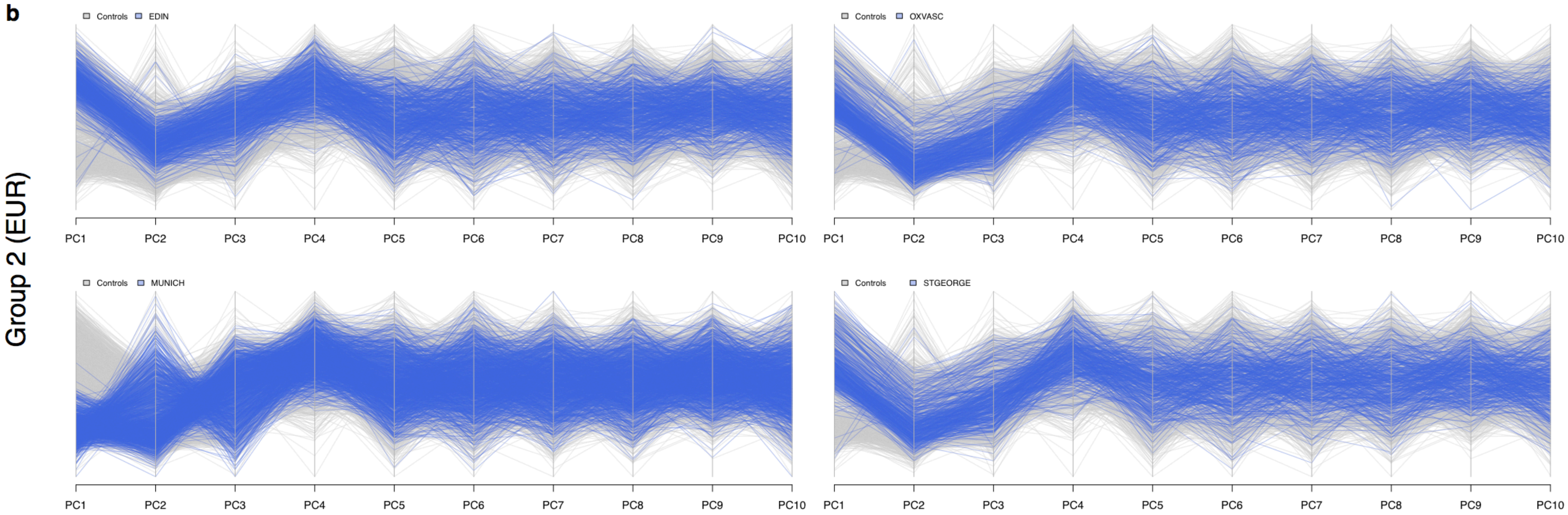
Supplementary Figure 15 (continued)



Supplementary Figure 16 | Parallel coordinates plots of cases and controls for each study stratum

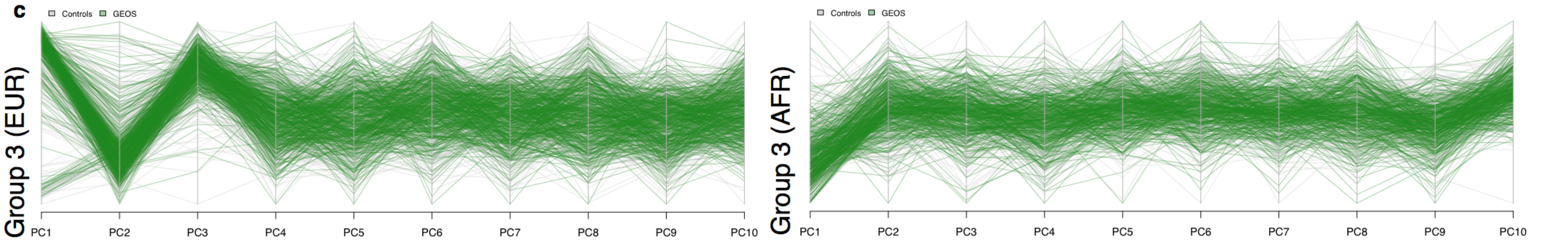
The top ten principal components (PCs) for groups of cases (bright colors) were compared to the top ten PCs for groups of controls (grey) to check for population stratification (**Section 9.3.4**). Group 10 (ASGC) was not checked because cases and controls were genotyped and analyzed together in a previous GWAS. Each line corresponds to one sample.

(a) Group 1 EUR. Controls are from HABC. Cases cohorts: BRAINS, GASROS, ISGS, SWISS.



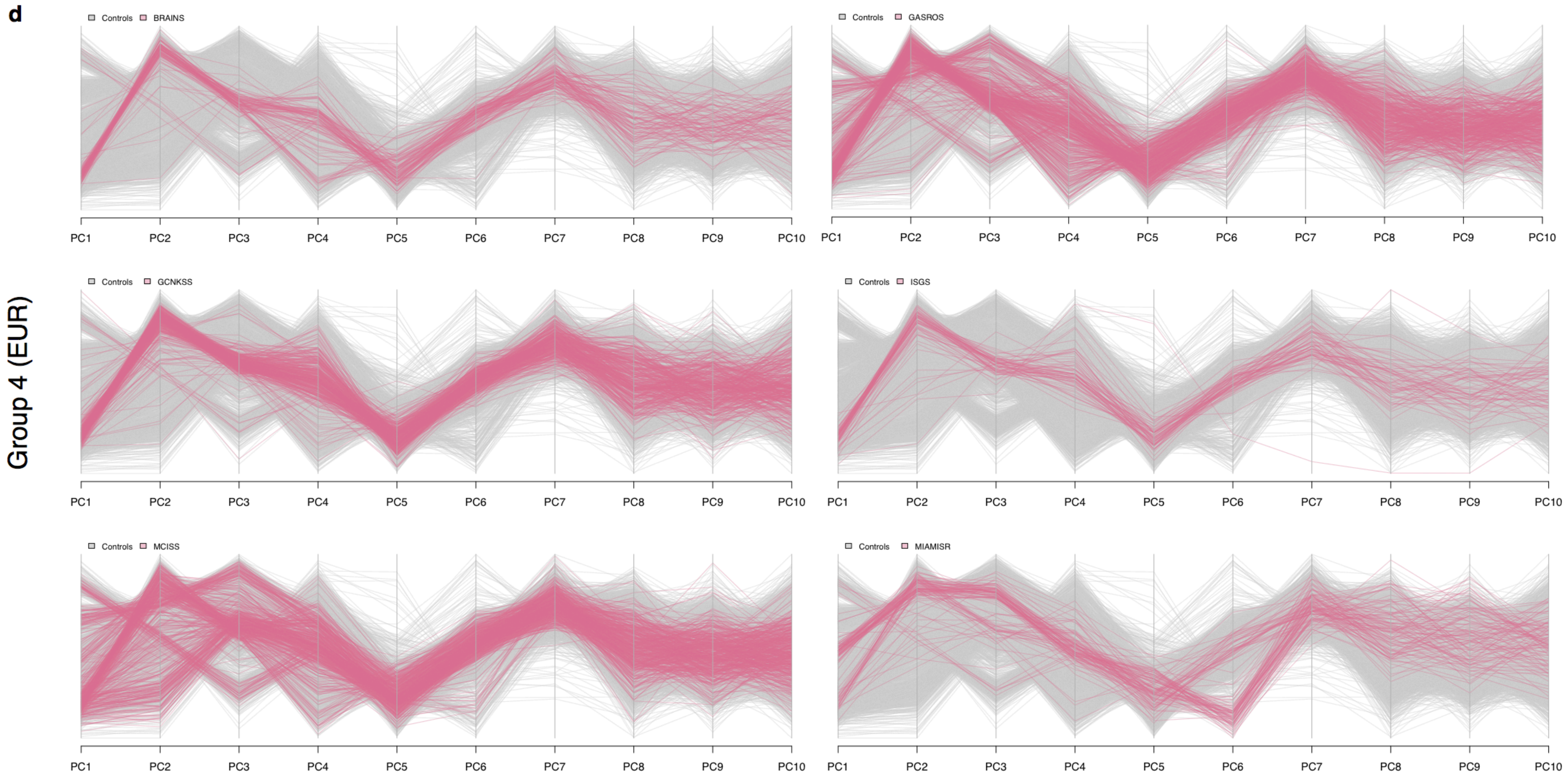
Supplementary Figure 16 (continued)

(b) Group 2 EUR: Controls are from KORA and WTCCC. Cases cohorts: ESS, MUNICH, OXVASC, STGEORGE.



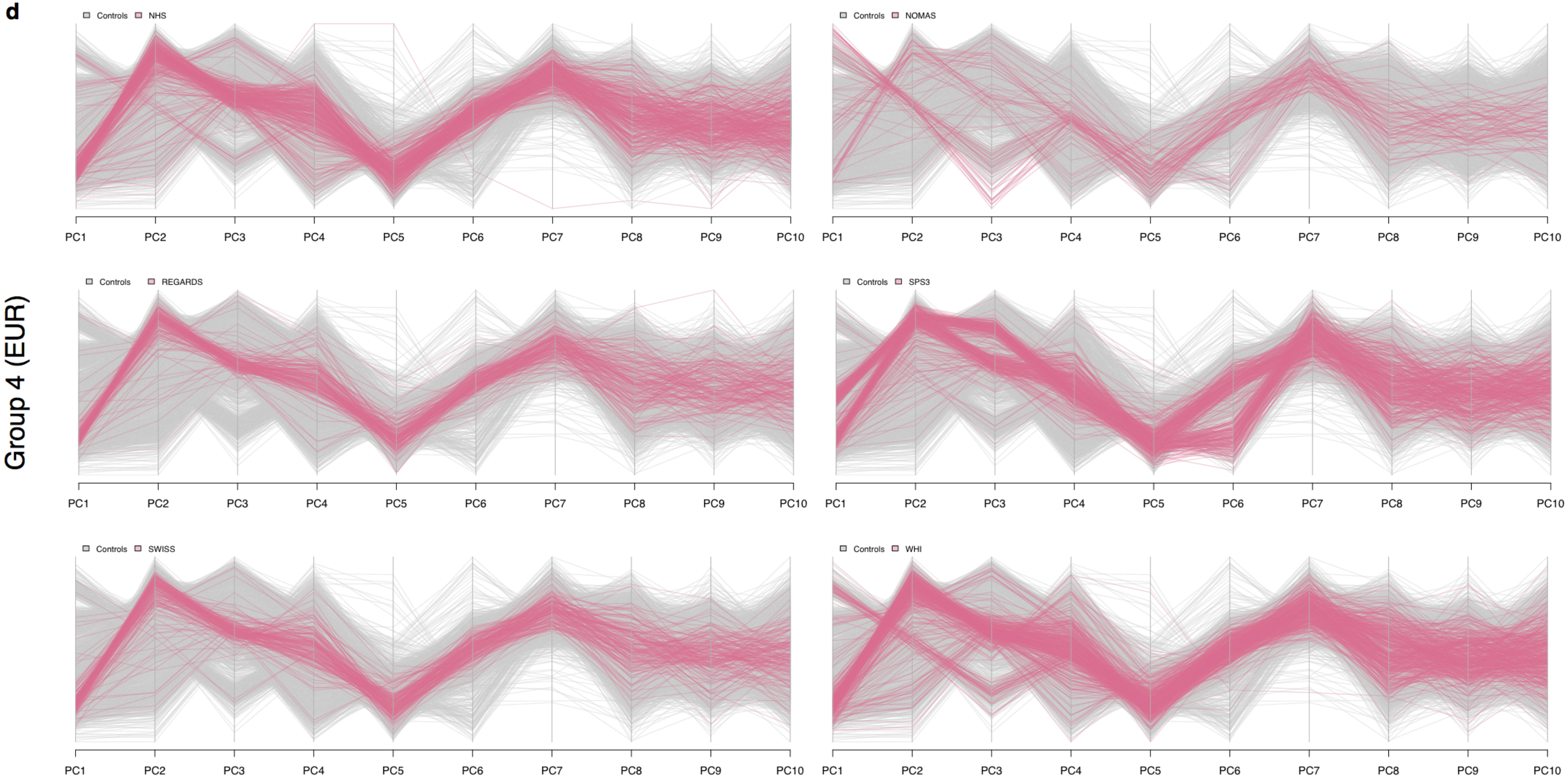
Supplementary Figure 16 (continued)

(c) Group 3 EUR and AFR: Controls and cases are from GEOS.

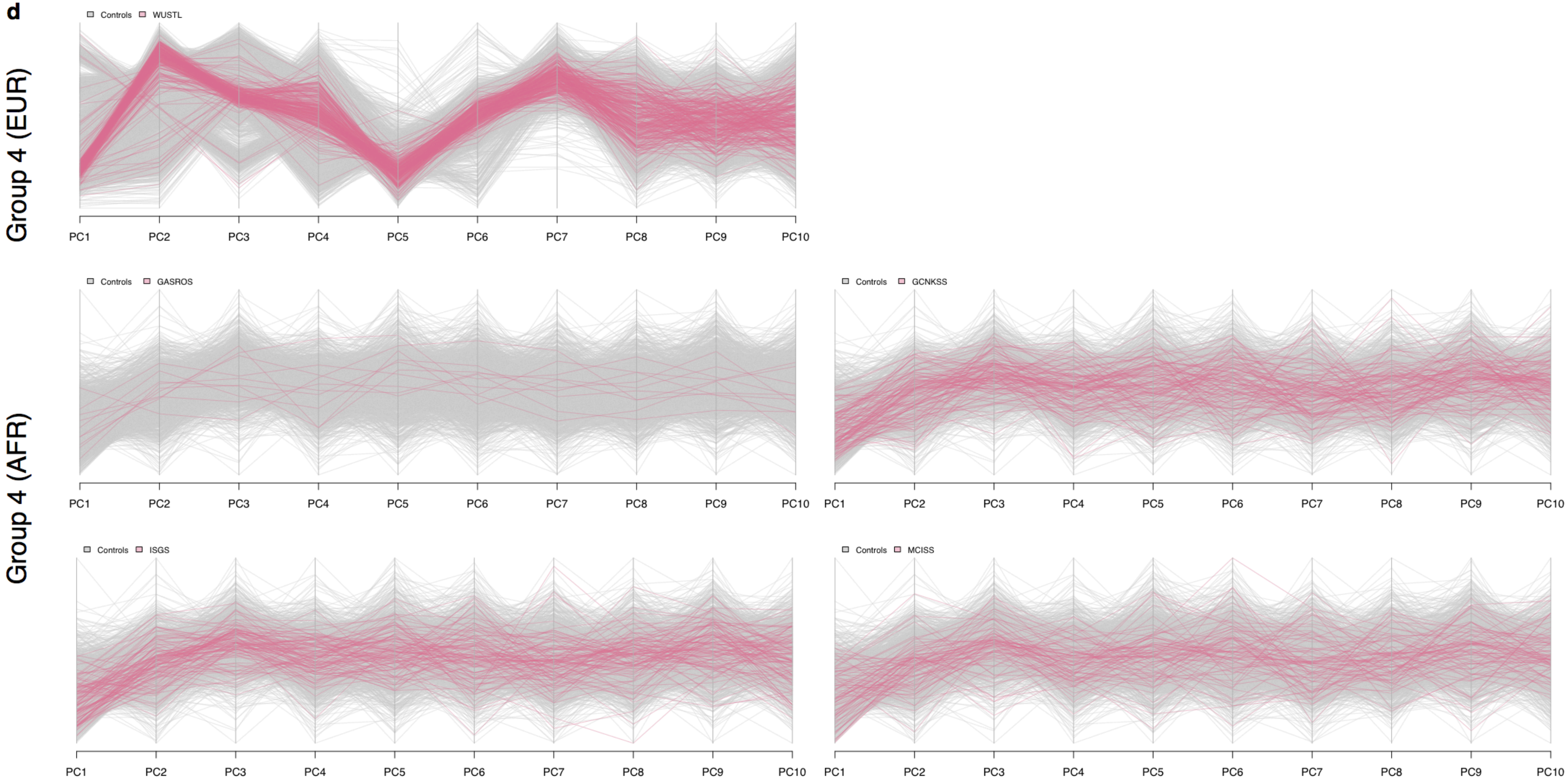


Supplementary Figure 16 (continued)

(d) Group 4 EUR: Controls are from HRS and OAI. Case cohorts: BRAINS, GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, and WUSTL.

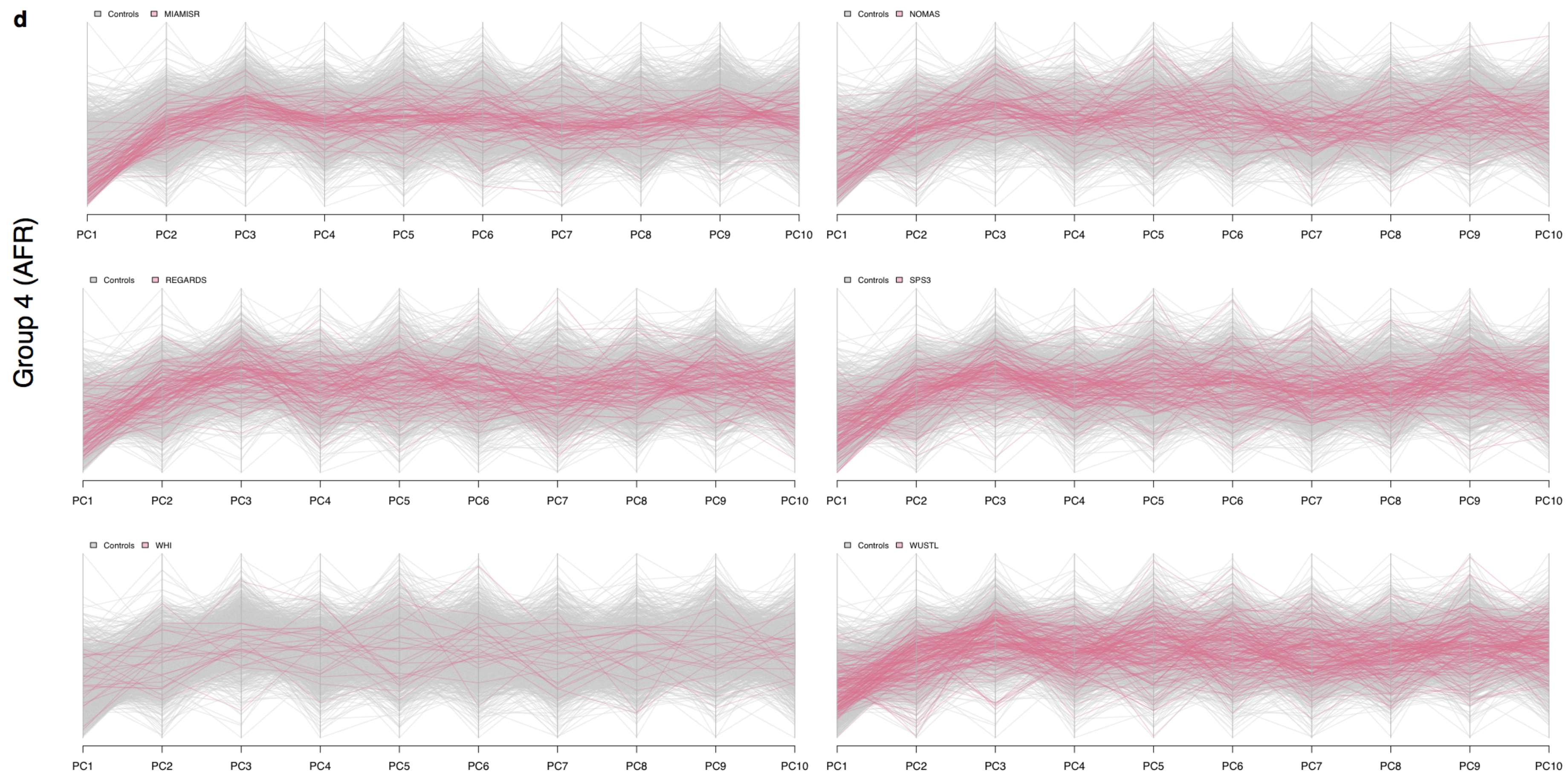


Supplementary Figure 16d (continued)

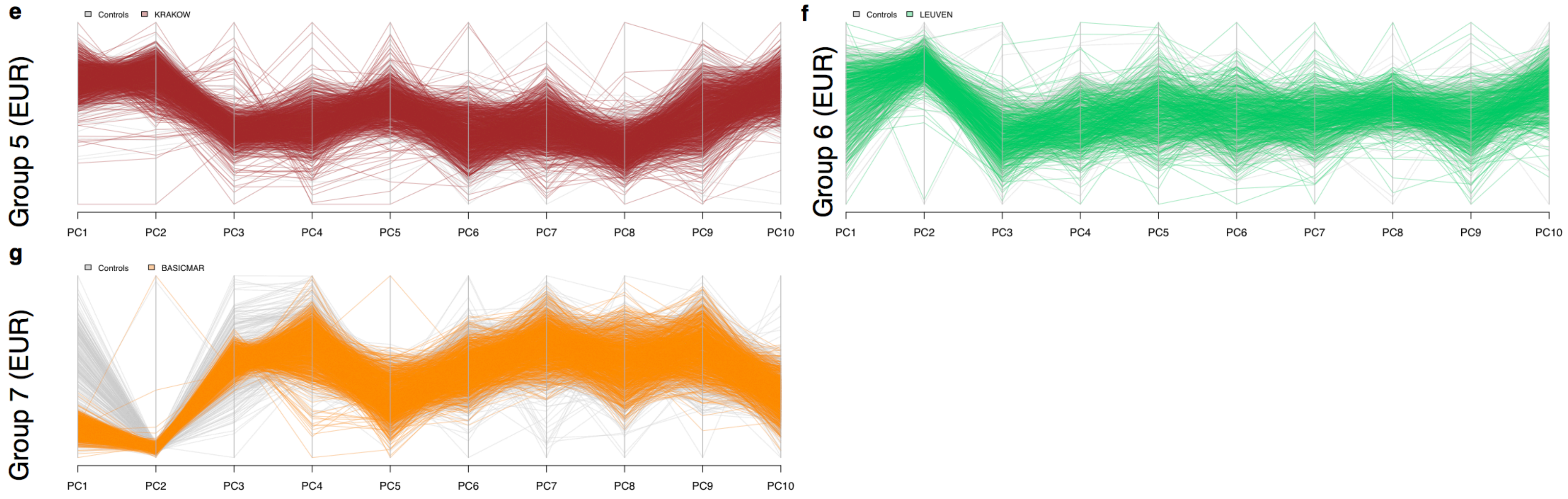


Supplementary Figure 16d (continued)

(d) Group 4 AFR: Controls are from HRS and OAI. Case cohorts: GASROS, GCKNSS, ISGS, MCISS, MIAMISR, NOMAS, REGARDS, SPS3, WHI, WUSTL.

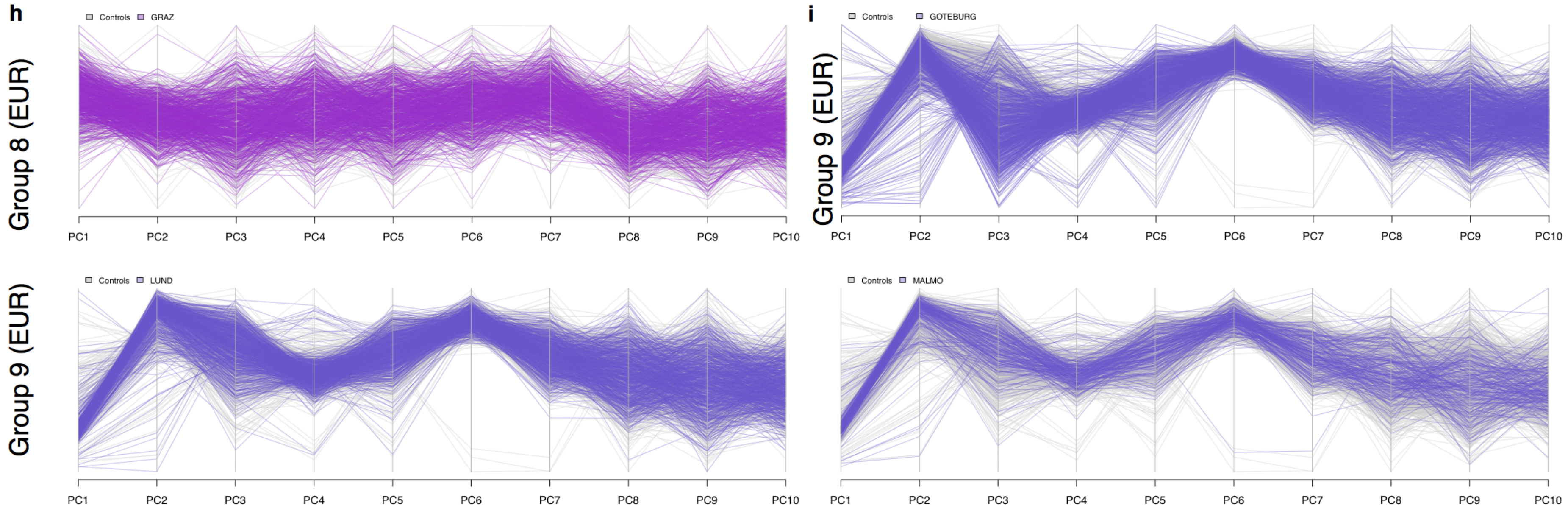


Supplementary Figure 16 (continued)



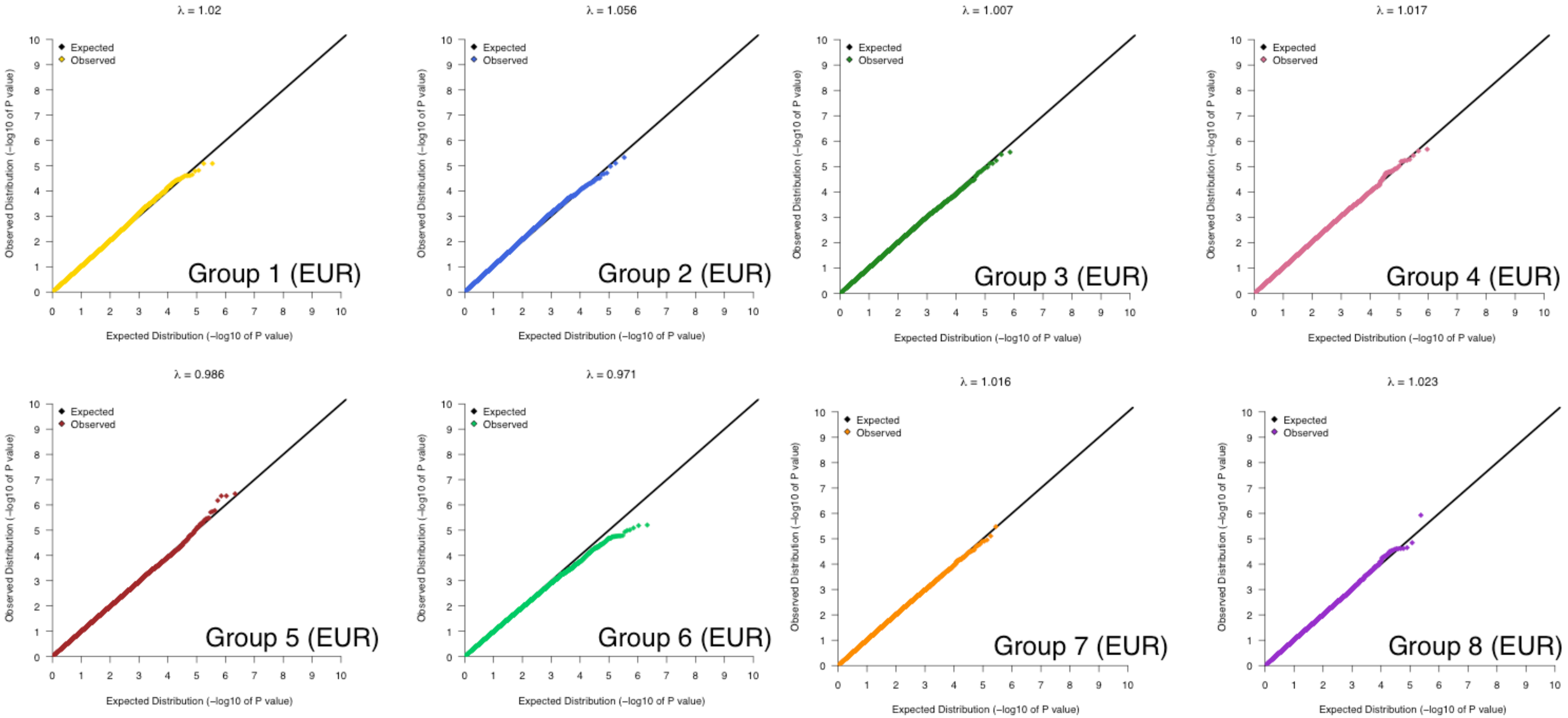
Supplementary Figure 16 (continued)

(e) Group5 EUR: Controls and cases are from KRAKOW.
(f) Group 6 EUR: Controls and cases are from LSGS.
(g) Group 7 EUR: Controls are from INMA and ADHD. Case group: BASICMAR.



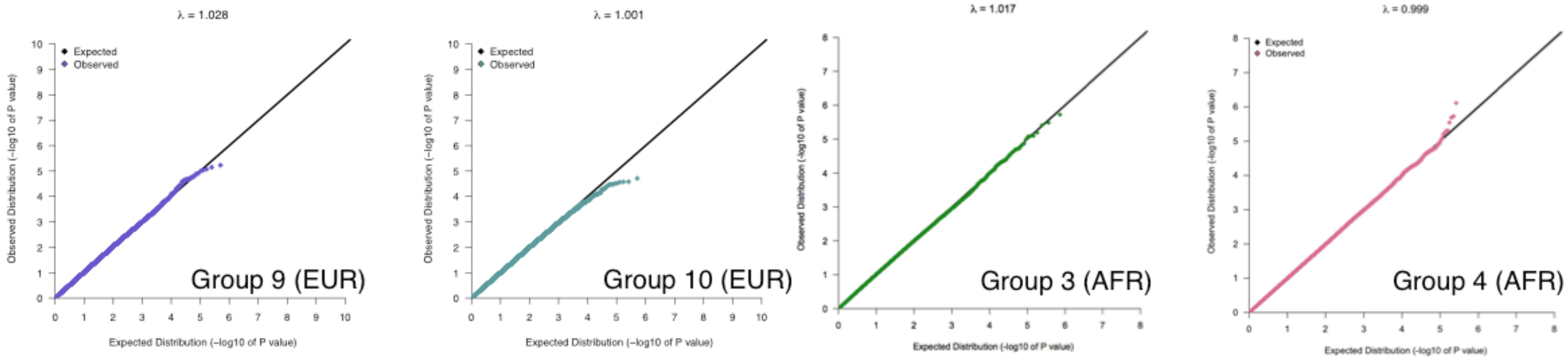
Supplementary Figure 16 (continued)

(h) Group 8 EUR: Controls and cases are from GRAZ.
(i) Group 9 EUR: Controls are from MDC. Case cohorts: MDC, LUND, SAHLSIS (here labeled as GOTEBURG).

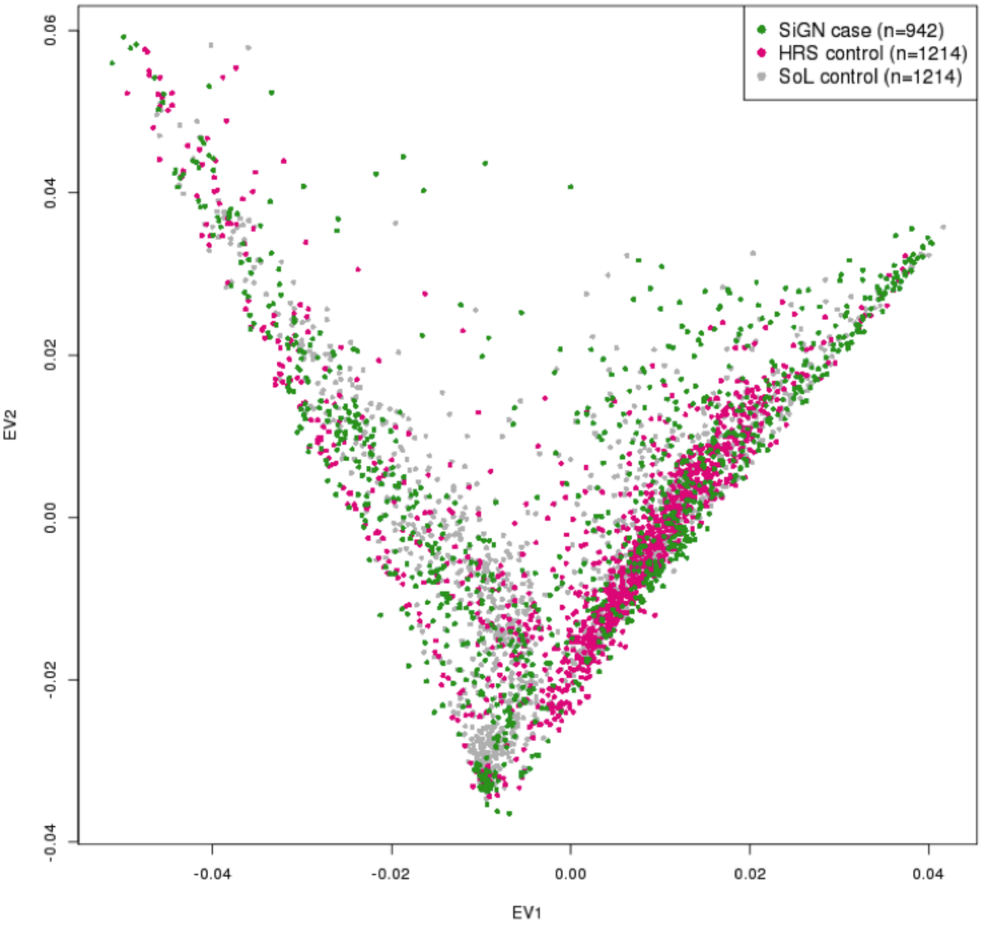


Supplementary Figure 17 | QQ plots of all stroke association testing in each stratum after sample and SNP quality control

Association testing was performed using logistic regression, adjusting for the top ten principal components and sex. A well-behaved lambda (approximately < 1.05) indicated successful quality control (**Section 9.5**). EUR, European-ancestry strata; AFR, African-ancestry strata.

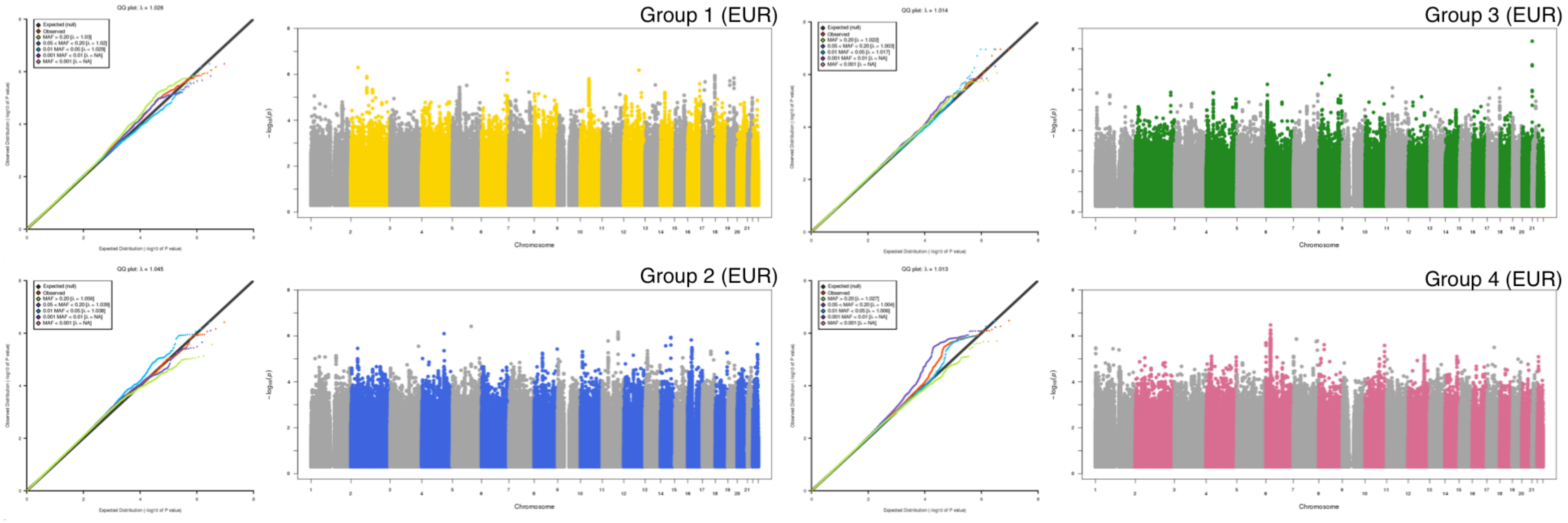


Supplementary Figure 17 (continued)



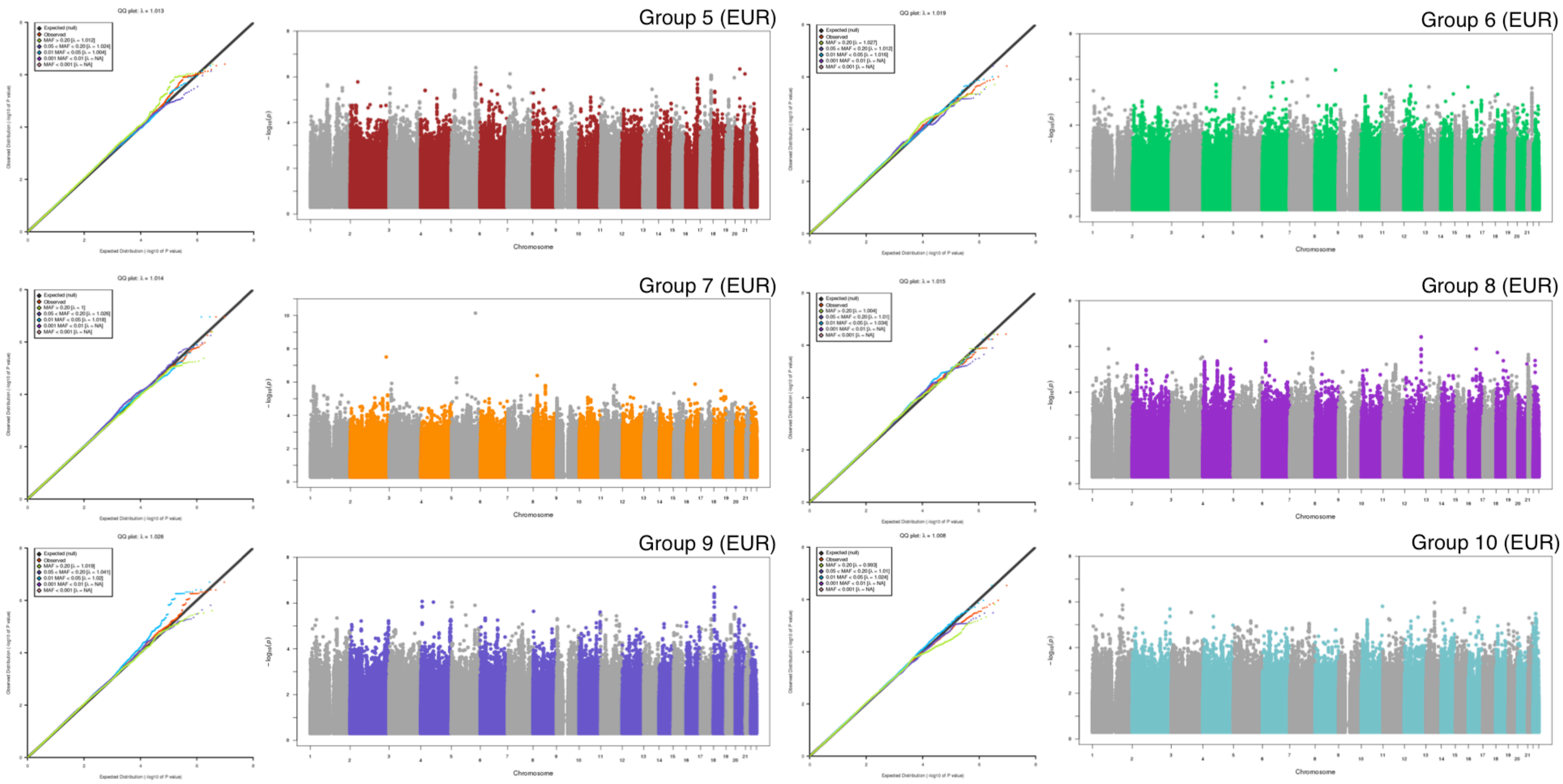
Supplementary Figure 18 | Final principal component analysis of unrelated Hispanic/Latino individuals in association tests

After all quality control was complete on Hispanic/Latino SiGN cases and HRS and HCHS/SOL controls, we calculated principal components in the cleaned set of samples. PC1 and PC2 are shown for the 3,371 HIS samples passing QC, color-coded by cohort. See **Section 9.5** for analysis details**.** EV, eigenvector.

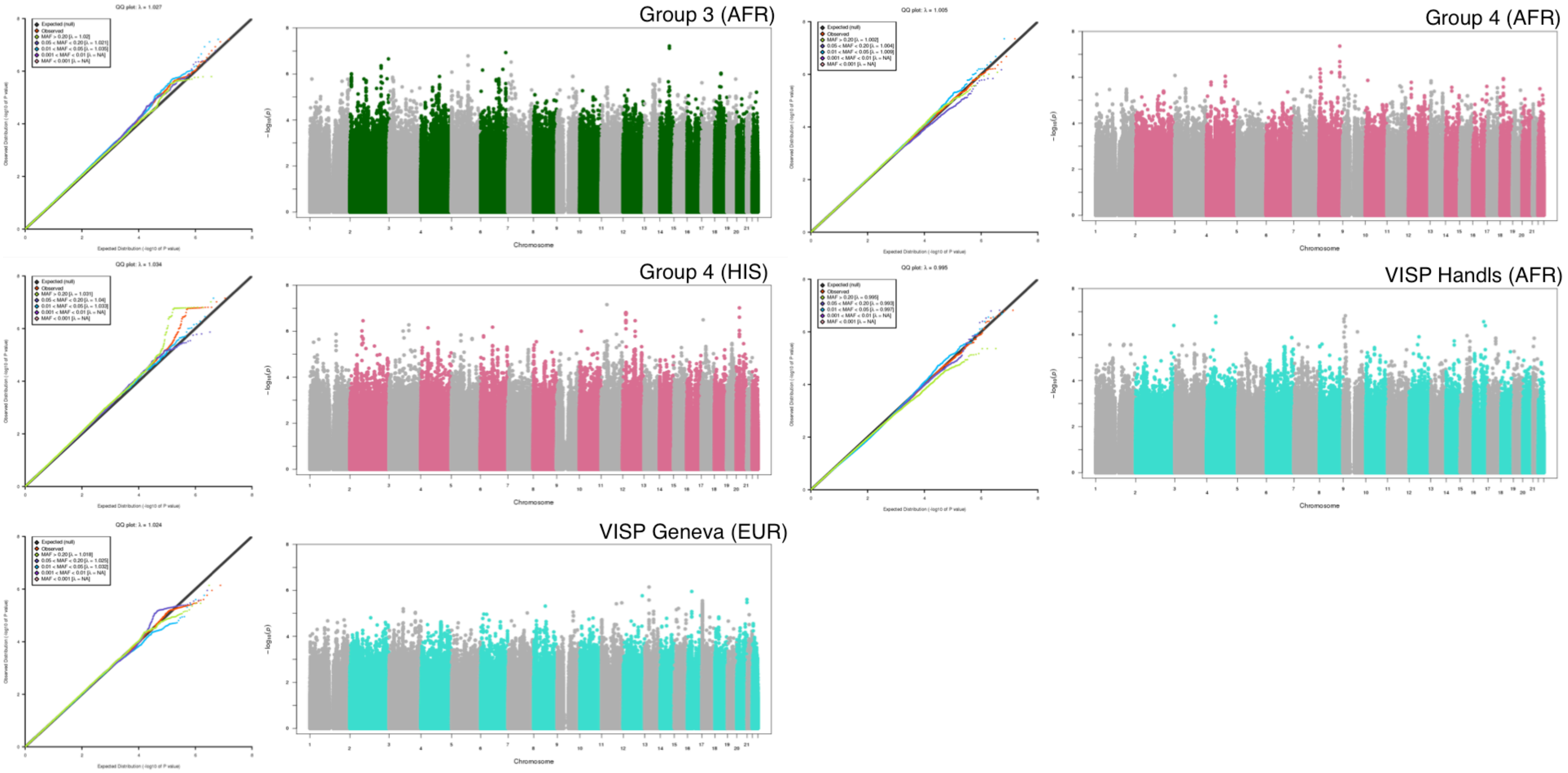


Supplementary Figure 19 | QQ and Manhattan plots of stage I (discovery) strata after imputation and quality control

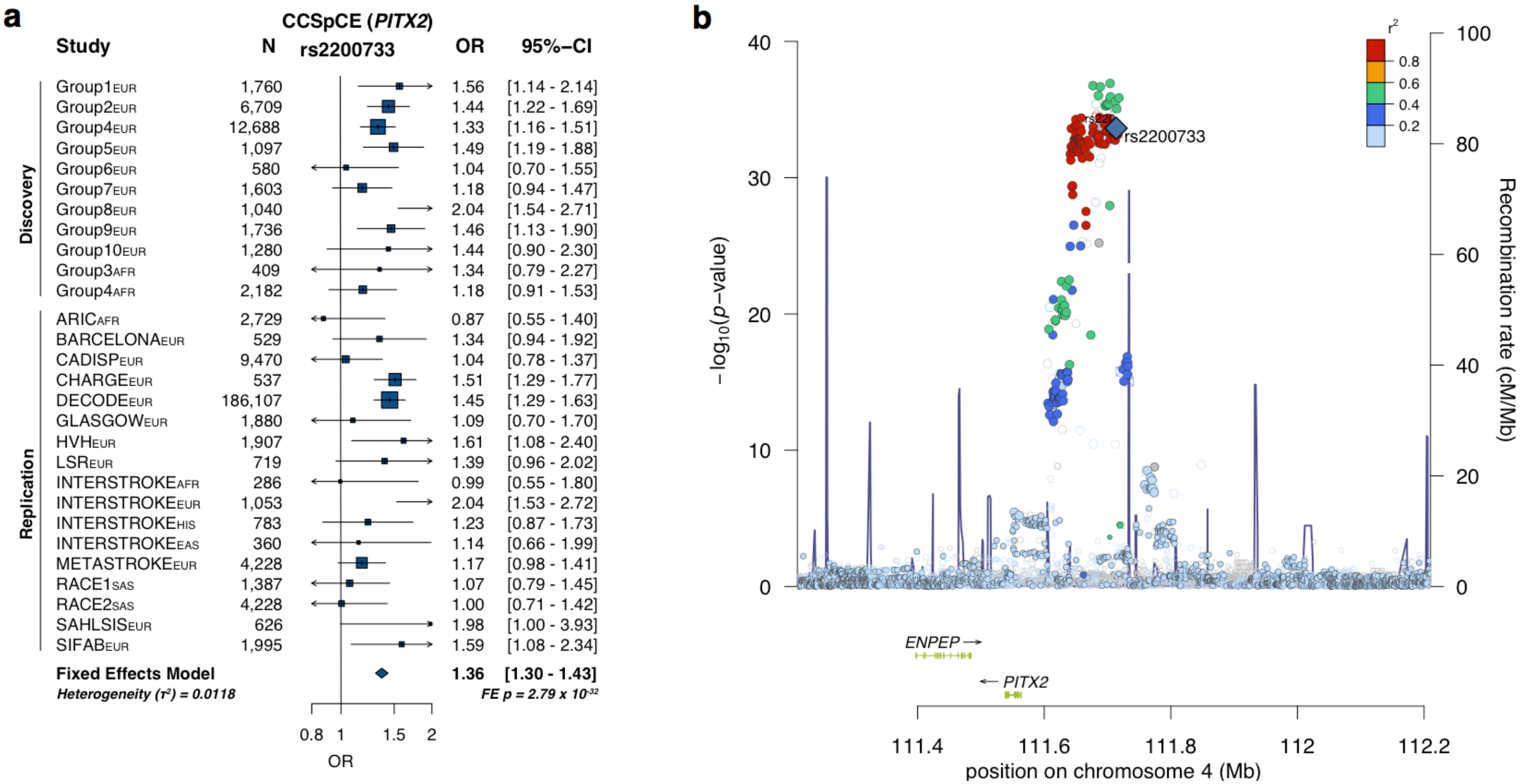
Results from genome-wide association testing in the all stroke (IS) phenotype are shown (**Section 12**). QQ plots show all SNPs (red) as well as SNPs stratified into frequency bins of > 20% (light green), 5 – 20% (purple), and 1 – 5% (light blue). Strata for 0.1 – 1%, and < 0.1% are also indicated in the legends but are absent because all SNPs with minor allele frequency < 1% were removed as a post-imputation QC filter. EUR, European ancestry; AFR, African ancestry.



Supplementary Figure 19 (continued)

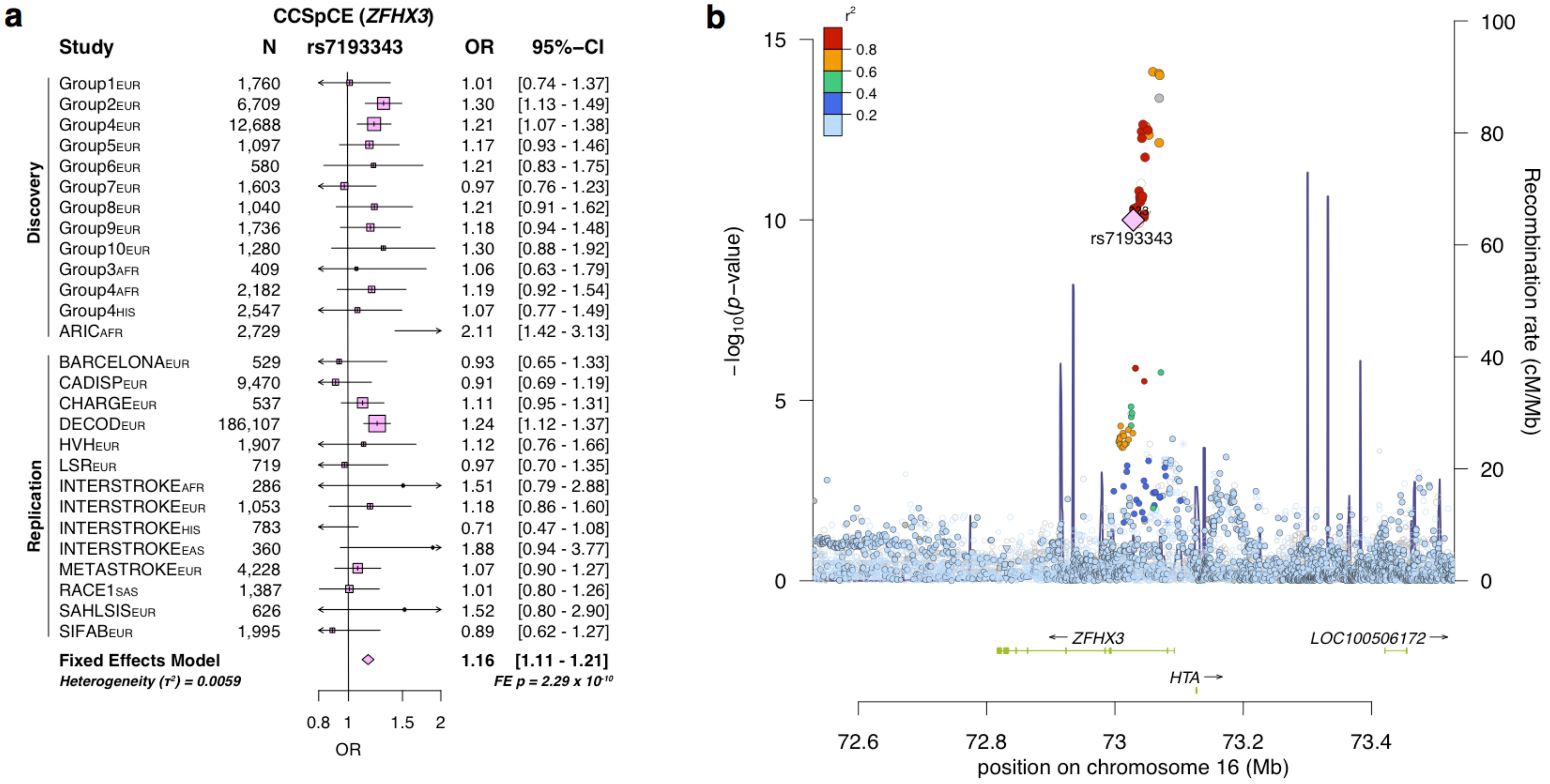


Supplementary Figure 19 (continued)



Supplementary Figure 20 | (a) Forest and (b) regional association plot of *PITX2* in analysis of the CCS Phenotypic cardioembolic (CCSpCE) subtype (Section 15)

The SNP shown is the original SNP shown to be associated to the subtype. OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic; EAS, East Asian ancestry; SAS, South Asian ancestry.

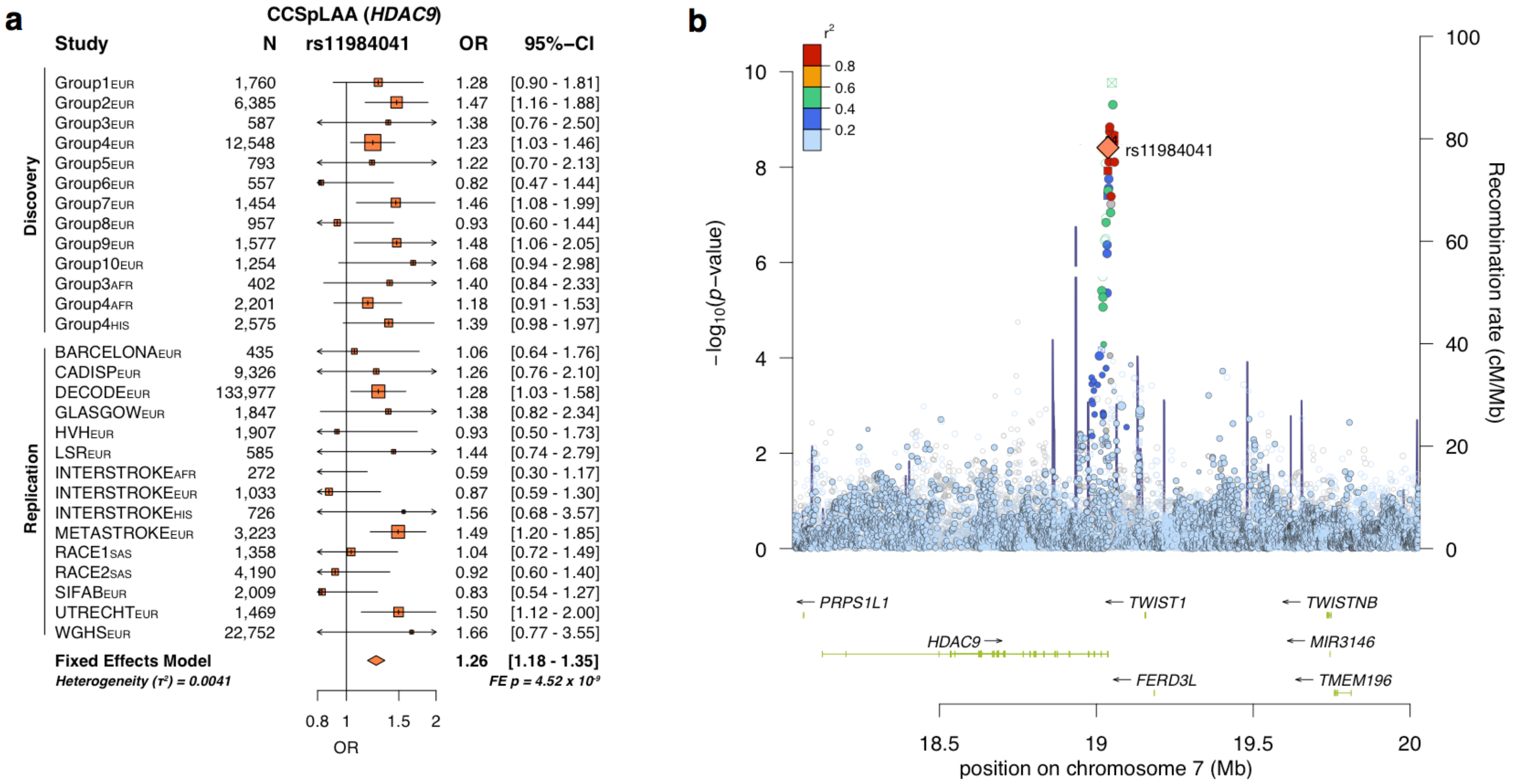


	Study	N	rs7193343	OR	95%–CI
Discovery	Group1EUR	1,760		1.01	[0.74 - 1.37]
	Group2EUR	6,709		1.30	[1.13 - 1.49]
	Group4EUR	12,688		1.21	[1.07 - 1.38]
	Group5EUR	1,097		1.17	[0.93 - 1.46]
	Group6EUR	580		1.21	[0.83 - 1.75]
	Group7EUR	1,603		0.97	[0.76 - 1.23]
	Group8EUR	1,040		1.21	[0.91 - 1.62]
	Group9EUR	1,736		1.18	[0.94 - 1.48]
	Group10EUR	1,280		1.30	[0.88 - 1.92]
	Group3AFR	409		1.06	[0.63 - 1.79]
	Group4AFR	2,182		1.19	[0.92 - 1.54]
	Group4HIS	2,547		1.07	[0.77 - 1.49]
	ARICAFR	2,729		2.11	[1.42 - 3.13]
Replication	BARCELONAEUR	529		0.93	[0.65 - 1.33]
	CADISPEUR	9,470		0.91	[0.69 - 1.19]
	CHARGEEUR	537		1.11	[0.95 - 1.31]
	DECODEUR	186,107		1.24	[1.12 - 1.37]
	HVHEUR	1,907		1.12	[0.76 - 1.66]
	LSREUR	719		0.97	[0.70 - 1.35]
	INTERSTROKEAFR	286		1.51	[0.79 - 2.88]
	INTERSTROKEEUR	1,053		1.18	[0.86 - 1.60]
	INTERSTROKEHIS	783		0.71	[0.47 - 1.08]
	INTERSTROKEEAS	360		1.88	[0.94 - 3.77]
	METASTROKEEUR	4,228		1.07	[0.90 - 1.27]
	RACE1SAS	1,387		1.01	[0.80 - 1.26]
	SAHLSISEUR	626		1.52	[0.80 - 2.90]
	SIFABEUR	1,995		0.89	[0.62 - 1.27]
	Fixed Effects Model			**1.16**	**[1.11 - 1.21]**

Heterogeneity (τ^2) = 0.0059 ***FE p = 2.29 x 10^{-10}***

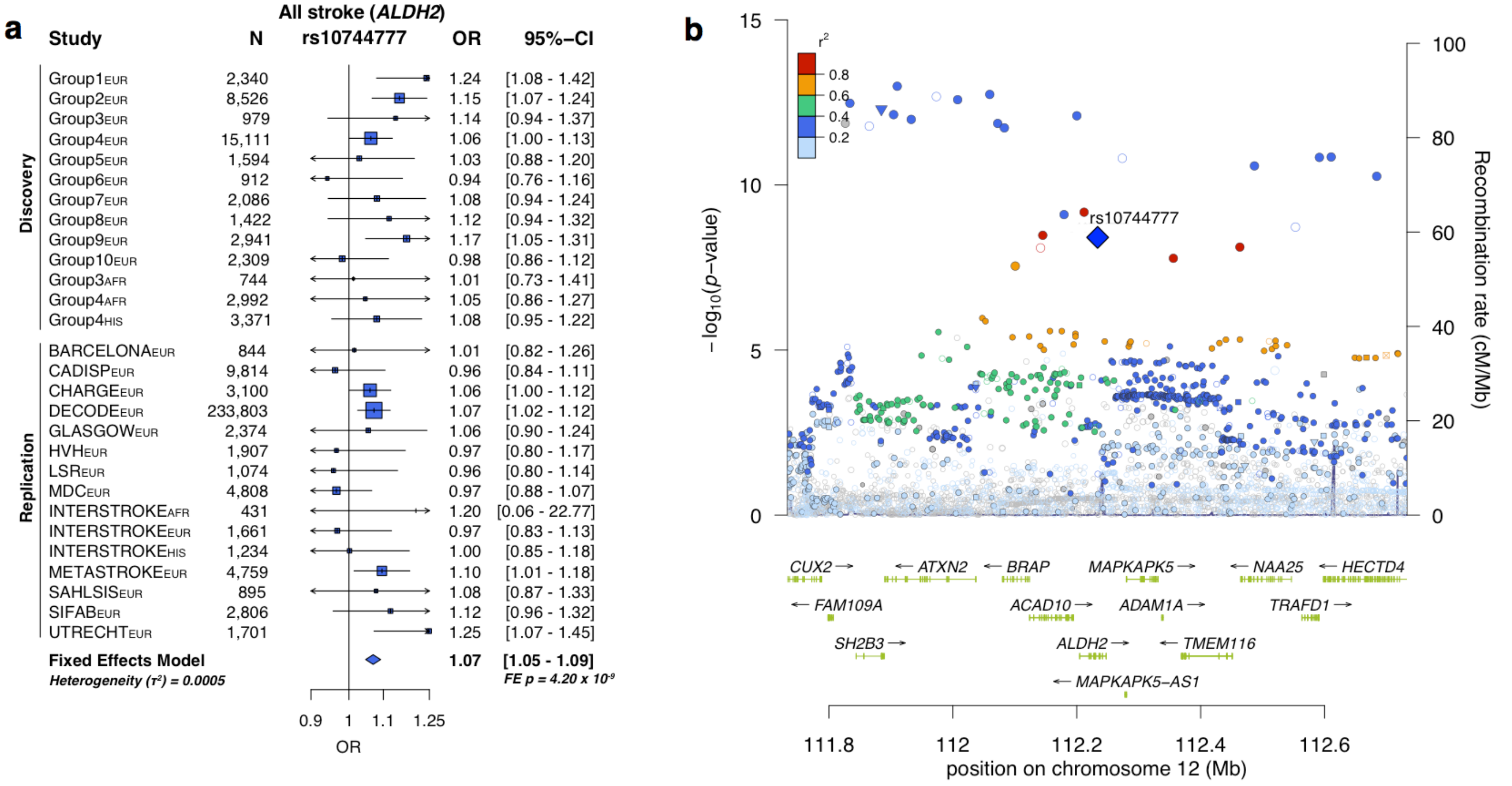
Supplementary Figure 21 | Forest and regional association plot of *ZFHX3* in the CCS Phenotypic cardioembolic (CCSpCE) subtype (Section 15)

The SNP shown is the original SNP shown to be associated to the subtype. OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic; EAS, East Asian ancestry; SAS, South Asian ancestry.



Supplementary Figure 22 | Forest and regional association plot of *HDAC9* in the CCS Phenotypic large artery atherosclerosis (CCSpLAA) subtype (Section 15)

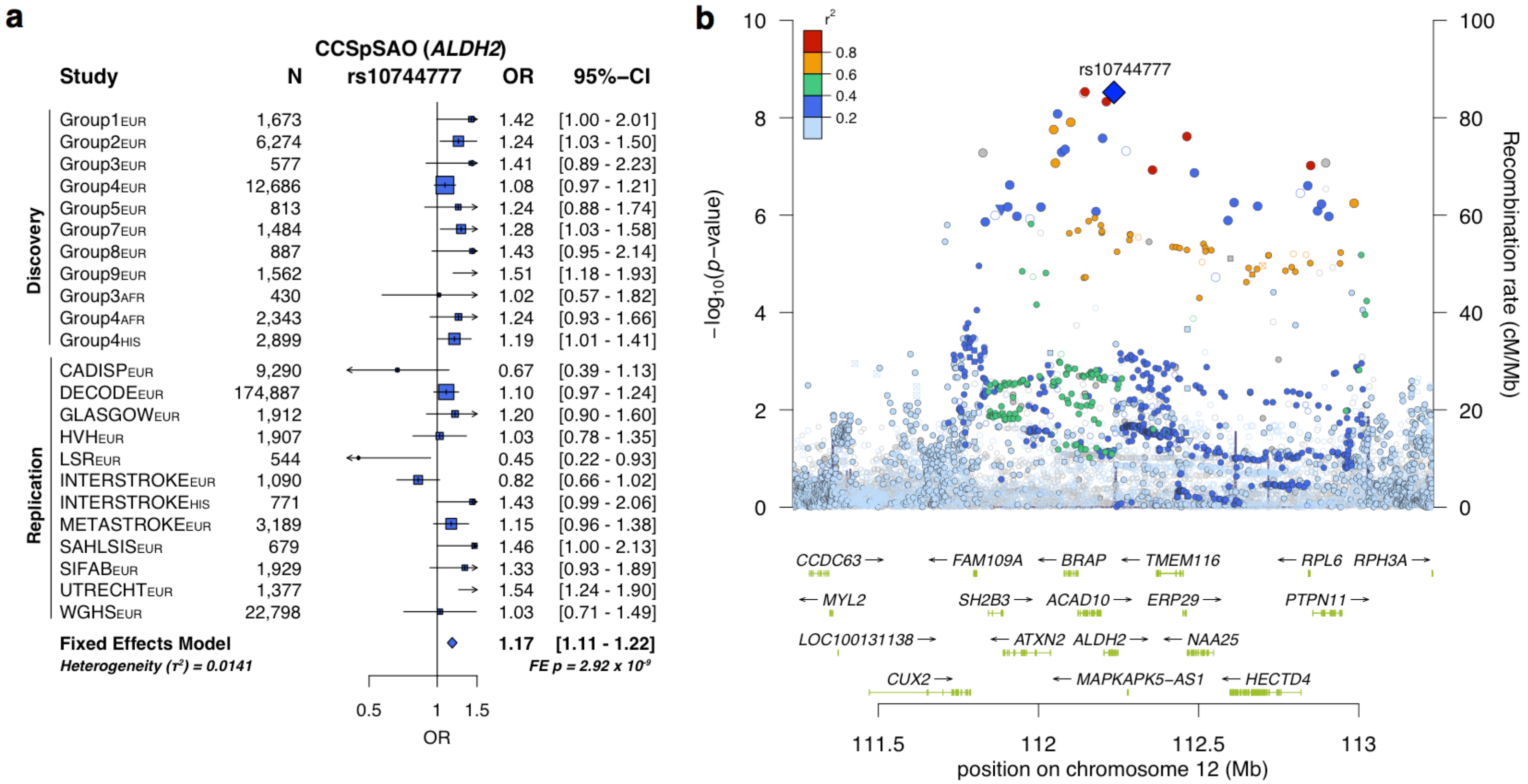
The SNP shown is the original SNP shown to be associated to the subtype. OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic; EAS, East Asian ancestry; SAS, South Asian ancestry.



Study	N	OR	95%–CI
Group1EUR	2,340	1.24	[1.08 - 1.42]
Group2EUR	8,526	1.15	[1.07 - 1.24]
Group3EUR	979	1.14	[0.94 - 1.37]
Group4EUR	15,111	1.06	[1.00 - 1.13]
Group5EUR	1,594	1.03	[0.88 - 1.20]
Group6EUR	912	0.94	[0.76 - 1.16]
Group7EUR	2,086	1.08	[0.94 - 1.24]
Group8EUR	1,422	1.12	[0.94 - 1.32]
Group9EUR	2,941	1.17	[1.05 - 1.31]
Group10EUR	2,309	0.98	[0.86 - 1.12]
Group3AFR	744	1.01	[0.73 - 1.41]
Group4AFR	2,992	1.05	[0.86 - 1.27]
Group4HIS	3,371	1.08	[0.95 - 1.22]
BARCELONAEUR	844	1.01	[0.82 - 1.26]
CADISPEUR	9,814	0.96	[0.84 - 1.11]
CHARGEEUR	3,100	1.06	[1.00 - 1.12]
DECODEEUR	233,803	1.07	[1.02 - 1.12]
GLASGOWEUR	2,374	1.06	[0.90 - 1.24]
HVHEUR	1,907	0.97	[0.80 - 1.17]
LSREUR	1,074	0.96	[0.80 - 1.14]
MDCEUR	4,808	0.97	[0.88 - 1.07]
INTERSTROKEAFR	431	1.20	[0.06 - 22.77]
INTERSTROKEEUR	1,661	0.97	[0.83 - 1.13]
INTERSTROKEHIS	1,234	1.00	[0.85 - 1.18]
METASTROKEEUR	4,759	1.10	[1.01 - 1.18]
SAHLSISEUR	895	1.08	[0.87 - 1.33]
SIFABEUR	2,806	1.12	[0.96 - 1.32]
UTRECHTEUR	1,701	1.25	[1.07 - 1.45]
Fixed Effects Model		**1.07**	**[1.05 - 1.09]**
Heterogeneity (τ^2) = 0.0005			***FE p = 4.20 x 10^{-9}***

Supplementary Figure 23 | (a) Forest and (b) regional association plot of 12q14.12 in all ischemic stroke (Section 15)

The SNP shown is the original SNP shown to be associated to all stroke. OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic; EAS, East Asian ancestry; SAS, South Asian ancestry.

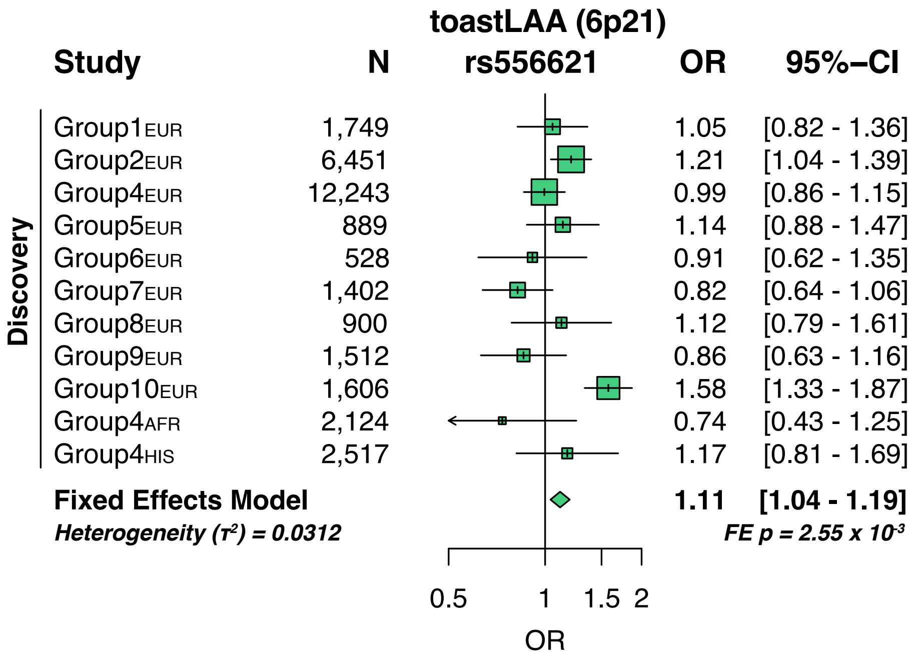


CCSpSAO (*ALDH2*)

Study	N	rs10744777	OR	95%–CI
Discovery				
Group1EUR	1,673		1.42	[1.00 - 2.01]
Group2EUR	6,274		1.24	[1.03 - 1.50]
Group3EUR	577		1.41	[0.89 - 2.23]
Group4EUR	12,686		1.08	[0.97 - 1.21]
Group5EUR	813		1.24	[0.88 - 1.74]
Group7EUR	1,484		1.28	[1.03 - 1.58]
Group8EUR	887		1.43	[0.95 - 2.14]
Group9EUR	1,562		1.51	[1.18 - 1.93]
Group3AFR	430		1.02	[0.57 - 1.82]
Group4AFR	2,343		1.24	[0.93 - 1.66]
Group4HIS	2,899		1.19	[1.01 - 1.41]
Replication				
CADISPEUR	9,290		0.67	[0.39 - 1.13]
DECODEEUR	174,887		1.10	[0.97 - 1.24]
GLASGOWEUR	1,912		1.20	[0.90 - 1.60]
HVHEUR	1,907		1.03	[0.78 - 1.35]
LSREUR	544		0.45	[0.22 - 0.93]
INTERSTROKEEUR	1,090		0.82	[0.66 - 1.02]
INTERSTROKEHIS	771		1.43	[0.99 - 2.06]
METASTROKEEUR	3,189		1.15	[0.96 - 1.38]
SAHLSISEUR	679		1.46	[1.00 - 2.13]
SIFABEUR	1,929		1.33	[0.93 - 1.89]
UTRECHTEUR	1,377		1.54	[1.24 - 1.90]
WGHSEUR	22,798		1.03	[0.71 - 1.49]
Fixed Effects Model			**1.17**	**[1.11 - 1.22]**
Heterogeneity (τ^2) = 0.0141			*FE p = 2.92 x 10⁻⁹*	

Supplementary Figure 24 | (a) Forest and (b) regional association plot of 12q14.12 small artery occlusion in the CCS Phenotypic (CCSpSAO) subtype (Section 15)

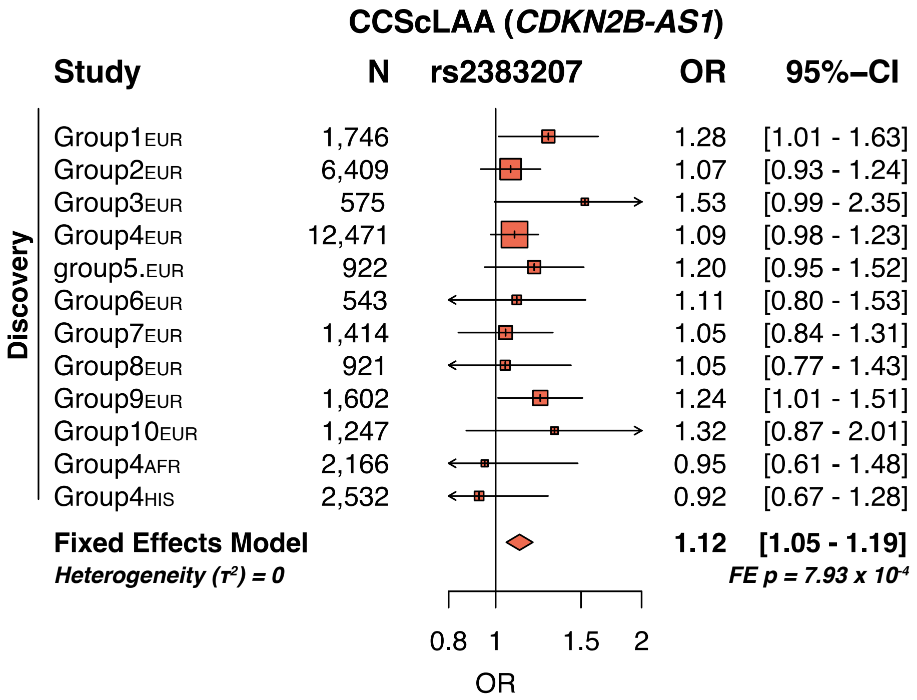
The SNP shown is the original SNP shown to be associated to all stroke that is now also associated to SAO in SiGN. OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic; EAS, East Asian ancestry; SAS, South Asian ancestry.



Supplementary Figure 25 | Forest plots for 6p21, failing to reach $p < 1 \times 10^{-6}$ after discovery (stage I)

Previously described loci failing to reach $p < 1 \times 10^{-6}$ (**Section 15.1**) after discovery and not pursued for replication included 6p21, previously implicated in large artery atherosclerosis (LAA). Data are shown for the subtyping method (TOAST) that yielded the most significant p-value.

OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic.



CCScLAA (*CDKN2B-AS1*)

Study	N	OR	95%–CI
Group1EUR	1,746	1.28	[1.01 - 1.63]
Group2EUR	6,409	1.07	[0.93 - 1.24]
Group3EUR	575	1.53	[0.99 - 2.35]
Group4EUR	12,471	1.09	[0.98 - 1.23]
group5.EUR	922	1.20	[0.95 - 1.52]
Group6EUR	543	1.11	[0.80 - 1.53]
Group7EUR	1,414	1.05	[0.84 - 1.31]
Group8EUR	921	1.05	[0.77 - 1.43]
Group9EUR	1,602	1.24	[1.01 - 1.51]
Group10EUR	1,247	1.32	[0.87 - 2.01]
Group4AFR	2,166	0.95	[0.61 - 1.48]
Group4HIS	2,532	0.92	[0.67 - 1.28]
Fixed Effects Model		**1.12**	**[1.05 - 1.19]**

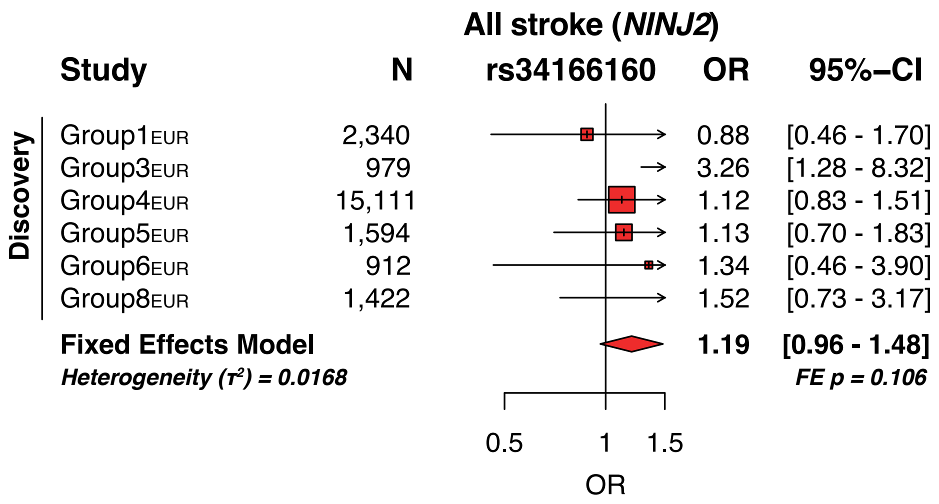
Heterogeneity (τ^2) = 0

FE p = 7.93 x 10^{-4}

Supplementary Figure 26 | Forest plots for *CDKN2B-AS1*, failing to reach p < 1 x 10^{-6} after discovery (stage I)

Previously described loci failing to reach p < 1 x 10^{-6} (**Section 15.1**) after discovery and not pursued for replication included *CDKN2B-AS1*, previously implicated in large artery atherosclerosis (LAA). Data are shown for the subtyping method (CCS Causative, CCSc) that yielded the most significant p-value.

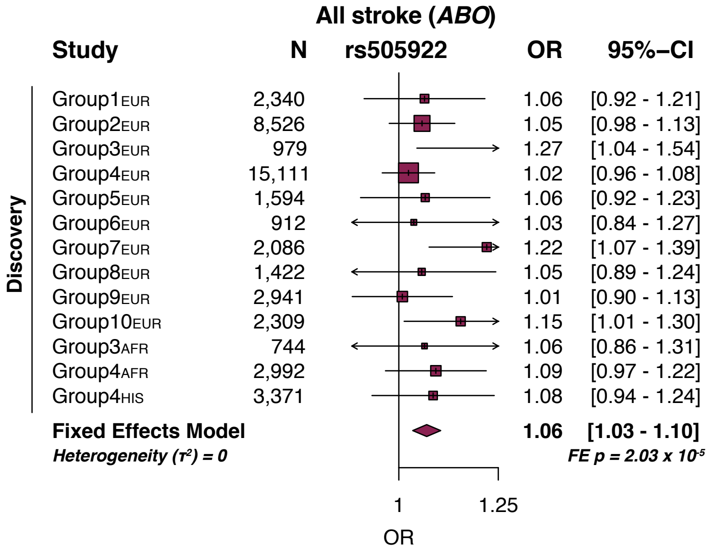
OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic.

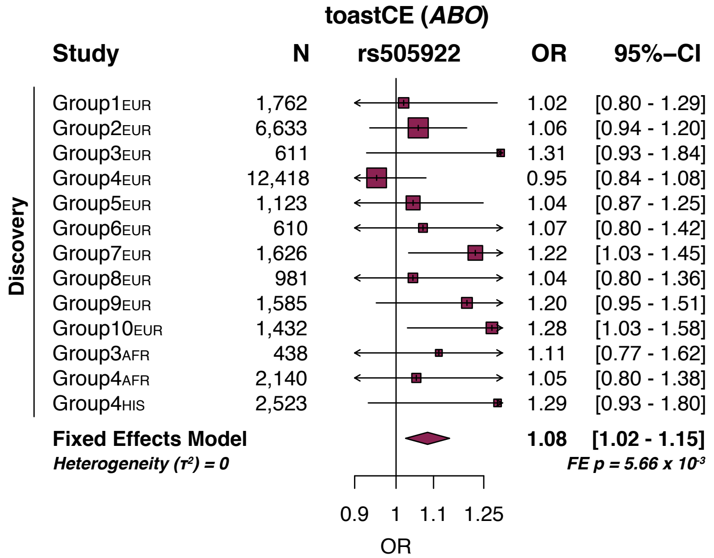


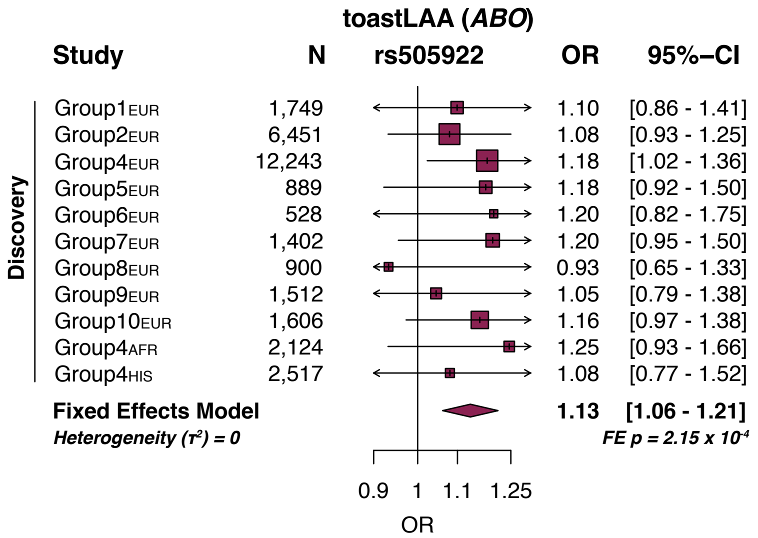
Supplementary Figure 27 | Forest plots for *NINJ2*, failing to reach p < 1 x 10^{-6} after discovery (stage I)

Previously described loci failing to reach p < 1 x 10^{-6} (**Section 15.1**) after discovery and not pursued for replication included *NINJ2*, previously implicated in all stroke.

OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic.







Supplementary Figure 28 | Forest plots for *ABO*, failing to reach p $< 1 \times 10^{-6}$ after discovery (stage I)

Previously described loci failing to reach p $< 1 \times 10^{-6}$ (**Section 15.1**) after discovery and not pursued for replication included *ABO*, previously implicated in (a) all stroke, (b) cardioembolic stroke (CE), and large artery atherosclerosis (LAA). Data are shown for the subtyping method (TOAST) that yielded the most significant p-value.

OR, odds ratio; FE p, fixed effects p-value; EUR, European ancestry; AFR, African ancestry; HIS, Hispanic.

II. Supplementary Tables

Cohort	Pop	IS		CE		LAA		SAO		UNDETER	
		Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
ARIC	AFR	263	2,466	51	2,678	--	--	--	--	--	--
CADISP	EUR	555	9,259	211	9,259	67	9,259	31	9,259	228	9,259
CHARGE	EUR	3,100	75,530	537	46,538	--	--	--	--	--	--
CHS	AFR	110	623	--	--	--	--	--	--	--	--
deCODE	EUR	5,291	228,512	1,304	184,803	496	133,481	596	174,291	1,781	221,490
GLASGOW	EUR	599	1,775	105	1,775	72	1,775	137	1,775	--	--
HVH	EUR	577	1,330	92	1,815	62	1,845	175	1,732	208	1,699
INTERSTROKE	AFR	192	239	47	239	33	239	47	239	17	239
	EAS	219	329	31	329	34	329	133	329	17	329
	EUR	812	849	204	849	184	849	241	849	158	849
	HIS	548	686	97	686	40	686	85	686	149	686
LUND	EUR	546	528	191	528	57	528	16	528	253	528
MALMO	EUR	1,304	3,504	--	--	--	--	--	--	--	--
METASTROKE	EUR	1,729	3,030	276	3,030	193	3,030	159	3,030	--	--
RACE1	SAS	1,218	1,158	229	1,158	200	1,158	192	1,158	--	--
RACE2	SAS	1,167	4,035	193	4,035	155	4,035	122	4,035	--	--
REGARDS	AFR	258	2,094	--	--	--	--	--	--	--	--
SAHLSIS	EUR	299	596	30	596	--	--	83	596	125	596
SIFAP	EUR	981	1,825	170	1,825	184	1,825	104	1,825	331	1,825
SWISS/ISGS	AFR	173	389	--	--	--	--	--	--	--	--
UTRECHT	EUR	556	1,145	--	--	324	1,145	232	1,145	--	--
BARCELONA	EUR	545	320	223	320	121	320	--	--	181	320
WGHS[1]	EUR	440	22,725	93	22,725	27	22,725	73	22,725	249	22,725
Total	**--**	**21,042**	**340,222**	**3,991**	**260,463**	**2,249**	**183,229**	**2,426**	**224,202**	**3,697**	**260,545**

Supplementary Table 1 | Stage II (replication) cohorts in SiGN

Cohorts with summary-level information available for SNPs selected from SiGN stage I for stage II follow-up. All cohorts used the TOAST subtyping method (**Section 14.2**). Populations (Pop): AFR, African American or other African-ancestry admixed samples; EAS, East Asian ancestry; EUR, European ancestry; SAS, South Asian ancestry. (Sub)types: IS, ischemic stroke; CE, cardioembolic; LAA, larger artery atherosclerosis; SAO, small artery atherosclerosis; UNDETER, undetermined.

SNP	Chromosome: Position	Alleles (risk/other)	Locus	Discovery samples	Trait	Meta-analysis results (excluding stage I samples)	
						OR [95% CI]	P-value
rs34166160	12: 732595	A/C	*NINJ2*	CHARGE	IS	1.20 [0.96 – 1.48]	0.106
rs11833579	12: 775199	G/A	*NINJ2*	CHARGE	IS	1.02 [0.95 – 1.01]	0.215
rs556621	6: 44594159	T/G	6p21	ASGC	LAA	1.04 [0.96 – 1.12]	0.304
rs2383207	9: 22115959	G/A	*CDKN2B-AS1*	ISGS	LAA	1.09 [1.02 – 1.17]	8.97 x 10^{-3}
rs505922	9: 136149229	C/T	*ABO*	WTCCC	IS	1.07 [1.03 – 1.10]	2.46 x 10^{-4}
					LAA	1.15 [1.07 – 1.24]	2.46 x 10^{-4}
					CE	1.09 [1.02 – 1.16]	7.10 x 10^{-3}

Supplementary Table 2 | Discovery (stage I) meta-analysis results for *NINJ2*, *ABO*, *CDKN2B-AS1*, and 6p21 after excluding samples initially used to discover each locus

NINJ2, *ABO*, *CDKN2B-AS1*, and 6p21 have been previously implicated by genome-wide association studies as conferring risk to ischemic stroke (IS), large artery atherosclerosis (LAA) stroke, and/or cardioembolic (CE) stroke. The samples used to discover the *NINJ2* locus (CHARGE) were not included in the SiGN stage I analysis. However, cohorts used to discover 6p21 (ASGC), *CDKN2B-AS1* (ISGS) and *ABO* (ESS, MUNICH, OXVASC, STGEORGE, KORA, WTCCC) were included in the stage I analysis phase. To investigate if there was additional independent evidence for association to these loci in SiGN, we first dropped the samples used in the initial GWAS that reported them (“Discovery samples”) and then re-performed the discovery meta-analysis. TOAST-subtyped cases were used for this analysis (**Section 15.1**).

a.

		CE			LAA			SAO			Undetermined				
		C	P	T	C	P	T	C	P	T	C1	C2	C3	P	T
CE	C	–	0.907	0.700	0.112	0.120	0.113	0.117	0.120	0.117	0.137	0.102	0.112	0.080	0.172
	P	0.887	–	0.698	0.161	0.214	0.138	0.150	0.177	0.140	0.217	0.212	0.119	0.084	0.227
	T	0.619	0.609	–	0.139	0.166	0.118	0.139	0.152	0.122	0.299	0.176	0.274	0.119	0.154
LAA	C	-0.230	-0.199	-0.201	–	0.828	0.678	0.116	0.128	0.126	0.139	0.103	0.118	0.090	0.237
	P	-0.229	-0.121	-0.178	0.811	–	0.611	0.177	0.225	0.183	0.221	0.213	0.130	0.102	0.271
	T	-0.210	-0.221	-0.240	0.591	0.502	–	0.121	0.143	0.111	0.190	0.144	0.148	0.108	0.111
SAO	C	-0.247	-0.242	-0.231	-0.216	-0.170	-0.187	–	0.943	0.751	0.151	0.108	0.129	0.098	0.224
	P	-0.251	-0.214	-0.220	-0.208	-0.116	-0.169	0.911	–	0.734	0.202	0.183	0.131	0.098	0.236
	T	-0.254	-0.272	-0.286	-0.213	-0.172	-0.229	0.625	0.583	–	0.286	0.220	0.210	0.171	0.135
Undetermined	C1	-0.352	-0.273	-0.135	-0.307	-0.197	-0.193	-0.330	-0.258	-0.118	–	0.713	0.751	0.516	0.620
	C2	-0.221	-0.091	-0.142	-0.193	-0.046	-0.114	-0.208	-0.110	-0.025	0.629	–	0.117	0.086	0.390
	C3	-0.226	-0.255	-0.030	-0.198	-0.203	-0.131	-0.212	-0.217	-0.125	0.644	-0.190	–	0.684	0.533
	P	-0.149	-0.168	-0.131	-0.130	-0.134	-0.082	-0.140	-0.143	-0.064	0.424	-0.125	0.659	–	0.429
	T	-0.207	-0.159	-0.309	-0.090	-0.067	-0.248	-0.145	-0.134	-0.295	0.406	0.226	0.291	0.292	–

Supplementary Table 3 | Correlation of genome-wide discovery analysis z-scores and phenotypes from CCS Causative (C), CCS Phenotypic (P), and TOAST (T) subtyping systems

(a) Upper triangle (blue): correlation (Pearson’s r) of genome-wide z-scores generated from stage I meta-analyses of C, P, and T subtypes (**Section 14.1**). Lower triangle (orange): correlation (Pearson’s r) of phenotypes generated by the three subtyping systems. Within-subtype correlations appear in darker blue and darker orange. Heatmaps of correlations are show in **Figure 1** (within subtypes) **Supplementary Figure 8**. CE, cardioembolic; LAA, large artery atherosclerosis; SAO, small artery occlusion; C1, CCSc undetermined; C2, CCSc incomplete and unclassified; C3, CCSc cryptogenic and CE minor.

b.

		CE			LAA			SAO			Undetermined				
		C	P	T	C	P	T	C	P	T	C1	C2	C3	P	T
CE	C	–	0.88	0.618	-0.228	-0.228	-0.207	-0.246	-0.251	-0.254	-0.337	-0.218	-0.224	-0.128	-0.206
	P		–	0.608	-0.193	-0.118	-0.213	-0.238	-0.211	-0.271	-0.269	-0.088	-0.246	-0.135	-0.159
	T			–	-0.197	-0.175	-0.235	-0.23	-0.219	-0.285	-0.132	-0.138	-0.029	-0.108	-0.309
LAA	C				–	0.811	0.591	-0.215	-0.207	-0.21	-0.282	-0.193	-0.198	-0.119	-0.088
	P					–	0.502	-0.17	-0.115	-0.17	-0.182	-0.046	-0.203	-0.121	-0.066
	T						–	-0.186	-0.168	-0.226	-0.175	-0.114	-0.131	-0.076	-0.24
SAO	C							–	0.911	0.622	-0.31	-0.206	-0.212	-0.124	-0.144
	P								–	0.582	-0.244	-0.109	-0.215	-0.125	-0.133
	T									–	-0.114	-0.025	-0.123	-0.055	-0.294
Undetermined	C1										–	0.567	0.586	0.305	0.399
	C2											–	-0.19	-0.116	0.218
	C3												–	0.606	0.282
	P													–	0.238
	T														–

(b) Correlations (Cohen’s kappa) within the primary subtypes generated by the three subtyping systems (C, P, and T). Dark blue highlights within-subtype correlation. CE, cardioembolic; LAA, large artery atherosclerosis; SAO, small artery occlusion; C1, CCSc undetermined; C2, CCSc incomplete and unclassified; C3, CCSc cryptogenic and CE minor.

Test	Performed
Computed tomography (CT) of the brain	82%
Magnetic resonance (MR) imaging of the brain	62%
CT-angiography	20%
MR-angiography	39%
Catheter angiography	6%
Intracranial vascular imaging	48%
Extracranial vascular imaging	86%
Transcranial Doppler	25%
Carotid-vertebral Doppler	70%
Electrocardiography	87%
Transthoracic echocardiography	60%
Transesophageal echocardiography	19%
Transthoracic or transesophageal echocardiography	68%
Prolonged Ambulatory Cardiac Monitoring	13%

Supplementary Table 4 | Diagnostic Investigations in SiGN

Percentages presented in this table are based on data from 13,757 patients with documentation of source work-up (**Section 5.3**). Intracranial vascular imaging included CT-angiography, MR-angiography, or catheter angiography. Extracranial vascular imaging included CT-angiography, MR-angiography, catheter angiography, or carotid-vertebral Doppler.

Cohort	Pop	Total N		Cardioembolic			Large artery atherosclerosis			Small artery occlusion			Other			Undetermined				
		Cases	Controls	C	P	T	C	P	T	C	P	T	C	P	T	C_1	C_2	C_3	P	T
BRAINS	EUR	267	0	53	63	22	28	21	65	25	28	90	29	32	0	131	14	117	9	38
HABC	EUR	0	1586	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GASROS	EUR	111	0	20	38	45	28	44	24	10	17	6	17	24	14	36	21	15	6	20
ISGS	EUR	351	0	54	66	102	99	103	69	35	40	48	22	27	19	141	99	42	52	113
SWISS	EUR	25	0	5	7	7	5	6	5	0	2	5	0	1	1	15	10	5	3	7
ESS	EUR	566	0	93	112	71	65	3	53	30	31	114	8	11	0	370	2	368	1	323
KORA	EUR	0	804	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MUNICH	EUR	1131	0	281	326	320	276	315	326	125	154	104	37	44	0	383	282	101	180	381
OXVASC	EUR	457	0	117	146	126	40	20	48	60	68	94	1	1	1	239	25	214	11	188
ST-GEORGE	EUR	418	0	144	171	162	74	93	70	60	67	59	2	3	0	138	57	81	0	127
WTCCC	EUR	0	5150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GEOS	AFR	383	361	43	48	77	34	41	27	71	69	78	56	66	24	179	100	79	43	181
	EUR	460	519	24	31	92	56	68	37	57	58	56	84	96	33	237	167	70	96	242

Supplementary Table 5 | Cohort sample distributions by (PCA- and hyperellispoid-based) ancestry for the CCS Causative, CCS Phenotypic, and TOAST subtyping systems

Case and control cohorts with counts of available CCS causative (C), CCS phenotypic (P), and TOAST (T) cases. Table splits indicate how cases and controls were grouped for analysis, with the exception of Hispanic (HIS) samples, which were pooled together as a single analysis group. Only summary results from the VISP cohort were provided for the discovery phase and only for the all stroke phenotype. The "other" subtypes were not analyzed due to low total sample counts. Population (Pop) corresponds to analysis strata (**Section 9.3**). C_1, CCS Causative Undetermined subtype; C_2, CCS Causative Undetermined, Incomplete or Unclassified subtype; C_3, CCS Causative Undetermined, Cryptogenic and CE minor subtype.

Supplementary Table 5 (continued)

Cohort	Pop	Total N		Cardioembolic			Large artery atherosclerosis			Small artery occlusion			Other			Undetermined				
		Cases	Controls	C	P	T	C	P	T	C	P	T	C	P	T	C_1	C_2	C_3	P	T
BRAINS	EUR	104	0	36	40	8	22	17	4	14	14	4	4	4	0	28	16	12	16	7
	HIS	6	0	1	1	0	1	2	1	1	2	0	0	0	0	3	1	2	0	0
GASROS	AFR	12	0	1	1	5	2	2	2	4	4	3	0	0	0	5	5	0	2	2
	EUR	417	0	61	80	133	93	103	71	58	63	40	37	46	38	168	149	19	54	80
	HIS	27	0	2	2	6	9	9	7	6	7	7	2	2	1	8	6	2	3	4
GCNKSS	AFR	113	0	23	28	22	15	23	14	37	42	29	2	2	3	36	22	14	18	45
	EUR	363	0	96	115	95	77	88	67	98	105	81	5	6	5	87	57	30	48	115
	HIS	6	0	3	3	3	1	1	1	1	1	1	0	0	0	1	1	0	1	1
HRS	AFR	0	1341	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	EUR	0	8619	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HIS	0	1214	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISGS	AFR	110	0	12	16	21	17	21	13	11	15	30	1	1	1	69	52	17	32	45
	EUR	64	0	12	16	16	17	18	13	4	5	8	5	8	4	26	19	7	10	23
	HIS	4	0	0	0	0	3	3	1	0	0	1	0	0	0	1	1	0	1	2
MCISS	AFR	85	0	17	29	22	16	21	12	16	22	16	4	4	3	32	25	7	3	32
	EUR	478	0	119	187	124	151	191	96	54	87	62	12	23	14	142	82	60	26	182
	HIS	56	0	9	14	11	17	21	11	5	11	9	3	4	5	22	14	8	4	20
MIAMSR	AFR	90	0	8	12	15	27	31	22	26	25	29	8	14	6	21	18	3	8	16
	EUR	108	0	27	31	39	20	28	26	17	17	17	8	9	10	36	27	9	15	10
	HIS	96	0	24	27	31	21	25	25	9	9	15	9	12	6	33	31	2	12	15
NHS	EUR	313	0	65	71	0	28	34	0	38	45	0	4	4	0	178	95	83	34	0
	HIS	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
NOMAS	AFR	87	0	15	19	10	16	21	14	28	28	22	1	2	0	27	22	5	7	22
	EUR	77	0	27	33	13	15	20	6	14	14	10	5	5	0	16	11	5	6	22
	HIS	194	0	33	36	22	33	41	27	60	53	41	5	5	1	63	53	10	26	59
OAI	AFR	0	681	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	EUR	0	3201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RE-GARDS	AFR	126	0	16	20	17	29	28	20	26	29	23	6	6	5	49	35	14	11	49
	EUR	170	0	32	43	32	42	40	32	36	37	30	1	7	2	59	37	22	19	49
	HIS	8	0	1	3	1	2	2	2	1	2	2	1	1	0	3	0	3	0	2
HCHS/SOL	HIS	0	1214	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SPS3	AFR	136	0	0	1	0	0	11	0	134	130	136	0	0	0	2	1	1	1	0
	EUR	345	0	1	10	0	0	19	0	339	334	345	0	0	0	5	0	5	0	0
	HIS	468	0	1	9	0	5	28	0	436	368	468	0	0	0	26	14	12	14	0

Supplementary Table 5 (continued)

Cohort	Pop	Total N		Cardioembolic			Large artery atherosclerosis			Small artery occlusion			Other			Undetermined				
		Cases	Controls	C	P	T	C	P	T	C	P	T	C	P	T	C_1	C_2	C_3	P	T
SWISS	AFR	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
	EUR	174	0	24	31	22	55	52	46	6	10	53	9	11	8	80	48	32	33	45
	HIS	6	0	2	2	2	0	0	0	1	1	2	1	1	1	2	0	2	0	1
WHI	AFR	30	0	3	3	6	4	3	4	16	15	13	0	0	0	7	5	2	4	7
	EUR	414	0	105	136	116	67	61	62	118	113	111	3	7	10	121	69	52	38	113
	HIS	10	0	2	3	2	1	1	2	2	2	1	1	1	0	4	3	1	0	5
WUSTL	AFR	180	0	27	31	0	17	17	0	12	11	0	8	11	0	116	16	100	4	0
	EUR	264	0	70	75	0	64	57	0	22	22	0	8	9	0	100	24	76	14	0
	HIS	5	0	1	1	0	1	1	0	0	0	0	0	1	0	3	0	3	0	0
KRAKOW	EUR	878	716	326	381	407	206	77	173	85	97	36	27	25	23	234	92	142	25	239
	HIS	2	1	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	1
LSGS	EUR	459	453	109	127	157	90	104	75	30	34	55	30	30	23	200	178	22	76	149
	HIS	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ADHD	EUR	0	411	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
INMA	EUR	0	807	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BASICMAR	EUR	868	0	336	385	408	196	236	184	250	266	276	0	0	0	86	31	55	13	0
	HIS	22	0	11	12	13	5	7	3	5	6	6	0	0	0	1	0	1	0	0
GRAZ	EUR	0	815	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GRAZ	EUR	607	0	183	225	166	106	142	85	67	72	74	20	24	18	231	111	120	34	114
	HIS	2	0	0	0	1	0	1	1	0	1	0	0	0	0	2	1	1	1	0
SAHLSIS	EUR	767	0	139	151	166	115	116	120	89	84	106	102	112	92	322	181	141	105	282
	HIS	16	0	1	1	1	2	2	6	3	3	2	4	4	3	6	4	2	2	4
LSR	EUR	601	0	207	223	0	125	99	0	112	116	0	22	25	0	129	41	88	14	0
	HIS	12	0	4	4	0	0	1	0	3	3	0	1	1	0	4	1	3	0	0
MDC	EUR	211	1362	0	0	57	0	0	30	0	0	77	0	0	3	0	0	0	0	44

Supplementary Table 5 (continued)

Cohort	Pop	Total N		Cardioembolic			Large artery atherosclerosis			Small artery occlusion			Other			Undetermined				
		Cases	Controls	C	P	T	C	P	T	C	P	T	C	P	T	C_1	C_2	C_3	P	T
ASGC	EUR	1109	1200	69	80	232	47	54	406	28	29	291	6	8	13	93	48	45	25	167
VISP/ Handls,	AFR	256	971	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
VISP/ Melanoma	EUR	1723	1047	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

a. CCS Causative

	Trait	CE	LAA	SAO	UNDETER	CRYPTCE	INCUNC
CCSc	CE	3,095	0	0	0	0	0
	LAA	0	2,488	0	0	0	0
	SAO	0	0	2,796	0	0	0
	UNDETER	0	0	0	4,756	2,424	2,332
	CRYPTCE	0	0	0	2,424	2,424	0
	INCUNC	0	0	0	2,332	0	2,332
CCSp	CE	3,095	145	90	370	0	370
	LAA	15	2,143	112	312	0	312
	SAO	2	26	2,638	217	0	217
	Cryptogenic	0	0	0	1,160	1,160	0
TOAST	CE	2,287	104	78	702	489	213
	LAA	34	1,610	53	277	127	150
	SAO	33	49	2,088	682	238	444
	UNDETER	212	387	314	2,355	1,269	1,086

b. CCS Phenotypic

	Trait	CE	LAA	SAO	Cryptogenic
CCSp	CE	3,726	354	178	0
	LAA	354	2,595	246	0
	SAO	178	246	2,889	0
	Cryptogenic	0	0	0	1,160
CCSc	CE	3,095	15	2	0
	LAA	145	2,143	26	0
	SAO	90	112	2,638	0
	UNDETER	370	312	217	1,160
TOAST	CRYPTCE	0	0	0	1,160
	INCUNC	370	312	217	0
	SAO	78	158	2,026	143
	UNDETER	462	465	360	779

c. TOAST

	Trait	CE	LAA	SAO	UNDETER
TOAST	CE	3,427	0	0	0
	LAA	0	2,406	0	0
	SAO	0	0	3,186	0
	UNDETER	0	0	0	3,593
CCSc	CE	2,287	34	33	212
	LAA	104	1,610	49	387
	SAO	78	53	2,088	314
	UNDETER	702	277	682	2,355
	CRYPTCE	489	127	238	1,269
	INCUNC	213	150	444	1,086
CCSp	CE	2,512	78	78	462
	LAA	175	1,464	158	465
	SAO	120	102	2,026	360
	Cryptogenic	48	67	143	779

Supplementary Table 6 | Cross-subtype classification of SiGN cases

The CCS and TOAST subtyping methods are only moderately correlated; cases classified as a single subtype by CCSc or CCSp may be classified differently by TOAST (**Section 8**). Columns indicate the set of cases being looked up across all available subtypes for CCSc, CCSp, and TOAST. For example, of the 3,427 cases classified as CE in TOAST, 2,287 were also classified as CE in CCSc, but 104 were classified as LAA by CCSc.

Section		SNPs Kept	SNPs Lost	Filter component
9.3.2	Start	2,302,224	-	No filter. All SNP probes (intersection of 3 arrays) are present
9.1	Cohort-specific	2,159,647	142,577	CIDR or Illumina technical filters
9.1	Cohort-specific	2,080,909	78,738	Project-specific quality filter
9.1	Cohort-specific	2.080,645	264	Positional duplicates (genomic position)
9.3.2b	Cross-array	2,053,805	26,840	≥ 1 discordant genotype calls in any cross-project duplicate pair
9.3.2b	Cross-array	2,053,191	614	≥ 5 discordant missing calls (SiGN-HCHS/SOL, HCHS/SOL-HRS pairs) ≥ 7 discordant missing calls (SiGN-HRS)
9.4.6b	HIS-specific	2,022,094	31,097	Differential missingness tests, cases vs. controls (Fishers $p < 10^{-3}$)
9.4.6b	HIS-specific	2,016,884	5,210	MCR > 1% in cases
9.4.6b	HIS-specific	2,013,028	3,856	MCR > 1% in controls
9.4.6b	HIS-specific	1,952,932	60,096	MCR > 1% in HRS
9.4.6b	HIS-specific	1,918,116	34,816	MCR > 1% in HCHS/SOL
9.4.6b	HIS-specific	1,913,845	4,271	Pseudo-association tests, controls vs. controls (LR $p < 10^{-3}$)
10.2	HIS-prephase	1,908,773	5,072	Illumina HumanOmni5Exome-4v1 annotation v.A to v.B updates
	HIS-descriptive	1,841,646	67,127	MAF = 0 among unrelated HIS individuals
	HIS-descriptive	1,425,794	415,852	MAF < 0.01

Supplementary Table 7 | Summary of recommended SNP filters after completion of CIDR genotyping and QC

The number of SNPs lost is given for sequential application of the filters in the order given.

QC Type	QC Filter Name	Threshold for removal	Notes
Sample	Missingness	> 10%	
Sample	Sex check	(1) X chromosome inbreeding > 0.8, phenotype info indicates female (2) X chromosome inbreeding < 0.2, phenotype info indicates male	3 samples from GEOS with known sex chromosome anomalies were left in the analysis
Sample	PCA	NA	Only for establishing homogenous and multiethnic cohorts
Sample	Relatedness	Pi-hat > 0.5 (Sibship or duplicate samples)	Only applied to cohorts with homogeneous ancestry
SNP	Missingness	> 10%	
SNP	A/T and C/G	All A/T and C/G SNPs removed	
SNP	Duplicate markers	Marker with higher genotype missingness dropped	

Supplementary Table 8 | Steps and filters for quality control of individual cohorts after initially receiving data

Quality control (QC) thresholds at this step (**Section 9.1.3 – 9.1.4**) were intentionally liberal as sample- and SNP-level QC were performed again on array- and ancestry-specific groups of cases and controls.

PCA: principal component analysis with sample projection onto reference samples from HapMap 3.

Analysis Group	Cohort	Cases or controls	Array	Country
1	BRAINS	Cases	650Q	U.S.A.
1	GASROS	Cases	610	U.S.A.
1	ISGS	Cases	610	U.S.A.
1	SWISS	Cases	610	U.S.A.
1	HABC	Controls	1M	U.S.A.
2	ESS	Cases	660	U.K.
2	MUNICH	Cases	660	U.K.
2	OXVASC	Cases	660	U.K.
2	STGEORGE	Cases	660	U.K.
2	KORA	Controls	550	Germany
2	WTCCC	Controls	660	U.K.
3	GEOS	Cases, controls	1M	U.S.A.
4	CIDR*	Cases	5M	U.S.A.
4	HRS	Controls	2.5M	U.S.A.
4	OAI	Controls	2.5M	U.S.A.
4	HCHS/SOL	Controls	2.5M	U.S.A.
5	KRAKOW	Cases, controls	5M	Poland
6	LSGS	Cases, controls	5M	Belgium
7	BASICMAR	Cases	5M	Spain
7	ADHD	Controls	1M	Spain
7	INMA	Controls	1M	Spain
8	GRAZ	Cases	5M	Austria
8	GRAZ	Controls	610	Austria
9	SAHLSIS	Cases	5M	Sweden
9	LSR	Cases	5M	Sweden
9	MDC	Cases, controls	750K Express/exome	Sweden
10	ASGC	Cases, controls	610	Australia

Supplementary Table 9 | Case and control cohorts after array matching (Section 9.2)

Shading indicates array-specific analysis group.

a. HapMap 3

Population	Description	N
ASW	African ancestry in Southwest USA	90
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	180
CHB	Han Chinese in Beijing	90
CHD	Chinese in Metropolitan Denver, Colorado	100
GIH	Gujarati Indians in Houston, Texas	100
JPT	Japanese in Tokyo	91
LWK	Luhya in Webuye, Kenya	100
MXL	People with Mexican ancestry in Los Angeles, California	90
MKK	Maasai in Kinyawa, Kenya	180
TSI	Toscani in Italia	100
YRI	Yoruba in Ibadan, Nigeria	180

b. 1000 Genomes (Phase I)

Population	Description	N
ASW	African ancestry in Southwest USA	61
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	85
CHB	Han Chinese in Beijing	97
CHS	Han Chinese South	100
CLM	Colombians in Medellin, Colombia	60
FIN	Finnish in Finland	93
GBR	British from England and Scotland	89
IBS	Iberians in Spain	14
JPT	Japanese in Tokyo	89
LWK	Luhya in Webuye, Kenya	97
MXL	people with Mexican ancestry in Los Angeles, California	66
PUR	Puerto Ricans in Puerto Rico	55
TSI	Toscani in Italia	98
YRI	Yoruba in Ibadan, Nigeria	88

c. Genome of the Netherlands

Population	Description	N
GoNL	Dutch trios and quartets	769 (499 unrelated)

Supplementary Table 10 | Reference data used for quality control and imputation

Description of populations represented in (a) HapMap 3 (SNP array data), (b) 1000 Genomes Phase 1 (sequencing data), and (c) Genome of the Netherlands (sequencing data), projects that were used for quality control and imputation of cases and controls (**Section 9**).

QC Type	QC Filter Name	Threshold for removal	Notes
Sample	Missingness	> 5%	
Sample	PCA	Outliers removed on a per-stratum basis	
Sample	Inbreeding	> 3 standard deviations from mean of distribution	
Sample	Relatedness	Kinship > 0.0625 (cousins or higher levels of relatedness)	KING method for AFR strata
SNP	Missingness	> 1%	Filter applied for cases only, controls only, and all samples
SNP	Frequency	< 1%	
SNP	HWE	$p < 1 \times 10^{-3}$	
SNP	Case/case and control/control cohort comparisons	$p < 1 \times 10^{-3}$	
SNP	Differential missingness	$p < 1 \times 10^{-3}$	
SNP	1KG comparison	$p < 1 \times 10^{-3}$	Association testing between SiGN cohorts and 1KG; only done for EUR strata
SNP	Cross-chip concordance mismatch	≥ 1 mismatch	

Supplementary Table 11 | Steps and filters for quality control of array-specific strata for European- and African-ancestry samples (Section 9.4)

PCA, principal component analysis; AFR, African-ancestry samples; HWE, Hardy-Weinberg Equilibrium; 1KG, 1000 Genomes Project Phase I data.

Cohort	1st Array	2nd Array	N	Overlapping SNPs
INMA	5M	Omni 1M	30	619,742
HRS	5M	Omni 2.5M	30	2,326,361
GRAZ	5M	Illumina 610	28	451,565
OAI	5M	Omni 2.5M	30	2,318,057
LSR (1)	5M	Omni Express 750K v1.0	25	690,460
LSR (2)	5M	Omni Express 750K v1.1	5	689,916

Supplementary Table 12 | Cross-study duplicate analysis to identify SNPs with genotyping discordance

A number of samples genotyped on the Illumina 5M as part of the CIDR genotyping effort had also been genotyped on another (smaller) array. These cross-study duplicates were used to analyze genotype concordance across different genotyping runs by examining SNPs in the intersect of the 5M and the 2nd array (**Section 9.4.1**). We removed SNPs with ≥ 1 discordant genotype from the analysis.

Filter type, exclude if	SiGN	HCHS/SOL	HRS
Sample MCR	> 2%	> 2%	> 2%
SNP MCR	> 2%	> 2%	> 2%
HWE among homogeneous	$p < 10^{-4}$ in KRAKOW or LSGS controls	meta-$p < 10^{-5}$, HIS subgroups	$p < 10^{-4}$, EUR or AFR samples
Mendel errors among	> 1 error, 24 HapMap trios	> 3 errors, 298 trios + 1,043 PO pairs	> 1 error, 25 HapMap trios
Duplicate discordance	> 2 among 248 pairs	> 2 among 291 pairs	> 4 among 423 pairs

Supplementary Table 13 | Summary of initial project-specific QC filters for Hispanic samples in SiGN, HCHS/SOL, and HRS

The thresholds applied for each type of filter in QC of Hispanic samples (**Section 9.1**). MCR, missing call rate; HWE, Hardy-Weinberg Equilibrium; EUR, European-ancestry; AFR, African-ancestry; PO, parent-offspring. Duplicate discordance refers to the number of genotype discordances among duplicate sample pairs.

Cohort	Array	Status	Pre-QC N	Post-QC N	Percent (%) kept
BRAINS	650Q	Cases	267	267	100
HABC	1M	Controls	2,802	1,586	56.6
GASROS	610	Cases	130	111	85.38
ISGS	610	Cases	373	351	94.1
SWISS	610	Cases	65	25	38.46
ESS	660	Cases	570	566	99.3
KORA	550	Controls	820	804	98.05
MUNICH	660	Cases	1,150	1,131	98.35
OXVASC	660	Cases	464	457	98.49
STGEORGE	660	Cases	423	418	98.82
WTCCC	660	Controls	5,186	5,150	99.31
GEOS	1M	Cases, controls	1,816	1,723	94.88
BRAINS	5M	Cases	114	100	87.72
GASROS	5M	Cases	468	456	97.44
GCNKSS	5M	Cases	499	482	96.59
HRS	2.5M	Controls	12,507[1]	11,842	94.7
ISGS	5M	Cases	187	178	95.19
MCISS	5M	Cases	630	619	98.25
MIAMSR	5M	Cases	299	294	98.33
NHS	5M	Cases	316	314	99.37
NOMAS	5M	Cases	363	358	98.62
OAI – AFR	2.5M	Controls	709	681	96.05
OAI – EUR	2.5M	Controls	3,302	3,201	96.94
REGARDS	5M	Cases	311	304	97.75
HCHS/SOL	2.5M	Controls	13,204	1,214	9.19[2]
SPS3	5M	Cases	962	949	98.65
SWISS	5M	Cases	271	181	66.79
WHI	5M	Cases	458	454	99.13
WUSTL	5M	Cases	455	449	98.68
KRAKOW	5M	Cases, controls	1,728	1,597	92.42
LSGS	5M	Cases, controls	949	913	96.21
ADHD	1M	Controls	435	411	94.48
INMA	1M	Controls	1,061	807	76.06
BASICMAR	5M	Cases	930	890	95.7
GRAZ	610	Controls	829	816	98.43
GRAZ	5M	Cases	639	607	94.99
SAHLSIS	5M	Cases	800	783	97.88
LSR	5M	Cases	642	641	99.84
MDC	750K, exome	Cases, controls	1,650	1,573	95.33
ASGC	610	Cases, controls	2,406	2,309	95.97

[1] N for HRS pre-SiGN QC but post CIDR QC

[2] Most HCHS/SOL samples were excluded to better match (by age and sex) the small number of HIS cases and were not per se QC failing

Supplementary Table 14 | Cohort sample sizes pre- and post-QC

Cases and controls were cleaned in two phases: (1) as an individual cohort (**Section 9.1**) and (2) as an array- and ancestry-specific group of merged cases and controls (**Section 9.3**). Shading indicates array-specific analysis group.

Study Stratum	Cohorts	Cases	Controls	SNPs	λ
1 – EUR	BRAINS, ISGS, GASROS, SWISS, HABC	754	1,586	350,285	1.020
2 – EUR	ESS, MUNICH, OXVASC, STGEORGE, WTCCC, KORA	2,572	5,954	338,380	1.056
3 – EUR	GEOS	460	519	748,158	1.007
4 – EUR	CIDR[1], HRS, OAI	3,291	11,820	948,227	1.017
5 – EUR	KRAKOW	878	716	2,233,797	0.986
6 – EUR	LSGS	459	453	2,218,479	0.971
7 – EUR	BASICMAR, ADHD, INMA	868	1,218	565,627	1.016
8 – EUR	GRAZ	607	815	236,884	1.023
9 – EUR	SAHLSIS, LSR, MDC	1,579	1,362	508,514	1.028
10 – EUR	ASGC	1,109	1,200	525,082	1.001
3 – AFR	GEOS	383	361	744,866	1.017
4 – AFR	CIDR[2], HRS, OAI	970	2,022	1,614,689	0.999
4 – HIS	CIDR[3], HCHS/SOL, HRS	942	2,429	1,908,773	1.033

[1] CIDR (EUR): BRAINS, GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, WUSTL

[2] CIDR (AFR): GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NOMAS, REGARDS, SPS3, SWISS, WHI, WUSTL

[3] CIDR (HIS): BASICMAR, BRAINS, GASROS, GCNKSS, SAHLSIS, GRAZ, ISGS, KRAKOW, LEUVEN, LUND, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, WUSTL

Supplementary Table 15 | Analysis strata sample sizes, SNPs available for downstream imputation, and genomic inflation (λ) from association testing in all stroke after QC (Section 9.5)

EUR, European ancestry; AFR, African ancestry; HIS, Hispanic.

	N	[16,40)	[40,45)	[45,50)	[50,55)	[55,60)	[60,65)	[65,67)	[67,107)
Females									
SiGN	942	0.01	0.02	0.03	0.04	0.05	0.06	0.02	0.20
HRS	1,214	0.01	0.01	0.02	0.08	0.12	0.08	0.05	0.26
HCHS/SOL all	10,363	0.13	0.06	0.10	0.10	0.08	0.06	0.02	0.04
HCHS/SOL subset	1,214	0.01	0.02	0.03	0.05	0.05	0.06	0.02	0.20
Males									
SiGN	942	0.01	0.02	0.04	0.08	0.09	0.08	0.03	0.20
HRS	1,214	0.00	0.01	0.00	0.04	0.08	0.04	0.03	0.18
HCHS/SOL all	10,363	0.11	0.04	0.07	0.06	0.05	0.04	0.01	0.02
HCHS/SOL subset	1,214	0.01	0.02	0.04	0.08	0.09	0.08	0.03	0.21

Supplementary Table 16 | Selection of controls in the Hispanic stratum

Cohort-specific proportions by age category (in columns, in years) and sex are shown. Table indicates how Hispanic controls were selected to match cases (**Section 9.5**).

Study Stratum	Cohorts	N	Imputed SNPs	Imputed SNPs post-QC
1 – EUR	BRAINS, ISGS, GASROS, SWISS, HABC	2,340	45,807,969	24,786,203
2 – EUR	ESS, MUNICH, OXVASC, STGEORGE, WTCCC, KORA	8,526	45,807,969	23,841,756
3 – EUR	GEOS	979	46,790,415	30,138,880
4 – EUR	CIDR[1], HRS, OAI	15,111	46,786,686	30,194,258
5 – EUR	KRAKOW	1,594	46,793,330	30,943,354
6 – EUR	LSGS	912	46,793,330	32,551,876
7 – EUR	BASICMAR, ADHD, INMA	2,086	46,782,781	28,197,093
8 – EUR	GRAZ	1,422	45,800,335	21,940,783
9 – EUR	SAHLSIS, LSR, MDC	2,941	46,786,453	27,792,416
10 – EUR	ASGC	2,309	46,789,601	27,893,216
3 – AFR	GEOS	744	38,835,942	25,449,196
4 – AFR	CIDR[2], HRS, OAI	2,992	38,835,942	26,925,041
4 – HIS	CIDR[3], HCHS/SOL, HRS	3,371	25,932,097	24,489,454

[1]CIDR (EUR): BRAINS, GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, WUSTL
[2]CIDR (AFR): GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NOMAS, REGARDS, SPS3, SWISS, WHI, WUSTL
[3]CIDR (HIS): BASICMAR, BRAINS, GASROS, GCNKSS, SAHLSIS, GRAZ, ISGS, KRAKOW, LEUVEN, LUND, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, WUSTL

Supplementary Table 17 | Analysis strata sample sizes, imputed SNPs available for genome-wide association analysis

Samples were prephased and then imputed (**Sections 10** and **11**) using either 1000 Genomes Phase I data (AFR and HIS strata) or a merge of the 1000 Genomes Phase I and GoNL data (EUR strata). After imputation, SNPs with info < 0.5 or out of Hardy-Weinberg equilibrium ($p < 10^{-6}$) were removed from the EUR and AFR strata.

Subtype Method	Trait	Abbreviation
–	All ischemic stroke	IS
CCS Causative (CCSc)	CE	CCScCEmajor
	LAA	CCScLAA
	SAO	CCScSAO
	Undetermined (1)	CCScUNDETER
	Undetermined (2)	CCScINCUNC
	Undetermined (3)	CCScCRYPTCE
CCS Phenotypic (CCSp)	CE	CCSpCEmajincl
	LAA	CCSpLAAmajincl
	SAO	CCSpSAOmajincl
	Undetermined	CCSpCryptoincl
TOAST	CE	toastCE
	LAA	toastLAA
	SAO	toastSAO
	Undetermined	toastUNDETER

Supplementary Table 18 | Traits analyzed in discovery

All traits analyzed (**Section 12**), including subtype-specific phenotypes determined by the CCS Causative, CCS Phenotypic, and TOAST classification systems. The four primary subtypes were large artery atherosclerosis (LAA), cardioembolic stroke (CE), small artery occlusion (SAO) and undetermined (UNDETER) and were available subtypes in each of the three subtyping methods (CCSc, CCSp, and TOAST). CCSc includes two additional undetermined subtypes in addition to the all undetermined (CCScUNDETER) type: CCScINCUNC, incomplete and inconclusive; CCScCRYPTCE, cryptogenic and CE minor. The CCSp system considers only cryptogenic strokes (CCSpCryptoincl) as undetermined.

Details of the subtyping in SiGN are provided in **Section 5**.

Analysis Stratum	Cohorts	Controls	Cases						
			IS	CE	LAA	SAO	UND-ETER	INC-UNC	CRYPTCE
1 (EUR)	BRAINS, GASROS, HABC, ISGS, SWISS	1,586	754	132	160	70	323	179	144
2 (EUR)	ESS, KORA, MUNICH, OXVASC, STGEORGE, WTCCC	5,954	2,572	635	455	275	1130	764	366
3 (AFR)	GEOS	361	383	43	34	71	179	79	100
3 (EUR)	GEOS	519	460	24	56	57	237	70	167
4 (AFR)	CIDR[1], HRS, OAI	2,022	970	122	144	310	364	163	201
4 (EUR)	CIDR[2], HRS, OAI	11,820	3,291	675	651	818	1046	412	634
4 (HIS)	CIDR[3], HRS, HCHS/SOL	2,429	942	95	103	532	181	52	129
5 (EUR)	KRAKOW	716	878	326	206	85	234	142	92
6 (EUR)	LSGS	453	459	109	90	30	200	22	178
7 (EUR)	ADHD, BASICMAR, INMA	1,218	868	336	196	250	86	55	31
8 (EUR)	GRAZ	815	607	183	106	67	231	120	111
9 (EUR)	SAHLSIS, LSR, MDC	1,362	1,579	346	240	201	451	229	222
10 (EUR)	ASGC	1,200	1,109	69	47	28	93	45	48
11 (AFR)	VISP (Handls)	971	256	--	--	--	--	--	--
11 (EUR)	VISP (Geneva)	1,047	1,723	--	--	--	--	--	--

[1] CIDR (EUR): BRAINS, GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, and WUSTL.

[2] CIDR (AFR): GASROS, GCNKSS, ISGS, MCISS, M IAMISR, NOMAS, REGARDS, SPS3, WHI, and WUSTL

[3] CIDR (HISP): BASIMAR, BRAINS, GASROS, SAHLSIS, KRAKOW, LUND, MCISS, MIAMSR, NHS, NOMAS, REGARDS, SPS3, SWISS, and WHI.

Supplementary Table 19 | Cases and Controls by analysis group and ancestry (post-QC) for all stroke and CCS Causative subtypes

Cases and controls were clustered into EUR, AFR, and HIS analysis groups (**Section 9.3**). If a study stratum contained < 40 cases for a certain phenotype, that phenotype was not analyzed for that study stratum (crossed out in the table above), as the distribution of the test statistic appeared systematically inflated in these strata and further QC failed to reduce the inflation (**Section 12.1**). IS, ischemic stroke; CE, cardioembolic; LAA, large artery atherosclerosis; SAO, small artery occlusion; UNDETER, undetermined; INCUNC, incomplete/unclassified; CRYPTCE, Cryptogenic and CE minor.

Analysis Stratum	Cohorts	Controls	Cases				
			IS	CE	LAA	SAO	Cryptoincl
1 (EUR)	BRAINS, GASROS, HABC, ISGS, SWISS	1,586	754	174	174	87	70
2 (EUR)	ESS, KORA, MUNICH, OXVASC, STGEORGE, WTCCC	5,954	2,572	755	431	320	192
3 (AFR)	GEOS	361	383	48	41	69	43
3 (EUR)	GEOS	519	460	31	68	58	96
4 (AFR)	CIDR[1], HRS, OAI	2,022	970	160	179	231	90
4 (EUR)	CIDR[2], HRS, OAI	11,820	3,291	868	728	866	313
4 (HIS)	CIDR[3], HRS, HCHS/SOL	2,429	942	118	146	468	63
5 (EUR)	KRAKOW	716	878	381	77	97	25
6 (EUR)	LSGS	453	459	127	104	34	76
7 (EUR)	ADHD, BASICMAR, INMA	1,218	868	385	236	266	13
8 (EUR)	GRAZ	815	607	225	142	72	34
9 (EUR)	SAHLSIS, LSR, MDC	1,362	1,579	374	215	200	119
10 (EUR)	ASGC	1,200	1,109	80	174	29	25
11 (AFR)	VISP (Handls)	971	256	--	--	--	--
11 (EUR)	VISP Geneva	1,047	1,723	--	--	--	--

[1]CIDR (EUR): BRAINS, GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, and WUSTL.

[2]CIDR (AFR): GASROS, GCNKSS, ISGS, MCISS, M IAMISR, NOMAS, REGARDS, SPS3, WHI, and WUSTL

[3]CIDR (HISP): BASIMAR, BRAINS, GASROS, SAHLSIS, KRAKOW, LUND, MCISS, MIAMSR, NHS, NOMAS, REGARDS, SPS3, SWISS, and WHI.

Supplementary Table 20 | Cases and Controls by analysis group and ancestry (post-QC) for all stroke and CCS Phenotypic subtypes

Cases and controls were clustered into EUR, AFR, and HIS analysis groups (**Section 9.3**). If a study stratum contained < 40 cases for a certain phenotype, that phenotype was not analyzed for that study stratum (crossed out in the table above), as the distribution of the test statistic appeared systematically inflated in these strata and further QC failed to reduce the inflation (**Section 12.1**). IS, ischemic stroke; CE, cardioembolic; LAA, large artery atherosclerosis; SAO, small artery occlusion; UNDETER, undetermined; INCUNC, incomplete/unclassified; CRYPTCE, Cryptogenic and CE minor.

Analysis Stratum	Cohorts	Controls	Cases				
			IS	CE	LAA	SAO	UN-DETER
1 (EUR)	BRAINS, GASROS, HABC, ISGS, SWISS	1,586	754	176	163	149	178
2 (EUR)	ESS, KORA, MUNICH, OXVASC, STGEORGE, WTCCC	5,954	2,572	679	497	371	1,019
3 (AFR)	GEOS	361	383	77	23	78	181
3 (EUR)	GEOS	519	460	92	37	56	242
4 (AFR)	CIDR[1], HRS, OAI	2,022	970	118	102	301	218
4 (EUR)	CIDR[2], HRS, OAI	11,820	3,291	598	423	761	646
4 (HIS)	CIDR[3], HCHS/SOL	2,429	942	94	88	552	114
5 (EUR)	KRAKOW	716	878	407	173	36	239
6 (EUR)	LSGS	453	459	157	75	55	149
7 (EUR)	ADHD, BASICMAR, INMA	1,218	868	408	184	276	0
8 (EUR)	GRAZ	815	607	166	85	74	114
9 (EUR)	SAHLSIS, LSR, MDC	1,362	1,579	223	150	183	326
10 (EUR)	ASGC	1,200	1,109	232	406	291	167
11 (AFR)	VISP (Handls)	971	256	--	--	--	--
11 (EUR)	VISP (Geneva)	1,047	1,723	--	--	--	--

[1] CIDR (EUR): BRAINS, GASROS, GCNKSS, ISGS, MCISS, MIAMISR, NHS, NOMAS, REGARDS, SPS3, SWISS, WHI, and WUSTL.

[2] CIDR (AFR): GASROS, GCNKSS, ISGS, MCISS, M IAMISR, NOMAS, REGARDS, SPS3, WHI, and WUSTL

[3] CIDR (HISP): BASIMAR, BRAINS, GASROS, SAHLSIS, KRAKOW, LUND, MCISS, MIAMSR, NHS, NOMAS, REGARDS, SPS3, SWISS, and WI.

Supplementary Table 21 | Cases and Controls by analysis group and ancestry (post-QC) for all stroke and TOAST subtypes

Cases and controls were clustered into EUR, AFR, and HIS analysis groups (**Section 9.3**). If a study stratum contained < 40 cases for a certain phenotype, that phenotype was not analyzed for that study stratum (crossed out in the table above), as the distribution of the test statistic appeared systematically inflated in these strata and further QC failed to reduce the inflation (**Section 12.1**). IS, ischemic stroke; CE, cardioembolic; LAA, large artery atherosclerosis; SAO, small artery occlusion; UNDETER, undetermined; INCUNC, incomplete/unclassified; CRYPTCE, Cryptogenic and CE minor.

Cohort	IS		CCSp: CE		CCSp: LAA		CCSp: SAO		CCSp: Cryptoincl	
	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}
1 – EUR	1.181	1.026	1.35	1.021	1.334	1.021	1.352	1.027	1.344	1.047
2 – EUR	1.127	1.045	1.206	1.021	1.212	1.038	1.256	1.014	1.294	1.085
3 – EUR	1.123	1.014	~~1.432~~	~~1.133~~	1.392	1.039	1.396	1.057	1.378	1.033
4 – EUR	1.14	1.013	1.277	1.012	1.276	1.006	1.209	0.993	1.214	1.009
5 – EUR	1.099	1.013	1.125	1.018	1.334	1.035	1.333	1.039	~~1.351~~	~~1.195~~
6 – EUR	1.116	1.019	1.307	1.036	1.344	1.045	~~1.451~~	~~1.111~~	1.412	1.068
7 – EUR	1.126	1.014	1.235	1.016	1.376	1.03	1.343	1.011	~~1.014~~	~~1.419~~
8 – EUR	1.109	1.015	1.231	1.031	1.283	1.033	1.375	1.052	~~1.435~~	~~1.108~~
9 – EUR	1.099	1.027	1.184	1.009	1.208	0.999	1.258	1.023	1.223	1.03
10 – EUR	1.103	1.008	1.375	1.036	1.373	1.038	~~1.333~~	~~1.095~~	~~1.306~~	~~1.142~~
11 – EUR	0.921	1.024	–	–	–	–	–	–	–	–
3 – AFR	1.086	1.027	1.391	1.085	1.427	1.134	1.312	1.063	1.415	1.091
4 – AFR	1.068	1.005	1.229	1.018	1.216	1.019	1.155	1.011	1.299	1.036
11 – AFR	0.968	0.995	–	–	–	–	–	–	–	–
4 – HIS	1.083	1.034	1.36	1.026	1.299	1.041	1.151	1.033	1.458	1.052

Cohort	CCSc: CE		CCSc: LAA		CCSc: SAO		CCSc: UNDETER		CCSc: INCUNC		CCSc: CRYPTCE	
	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}
1 – EUR	1.359	1.025	1.335	1.013	1.351	1.036	1.297	1.021	1.344	1.02	1.35	1.025
2 – EUR	1.223	1.026	1.228	1.035	1.264	1.011	1.189	1.025	1.229	1.023	1.229	1.037
3 – EUR	~~1.402~~	~~1.176~~	1.387	1.045	1.395	1.063	1.191	1.025	1.391	1.047	1.278	1.035
4 – EUR	1.275	1.01	1.276	1.005	1.204	0.991	1.272	1.006	1.218	1.004	1.27	1.01
5 – EUR	1.146	1.02	1.216	1.03	1.341	1.043	1.209	1.027	1.293	1.022	1.359	1.057
6 – EUR	1.338	1.044	1.373	1.052	~~1.426~~	~~1.14~~	1.193	1.031	~~1.378~~	~~1.236~~	1.221	1.035
7 – EUR	1.261	1.014	1.393	1.018	1.351	1.012	1.398	1.045	1.34	1.04	~~1.243~~	~~1.119~~
8 – EUR	1.263	1.034	1.311	1.032	1.381	1.052	1.211	1.014	1.32	1.032	1.313	1.032
9 – EUR	1.193	1.009	1.215	1.001	1.247	1.015	1.126	1.002	1.204	1.009	1.17	1.009
10 – EUR	1.382	1.034	1.373	1.046	~~1.328~~	~~1.102~~	1.358	1.032	1.372	1.051	1.371	1.044
11 – EUR	–	–	–	–	–	–	–	–	–	–	–	–
3 – AFR	1.42	1.102	~~1.46~~	~~1.168~~	1.307	1.064	1.12	1.033	1.266	1.061	1.224	1.053
4 – AFR	1.269	1.027	1.232	1.015	1.157	1.01	1.142	1.009	1.23	1.024	1.198	1.013
11 – AFR	–	–	–	–	–	–	–	–	–	–	–	–
4 – HIS	1.405	1.051	1.368	1.056	1.137	1.034	1.257	1.019	1.46	1.062	1.341	1.041

Cohort	TOAST: CE		TOAST: LAA		TOAST: SAO		TOAST: UNDETER	
	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}	λ	λ_{QC}
1 – EUR	1.349	1.028	1.33	1.015	1.358	1.032	1.343	1.018
2 – EUR	1.217	1.031	1.214	1.032	1.261	1.016	1.201	1.03
3 – EUR	1.387	1.05	1.403	1.095	1.404	1.04	1.194	1.024
4 – EUR	1.273	1.011	1.253	1.001	1.196	0.993	1.263	1.013
5 – EUR	1.12	1.02	1.238	1.018	1.372	1.098	1.191	1.011
6 – EUR	1.242	1.024	1.402	1.049	1.424	1.036	1.254	1.033
7 – EUR	1.225	1.014	1.402	1.026	1.33	1.008	–	–
8 – EUR	1.271	1.026	1.343	1.041	1.366	1.05	1.33	1.043
9 – EUR	1.173	1.008	1.191	1.019	1.236	1.005	1.141	1.017
10 – EUR	1.269	1.016	1.185	1.021	1.258	1.019	1.334	1.029
11 – EUR	–	–	–	–	–	–	–	–
3 – AFR	1.285	1.062	1.525	1.326	1.29	1.067	1.127	1.039
4 – AFR	1.269	1.021	1.281	1.03	1.159	1.008	1.192	1.015
11 – AFR	–	–	–	–	–	–	–	–
4 – HIS	1.407	1.045	1.402	1.054	1.117	1.02	1.345	1.036

Supplementary Table 22 | Post-imputation genomic inflation in study strata before and after QC

A GWAS was run in each stratum for each of the 15 phenotypes (unless the case count in a stratum was < 40, diagonal lines). Genomic inflation after imputation (λ) was higher post-imputation; removing SNPs with minor allele frequency < 1% removed the excess inflation (**Section 12.2**). λ: genomic inflation before QC, λ_{QC}: genomic inflation after QC. Dashes indicate subtypes with no cases for that stratum.

Phenotype	Subtyping	Cases	Controls
Ischemic stroke	--	16,851	31,259
Cardioembolic	CCSc	3,071	28,722
	CCSp	3,695	28,722
	TOAST	3,427	29,241
Large Artery Atherosclerosis	CCSc	2,454	28,880
	CCSp	2,715	29,241
	TOAST	2,346	28,961
Small artery occlusion	CCSc	2,736	27,588
	CCSp	2,734	27,588
	TOAST	3,147	28,525
Undetermined	CCSc (UNDETER)	4,755	25,292
	CCSc (INCUNC)	2,310	28,788
	CCSc (CYPTCE)	2,392	28,023
	CCSp	1,062	25,292
	TOAST	3,593	28,023

Supplementary Table 23 | Cases and controls for discovery meta-analyses in ischemic stroke and its subtypes (Section 13)

Trait	SNP	Alleles	RAF	Primary meta-analysis		Secondary meta-analysis		Ratio (primary/secondary)
			EUR | AFR | AMR	OR (C)	P (C)	OR (C)	P (C)	log(OR) (C)
				OR (P)	P (P)	OR (P)	P (P)	log(OR) (P)
				OR (T)	P (T)	OR (T)	P (T)	log(OR) (T)
IS	rs10744777	T/C	0.667 | 0.045 | 0.515	1.10	3.07×10^{-8}	1.10	1.31×10^{-8}	0.99
IS	rs2634074	T/A	0.205 | 0.477 | 0.413	1.10	2.56×10^{-7}	1.10	5.88×10^{-8}	0.97
IS	rs2107595	A/G	0.157 | 0.219 | 0.218	1.10	7.74×10^{-7}	1.09	2.68×10^{-6}	1.07
IS	rs12425791	G/A	0.202 | 0.086 | 0.247	1.01	5.95×10^{-1}	1.01	5.06×10^{-1}	1.00
IS	rs505922	C/T	0.351 | 0.326 | 0.235	1.07	2.03×10^{-5}	1.07	1.72×10^{-5}	1.00
CE	rs2200733	T/C	0.120 | 0.215 | 0.256	1.39	1.24×10^{-16}	1.42	1.09×10^{-20}	0.94
				1.39	3.26×10^{-19}	1.41	2.52×10^{-23}	0.95
				1.37	1.02×10^{-16}	1.39	7.29×10^{-20}	0.96
CE	rs7193343	T/C	0.174 | 0.240 | 0.189	1.17	1.12×10^{-5}	1.19	1.58×10^{-6}	0.95
				1.19	2.93×10^{-7}	1.20	3.08×10^{-8}	0.95
				1.17	1.45×10^{-5}	1.16	1.71×10^{-5}	1.03
CE	rs505922	C/T	0.351 | 0.326 | 0.235	1.04	0.19	1.04	0.19	1.00
				1.04	0.16	1.04	0.17	1.03
				1.08	5.66×10^{-3}	1.08	5.40×10^{-3}	1.01
LAA	rs11984041	T/C	0.093 | 0.224 | 0.067	1.30	8.46×10^{-8}	1.27	4.78×10^{-7}	1.10
				1.29	3.50×10^{-8}	1.28	8.58×10^{-8}	1.05
				1.30	3.62×10^{-7}	1.28	7.23×10^{-7}	1.06
LAA	rs556621	T/G	0.291 | 0.081 | 0.407	1.04	3.18×10^{-1}	1.04	0.23	0.86
				1.02	6.36×10^{-1}	1.03	0.46	0.64
				1.11	2.55×10^{-3}	1.11	1.8×10^{-3}	0.99
LAA	rs2383207	G/A	0.499 | 0.045 | 0.413	1.12	4.34×10^{-4}	1.12	3.2×10^{-4}	1.00
				1.11	7.93×10^{-4}	1.11	7.7×10^{-4}	1.03
				1.09	8.13×10^{-3}	1.10	5.2×10^{-3}	0.97
LAA	rs505922	C/T	0.351 | 0.326 | 0.235	1.09	6.93×10^{-3}	1.09	9.4×10^{-3}	1.08
				1.11	1.29×10^{-3}	1.10	2.9×10^{-3}	1.11
				1.14	2.15×10^{-4}	1.13	2.6×10^{-4}	1.05
LAA	rs12122341	G/C	0.257 | 0.088 | 0.195	1.20	3.38×10^{-7}	1.21	3.88×10^{-8}	0.96
				1.21	4.50×10^{-8}	1.22	8.02×10^{-9}	0.98
				1.15	1.61×10^{-4}	1.16	4.16×10^{-5}	0.95
UND	rs74475935	G/C	0.002 | 0.018 | 0.006	5.17	3.69×10^{-9}	-	-	-
				8.68	5.94×10^{-11}	-	-	-
				2.18	1.58×10^{-2}	-	-	-

Supplementary Table 24 | Comparison of primary and secondary discovery meta-analysis results

Comparison of primary and secondary discovery analyses, performed by two independent analysts (**Section 13.4**). Analyses for associated SNPs were highly concordant; differences are the result of slightly different QC filters applied in each analysis. Numbers are rounded to two decimal places. Alleles are ordered as risk/protective. Rs74475935 was filtered out of the secondary analysis.

Phenotype	Subtyping	Cases	Controls
Ischemic stroke	--	37,893	371,481
Cardioembolic	CCSc	7,062	289,185
	CCSp	7,686	289,185
	TOAST	7,418	289,704
Large Artery Atherosclerosis	CCSc	4,703	212,109
	CCSp	4,964	212,470
	TOAST	4,595	212,190
Small artery occlusion	CCSc	5,162	251,790
	CCSp	5,160	251,790
	TOAST	5,573	252,727
Undetermined	CCSc (UNDETER)	8,452	285,837
	CCSc (INCUNC)	6,007	289,333
	CCSc (CYPTCE)	6,089	288,568
	CCSp	4,759	285,837
	TOAST	7,290	288,568

Supplementary Table 25 | Cases and controls for joint (stage I and stage II) meta-analyses in ischemic stroke and its subtype (Section 14.3)

Trait	SNP	Chromosome: Position	Risk/ other allele	RAF (%)	Nearest Gene	Effective sample size	Joint meta-analysis (Stage I + Stage II)			Directions	Average info score	Tau Squared
				EUR | AFR | AMR			Cases-C	OR [95% CI]	P			
							Cases-P		P			
							Cases-T		P			
LAA	rs12122341	1:115655690	G/C	25.7 | 8.8 | 19.5	*TSPAN2*	203,533	4,703	1.18 [1.12 – 1.25]	8.32 x 10^{-9}	+--+-++++++-++.+-++-+.+++--.	0.850	0.015
						204,010	4,964	1.19 [1.12 – 1.26]	1.30 x 10^{-9}	+--+++++++++-++.+-++-+.+++--.	0.853	0.009
						202,932	4,595	1.15 [1.08 – 1.22]	2.70 x 10^{-6}	+-+-++++++-++.+-++-+.+++--.	0.845	0.013
UND	rs74475935	16:15961249	G/C	0. 2 | 1.8 | 0.6	*ABCC1*	18,897	5,861	4.63* [2.77 – 7.72]	4.70 x 10^{-11}	........++.+..-.-+...-.	0.569	0.000
						18,735	4,531	6.89 [3.80 – 12.47]	1.85 x 10^{-10}	......++.+..-.-+...-.	0.569	0.065
						26,140	7,062	2.11 [1.20 – 3.70]	9.22 x 10^{-3}	.........++.+..-.-+...-+.	0.569	0
IS	rs10744777	12:111795214	T/C	66.7 | 4.5 | 5.2	*ALDH2*	311,332	37,893	1.07 [1.5 – 1.09]	4.20 x 10^{-9}	++++-+++-+++++.-..++-+.-+--....++.++	0.953	0.001
CE	rs2200733	4:110789013	T/C	12.0 | 2.2 | 2.6	*PITX2*	238,051	7,062	1.37 [1.30 – 1.45]	1.04 x 10^{-29}	++++-++++++.++.+++-++++.+++++	0.898	0.011
						238,651	7,686	1.37 [1.30 – 1.45]	2.79 x 10^{-32}	++++-++++++.++.+++-++++.+++++	0.898	0.010
						238,928	7,418	1.36 [1.29 – 1.44]	8.05 x 10^{-30}	++++-++++++.++.+++-++++.+++++	0.901	0.016
CE	rs7193343	16:72995261	T/C	17.4 | 2.4 | 18.9	*ZFHX3*	237,163	7,062	1.17 [1.10 – 1.22]	7.28 x 10^{-9}	-++++-+++++++-.+.++++--.+.+--	0.880	0.009
						237,779	7,686	1.17 [1.11 – 1.23]	2.29 x 10^{-10}	+++++-+++++++-.+.++++--.+.+--	0.880	0.008
						238,036	7,418	1.16 [1.10 – 1.22]	8.88 x 10^{-9}	-+++++-+++-+++-.+.++++--.+.+--	0.885	0.011
LAA	rs11984041	7:18992312	T/C	9.3 | 2.2 | 6.7	*HDAC9*	211,426	4,703	1.23 [1.15 – 1.33]	1.10 x 10^{-8}	+++++-+-+++++++--.-++.+--+++	0.948	0.012
						211,916	4,964	1.24 [1.15 – 1.33]	4.52 x 10^{-9}	++++-+-++++++++--.-++.+--+++	0.948	0.002
						210,829	4,595	1.23 [1.14 – 1.33]	4.48 x 10^{-8}	++++++-+++++++--.-++.+--+++	0.946	0.007
SAO	rs10744777	12:111795214	T/C	66.7 | 4.5 | 5.2	*ALDH2*	244,950	5,162	1.16 [1.10 – 1.22]	2.77 x 10^{-8}	+++++++++++-+++..-+-...++++	0.942	0.016
						245,036	5,160	1.17 [1.11 – 1.23]	2.92 x 10^{-9}	+++++++++++-+++..-+-...++++	0.942	0.016
						246,281	5,573	1.13 [1.07 – 1.18]	1.62 x 10^{-6}	+++++++++++-+++..-+-...++++	0.944	0.016

*Reported for the CCS Causative CRYPTCE undetermined phenotype

Supplementary Table 26 | Detailed summary-statistics of genome-wide significant loci after combined meta-analysis of stage I (discovery) and stage II (replication) data

Results after meta-analysis of stage I (discovery) and stage II (replication) data are shown (**Section 15**). Directions indicate the consistency of the sign (+/-) of the beta for the SNP across the various strata. Average info score is the average imputation quality score across all of the contributing strata. Risk allele frequencies are from 1000 Genomes Phase I. Results are shown for all three subtyping methods: CCS Causative (C), CCS Phenotypic (P) and TOAST (T).

Trait	Subtype	Stage I lambda (λ)	Joint lambda (λ)
All stroke	--	1.004	1.005
Cardioembolic	CCSc	0.953	0.953
	CCSp	0.963	0.963
	TOAST	0.964	0.964
Large artery atherosclerosis	CCSc	0.993	0.993
	CCSp	0.962	0.962
	TOAST	0.977	0.977
Small artery occlusion	CCSc	0.981	0.981
	CCSp	0.979	0.979
	TOAST	0.987	0.987
Undetermined	CCSc(1)	0.998	0.998
	CCSc(2)	0.978	0.978
	CCSc(3)	0.983	0.983
	CCSp	0.936	0.936
	TOAST	0.980	0.980

Supplementary Table 27 | Genomic inflation (lambda) of all analyses for analysis of discovery (stage I) data and joint analysis of the discovery and replication data

Lambdas are shown for all stroke (IS) as well as for each subtype as determined by the CCS Causative (CCSc), CCS Phenotypic (CCSp) and TOAST subtyping methods. Note that discovery samples had CCSc, CCSp, and TOAST subtypes available; replication samples only had TOAST subtypes. CCSc has three undetermined subtypes: (1) all undetermined, (2) incomplete and unclassified, and (3) cryptogenic and CE minor.

III. Supplementary Note

1. Introduction

In the Supplementary Note, we provide a detailed description of the Stroke Genetics Network (SiGN) study of ischemic stroke and its subtypes. Stroke cases collected around the world were genotyped and phenotyped, and then merged with publicly-available controls, enabling the largest multi-stage genome-wide association study of ischemic stroke to date (**Supplementary Figure 1**)

The Supplementary Note is an in-depth report of the various steps involved in the project with the aim of providing a comprehensive and transparent description of the full project for future reference.

2. Organizational structure of the SiGN study

SiGN is a National Institute of Neurological Disorders and Stroke (NINDS) funded project (U01NS69208) that consists of 26 Genetic Research Centers (GRCs): 13 from the United States, 12 from Europe and 1 from Australia. The GRCs represent centers that had DNA samples or pre-existing GWAS data from ischemic stroke cases, and that agreed to characterize all cases for stroke subtype using the web-based Causative Classification System (CCS).(1) An additional study, the Vitamin Intervention for Stroke Prevention (VISP), was not able to phenotype their cases with CCS but contributed previously genotyped ischemic stroke cases to the all ischemic stroke discovery analysis. Nine control-only studies collaborated with SiGN: 4 from the United States and 5 from Europe.

The CCS is a single standardized protocol requiring detailed clinical and imaging information. Informed consent for data sharing was a requirement for inclusion as a GRC. Participating GRCs included first-ever, recurrent stroke, cohort, case-control, population-based, and hospital-based studies of ischemic stroke.

To maximize power, cases were preferentially genotyped and controls genotyped only where a suitable publicly available control population was not available. In addition, given the importance of subtyping, cases with complete diagnostic testing with investigations for cardiac and large vessel mechanisms of stroke were preferentially selected for genotyping compared to cases that had incomplete testing. Data from these case and control populations were used for the discovery analyses. Eighteen studies with pre-existing phenotype and genetic data on ischemic stroke cases and controls collaborated with SiGN for the purpose of replication. For these replication studies, stroke phenotyping was generally performed using the TOAST phenotyping system.(2)

Supplementary Figure 10 shows the administrative structure of SiGN. The Scientific Steering Committee leads SiGN. Its members include co-principal investigators, the Analysis Committee, and NINDS staff. The Scientific Steering Committee is responsible for scientific direction and policy decisions. It also oversees the Publications and Data Access Committee, which develops guidelines for publication and authorship, prioritizes analysis resources for manuscript proposals, and recommends approval of proposals and manuscripts to the Scientific Steering Committee.

The study has 4 cores: Administrative, Data Management, Imaging, and Genotyping.

(a) The Administrative Core and Data Management Core monitor study progress, maintain efficient interactions among the cores and the participating genetic research centers (GRCs), ensure regulatory and policy compliance, and are responsible for submitting genotype and phenotype data to dbGaP.

(b) The Imaging Core is a centralized repository for clinically obtained MRI data from the GRCs and provides critical information to the Phenotype Committee on stroke subtyping.

(c) The Genotyping Core is the NINDS-designated Center for Inherited Disease Research (CIDR, Baltimore, MD). The Genotyping Core performs quality control of the submitted DNA, as well as initial quality control of the GWAS and exome-enriched genotyping.

(d) The Phenotype Committee is responsible for training and quality assurance of ischemic stroke subtyping at the GRCs and advises the Analysis Committee on stroke subtyping issues.

The Analysis Committee, composed of genetic epidemiologists and statistical geneticists from 4 different institutions, advises the Scientific Steering Committee on design issues and is responsible for the analyses of phenotypic and genetic data. The Data Management Core also works closely with the Analysis Committee in the preparation of publications.

The Biostatistics Department Genetics Coordinating Center (GCC) at the University of Washington in Seattle provides more extensive quality control of the genotype data through a subcontract with CIDR.

CIDR, the UW-GCC, and the Analysis Committee jointly decided on the design of the study, including choice of controls, and selection and use of within- and cross-study duplicates.

3. Studies for stage I analyses

The following section contains cohort descriptions for the cases and controls included in the stage I (discovery) phase of SiGN.

Commonly used acronyms and terms in this section:

(a) CCS: Causative Classification System for Ischemic Stroke
(b) MELAS: Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes
(c) CT: Computed tomography
(d) CTA: computed tomography angiography
(e) DWI: diffusion weighted MRI in acute stroke
(f) ECG or EKG: Electrocardiogram
(g) MRA: magnetic resonance angiography
(h) MRI: magnetic resonance imaging
(i) NDCI: Non-disabling cerebral infarction
(j) NIHSS: National Institutes of Health Stroke Score
(k) TIA: transient ischemic attack
(l) TOAST: Trial of Org 10172 in Acute Stroke Treatment
(m) WHO: World Health Organization

3.1 Cases for stage I (discovery) analysis

3.1.1 Australian Stroke Genetics Collaborative (ASGC)

The ASGC continues to enroll cases that are first-in-a-lifetime ischemic strokes admitted to the Acute Stroke Units within the Hunter Stroke Service (John Hunter and Newcastle Mater Hospitals), the Central Coast Stroke Service (Gosford and Wyong Hospitals), Queen Elizabeth and the Royal Adelaide Hospitals, and the Royal Perth Hospital. These seven acute care hospitals are the principal referral hospitals in their respective regions and have academic support and acute and rehabilitation stroke care units.

Recruitment into the three regional cohorts began in 1998. Stroke is defined according to the World Health Organization (WHO)(3) clinical criteria as rapidly developing clinical symptoms and/or signs of focal, and at times global loss (applied to patients in deep coma and to those with subarachnoid hemorrhage) of cerebral function, with symptoms lasting more than 24 hours or leading to death, with no apparent cause other than that of vascular origin.(3) Investigators confirm the diagnosis of ischemic stroke by head-computed tomography (CT) and/or brain magnetic resonance imaging (MRI).

The only inclusion criterion is first-ever ischemic stroke cases. Exclusion criteria consist of hemorrhagic stroke, transient ischemic attack (TIA), those not able to undergo baseline brain imaging, history of previous stroke, and lack of informed consent. All cases of ischemic stroke have been classified using the Trial of Org 10172 Acute Stroke Treatment (TOAST) system.(2) Cases in this study were also classified using the Causative Classification of Stroke (CCS) system.(1)

3.1.2 BASe de datos de ICtus del hospital del MAR (BASICMAR)

BASICMAR is an ongoing prospective study of all acute strokes assessed since 2005 at the IMIM-Hospital Universitari del Mar (Barcelona, Spain). It includes both first-ever and recurrent strokes. There were no exclusion criteria regarding age or race-ethnicity of the individuals. All patients had an electrocardiogram (ECG), a blood analysis, and neuroimaging at the acute stage. Additional diagnostic procedures were s performed when clinically indicated. A follow-up of three months after stroke was completed for all survivors.

Ischemic stroke etiologic subtypes were classified according to TOAST criteria.(2) For this study, only individuals of European origin with ischemic stroke were selected from BASICMAR, with eligible events defined as a clinical syndrome of any duration associated with a radiographically proven acute infarct, without radiographic evidence of a demyelinating or neoplastic disease or other structural disease including primary intracerebral hemorrhage.

3.1.3 BRAINS Bio-Repository of DNA in stroke

BRAINS is an ongoing, hospital-based study that seeks to establish a high quality biobank resource. Cases participating in the SiGN study were of European descent and recruited within the United Kingdom between September 25, 2009 and August 4, 2011. Extensive phenotype information is collected including subtype of stroke, past and family cardiovascular history, blood pressure data, MRI or CT brain imaging, carotid anatomy and blood tests (including cholesterol).

All hospital admitted participants over the age of 18 years with first-ever or recurrent stroke that provided informed consent (or caregivers on their behalf) were recruited. Participants must have image-positive lesions. Exclusion criteria are mainly for those that were brain image-negative, even if the clinical presentation is that of stroke. There are no eligibility criteria based on stroke severity or participation in a treatment trial. Inability to obtain consent results in mandatory exclusion. Additional information about BRAINS can be found in prior publications.(4,5)

3.1.4 Edinburgh Stroke Study (ESS)

Between 2002 and 2005, consecutive consenting patients with stroke who were admitted to or seen as outpatients at the Western General Hospital, Edinburgh were prospectively recruited from stroke centers in Edinburgh, Scotland, U.K. There were no exclusion criteria for cases based on age, stroke severity, or inclusion in other clinical research studies. Cases in this study were of European origin, with a clinically evident stroke, demonstrated by brain imaging (CT or MRI) to be ischemic. An experienced stroke physician assessed each participant as soon as possible after stroke onset, prospectively recording demographic and clinical details, including vascular risk factors and results of brain imaging and other investigations. Ischemic subtypes were assigned according to the TOAST criteria(2) and, subsequently, using CCS, specifically for the purposes of the SIGN study.(1)

ESS cases were collected as part of the WTCCC2 effort (**Section 3.2.15**). All WTCCC2 cases were genotyped as part of the WTCCC2 Ischemic Stroke study using the Illumina Human660W-Quad array. Quality control procedures in the WTCCC2 excluded SNPs not genotyped on all case and control collections and SNPs with Fisher information measure < 0.98, genotype call rate < 0.95, MAF < 0.01 or Hardy-Weinberg P-value < 1 x 10^{-20} in either the case or control collections. Samples were excluded if identified as outliers on call rate, heterozygosity, ancestry and average probe intensity based on a Bayesian clustering algorithm. Samples were also removed if they exhibited discrepancies between inferred and recorded sex or were shown to have cryptic relatedness with other WTCCC2 samples (pairwise identity-by-descent > 0.05).

3.1.5 Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS)

The GCNKSS is a population-based epidemiologic study of stroke in blacks and whites that is designed to measure temporal trends and racial differences in incidence of stroke. The catchment area includes two southwestern Ohio, U.S.A., counties (Hamilton, which includes the city of Cincinnati, and Clermont to the east) and three Northern Kentucky, U.S.A., counties (Boone, Kenton, and Campbell) to the south of Cincinnati across the Ohio River.

As part of the GCNKSS, for calendar years 1999 and 2005, prospective cohorts of first-ever and recurrent ischemic stroke cases were assembled using “hot pursuit” methodology at all local hospitals in the region (18 in 1999, and 17 in 2005), except for one hospital that is solely devoted to treating pediatric cases. Participants remained eligible if they were in a treatment trial, but participation in a treatment trial was not required for enrollment. Subjects with all degrees of severity of stroke were eligible, and no particular racial group was intentionally oversampled (about 80% were white participants and 20% black).

Study research nurses prospectively screened inpatient admission and emergency department logs to identify acute ischemic stroke patients. When a case was identified and the treating physician had given permission to approach the patient, a study nurse asked the subject or a proxy (the most closely related competent individual, preferably a person living with the subject prior to the stroke) to consent to participate in the cohort. After consent was granted, a study nurse performed an extensive interview, and a blood sample was obtained for genetic analysis. In addition, a study nurse abstracted information about the individual, the subject's medical history, the stroke event, and imaging studies from the hospital chart. A study physician reviewed every abstract, along with the imaging studies, to verify that an acute stroke had occurred, and to classify the event according to TOAST(2) and CCS criteria.(1)

3.1.6 Genetics of Early Onset Stroke (GEOS)

The GEOS study is a population-based case-control study designed to identify the genetic determinants of early-onset ischemic stroke and to characterize interactions of stroke-associated genes with environmental risk factors. Cases with a first-ever ischemic stroke were identified by discharge surveillance from one of 59 hospitals in the U.S. greater Baltimore-Washington area and by direct referral from regional neurologists. Cases and controls were recruited in three different time periods:

(a) Stroke Prevention in Young Women-1 (SPYW-1), conducted from 1992-1996,
(b) Stroke Prevention in Young Women-2 (SPYW-2), conducted from 2001-2003, and
(c) Stroke Prevention in Young Men (SPYM), conducted from 2003-2007.

SPYW-1 included cases aged 15 – 44 years recruited within one year of stroke and was designed with a 1:2 case-to-control ratio. SPYW-2 and SPYM included cases aged 15 – 49 recruited within three years of stroke and was designed with a 1:1 case-to-control ratio. Control participants without a history of stroke were identified by random-digit dialing. Controls were balanced to cases by age and region of residence in each study and were additionally balanced for ethnicity in SPYW-2 and SPYM. The number of cases and controls recruited in each study is as follows: 115 cases and 198 controls from SPYW-1, 234 cases and 209 controls from SPYW-2, and 478 cases and 500 controls from SPYM.

The abstracted hospital records of potential cases were reviewed and adjudicated for ischemic stroke, ischemic stroke subtype, and modified Rankin Scale(6) at discharge by a pair of vascular neurologists according to previously published procedures(7,8) with disagreements resolved by a third vascular neurologist. Stroke was defined according to the criteria of the WHO(3) and ischemic stroke was defined based on the criteria of the National Institute of Neurological Disorders and Stroke Data Bank.(9) Cases had a head CT and/or brain MRI that was consistent with cerebral infarction. Visualization of the infarct was not required, only that no alternative etiology was identified. The ischemic stroke subtype classification system retains information on all probable and possible causes, and is reducible to the more widely used TOAST[1] system that assigns each case to a single category. Cases were subsequently subtyped using the CCS.(1)

Ischemic strokes with the following characteristics were excluded from participation: stroke occurring as an immediate consequence of trauma, stroke within 48 hours after a hospital procedure, stroke within 60 days after the onset of a non-traumatic subarachnoid hemorrhage, and cerebral venous thrombosis. Additional exclusions for genetic analyses modified from(10) were as follows: known single-gene or mitochondrial disorders recognized by a distinctive phenotype (e.g. cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), mitochondrial encephalopathy with lactic acidosis and stroke-like episodes (MELAS), homocystinuria, Fabry disease, or sickle cell anemia); mechanical aortic or mitral valve at the time of index stroke; untreated or actively treated bacterial endocarditis at the time of the index stroke; neurosyphilis or other central nervous system infections; neurosarcoidosis; severe sepsis with hypotension at the time of the index stroke; cerebral vasculitis by angiogram and clinical criteria; post-radiation arteriopathy; left atrial myxoma; major congenital heart disease; and cocaine use in the 48 hours prior to stroke onset. There were no exclusions based on race or ethnicity, stroke severity, or participation in clinical trial research.

Demographic variables, including age, ancestry, ethnicity, and established stroke risk factors, were collected during a standardized interview. Risk factors included history of hypertension, diabetes mellitus, myocardial infarction, and current smoking status and/or oral contraceptive use, defined as use within one month prior to event for cases and at a comparable reference time for controls.

3.1.7 GRAZ

Between 1994 and 2003, subjects with first-ever and recurrent ischemic strokes admitted to the stroke unit of the Department of Neurology, Medical University of Graz (Graz, Austria) were included. All race-ethnic groups were eligible and there was no intentional oversampling. All age groups were allowed, though only subjects above the age of 18 were admitted to our department. Ischemic stroke was defined as an episode of focal neurological deficits with acute onset and lasting > 24 hours. There were no selection criteria based on stroke severity. Those individuals in treatment trials were excluded. 685 subjects were eligible to participate in this study (278 women, 407 men). All cases were Caucasian. Mean age was 68.9 ± 13.8 years with an age range from 19 – 101 years.

In addition to a standardized protocol including a laboratory examination and carotid ultrasound or magnetic resonance angiography and ECG, 304 subjects underwent neuroimaging by CT and 381 by MRI. More extensive cardiac examination, including transesophageal echocardiography or transthoracic echocardiography and Holter, was performed in subjects with suspected cardiac embolism. Stroke subtypes were assessed according to modified TOAST criteria[1] and were conducted by trained stroke neurologists.

3.1.8 Ischemic Stroke Genetics Study (ISGS)

The ISGS is a multicenter prospective, hospital-based inception cohort study of first-ever ischemic stroke. Enrollment for ISGS began in December 2002 and was completed in July 2007. During this time, 656 cases of first-ever ischemic stroke and 648 stroke-free controls were enrolled across the 5 centers (Mayo Clinic, Jacksonville, FL, U.S.A.; Mayo Clinic, Rochester, MN, U.S.A.; University of Virginia, Charlottesville, VA, U.S.A.; Shands Hospital, Jacksonville, FL, U.S.A.; Grady Hospital, Atlanta, GA, U.S.A.).

All cases required meeting the WHO definition for stroke(3) and head imaging, by either head MRI or CT, confirmed no alternative cause for the stroke symptoms other than focal cerebral ischemia. All participants had to be over age 18 years. There were no eligibility criteria based on stroke severity or enrollment status in a treatment trial. Cases were excluded if they had CADASIL, MELAS, homocystinuria, or sickle cell anemia or if their stroke was due to vasculitis, vasospasm due to subarachoid hemorrhage, mechanical aortic valve or mechanical mitral valve, or occurred within 30 days of a vascular surgical procedure.

Baseline assessment of patients included standardized assessment of demographics; medical history; vital signs; results of baseline blood tests, pre-stroke functional status per modified Rankin Scale; and National Institutes of Health Stroke Score (NIHSS) by certified examiner.(6) Functional outcomes at 90 days post stroke onset were assessed using telephonic structured interview to obtain Oxford Handicap Scale, Glasgow Outcome Scale and Barthel Index.(6) To minimize center-to-center variability, a single vascular neurologist (Robert D. Brown, Jr., MD) reviewed all available records of every ischemic stroke case for purposes of classification by etiology and syndrome using the TOAST criteria,(2) along with the Baltimore-Washington criteria(7) and the Oxford Community Stroke Project criteria. Medical records were received from the five centers and were stripped of personal identifiers, coded with study ID, and compiled in standard fashion. A separate neurologist independently reclassified all cases using the CCS system.(1)

3.1.9 KRAKOW

All consecutive subjects with ischemic stroke (fulfilling WHO criteria(3)) who were admitted to the Stroke Unit at the Jagiellonian University (Krakow, Poland) and who provided informed consent were included in the study. The Stroke Unit serves as a stroke emergency center for one district of Krakow, Poland (200,000 inhabitants) and as a referral center for South East Poland (up to 15% of all admissions). For this on-going, prospective single-center, hospital-based study participants with first ever or recurrent strokes were recruited from January 22, 2002 to September 9, 2010. The local research ethics committee approved the study.

Participants in treatment trials were excluded. All subjects were of European origin. Stroke severity was not a criterion for inclusion or exclusion. All cases had performed clinically relevant diagnostic workup, including brain imaging with CT (100%) and/or MRI (up to 20%) as well as ancillary diagnostic investigations including duplex ultrasonography of the carotid and vertebral arteries (approximately 90%), and transthoracic echocardiography (approximately 70%). Magnetic resonance angiography (MRA), computed tomographic angiography (CTA), and ambulatory ECG monitoring, transesophageal echocardiography and blood tests for hypercoagulability were performed.

Stroke cases were classified into etiologic subtypes according to TOAST.(2) All cases were phenotyped independently by two experienced stroke neurologists with review of original imaging. Cases were subsequently classified additionally using the CCS system.(1)

3.1.10 Leuven Stroke Genetics Study (LSGS)

Cases of European descent with cerebral ischemia, defined as a clinical stroke with imaging confirmation or a TIA with a new ischemic lesion on diffusion-weighted imaging, who were admitted to the Stroke Unit of the University Hospitals (Leuven, Belgium) were enrolled in the LSGS between 2005 and 2009. All participants from the LSGS study underwent brain imaging (MRI in 91% of patients, CT in the remainder) and a standardized protocol including lab examination, carotid ultrasound or CTA and cardiac examination (echocardiography and ambulatory ECG monitoring) in all patients.

Based on clinical presentation and results from the diagnostic work-up, cases were classified into ischemic stroke etiologic subtypes according to modified TOAST criteria(2) by a single reviewer. The reviewer had access to all information and imaging.

Large-vessel disease was defined as either occlusive or significant stenosis (corresponding to > 50% luminal diameter reduction according to North American Symptomatic Carotid Endarterectomy Trial (NASCET) criteria(11)) of a clinically relevant pre-cerebral or cerebral artery, presumably due to atherosclerosis. Carotid ultrasound was used as a screening tool, and in principle, additional imaging with CTA or MRA was performed when a high-grade stenosis was identified. In case CTA was used as the primary imaging modality, stenosis was confirmed by carotid ultrasound. In case of posterior circulation infarcts on imaging, CTA or MRA was used as the primary imaging modality to determine the degree of stenosis. Probable causes of cardiac embolism were excluded. The presence of a patent foramen ovale was not considered a cardiac source in this context. Intracranial atherosclerosis was considered present only if repeat imaging after at least one week revealed a similar degree of stenosis or persistent occlusion. If not, the findings were interpreted as an embolism from a proximal source.

Small-vessel disease was defined as a symptomatic infarct of < 20 mm on DWI in areas supplied by single, small penetrating branches from middle cerebral artery, posterior cerebral artery or basilar artery in the absence of both a cardioembolic source and significant stenosis/occlusion due to atherosclerosis of an appropriate major brain artery.

Cardioembolic stroke was defined as ischemic stroke in the presence of atrial fibrillation, sick sinus syndrome, myocardial infarction in the past four weeks, cardiac thrombus, infective endocarditis, atrial myxoma, prosthetic mitral or aortic valve, valvular vegetations, left ventricular akinetic segment, dilated cardiomyopathy, or patent foramen ovale or atrial septal aneurysm. Significant stenosis/occlusion due to atherosclerosis of an appropriate pre-cerebral or cerebral artery should be excluded.

Other determined cause of stroke included those with arterial dissection, vasculitis, hematologic disorders, monogenic syndromes and complications of cardiovascular procedures. Dissection was diagnosed by typical findings on contrast-enhanced MRA and T1-fat suppressed MRI. Cryptogenic stroke was defined when no cause was identified despite an extensive evaluation. Strokes associated with significant aortic arch atheroma with plaques of ≥ 4 mm were also considered cryptogenic strokes. In addition to this primary classification, cases were reclassified using CCS.(1)

3.1.11 Lund Stroke Register (LSR)

The LSR is an ongoing study including consecutive subjects with first-ever stroke since March 1, 2001 from the local uptake area of Skåne University Hospital, Lund (Sweden). Stroke was defined using the WHO criteria.(3) Subjects aged 18 years or older with stroke caused by cerebral infarct, intracerebral hemorrhage or subarachnoid hemorrhage are included. Cases are included regardless of stroke severity, race-ethnic group belonging, or participation in any treatment trial. Those with iatrogenic or traumatic stroke are excluded.

In the discovery phase of the SiGN study, subjects from LSR with first-ever ischemic stroke between March 1, 2001 and February 28, 2010 were included if they or their next of kin provided informed consent. Age over 90 years was set to 90 years to maintain anonymity. Every participant underwent CT, MRI, or autopsy of the brain; and ECG. Echocardiography, ultrasound, CTA or MRA of cerebral arteries was performed when judged clinically relevant. The subtype of ischemic stroke was determined using CCS.(1)

For the replication phase of SiGN, LSR individuals not included in the SiGN discovery phase participated after genotyping in the South Swedish genome-wide association study as follows: first-ever ischemic stroke cases recruited in 2006 and 2010 to 2012, and age- and sex-matched LSR control subjects without stroke recruited in 2001 to 2002 and 2006 to 2007 from the same geographical area with use of the official Swedish population register.

3.1.12 Malmo Diet and Cancer (MDC) Study

The MDC study is a population-based prospective cohort study. A total of 30,447 individuals, 45 to 73 years old, 60% women, attended a baseline examination between February 1991 and September 1996. Between 1992 and 1994, a total of 6,103 randomly selected subjects attended an extended baseline examination with the purpose of studying the epidemiology of cardiovascular diseases (the MDC-cardiovascular cohort, MDC-CC). At the baseline examination, 23% of the participants were smokers, 16% used anti-hypertensive medication, 14% were obese (body mass index > 30 kg/m^2), 88% were born in Sweden and > 99% were born in Europe.

Genotyping was performed using the Illumina Infinium Omni5 platform with exome content. Incidence of stroke was monitored prospectively from the baseline examination in 1992 to 1994 until December 31, 2008. The case-finding procedures included a broad search among patients with neurological symptoms that could indicate stroke. Stroke was defined according to the WHO criteria.(3) By definition, patients with transient ischemic attacks are excluded. The stroke subtypes are coded according to International Classification of Diseases revision 9. Cerebral infarction (International Classification of Diseases code 434) is diagnosed when CT, MRI, or autopsy verifies the infarction in location corresponding to the focal neurology or excludes hemorrhage and nonvascular disease. The ischemic strokes were retrospectively classified into etiological subtypes by review of hospital records. A board-certified neurologist with expertise in cerebrovascular diseases and a specialized research nurse reviewed the records. The TOAST(2) and CCS criteria(1) were applied.

3.1.13 Middlesex County Ischemic Stroke Study (MCISS)

The MCISS was initiated as a prospective hospital based stroke registry at the New Jersey Neuroscience Institute (Edison, NJ, U.S.A.). All cases over age 18 years were included, and no specific ethnic/racial group was targeted or excluded. From 2000 to 2009, 1,139 subjects with ischemic strokes were enrolled in this registry. There was no selection criterion based upon stroke severity, and both first-ever and recurrent strokes were included. Cases that were participants in treatment trials were not excluded.

The major race/ethnic groups are Whites (67.2%), African Americans (14.3%), Asian Indians (8.2%), Hispanic (5.5%) and others (4.8%, Chinese and other Asians). All subjects with clinical suspicion of a stroke were admitted through the emergency room to a dedicated stroke unit supervised by a vascular neurologist. After a history and neurological examination, a standardized series of investigations were performed: complete blood count and differential, comprehensive metabolic panel, electrolytes, blood urea nitrogen, creatinine, lipid panel (total cholesterol, low-density lipoprotein, high-density lipoprotein, triglyceride levels, homocysteine levels, a cerebral MRI/MRA (if the MRI could not be performed, a head CT scan was done), carotid duplex ultrasound, ECG and an echocardiogram. The diagnosis of cerebral infarct was confirmed by the imaging studies.

The epidemiological and clinical data on these participants was collected prospectively. Two independent investigators (one of which was a board-certified neurologist with expertise in vascular neurology) reviewed the data, and all strokes were classified into etiological subtypes using TOAST criteria.(2) In addition, the Oxfordshire stroke classification(12) was applied, and the vascular distribution of stroke was tabulated. All procedures, including the generation of the databases and recruitment of the stroke subjects, were conducted following Institutional Review Board policies and procedures at the New Jersey Neuroscience Institute/JFK Hospital.

3.1.14 Miami Stroke Registry and Biorepository (MIAMISR)

The MIAMISR at the University of Miami/Jackson Memorial Hospital (Miami, FL, U.S.A.) is an ongoing prospective hospital registry of consecutive patients subjects with prevalent stroke (ischemic and hemorrhagic) and TIA with available neuroimaging (CT or MRI) who provide informed consent. There are no specific exclusion criteria with the respect to age, stroke severity, disability or participation in treatment trials. It was established in November of 2008 in order to investigate stroke type, ischemic stroke subtypes, stroke genetics and stroke outcomes in diverse ethnic population of Miami.

The stroke population is predominately Hispanic (63%), with Cuba (32%), Nicaragua (4.8%), Colombia (4.8%), and Puerto Rico (4.1%) contributing the most subjects. Jackson Memorial Hospital is a 1,550-bed county hospital affiliated with the University of Miami with approximately 900 stroke and TIA admissions per year. Demographic and clinical data along with blood samples for genetic and other research have been collected prospectively during the hospitalizations. Follow-up information was obtained at 90 days by telephone interview or in person. Trained research staff obtained written informed consent from the stroke patients or the health care proxy when available for participation in MIAMISR.(13)

3.1.15 Genes Affecting Stroke Risk and Outcome Study (MGH-GASROS)

MGH-GASROS enrolled ischemic stroke subjects as part of a single-center prospective cohort study of consecutive patients with ischemic stroke aged ≥18 years admitted to the Massachusetts General Hospital Stroke Unit (Boston, MA, U.S.A.) between 2003 and 2011 after presenting to the emergency department within 24 hours of symptom onset. Ischemic stroke was defined as a clinical syndrome of any duration associated with a radiographically proven acute infarct consistent with a vascular pattern of involvement and without radiographic evidence of a demyelinating or neoplastic disease or other structural disease, including vasculitis, subacute bacterial endocarditis, vasospasm due to subarachnoid hemorrhage or cocaine abuse, or primary intracerebral hemorrhage.

Diagnosis of acute cerebral ischemia was confirmed for all subjects in the present study by admission diffusion weighted imaging (DWI) completed within 48 hours after symptom onset. Vascular and critical care neurologists subtyped ischemic strokes by systematic medical record review using the TOAST(2) and CCS systems.(1) Controls were matched to cases on the basis of age, sex and race/ethnicity and drawn from stroke-free individuals who received care at primary care practices within Massachusetts General Hospital.

3.1.16 MUNICH

Subjects with first-ever or recurrent ischemic stroke were recruited consecutively from a single dedicated stroke unit (Klinikum Groβhadern, Ludwig-Maximilians-University of Munich, Germany) from 2002 onward. All participants were over the age of 18 years and of European descent. Brain imaging was performed in all cases, with most patients (> 80%) undergoing MRI, including DWI. Diagnosis of ischemic stroke was based on neurological symptoms in combination with a documented acute infarct on neuroimaging. Subjects were not excluded based on stroke severity or whether they were enrolled in a treatment trial. Diagnostic workup included ECG and duplex ultrasonography of the extracranial carotid arteries in all cases. Transcranial ultrasonography, CTA and/or MRA, transthoracic and transesophageal echocardiography, and ambulatory ECG were performed if clinically indicated.

QC was identical for all WTCCC cohorts, as described in **Section 3.1.4**.

3.1.17 Nurses' Healthy Study (NHS)

The NHS cohort consists of 121,700 female registered nurses aged 30 – 55 years who were residing in 11 U.S. states and who were enrolled in 1976 through responding to a mailed questionnaire on their medical history and lifestyle practices. They have been followed with biennial mailed questionnaires collecting information on disease risk factors and health status.

From 1989 – 1990, blood samples were collected from 32,826 participants. Among these participants, we prospectively identified incident strokes and confirmed ischemic stroke cases by medical record review. Clinical symptoms consistent with stroke and exclusion of alternate etiologies were required for classification of stroke. Virtually all cases had imaging, but confirmation on CT or MRI was not required. No participants were excluded based on race/ethnicity. Neither stroke severity nor enrollment in a treatment trial was part of the eligibility criteria.

3.1.18 Northern Manhattan Study (NOMAS)

NOMAS is an ongoing population-based study designed to determine stroke incidence, risk factors and outcome in an urban multiethnic population.(14) NOMAS started in 1993 as a case-control study of index ischemic stroke cases admitted to the Columbia University Presbyterian Medical Center (New York, NY, U.S.A.) and affiliated hospitals and matching community controls (Northern Manhattan Stroke Study, NOMASS) and continued as a prospective stroke incidence study by following up controls in 1997 (NOMAS). Demographic and clinical data were collected prospectively during the hospitalizations and annually by phone or in person.

Genetic samples were derived from two sources:

(a) the population-based case-control study conducted from 1993-98 (NOMASS) and
(b) the ongoing prospective cohort study (NOMAS).

First-ever ischemic stroke cases were identified for the case-control study by screening of patient admissions, discharge codes, and referrals for neuroimaging at 15 acute care hospitals in the defined study area and multiple approaches to monitor for non-hospitalized cases.

Incident ischemic stroke cases were identified from the prospective cohort study through follow-up visits and scheduled telephone contacts. Ischemic stroke cases from both sources were followed at 6 months by telephone and then annually afterwards in order to assess functional status and other outcomes. The administrative coordinating center of NOMAS moved from New York to Miami in 2007. The Institutional Review Boards of both institutions, Columbia University and the University of Miami (Miami, FL, U.S.A.), approved the study.

3.1.19 Oxford Vascular Study (OXVASC)

OXVASC is an on-going population-based study of the incidence and outcome of cerebrovascular, cardiovascular and peripheral vascular events since April 1, 2002. The OXVASC study population comprises all 91,105 individuals, irrespective of age, registered with 101 general practitioners in 9 general practices in Oxfordshire, UK. Multiple overlapping methods of "hot" and "cold" pursuit are used to achieve near complete ascertainment of as many cases as possible. All subjects are consented and seen by study physicians as soon as possible after their initial presentation.

In the SiGN study, cases of all ethnic groups from OXVASC with any ischemic stroke between April 1, 2002 and August 31, 2010 were included if they consented to have research DNA samples extracted. Ischemic stroke was defined as an episode of focal neurological deficits with acute onset lasting > 24 hours or until death, with no apparent non-vascular cause, and no signs of primary haemorrhage on brain imaging. An infarct did not need to be seen on CT or MRI to be included in this study. Cases were not excluded if they were of a treatment trial or for their stroke severity.

Demographic data, major vascular risk factors (hypertension, diabetes, smoking, hyperlipidemia, prior TIA and history of coronary disease or peripheral vascular disease), and symptomatology were recorded in all patients. Cases routinely had brain imaging (CT or MRI), vascular imaging (carotid Doppler or CTA /MRA or digital subtraction angiography), and 12-lead ECG. Echocardiography and 24-hour ambulatory ECG monitoring were done in selected patients.

A senior neurologist subsequently reviewed all cases, and stroke etiology was classified according to modified TOAST criteria.(2) Risk factors such as hypertension and diabetes were not included in the criteria. The subjects were classified as undetermined stroke only if the diagnostic workup was complete (any form of brain imaging plus ECG and any form of vascular imaging), but no clear etiology was found. Those with incomplete investigation were classified as unknown stroke while stroke of multiple causes was classified separately.

QC was identical for all WTCCC cohorts, as described in **Section 3.1.4**.

3.1.20 Reasons for Geographic and Racial Differences in Stroke (REGARDS)

The REGARDS study is a U.S. national, population-based, longitudinal cohort of 30,239 African American and white adults aged ≥ 45 years, recruited January 2003 to October 2007 with ongoing follow-up. Suspected stroke is queried every six months and triggered by participant self-report of stroke, stroke symptom(s), hospitalization, or proxy report of death. Stroke severity and participation in a treatment trial did not limit inclusion in this study.

Medical records for these reported events are retrieved and reviewed by at least two members of a committee of stroke experts with disagreements resolved by a third adjudicator. A symptom-based approach, independent of neuroimaging outcome, is used to confirm events using the WHO definition of stroke.(3) An infarct did not need to be seen on brain imaging to be included in this study. Ischemic stroke subtype classification is conducted using the TOAST system.(2,15)

3.1.21 Sahlgrenska Academy Study on Ischemic Stroke (SAHLSIS)

SAHLSIS is a case-control study of ischemic stroke based in Gothenburg, Sweden.(16) Adult subjects who presented with first-ever or recurrent acute ischemic stroke before 70 years of age were recruited consecutively at stroke units in western Sweden from 1998 to 2012. All participants were of European origin. Patients were not excluded based on stroke severity or whether they were enrolled in a treatment trial. All participants underwent ECG and neuroimaging at the acute stage (all by CT and 58% also by MRI). Additional diagnostic work-up was performed when clinically indicated. Inclusion criteria was ischemic stroke which was defined as an episode of focal neurological deficits with acute onset and lasting > 24 hours or until death, with no apparent non-vascular cause, and no signs of primary hemorrhage on brain imaging. Subjects were excluded if they had a diagnosis of cancer at advanced stage, infectious hepatitis or human immunodeficiency virus. Ischemic stroke was assigned according to modified TOAST criteria.(17) Cases in this study were also classified using the CCS system.(1)

3.1.22 Secondary Prevention of Small Subcortical Strokes (SPS3)

The SPS3 trial (NCT00059306, **Online Source [1]**) is a randomized, multicenter, Phase 3 trial of antiplatelet therapy and antihypertensive therapy. Participants are randomized to aspirin alone or the combination of aspirin and clopidogrel. Participants are also randomized to two groups of blood pressure control: either to a target systolic blood pressure of 130 – 149 mm Hg or < 130 mm Hg. Principal eligibility criteria include man or woman at least 30 years of age with clinical evidence of small subcortical stroke and brain MRI evidence of small subcortical infarct. Subjects were required to not have evidence of ipsilateral symptomatic cervical carotid stenosis or high-risk cardioembolic sources for embolism. Further details of eligibility criteria have been published.(18)

Primary outcomes include ischemic and hemorrhagic stroke. DNA samples were collected from 38% (1,139/3,020) of participants in the trial. These samples were obtained from 46% (37/81) participating centers across the U.S., Canada, Spain, Mexico, Chile, Ecuador and Peru. No additional eligibility criteria were necessary beyond informed consent for participating in the DNA sub-study. A total of 0.9% (10/1,139) of DNA donors gave sample at time of randomization, with the remainder donating at a later time point in follow-up.

3.1.23 St. George's Hospital (STGEORGE)

First-ever and recurrent ischemic stroke cases of European descent attending a cerebrovascular service were recruited from 1995 to 2008. All cases were phenotyped by one experienced stroke neurologist with review of original imaging. All participants had clinically relevant diagnostic workup performed, including brain imaging with CT and/or MRI as well as ancillary diagnostic investigations including duplex ultrasonography of the carotid and vertebral arteries or MRA/CTA, blood tests, and ECG, and where clinically indicated echocardiography and ambulatory ECG monitoring was performed. Cases were enrolled only if a symptomatic acute infarct was detected on head imaging. Participants had to be over the age of 18 years and have provided informed consent. No case was excluded for participation in a treatment trial or because of stroke severity. An algorithm was established to use the clinical trials database to automatically populate the web-based CCS tool to generate CCS stroke subtype diagnoses.(1)

QC was identical for all other WTCCC cohorts, as described in **Section 3.1.4**.

3.1.24 Siblings with Ischemic Stroke Study (SWISS)

SWISS is a prospective, hospital-based affected sibling pair study of ischemic stroke. Enrollment for SWISS began in December 2000 and was completed in February 2011. DNA samples were collected from 312 ischemic stroke-affected sibling pairs. During this time, 1,026 cases with first-ever or recurrent ischemic stroke were enrolled across 70 centers in North America (66 in the U.S. and 4 in Canada).

All probands required at least one living sibling with a history of stroke and required meeting the WHO definition for stroke(3) with head imaging, by either head MRI or CT, confirming no alternative cause for the stroke symptoms other than focal cerebral ischemia. Probands were excluded if they had CADASIL, MELAS, homocystinuria, or sickle cell anemia or if their stroke was due to vasculitis, vasospasm due to subarachoid hemorrhage, mechanical aortic valve or mechanical mitral valve, or occurred within 30 days of a vascular surgical procedure.

Baseline assessment of cases included standardized assessment of demographic and medical history. Siblings were recruited primarily using proband-initiated contact. Stroke-affected siblings were screened using the Questionnaire for Verifying Stroke-free Status (QVSS).(19) Eligibility criteria for affected siblings were the same as for probands. The Stroke Verification Committee, composed of two vascular neurologists, confirmed ischemic stroke status in affected siblings by medical record review.

For probands, the center principal investigator classified ischemic stroke using the original TOAST classification system.(2) Center principal investigators were neurologists certified in TOAST classification using stroke vignette training and certification process. The Stroke Verification Committee classified all affected siblings using TOAST based on medical record review. The Committee received medical records stripped of personal identifiers, coded with study identification number, and compiled in standard fashion. All participants gave written informed consent for participation in the study, and the local institutional review boards of each individual clinical center and the Mayo Clinic institutional review board approved the study.

3.1.25 Vitamin Intervention for Stroke Prevention (VISP)

The VISP trial was a multicenter, randomized double-blind controlled clinical trial that enrolled subjects aged 35 years or older with homocysteine levels above the 25th percentile at screening and a non-disabling cerebral infarction (NDCI) within 120 days of randomization.(20) Non-disabling cerebral infarction (NDCI) was defined as an ischemic brain infarction not due to embolism from a cardiac source, characterized by the sudden onset of a neurological deficit. The deficit must have persisted for at least 24 hours, or, if not, an infarction in the part of the brain corresponding to the symptoms must have been demonstrated by CT or MRI.

The trial was designed to determine if daily intake of a multivitamin tablet with high-dose folic acid, vitamin B_6 and vitamin B_{12} reduced recurrent cerebral infarction (primary endpoint), and nonfatal myocardial infarction or mortality (secondary endpoints). Subjects were randomly assigned to receive daily doses of the high-dose formulation (N = 1,827), containing 25mg pyridoxine (B_6), 0.4mg cobalamin (B_{12}), and 2.5mg folic acid; or the low-dose formulation (N = 1,853), containing 200µg pyridoxine, 6µg cobalamin and 20µg folic acid. Enrollment began in August 1997 and ended in December 2001, with 3,680 participants enrolled, from 55 clinic sites across the US and Canada and one site in Scotland. All participants provided written informed consent, and all local governing institutional review boards approved the trial. A subset of VISP participants provided separate consent for genetic analyses.

3.1.26 Women's Health Initiative Observational Study (WHI-OS)

The Women's Health Initiative Observational Study (WHI-OS) is a long-term follow-up study of post-menopausal women to identify and assess the effects of biological, genetic and lifestyle risk factors for cancer, cardiovascular disease, osteoporosis and other diseases of older women. The cases submitted here came from a case-control ancillary study nested within the WHI-OS of the first 972 strokes occurring after WHI-OS baseline. This case-control study was the Hormones and Biomarkers Predicting Stroke Study (HaBPS), conducted to examine blood biomarkers in relation to stroke. Forty clinical centers throughout the United States enrolled 93,676 women ages 50 to 79 years at baseline into the parent study, the WHI-OS, between September 1993 and February 28, 1997. Follow-up for clinical events and exposures is ongoing.

Recruitment into WHI-OS was mostly through mass mailings to age-eligible women from large mailing lists such as voter registration, driver's license, Health Care Financing Administration, or other insurance lists. Recruitment of minorities and older women was a particular study objective. Women were either specifically recruited for the Observational Study or entered it because they were ineligible or unwilling to be randomized into the Women's Health Initiative Clinical Trials of hormone therapy and/or dietary modification. Exclusions from WHI-OS were participation in other randomized trials, predicted survival of < 3 years, alcoholism, drug dependency, mental illness, dementia, or other conditions making them unable to participate in the study. Exclusions for the HaBPS case-control study of biomarkers of stroke were women with prior history of myocardial infarction or stroke or those who did not have adequate blood sample for biomarker assays.

Strokes were first identified through annual mail and/or telephone follow-up, and participant or third-party reports of overnight hospitalizations which were further investigated by obtaining laboratory results, medical records, and available imaging study reports. Trained local physician adjudicators assigned a diagnosis according to standard criteria. Locally adjudicated strokes were sent for central adjudication by three neurologists. Two neurologists adjudicated each potential case, and disagreements were resolved by conference call consensus of the three neurologists.

Only centrally confirmed ischemic strokes that required hospitalization were used in this study. TIAs and hemorrhagic strokes (determined on review of reports of brain imaging studies) were excluded. Ischemic stroke was defined as the rapid onset of a persistent neurologic deficit attributed to a vessel occlusion lasting more than 24 hours and without evidence for other causes. The deficit must have lasted > 24 hours unless death supervened or there was a lesion compatible with acute stroke demonstrated on CT or MRI scan. Ischemic strokes were also centrally classified by TOAST(2) and CCS criteria.(1)

3.1.27 Washington University St. Louis (WUSTL) Study

The WUSTL patient collection included ischemic stroke cases admitted to Barnes-Jewish Hospital/Washington University Medical Center (St. Louis, MO, U.S.A.) for genetic studies starting from August 1, 2008. Participants were identified for the genetic studies by screening admissions at our tertiary care hospital (both in the Emergency Department and on the Inpatient Stroke Service) without regard to age, race or ethnicity, including both first-ever and recurrent strokes. Subjects were retained in the study if their discharge diagnosis was ischemic stroke (without requirement for the stroke to be visualized on CT or MRI). Demographic and clinical data were collected prospectively during the hospitalization and at 90 days, by phone or in person.

Genetic samples were derived from subjects enrolled in 3 different studies:

(a) Acute tPA pharmacogenomics study (Ischemic stroke cases who received tPA and were admitted to BJH/Washington University; serial NIHSS scores,(6) and data on hemorrhagic transformation was collected)

(b) Recovery Genomics after Ischemic Stroke Study (ReGenesIS, Ischemic stroke cases with NIHSS > 3 points without underlying chronic neurological disease, and expected survival up to 3 months after stroke), and

(c) the Cognitive Recovery and Rehabilitation Group (CRRG) Registry (all ischemic stroke cases admitted to BJH/Washington University who consent to entering their clinical data into a stroke registry, and the collection of blood for genetic analysis).

Cases that were part of a treatment trial were excluded from the tissue plasminogen activator pharmacogenomics and ReGenesIS study, but not the CRRG registry.

3.2 Controls for stage I (discovery) analysis

3.2.1 Attention-deficit Hyperactivity Disorder (ADHD)

The Vall d'Hebron Research Institute (VHIR) cohort included 435 blood donors of Caucasian origin recruited from 2004 to 2008 at the Hospital Universitari Vall d'Hebron (Barcelona, Spain) to identify loci conferring susceptibility to Attention-Deficit Hyperactivity Disorder. Seventy-six percent of participants were male (N = 330) and the average age at assessment was 43.8 years (s.d. = 14.3). Genome-wide genotyping was performed with the Illumina HumanOmni1-Quad BeadChip platform. The study was approved by the ethics committee of the institution and informed consent was obtained from all participants in accordance with the Declaration of Helsinki.

3.2.2 Australian Stroke Genetics Collaborative (ASGC)

ASGC controls were participants in the Hunter Community Study (HCS), a population-based cohort of individuals aged 55–85 years, predominantly of European ancestry and residing in the Hunter Region in New South Wales, Australia. Detailed recruitment methods for the HCS have been previously described.(21) Briefly, participants were randomly selected from the New South Wales State electoral roll and were contacted by mail between 2004 and 2007. Consenting participants completed five detailed self-report questionnaires and attended the HCS data collection center, at which time a series of clinical measures were obtained. A total of 1,280 HCS participants were genotyped for the current study.

3.2.3 Genetics of Early Onset Stroke (GEOS)

See **Section 3.1.6**

3.2.4 GRAZ

The Austrian Stroke Prevention Study (ASPS) is a single-center, prospective follow-up study on the cerebral effects of vascular risk factors in the normal elderly population of the city of Graz, Austria. Details have been described elsewhere.(22,23)

Briefly, 2007 participants of European descent aged between 45 and 90 were randomly selected from the official community register. Individuals were excluded from the study if they had a history of neuropsychiatric disease, including previous stroke, transient ischemic attacks, and dementia, or an abnormal neurologic examination determined on the basis of a structured clinical interview and a physical and neurologic examination. During 2 study periods between September 1991 and March 1994 and between January 1999 and December 2003 an extended diagnostic work-up including MRI and neuropsychological testing was done in 1,076 individuals. A total of 815 genotyped samples were included in the current study.

3.2.5 Health ABC (HABC)

The Health Aging and Body Composition (Health ABC) Study is a National Institute on Aging (NIA)-sponsored cohort study of the factors that contribute to incident disability and the decline in function of healthier older persons, with a particular emphasis on changes in body composition in old age. Between April 15, 1997 and June 5, 1998, the Health ABC study recruited 3,075 70 – 79 year old community-dwelling adults (41% African American), who initially had no indications of disability related to mobility and activities of daily living. The key components of Health ABC include a baseline exam, annual follow-up clinical exams, and phone contacts every six months to identify major health events and document functional status between clinic visits.

The core yearly examination for Health ABC includes measurement of body composition by dual energy x-ray absorptiometry (DXA), walking ability, strength, an interview that includes self-report of limitations and weight, and a medication survey. At baseline, visceral adiposity was measured by computerized tomography (CT). Provision has been made for banking of blood specimens and extracted DNA (HealthABC repository). The overall goal of this project is to identify genetic determinants of visceral adiposity.

3.2.6 The Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) was initiated in 2006 to investigate the prevalence and risk factors affecting several health conditions, including heart, lung and blood disorders, kidney and liver function, diabetes, cognitive function, dental conditions and hearing disorders.(24,25) Participants aged 18 – 74 self-identified as Hispanic or Latino, with substantial representation of Mexican, Puerto Rican, Dominican, Cuban, Central and South American groups. They were recruited from four field centers in the United States: San Diego, CA; the Bronx, NY; Chicago, IL; and Miami, FL. 12,803 study participants consented to genetic studies and will be included in the HCHS/SOL dbGaP posting.

Genotyping of the HCHS/SOL participants was performed at Illumina Microarray Services using the SOL HCHS Custom 15041502 array (annotation version “B3”, genome build 37), which includes 2,575,443 variants (of which 2,427,090 are in common with the Illumina HumanOmni2.5 and 148,353 are custom content).

3.2.7 Health and Retirement Study (HRS)

The University of Michigan Health and Retirement Study (HRS) is a longitudinal panel study that surveys a representative sample of more than 20,000 Americans over the age of 50 every two years. Supported by the National Institute on Aging (NIA U01AG009740) and the Social Security Administration, the HRS explores the changes in labor force participation and the health transitions that individuals undergo toward the end of their work lives and in the years that follow.

Since its launch in 1992, the study has collected information about income, work, assets, pension plans, health insurance, disability, physical health and functioning, cognitive functioning, and health care expenditures (**Online Sources [2]**).

HRS is intended to be a nationally representative sample with 2:1 oversampling of minority groups including African-American and Hispanic/Latino populations.(26) In Phases I – II, 12,507 study participants were included in the dbGaP posting.

Genotyping of the HRS Phase I – II participants was performed at CIDR using the Illumina HumanOmni2.5-4v1 array (annotation version "D", genome build 37) and released a total of 2,443,179 variants.

3.2.8 INfancia y Medio Ambiente (INMA)

The INfancia y Medio Ambiente (Environment and Childhood) project is a research project comprising a Spanish population-based birth cohort created to study the role of the environmental pollutants during pregnancy and first stages of life and their effects on childhood growth and development. The cohort was established between 2003 and 2008 from mothers enrolled in four regions within Spain and included their infants (**Online Sources [3]**).

3.2.9 MONICA/KORA Augsburg Study

For the German MUNICH discovery samples and the Stroke in Young Fabry Patients (SIFAP) replication samples, independent control groups were selected from Caucasians of German origin participating into the population KORAgen study (**Online Sources [4]**).This survey represents a sex- and age stratified random sample of all German residents of the Augsburg area and consists of individuals 25 – 74 years of age, with about 300 subjects for each 10-year increment. All controls were free of a history of stroke or transient ischemic attack.

KORA samples were genotyped on the Illumina Human 550k platform. QC was identical for all WTCCC cohorts, as described in **Section 3.1.4**.

3.2.10 KRAKOW

See **Section 3.1.9** for information on cases.

The control group included unrelated subjects taken from the population of southern Poland. Control subjects had no apparent neurological disease based on the findings in a structured questionnaire and a neurological examination. Local research ethics committees approved the study and informed consent was obtained from all participants.

3.2.11 Leuven Stroke Genetics Study (LSGS)

See **Section 3.1.10** for information on cases

Control individuals were recruited in the same population amongst healthy individuals, spouses of patients suffering from neurological diseases (amyotrophic lateral sclerosis, ischemic stroke or multiple sclerosis), and from the Leuven University Gerontology Database as previously described.(27)

3.2.12 Malmo Diet and Cancer (MDC) Study

See **Section 3.1.12**

Controls from this prospective study were used for discovery samples of the Lund Stroke Register (LSR) and Sahlgrenska Academy Study on Ischemic Stroke (SAHLSIS).

3.2.13 Osteoarthritis Initiative (OAI)

The OAI is a publicly and privately funded prospective longitudinal cohort with a primary objective of identifying risk factors for incidence and progression of tibiofemoral knee OA. The OAI utilized a focused population-based recruitment to enroll 4,674 men and women between the ages of 45 – 79 years who either had radiographic symptomatic knee OA or who were without radiographic symptomatic OA in both knees but were considered high risk for OA because they had two or more known risk factors for knee OA. Subjects were recruited into the baseline phase of the OAI at multiple sites throughout the US between 2004 and 2006. All subjects were invited back for follow-up examinations to assess incidence or progression of OA annually, for up to 5 years.

Phenotype data from the baseline and follow-up examinations are available for public access from the Osteoarthritis Initiative (OAI) database (**Online Sources [5]**).

The Genetic Components of Knee Osteoarthritis (GeCKO) Study was initiated in 2009 as a genetic ancillary study to perform a genome-wide association study to identify genetic variants associated with radiographic osteoarthritis. This study included 4,482 individuals participating in the parent OAI study genotyped on the Illumina HumanOmni2.5M

3.2.14 Vitamin Intervention for Stroke Prevention (VISP)

All VISP participants are stroke cases, therefore, for cases of European descent, we obtained GWAS data from (dbGAP) for 1047 external controls from the High Density SNP Association Analysis of Melanoma: Case- Control and Outcomes Investigation (Study Accession: phs000187.v1.p1).(28,29) These samples were genotyped on the Illumina HumanOmni1-Quad.

For case-control analyses using cases of African ancestry, we used controls from the Healthy Aging in Neighborhoods of Diversity across the Life Span Study (HANDLS). HANDLS is an interdisciplinary, community-based, prospective longitudinal epidemiologic study examining the influences of race and socioeconomic status (SES) on the development of age-related health disparities among socioeconomically diverse African Americans and whites in Baltimore, MD, USA.(30) This study assessed physical parameters over a 20-year period while evaluating genetic, biologic, demographic, and psychosocial influences.

HANDLS recruited 3,722 participants (2,200 African Americans (59%) and 1,522 whites (41%)) from Baltimore, MD. Genotyping was focused on a subset of participants self-reporting as African American and was performed at the Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health. Genotype data (for up to 907,763 SNPs) were generated for 1,024 participants using either Illumina 1M and 1M duo arrays (N = 709), or a combination of 550K, 370K, 510S and 240S to be roughly equally to the coverage on the Illumina 1M.

Inclusion criteria for genetic data in HANDLS includes concordance between self-reported sex and sex estimated from X chromosome heterogeneity, > 95% call rate per participant (across all equivalent arrays), concordance between self-reported African ancestry and ancestry confirmed by analyses of genotyped SNPs, and no cryptic relatedness to any other samples at a level of proportional sharing of genotypes > 15% (effectively excluding 1st cousins and closer relatives from the set of probands used in analyses).

HANDLS study controls were used for the VISP and SWISS-ISGS African American case-control analyses (with no overlap across studies).

3.2.15 Wellcome Trust Case-Control Consortium (WTCCC)

The WTCCC2 samples were genotyped as part of the WTCCC2 ischemic stroke study. Stroke cases included samples recruited by investigators at St. George's University London (SGUL, London, U.K), the University of Oxford in the UK (OXVASC), the Department of Neurology, Klinikum Großhadern, Ludwig-Maximilians- University, Munich (Munich, Germany), and the University of Edinburgh collection. For further description of the WTCC2 cases, please see **Sections 3.1.4** (Edinburgh (ESS)), **3.1.16** (Munich), **3.1.19** (OXVASC), and **3.1.23** (SGUL or STGEORGE).

Controls for the UK samples were drawn from shared WTCCC controls obtained from the 1958 Birth Cohort. This is a prospectively collected cohort of individuals born in 1958 (**Online Sources [6]**), and ascertained as part of the national child development study (**Online Sources [7]**). Data from this cohort are available as a common control set for a number of genetic and epidemiological studies.

For the German cases, controls were Caucasians of German origin participating into the population KORAgen study (**Online Sources [8]**). For more information, see **Section 3.2.9**.

QC was identical for all WTCCC cohorts, as described in **Section 3.1.4**.

4. Studies for stage II analyses

4.1.1 Cervical Artery Dissection and Ischemic Stroke Patients (CADISP)

The CADISP study includes subjects with ischemic stroke without cervical artery dissection. They were recruited from the same centers as cervical artery dissection patients for a specificity analysis. These were cases with a diagnosis of ischemic stroke, for whom dissection had been formally ruled out according to CADISP inclusion criteria. Non-dissection ischemic stroke cases were frequency-matched on age (by 5-year intervals) and sex with dissection patients.

A total of 658 non-dissection ischemic stroke cases were included. We excluded 19 cases due to unavailability of geographically matched healthy controls, or due to non-European origin; of the remaining 639 ischemic stroke cases, 613 individuals had good quality DNA available and were genotyped at the Centre National de Génotypage CNG. Of these, a total of 555 non-dissection ischemic stroke cases aged < 60 years, who were successfully genotyped and met genotyping quality control criteria, were available for the SiGN analysis.

4.1.2 Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE)

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium was formed to facilitate genome-wide association study (GWAS) meta-analyses among multiple large and well-phenotyped longitudinal cohort studies.(31) The CHARGE consortium performed a GWAS of incident stroke on 84,961 participants of European origin who belonged to 18 community-based prospective cohort studies and sub-studies conducted in Europe and the USA:

(a) Age, Gene/Environment Susceptibility (AGES) - Reykjavik Study
(b) Atherosclerosis Risk in Communities (ARIC) study
(c) Cardiovascular Health Study (CHS)
(d) Framingham Heart Study (FHS)
(e) FINRISK surveys (CoreExome, Corogene and PredictCVD)
(f) Health, Aging, and Body Composition (Health ABC) Study
(g) Rotterdam Study I and II
(h) Study of Health in Pomerania (SHIP)
(i) Women's Genome Health Study (WGHS)
(j) Multi-Ethnic Study of Atherosclerosis (MESA)
(k) PROspective Study of Pravastatin in the Elderly at Risk (PROSPER)
(l) TWINGENE; Uppsala Longitudinal Study of Adult Men (ULSAM)
(m) 3C-Study (Dijon and Bordeaux-Montpellier)

All participants were free of stroke at baseline and 4,348 of them sustained an incident stroke during 13 years of follow-up on average. Stroke was defined as a focal neurologic deficit of presumed vascular origin with a sudden onset and lasting for at least 24 hours, or until death if the participant died less than 24 hours after the onset of symptoms. Strokes were classified as ischemic (N = 3,100), hemorrhagic (N = 277), or unknown type (N = 971) based on clinical and imaging criteria. Ischemic strokes were further subdivided into cardioembolic (N = 602) and non-cardioembolic subtypes (N = 1,770). The subtyping algorithms varied between studies but most studies with ischemic stroke subtype information had categories broadly conforming to TOAST subtypes.

Imputed genotypes based on the 1000 Genomes "All" reference panel (phase I, version 3)(32) were used for association testing by all participating studies except one study which used imputed genotypes based on HapMap3 panel.(33) The primary analyses were genome-wide multivariable Cox regressions testing the association of genetic variants with incident stroke and with incident ischemic stroke under an additive genetic model, adjusting for sex, age, and when relevant, principal components of population stratification, study site or familial structure. Cox regression analyses were run at each study site, after the proportional hazards assumption was verified, using as the end point time to event, namely time between the date of DNA draw and occurrence of first stroke.

Participants known to be stroke-free were right censored at death or at the time of their last follow-up examination or health status update. For all analyses, participants with subarachnoid hemorrhage were censored at time of event. For analyses of association with incident ischemic stroke, participants were also right-censored when they had an alternative type of stroke, hemorrhagic or unknown. In secondary analyses we tested for associations with incident ischemic stroke subtypes and incident intracerebral hemorrhage. For these analyses participants were also right-censored when they had an alternative type of stroke or alternative subtype of ischemic stroke.

Genetic variant and allele names were harmonized across all studies, duplicate markers and markers not present in the 1000G pI v3 reference panel were removed prior to meta-analysis. Only genetic variants with absolute value of regression coefficient < 5, standard error > 0 and < 10,000, and effective allele count > 10 were retained for analysis. Effective allele count was defined as twice the product of minor allele frequency, imputation accuracy (r^2), and number of cases. We also restricted our analyses to common variants with a minor allele frequency > 0.05. Moreover, genetic variants available in less than five studies or in less than 50% of the total number of cases were discarded. Genomic control was applied before and after meta-analysis. Meta-analysis was performed using fixed effect inverse variance weighting as implemented by METAL.(34)

4.1.3 deCODE

Cases, irrespective of age, were identified from a registry of individuals diagnosed with ischemic stroke or TIA at Landspitali University Hospital (Reykjavik, Iceland), the only tertiary referral center in Iceland, from 1993 – 2006. The ischemic stroke or TIA diagnoses were based on WHO criteria and either CT or MRI, and were clinically confirmed by neurologists.

Eligible cases who survived the stroke were invited to participate the genetic study, either by attending a recruitment center for deCODE's genetic studies, or they were visited at their home by a study nurse. Control subjects were participants from a large variety of genetic programs at deCODE. Individuals with confirmed stroke (identified by cross-matching with hospital lists) who had participated in genetic studies other than those of cardiovascular diseases (CVD) (but not participated in CVD studies) were excluded as controls.

4.1.4 Glasgow, Scotland ImmunoChip Study Samples

Cases with ischemic stroke attending the cerebrovascular service of the Western Infirmary (Glasgow, Scotland) were recruited between 1990 and 2004 as part of an on-going study of genetic and circulating biomarkers in stroke. All subjects underwent brain imaging and extracranial carotid ultrasonography according to a standard clinical protocol. Cases were classified into ischemic stroke subtypes using TOAST criteria(2) by a team of experienced stroke physicians with review of original brain imaging. Cases of undetermined etiology were excluded from genotyping. The West Ethics Committee approved the study.

Controls for the UK samples were drawn from shared Wellcome Trust Case Control Consortium (WTCCC) controls obtained from the 1958 Birth Cohort. This is a prospectively collected cohort of individuals born in 1958 (**Online Sources [6]**), and ascertained as part of the National Child Development Study (**Online Sources [7]**). Data from this cohort are available as a common control set for a number of genetic and epidemiological studies.

4.1.5 Heart and Vascular Health (HVH) Study

The setting for this study was Group Health (GH), a large integrated health care system in western Washington State, U.S. Data were utilized from an ongoing case-control study of incident myocardial infarction and stroke cases with a shared common control group. Methods for the study have been described previously(35–37) and are briefly summarized below. The human subjects committee at GH approved the study, and all study participants provided written informed consent.

All study participants were GH members and aged 30 to 79 years. MI and stroke cases were identified from hospital discharge diagnosis codes and were validated by medical record review. Controls were a random sample of GH members frequency matched to MI cases on age (within decade), sex, treated hypertension, and calendar year of identification. The index date for controls was a computer-generated random date within the calendar year for which they had been selected. For stroke cases, the index date was the date of admission for the first acute stroke. Participants were excluded if they were recent enrollees at GH, had a history of prior stroke, or if the incident event was a complication of a procedure or surgery.

Trained medical record abstractors collected eligibility and risk factor information from a review of the GH medical record using only data available prior to the index date and through a telephone interview. Medication use was ascertained using computerized GH pharmacy records. A venous blood sample was collected from all consenting subjects, and DNA was extracted from white blood cells using standard procedures. Diagnostic criteria for ischemic stroke were adopted from the Cardiovascular Health Study.(38) These criteria included:

(a) rapid onset of neurologic deficit or subarachnoid hemorrhage,
(b) deficit persisting for longer than 24 hours unless computed tomography or magnetic resonance imaging show evidence of permanent damage, and
(c) no underlying brain trauma, tumor, or infection to cause symptoms.

These analyses were limited to ischemic stroke cases, namely those satisfying 1 or more of the following criteria:

(a) focal deficit, without evidence of blood on CT or MRI,
(b) focal deficit, with mottled appearance in the appropriate location on CT, or
(c) surgery or autopsy evidence of infarction.

Among ischemic strokes, the subtypes were defined as follows:

(a) Lacunar stroke required either:

- CT/MRI demonstrates a deep area of infarction (decreased density) less than 2 cm. across, or
- A normal CT, but the clinical syndrome is typical of a lacunar infarction, that is: a pure motor stroke, a pure sensory stroke, hemiparesis plus ataxia, or dysarthria plus a clumsy hand.

(b) Embolic stroke required either

- a recognized source of emboli such as atrial fibrillation, endocarditis, mitral stenosis, thrombus in heart, recent myocardial infarct or cardiac surgery, or
- a mottled appearance consistent with infarction on the CT.

(c) Atherosclerotic infarction required both

- evidence of large vessel atherosclerosis by carotid ultrasound or angiography and
- no apparent source of cardiac emboli or evidence of lacunar infarction

For analysis of the stroke subtypes, other stroke cases (excluding the one being analyzed) were used as controls. Analyses were adjusted for matching covariates, the first principal component, sex, age, hypertension status, and index year.

4.1.6 The Heart Protection Study (HPS)

HPS was a large randomized trial involving individuals at increased risk of vascular events. Between 1994 and 1997, 20,536 men and women aged 40 – 80 years were recruited from 69 collaborating hospitals in the United Kingdom (with ethics committee approval). Participants were eligible for inclusion provided they had non-fasting blood total cholesterol concentrations of at least 135 mg/dL (3.5 mmol/L) and either a previous diagnosis of coronary disease, ischemic stroke, other occlusive disease of non-coronary arteries, diabetes mellitus, or treated hypertension for men 65 years and older. None of them was on statin therapy.

At the initial screening visit, all participants provided written consent and began a “run-in” phase involving 4 weeks of placebo followed by 4 to 6 weeks of 40 mg simvastatin daily, after which compliant and eligible individuals were randomly allocated 40 mg simvastatin daily or matching placebo for approximately 5 years. Individuals entering HPS with a clinical diagnosis of ischemic stroke were used as cases in the METASTROKE study. Individuals entering HPS with pre-existing diabetes but no history of cerebrovascular disease, coronary heart disease or peripheral vascular disease were used as controls.

DNA was extracted from stored white cells and genotyping was carried out at the Centre National de Génotypage in Evry, France. Genotypes were measured using the Illumina 610K Quad panel, called using Illumina BeadStudio software, and imputed with reference to HapMap2 CEU release 22 (build 36) using MACH. Single nucleotide polymorphisms with < 97.5% call rate, significant deviation from Hardy-Weinberg equilibrium ($p < 1x10^{-6}$) or low minor allele frequency (< 0.01) were excluded. Genotype data were available for 578 stroke cases and 468 controls after quality control exclusions for discrepant sex, repeated samples and non-European ancestry.

4.1.7 INTERSTROKE

INTERSTROKE is an international, multicentered, case-control study of stroke investigating the global burden of risk factors across various regions and ethnic groups around the world. A detailed report of the study design has been published.(39) Briefly, stroke cases had acute first stroke (within 5 days of symptoms onset and 72 hours of hospital admission) in whom neuroimaging (CT or MRI) was performed. The TOAST classification system was used to define ischemic stroke subtypes. Cases were excluded if:

(a) they were unable to communicate due to severe stroke without a valid surrogate respondent (e.g. first-degree relative or spouse),
(b) they were hospitalized for acute coronary syndrome/myocardial infarction, or
(c) stroke was attributed to non-vascular causes (e.g. tumor).

Controls were selected from the community and had no history of stroke. All samples were genotyped at the Genetic Molecular Epidemiology Laboratory in Hamilton, Ontario, Canada.

4.1.8 Lund Stroke Register (LSR)

See **Section 3.1.11**

4.1.9 Malmo Diet and Cancer (MDC) study

See **Section 3.1.12**

4.1.10 Milano

Milano includes consecutive Italian subjects referred to Besta Institute (Milan, Italy) from 2000 to 2009 with stroke and included in the Besta Cerebrovascular Diseases Registry (CEDIR). Ischemic stroke cases, first ever or recurrent, confirmed on brain imaging, were selected for this study. All cases were of self-reported Caucasian ancestry and had a clinically relevant diagnostic workup performed. All cases were phenotyped by an experienced stroke neurologist according to TOAST criteria, based on relevant clinical imaging and available information on cardiovascular risk factors. Controls are Italian individuals enrolled within the PROCARDIS Study, with no personal or sibling history of coronary heart disease before age 66 years.

Italian cases were genotyped using Illumina Human610-Quad v1_B or Human660W-Quad v1_A chips. Italian controls were genotyped with the Illumina HumanHap610-Quad chip. PCA with HapMap 3 on the Italian cases showed that Italian PROCARDIS controls had similar ancestry to the cases. All samples had a genotype call rate > 95%. Samples were excluded due to unexpected duplicates or evidence of non-European ancestry based on principal component analysis. Quality control procedures excluded SNPs with minor allele frequency < 0.01 or Hardy-Weinberg $p < 5 \times 10^{-6}$ in either the case or control collections.

4.1.11 Risk Assessment of Cerebrovascular Events (RACE) Study, Pakistan

RACE is a retrospective case-control study designed to identify and evaluate genetic, lifestyle and biomarker determinants of stroke and its subtypes in Pakistan. Samples were recruited from six hospital centers in Pakistan. Cases were eligible for inclusion in the study if they:

(a) were aged at least 21 years,
(b) presented with a sudden onset of neurological deficit affecting a vascular territory with sustained deficit at 24 hours verified by medical attention within 72 hours after onset (onset is defined by when the patient was last seen normal and not when found with deficit),
(c) the diagnosis was supported by CT/MRI, and
(d) presented with a Modified Rankin Score of < 2 prior to the stroke.

TOAST(2) and Oxfordshire(12) classification systems were used to sub-phenotype all stroke cases.

Control participants were individuals enrolled in the Pakistan Risk of Myocardial Infarction Study (PROMIS), a case-control study of acute MI based in Pakistan. Controls in PROMIS were recruited following procedures and inclusion criteria as adopted for RACE cases. In order to minimize any potential selection biases, PROMIS controls selected for this stroke study were frequency matched to RACE cases based on age and sex and were recruited in the following order of priority:

(a) non-blood related or blood related visitors of subjects in the out-patient department,
(b) non-blood related visitors of stroke cases.
(c) subjects in the out-patient department presenting with minor complaints.

4.1.12 Sahlgrenska Academy Study on Ischemic Stroke (SAHLSIS)

See **Section 3.1.21**. The replication sample from this study consisted of population-based controls from SAHLSIS,(16) as well as cases from this study not included in the SiGN discovery phase that had been genotyped as part of the South Swedish GWAS.

4.1.13 The Sea Islands Genetics Network – the REasons for Geographic And Racial Differences in Stroke (REGARDS) sub-study (SIGNET-REGARDS)

The Sea Islands Genetics Network (SIGNET) study consists of:

(a) REasons for Geographic And Racial Differences in Stroke (REGARDS)
(b) the Sea Islands Genetic African American Registry (Project SuGAR)
(c) a COBRE for Oral Health study (COBRE), and
(d) the Systemic Lupus Erythematosus in Gullah Health study (SLEIGH).

All subjects are African Americans (AA), and all provided written informed consent.

All SIGNET samples (N = 4,298) were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Imputation was performed using MACH (version 1.0.16) to impute all autosomal SNPs using the CEU + YRI reference panel (as supplied by Goncalo Abecasis) from build 36 (2,318,207 SNPs in total; CEU, Europeans living in Utah with Northern and Western European ancestry; YRI, Yorubans in Ibadan, Nigeria).

The REGARDS study design is detailed under **Section 3.1.20**.(40)

For SIGNET, we selected all AA REGARDS type 2 diabetes (T2D) cases recruited from SC, GA, NC, and AL, and an equivalent number of race, sex, and age-strata matched diabetes-free controls. We also included all participants not already included that were current residents of the 15-county "Low Country" region of South Carolina (SC) and Georgia (GA) (SC counties Beaufort, Berkeley, Charleston, Colleton, Dorchester, Georgetown, Hampton, Horry, Jasper; GA counties Bryan, Camden, Chatham, Glynn, Liberty, McIntosh). The subset of REGARDS participants genotyped under SIGNET are referred to as SIGNET-REGARDS. GWAS genotyping was completed among 2,398 SIGNET-REGARDS AA participants, including 1,149 with diabetes and 1,249 without diabetes.

4.1.14 Stroke in Young Fabry Patients (SIFAP)

The SIFAP study is a multicenter study carried out to determine the frequency of Fabry disease in an unselected group of young adult patients with acute cerebrovascular events defined as having had an acute ischemic stroke or transient ischemic attack less than three months before enrollment into the study. The study has been described previously,(41,42) and is briefly summarized here.

First-ever (80.5%) and recurrent ischemic strokes were included. MRI was a mandatory procedure but, in the case of negative or missing MRI, a qualified stroke neurologist could confirm the clinical diagnosis. For this project, ischemic stroke cases recruited from 15 sites throughout Germany and determined not to have Fabry Disease were included in the analysis. All were of European ancestry and had age of first stroke of 18 – 55 years. The diagnosis of Fabry disease was based in males as well as in females in the first level on the sequencing data of the entire exon structure including promoter of the α-galactosidase gene. In cases where a mutation was detected, biochemical analysis was done. Stroke cases from SIFAP were genotyped at CIDR (Baltimore, MD) using the Illumina Human Omni 2.5M-Quad array. Only those cases without Fabry disease were selected for genotyping.

Controls free of cardiovascular diseases were selected from the KORA Study previously genotyped at CIDR in the same platform. The Cooperative Health Research in the Region of Augsburg (KORA) study is a population-based study of cardiovascular and metabolic traits carried out in the region of Augsburg, Southern Germany. A subset of control subjects (N = 28) was re-genotyped together with cases to provide cross-set duplicates. This joint clustering was used to minimize possible artifactual differences in allelic frequency between cases and controls due to genotyping at different times, and the cross-set duplicates were used to detect such artifacts that may have occurred.

4.1.15 Utrecht, The Netherlands ImmunoChip Study Samples/PROMISe Study (UTRECHT)

The PROMISe study has been described elsewhere.(43,44) Subjects were included with non-disabling cerebral ischemia of arterial origin, who were referred to the University Medical Center (Utrecht, The Netherlands) and were included in the Second Manifestations of Arterial Disease study,(45) or the Utrecht Stroke Database. Subjects with non-atherosclerotic causes of cerebral ischemia or with suspected source of cardiac embolism were excluded from this study. Included cases were classified as large artery atherosclerosis or small artery occlusion.

All cases were genotyped at the Genetic Laboratory at the Erasmus Medical Center (Rotterdam, the Netherlands). Dutch individuals genotyped separately from the cases (at the Department of Human Genetics, University Medical Center, Groningen, The Netherlands) were used as controls, as described in a prior study of celiac disease.(46) Written informed consent was obtained from all subjects with approval from the ethics committee or institutional review board of all participating.

4.1.16 VHIR-FMT-Barcelona

The Barcelona cohort is a subset of Caucasian ischemic stroke subjects that were enrolled as a part of GODs project (Genetic contribution to functional Outcome and Disability after Stroke) with demonstration of acute ischemic stroke in a neuroimaging study during the first 7 days after stroke. We included cases with a first-ever and with a recurrent stroke. We did not include lacunar strokes due to the study was focused in disability after stroke of non-lacunar patients. We did not use age or stroke severity as exclusion criteria. Participants were not part of a treatment trial.

We classified cases using the TOAST criteria. Control subjects were healthy subjects without history of ischemic stroke or familiar history of ischemic stroke. In addition, they were subjects without history of myocardial infarction. All the samples (cases and controls) were genotyped using the Infinium HumanExome BeadChip Kit (Illumina). Written informed consent was obtained from all subjects with approval from the ethics committee of all participating institutions.

4.1.17 Women's Genome Health Study (WGHS)

The WGHS study(47) is a subset of the Women's Health Study (WHS),(48) which consists of healthy female participants who were randomized either to an aspirin intervention arm or placebo. DNA was collected at enrollment in the WHS. Since study enrollment, all WHS participants have been followed prospectively for the occurrence of common clinical outcomes. For the primary trial endpoints of cardiovascular disease and cancer, full medical records are obtained for reported endpoints and reviewed by an endpoints committee of physicians unaware of randomized treatment assignment. Stroke subtypes were classified according to TOAST criteria.(2) Written informed consent was obtained from all subjects with approval from the ethics committee or institutional review board of all participating institutions.

4.1.18 WHI Hormone Trial HT (WHI-HT)

WHI-HT consisted of two separate clinical trials in postmenopausal women ages 50 – 79 years at baseline:

(a) a trial of combined estrogen and progestin (Estrogen plus Progestin or E+P) in women who had an intact uterus at baseline (N = 16,608) and

(b) a trial of estrogen (Estrogen Alone or E-Alone) in women who had a prior hysterectomy at baseline (N = 10,739).

Postmenopausal women who gave written informed consent were enrolled in the WHI at 40 clinical centers in the United States. Exclusions for safety reasons included prior diagnosis of breast cancer or other cancers within the past 10 years (except nonmelanoma skin cancer). Women with systolic blood pressure (SBP) of 200 mm Hg or higher or diastolic blood pressure (DBP) of 105 mm Hg or higher were advised to see their physician within a specified period depending on blood pressure level and were temporarily excluded from the clinical trials until their blood pressure was determined to be under control.

Stroke diagnosis requiring and/or occurring during hospitalization was based on rapid onset of a neurological deficit attributable to an obstruction or rupture of an arterial vessel system. Hospitalized incident stroke events were identified by semiannual questionnaires and adjudicated following medical record review, which occurred both locally and centrally. Ischemic strokes were further classified by the central neurologist adjudicators according to the Oxfordshire(12) and Trial of Org 10172 Acute Stroke Trial (TOAST)(2) criteria to examine stroke subtypes. The TOAST classification focuses on the presumed underlying stroke mechanism and requires detailed investigations (such as brain computed tomography, magnetic resonance imaging, angiography, carotid ultrasound, and echocardiography).

WHI-GARNET participants were genotyped using the Illumina Omni-quad chip at the Broad Institute, and imputation using 1000 Genomes as a reference was performed at the GARNET Coordinating Center (University of Washington) using BEAGLE. All SNPs passed QC and were in Hardy-Weinberg equilibrium. Association testing for typed or imputed SNPs was performed using PLINK.

5. Phenotyping

5.1 Causative Classification System for Ischemic Stroke (CCS)

SiGN used the web-based CCS system for stroke subtyping (available online, **Online Source [9]**). The details of CCS were published elsewhere.(1,49,50) We have chosen CCS because it is a rule- and evidence-based classification system. Other advantages of CCS include high reliability, web-based interface, capacity to retain and standardize individual data points that underlie subtype classification such as carotid stenosis or patent foramen ovale. The CCS allows separate exploration of the genetic bases for phenotypic or causative subtypes in addition to overall ischemic stroke.

5.1.1 CCS Causative Classification

Causative subtyping requires integration of multiple aspects of ischemic stroke evaluation in a probabilistic and objective manner. We grouped cardiac pathologies with uncertain risk of stroke (minor sources) into the cryptogenic category leading to generation of 6 causative subtypes. Detailed definitions for subtypes were published elsewhere(1,49,50) and can be viewed online (**Online Source [10]**).

5.1.2 CCS Phenotypic Classification

Phenotypic subtypes referred to abnormal test findings categorized in major etiologic groups without weighting towards the most likely cause in the presence of multiple causes. For this reason, phenotypic subtypes were not mutually exclusive. There were 4 main phenotypic categories:

(a) large artery atherosclerosis (LAA)
(b) cardiac embolism (CE)
(c) small artery occlusion (SAO), and
(d) other-uncommon causes.

There were 4 possible states for LAA and CE (major, minor, absent, incomplete evaluation), 3 for SAO (major, absent, incomplete evaluation), and 2 for other-uncommon causes (major and absent), giving rise to a total of 96 phenotypic categories. We collapsed these 96 categories into the following 7 subtypes: LAA-major, CE-major, SAO-major, other-major, no major etiology (cryptogenic), multiple competing major etiologies, and incomplete investigation.

5.2 Trial of Org 10172 in Acute Stroke Treatment (TOAST)

The majority (85.5%) of the discovery sample had TOAST(2) subtypes prior to the beginning of SiGN and nearly all of the replication sample was categorized using TOAST. TOAST subtypes were determined locally by site investigators following individual study protocols without benefit of central oversight. Of note, TOAST subtypes in the discovery sample were determined using the same data sources that were available for the CCS classifications. TOAST and CCS classifications were completed by different physicians and at different time points in the majority of study sites but using the same study or site-specific case report forms. CCS adjudicators in the discovery sample were required to confirm that they were fully blind to TOAST results before they began to enter patient data into CCS. TOAST subtypes included LAA, CE, SAO, other-uncommon causes, and undetermined causes.(2) Unlike CCS, this 5-subtype TOAST classification did not allow separate categorization of cryptogenic strokes. TOAST and CCS have been shown to be moderately correlated.(51)

5.3 Phenotyping of the stage I discovery sample

Etiologic stroke classification in SiGN started in July 2010 and ended in August 2014. The discovery sample included cases with imaging-confirmed ischemic stroke from 17 US sites, 1 Australian site, and 13 European sites from 8 countries. Recruitment to contributing studies in the discovery dataset occurred during a 23-year period between 1989 and 2012.

The subtyping was based on diagnostic investigations performed during the routine clinical care of ischemic stroke patients. Frequencies of completion of selected diagnostic investigation are shown in **Supplementary Table 4**. Study-specific case report forms and un-abstracted medical records served as data source for subtyping. Data sources varied in length and detail among the study sites. Subtype assignments were done based on data available at the time of discharge in the majority although post-discharge test results were used when available.

For the purpose of SiGN, we customized CCS by generating a confidential, password protected data collection platform. We also modified the online CCS form separating the single data entry field for small artery occlusion (SAO) in the original CCS into two separate data entry fields: one to indicate the presence of a typical lacunar infarct on neuroimaging and the second to rule out accompanying parent artery disease at the origin of the penetrating artery supplying the site of the lacunar infarct. Thus, it became possible to collect phenotypic data on lacunar infarcts for which vascular imaging for parent artery disease was not available. Other modifications included addition of a yes/no checkbox for each individual stroke test to provide systematic collection of stroke work-up data across the participating study sites and the addition of a confirmation box to the end of each section in the classification form to confirm that the user reviewed all the data entry fields but none applied. No modification was made in the decision-making code of the CCS; both customized and publicly available CCS algorithms provide the same subtype for each given test condition.

A centralized Phenotype Committee of 4 expert stroke neurologists met weekly to monitor data quality and site performance. The Phenotype Committee members provided training to adjudicators/re-adjudicators regarding data entry, data submission, and archiving at scheduled study meetings and via webinars. A total of 52 adjudicators (13 stroke neurologists, 17 stroke fellows, 13 neurology residents, 9 non-neurologists) performed stroke subtyping. Each adjudicator had to complete an interactive online training module and was required to pass an on-line certification examination available at the CCS website before they performed CCS subtyping. Each adjudicator also participated in a 120-minute interactive training webinar developed by the SiGN Phenotype Committee. All data entered into CCS, as well as the system output, were saved in a confidential SiGN database. In addition to subtype-related data, each study site provided baseline variables such as age, sex, race, and vascular risk factors, using a structured data collection form.

The Phenotype Committee blindly re-adjudicated a randomly 10% of cases recruited from the US and Australian studies for quality control. Similarly, 10% of cases from European studies were blindly re-adjudicated by European investigators (N = 20). Re-adjudicators used the same data source available to adjudicators to determine the CCS subtypes. Based on data from 1509 paired rating in the SiGN study, the crude agreement between 52 adjudicators and 24 re-adjudicators was 80% (kappa = 0.75; 95% CI: 0.72 – 0.77) for the causative CCS and 81% (kappa = 0.75; 95% CI: 0.72 – 0.78) for the phenotypic CCS.(52)

5.4 Phenotyping of the stage II replication sample

Because we included all available CCS cases in the discovery phase, the independent replication sample included cases with TOAST subtyping and their matched controls. TOAST subtypes were determined locally by site investigators following individual study protocols. Some, but not all, sites had central quality control procedures. Since only a small minority of sites subcategorized the undetermined group into cryptogenic cases, case with multiple etiology, and cases with insufficient workup, these were collapsed into the undetermined category for analysis purposes.

6. Genotyping

6.1 Pre-existing genotyped cohorts

The bulk of cases and controls were drawn from previously genotyped cohorts, described in **Section 3.2**. All cases and controls were genotyped on an Illumina platform, ranging in size from the Illumina 550 (~550,000 sites) to the Illumina 5M (~5,000,000 sites). The genotyping array for each cohort is listed in **Table 1** of the main text and additional genotyping information is provided in the cohort descriptions in **Section 3**.

6.2 Center for Inherited Diseases Research (CIDR) genotyping

In addition to previously genotyped cases and controls, newly genotyped cases from 13 sites and a small number of controls from 2 sites were also included in the discovery stage. To boost power for discovery, new genotyping focused primarily on cases.

A total of 10,966 study participants and 118 investigator-provided controls from multiple racial/ethnic groups represented in the SiGN were genotyped and posted to dbGaP. The study participants include 9,721 stroke cases and 1,245 controls. The 118 investigator controls refer to individuals previously genotyped in other studies. These investigator-provided controls (along with HapMap controls) were used to assess genotype concordance and homogeneity with the earlier genotyping.

The genotyping was performed at the Center for Inherited Disease Research (CIDR) using the Illumina HumanOmni5Exome-4v1 array and calling algorithm GenomeStudio version 2011.1, Genotyping Module 1.9.4, and GenTrain version 1.0. CIDR used annotation version “A”, genome build 37. The final data release contained 4,511,703 variants.

Earlier versions of Illumina annotations incorrectly annotated chromosome information for many SNPs designated as “X” or “Y” rather than as “XY”. These SNPs occur in pseudo-autosomal (PAR1, PAR2) regions or in the X-translocated region (XTR). The annotation was corrected prior to genotype calling for SiGN.

7. Collection of publicly-available controls

Several previously genotyped studies included in the discovery phase of SiGN included internal controls genotyped along with cases (ASGC, **Section 3.2.2**; GEOS, **Section 3.2.3**; KRAKOW, **Section 3.2.10**; LSGS, **Section 3.2.11**; MDC, **Section 3.2.12**, and VISP, **Section 3.2.14**). For the remaining cases, publicly-available controls were selected to be matched to the SiGN cases. The selection process for these control cohorts is described in this section.

ADHD (**Section 3.2.1**) and *INMA* (**Section 3.2.8**)

The ADHD and INMA samples were obtained from Spain and genotyped on the Illumina 1M. These controls were chosen to match cases from also obtained from Spain and genotyped at CIDR on the Illumina 5M.

GRAZ (**Section 3.2.4**)

Controls from Graz, Austria who had previously been genotyped on the Illumina 660 were obtained from the PI of the Graz stroke study. These controls were included to match the GRAZ cases genotyped on the Illumina 5M as part of the CIDR genotyping effort.

HABC (**Section 3.2.5**)

Collected around the United States and genotyped on the Illumina 1M Duo (v3, B), the HABC controls were selected to match European-ancestry cases genotyped on the smaller Illumina arrays represented in the case cohorts (Illumina 550, Illumina 660)

HRS (**Section 3.2.7**), *OAI* (**Section 3.2.13**), and *HCHS/SOL* (**Section 3.2.6**)

The HRS, OAI, and HCHS/SOL samples were selected as controls for the newly genotyped CIDR cases for two primary reasons:

(a) Each of these control cohorts was genotyped on the Illumina 2.5M or 2.5M plus custom content. Such an array would allow for retention of maximal SNP content in the CIDR cases, which were genotyped on the Illumina 5M.

(b) All three control cohorts contained samples drawn from ancestries (European and admixed) that were likely to cluster well with the CIDR cases.

KORA (**Section 3.2.9**) and *WTCCC* (**Section 3.2.15**)

KORA and WTCCC were original controls in the WTCCC2 ischemic stroke genome-wide association study (GWAS) and were kept as controls for the four case cohorts in the WTCCC2 GWAS that were also used in the discovery phase of SiGN. Those cohorts are: ESS (**Section 3.1.4**), MUNICH (**Section 3.1.16**), OXVASC (**Section 3.1.19**), and STGEORGE (**Section 3.1.23**).

LSR (**Section 3.2.11**)

LSR samples were selected as controls for the LSR cases that had been genotyped as part of the CIDR genotyping effort.

8. Distribution of phenotypes across study sites

The distribution of cases and controls (post-QC and thus examined in the meta-analyses) across the various contributing cohorts and within subtypes (both CCS and TOAST) are depicted in **Supplementary Figure 2** and listed in **Supplementary Table 5**.

Of the 14,872 cases contributing to the discovery meta-analyses (not including VISP, for which only summary-level results were received), 6,855 (46.1%) were female.

8.1 Distribution of risk factors for stroke in the SiGN stage I discovery cases

Of the 14,661 cases with data available on the age at stroke event, the average age at event was 66.54 years; the full age distribution is shown in **Supplementary Figure 11a**. The average age of event among females was 68.97 years and the average of event among males was 64.46 years (**Supplementary Figure 11b**).

Of the 14,300 cases with data available on hypertension status, 9,598 (67.12%) were currently hypertensive (47.98% female). 4,702 cases identified as not or never being hypertensive (42.88% female) (**Supplementary Figure 11c**).

Diabetes mellitus data was also available on 14,300 cases. 3,499 cases (24.47%) currently had diabetes mellitus and of the cases with diabetes mellitus, 44.47% were female. Three male samples formerly had diabetes mellitus and the rest of the cases (N = 10,798) did not (**Supplementary Figure 11d**).

Information about atrial fibrillation was available for 13,843 cases. 3,015 cases had atrial fibrillation, while the rest did not. Of the cases with atrial fibrillation, 1,642 (54.46%) were female (**Supplementary Figure 11e**).

Information about cardiovascular disease (CAD) was available for 13,555 cases. 2,729 cases (20.13%) were annotated as currently have CAD. Of these, 41.22% were female. An additional 10,811 were annotated as not having CAD (47.81% female), and 11 cases were annotated as never having CAD (36.26% female) (**Supplementary Figure 11f**).

Finally, 13,891 cases had smoking status available. Of these cases, 3,241 were current smokers (37.43% female), 3,602 were former smokers (37.34% female), and 7,048 had never been smokers (55.28% female) (**11g**).

It is important to note that while age, hypertension, smoking status, CAD, and diabetes mellitus information was known for the bulk of cases, it was not known for the majority of the (publicly-available) controls, and thus could not be used as covariates in genome-wide association testing.

8.2 Subtype assignments in the SiGN stage I discovery cases

Samples could be assigned to five major subtypes:

(a) cardioembolic (CE)
(b) large artery atherosclerosis (LAA)
(c) small artery occlusion (SAO), and
(d) undetermined, and
(e) other.

In SiGN, 13,757 cases had an assigned subtype for the CCS Causative (CCSc) subtyping method; 10,209 cases had an assigned subtype for the CCS Phenotypic (CCSp) method; 13,002 cases had an assigned subtype for the TOAST method.

A sample could be assigned to one and only one CCSc subtype such that, for example, if a sample was subtyped as having a cardioembolic stroke by CCSc, it was not annotated as having any other CCSc subtype. There are three “undetermined” classifications in the CCSc subtyping system: all undetermined (UNDETER), cryptogenic and CE minor (CRYPTCE), and incomplete and unclassified (INCUNC). If a sample was assigned to the CCSc all undetermined (UNDETER) classification, it was then also assigned to either CRYPTCE or INCUNC, which were two mutually exclusive subsets of UNDETER.

For the CCSp subtyping method, classifications were *not* unique. Cross-subtype classifications in CCSp were as follows:

(a) CE and LAA: N = 354
(b) CE and SAO: N = 178
(c) CE and undetermined (cryptogenic): N = 0
(d) LAA and SAO: N = 246
(e) LAA and cryptogenic: N = 0
(f) SAO and cryptogenic: N = 0

Like CCSc, TOAST only assigned a case to a single subtype and there were no cross-subtype classifications.

The CCS and TOAST subtyping systems have only moderate correlation when assigning subtypes to identical samples.(51) Consequently, a sample subtyped as having an LAA stroke by CCSc may be subtyped differently by TOAST. A tabulation of samples annotated as having more than one subtype is provided in **Supplementary Table 6**.

9. Data cleaning

To begin stage I of the study, sample and SNP quality control (QC) were performed sequentially in a number of phases to accommodate the design of the SiGN study, which included many separately-genotyped case and control cohorts and representation of numerous ancestries. The QC phases included analyses:

(a) on newly-genotyped cases and controls,
(b) on individual previously-genotyped cohorts,
(c) after cases and controls had been matched together based on genotyping array and ancestry,
(d) after imputation had been performed (SNPs only), and
(e) after genome-wide association testing (SNPs only).

The full details of the QC performed on the data, both on samples and SNPs, is provided in this section.

9.1 Individual cohort quality control

9.1.1 Quality Control of CIDR genotyping for SiGN

Samples were analyzed for missingness (genotyping callrate), discrepancies between annotated sex versus genetic sex, relatedness, and verification of expected and unexpected duplicates. Any samples with unresolved identity issues were identified during this process and flagged for exclusion from downstream association testing.

An analysis was also performed to detect large chromosomal anomalies, and genotypes within identified regions larger than 5Mb were set to missing for that region and that individual if there was indication of genotyping errors.(53–56) Further variant-level QC were conducted to detect batch effects, missing call rates, Hardy-Weinberg equilibrium, to detect genotyping artifacts rather than population structure within groups with homogeneous PCs, Mendelian errors based on HapMap trios included in genotyping, and genotype discordance between duplicate samples (**Supplementary Table 7**).

These analyses were performed using the R packages GWASTools(57) and SNPRelate(58) unless indicated otherwise and the methods follow previously-described QC procedures.(59)

9.1.2 Data liftover to hg19

After the CIDR data had been cleaned, it and all other (previously genotyped) cohorts contributing to the discovery phase of SiGN were placed at one central location for additional QC, imputation, and association testing. Upon centralizing all of the discovery cohorts, we first ensured that all data was either genotyped on or lifted over to the same genome build.

All cohorts with genotyping information from dbSNP build 136 (hg18) (**Online Source [11]**) were lifted over to dbSNP build 137 (hg19) by updating chromosomal positions. All marker names across all cohorts were made consistent with dbSNP 138 based on chromosome, position, and alleles. Once the liftover was complete, chromosome, position, marker name and alleles were all checked for consistency across cohorts.

9.1.3 Sample QC of individual cohorts

We initially performed sample-level QC in each of the individual cohorts. A full description of filter names and thresholds is provided in **Supplementary Table 8**. Most cohorts were comprised of either only cases or only controls, with the exception of KRAKOW and LSGS (cases and controls genotyped together as part of the CIDR effort), MDC, and ASGC.

Initial cohort-specific QC of HRS and HCHS/SOL is described in **Section 9.3.2**. For other cohorts, initial QC is described below in **Sections 9.1.3** and **9.1.4**.

a. Missingness

Sample-level missingness (*--missing*) was calculated in PLINK(60) and samples with high (> 10%) missingness were removed.

b. Sex discrepancies

Samples with mismatching sex information between genotypes and provided phenotype information were also excluded. Heterozygosity was calculated across the X chromosome for each sample using PLINK(60) (*--check-sex*). If the coefficient was > 0.8 (indicating a male sample) and the phenotype information indicated a female sample, the sample was removed. If the coefficient was < 0.2 (indicating a female sample) and the phenotype information indicated a male sample, the sample was removed.

For three samples, the cohort had established sex-chromosome anomalies explaining the heterozygosity and phenotype information discordance, and the samples were kept in the analysis.

c. Principal component analysis (PCA)

We used PCA to check for ancestral homogeneity within each cohort. We performed PCA by first selecting a high-quality set of SNPs using PLINK(60) defined as:

(a) missingness < 0.01% (*--geno*)
(b) minor allele frequency > 5% (*--maf*)
(c) ld-pruned r2 threshold of 0.2 (*--indep-pairwise 50 5 0.2*)
(d) excluding the MHC region (chr6, position 24,092,021 – 38,892,022) and LCT (chr2, position 129,883,530 – 140,283,530) due to their high stratification across different ancestral populations
(e) excluding the inversions on chromosome 8 (position 6,612,592 – 13,455,629) and chromosome 17 (position 40,546,474 – 44,644,684)
(f) autosomal SNPs only

Once these SNPs were extracted from the data, they were then merged with genotypes available from HapMap Phase 3(61) (HM3) using PLINK. HM3 contains 10 different populations available for analysis:

(a) CEU: Europeans of Northern and Western European ancestry living in Utah (European)
(b) TSI: Tuscans in Italy (European)
(c) YRI: Yorubans in Ibadan, Nigeria (African)
(d) MKK: Massai in Kinyawa, Kenya (African)
(e) LWK: Luhye in Webuye, Kenya (African)
(f) CHD: Chinese in Denver (East Asian)
(g) CHB: Chinese in Beijing (East Asian)
(h) JPT: Japanese in Tokyo (East Asian)
(i) MXL: Mexicans in Los Angeles, California (Admixed)
(j) GHI: Gujarati Indians in Houston, Texas (Admixed)

Using EIGENSTRAT,(62) we calculated principal components (PCs) for the HM3 samples and the SiGN data were projected onto these PCs. If a cohort contained multiple ancestry groups, no further sample QC was performed at this point. If a cohort contained samples from a single ancestry, a relatedness check was performed using PLINK (next section).

d. Relatedness

We used PLIINK(60) (*--genome*) to calculate identity-by-descent and identify related pairs of individuals. Duplicate samples and sibships (pi-hat ≥ 0.5) were identified and the sample from each pair with higher missingness was removed from the data. If a sample appeared in more than one pair, it was preferentially removed so as to minimize sample loss.

Further sample QC was not performed until after array- and ancestry-specific groups of cases and controls had been established.

9.1.4 SNP QC of individual cohorts

We also performed an initial SNP-level QC on the individual cohorts (**Supplementary Table 8**).

a. Missingness

SNP missingness was calculated using PLINK(60) (*--missing*) and SNPs with high (> 10%) missingness were excluded.

b. A/T and C/G SNPs

Additionally, all A/T and C/G SNPs, which comprised a small percentage of all the SNP data, were excluded at this point to avoid downstream issues related to strandedness. For A/T and C/G SNPs with frequency of 40 – 60%, it can be difficult to determine which strand the allele is on, and all alleles must be on the positive strand for imputation.

c. Duplicate markers

If a cohort contained duplicate markers, the marker with the better genotyping call rate was kept and the other removed from the analysis.

9.2 Matching of cases and controls

To construct well-matched groups of cases and controls for downstream QC, imputation, and association testing, we first matched control groups to case groups based on genotyping array and cohort or region, where relevant.

All cohorts were genotyped on an Illumina array, though the number of SNPs on the array varied widely, from 500,000 – 5,000,000. We matched case and control groups so as to maximize the number of SNPs kept for downstream analyses.

The matched case and control groups appear in **Supplementary Table 9**. In short, the groups were composed as follows (with genotyping arrays indicated in parentheses):

Group 1
Cases: BRAINS (650Q), ISGS (600), MGH-GASROS (600), SWISS (600)
Controls: HABC (1M Duo)

All case cohorts were of European ancestry so the non-European-ancestry samples in HABC were excluded at this point based on PCA projection using HapMap 3(61) as a reference. Any HABC samples with a PC1 or PC2 value +/- 10 s.d. beyond the average PC1 and PC2 values for the HapMap 3 European-ancestry (CEU, TSI) samples were removed.

Group 2
Cases: ESS (660), MUNICH (660), OXVASC (660), STGEORGE (660)
Controls: KORA (550), WTCCC (660)

Group 2 is a reconstruction of the Wellcome Trust stroke analysis. All samples were of European ancestry.

Group 3
Cases: GEOS (1M)
Controls: GEOS (1M)

GEOS comprised its own analysis group as both cases and controls were genotyped on the 1M array. The GEOS cohort had samples of both European ancestry and African-European admixed ancestry.

Group 4
Cases: CIDR (5M), excluding KRAKOW, LSGS, BASICMAR, GRAZ, LSR, and SAHLSIS
Controls: HRS (2.5M), OAI (2.5), HCHS/SOL (2.5M + custom content)

CIDR, HRS and OAI all contained samples of multiple ancestries. HCHS/SOL was selected specifically as controls for Hispanic cases.

Group 5
Cases: KRAKOW (5M)
Controls: KRAKOW (5M)

The KRAKOW cases and controls were genotyped together as part of the CIDR genotyping effort.

Group 6
Cases: LSGS (5M)
Controls: LSGS (5M)

The LSGS cases and controls were genotyped together as part of the CIDR genotyping effort.

Group 7
Cases: BASICMAR (5M)
Controls: ADHD (1M), INMA (1M)

All cases and controls originate from Spain. Some of the BASICMAR cases were later identified as clustering better with non-European-ancestry samples by PCA and were pooled with Group 4 Hispanic cases and controls for analysis (**Supplementary Table 5**).

Group 8
Cases: GRAZ (5M)
Controls: GRAZ (660)

The GRAZ cases were genotyped as part of the CIDR genotyping effort. Older, pre-existing controls from GRAZ were selected as controls because the samples originated from the same cohort. Some of the GRAZ cases were identified as clustering better with non-European-ancestry samples by PCA and were pooled with Group 4 Hispanic cases and controls for analysis (**Supplementary Table 5**).

Group 9
Cases: LSR (5M), SAHLSIS (5M), MDC (Express 750K + Exome)
Controls: MDC (Express 750K + Exome)

All of the samples originating from Sweden were pooled together to form Group 9. Some of the LSR and SAHLSIS cases were identified as clustering better with non-European-ancestry samples by PCA and were pooled with Group 4 Hispanic cases and controls for analysis (**Supplementary Table 5**).

Group 10
Cases: ASGC (610)
Controls: ASGC (610)

The ASGC cases and controls were genotyped together and were therefore kept together for analysis

9.3 PCA- and hyperellipsoid-based ancestry-group assignment

After case-control matching was complete, we proceeded with determining analysis groups within each case-control pool of samples. To identify cases and controls that were well matched based on genetic ancestry, we performed principal component analysis and hyperellipsoid clustering within each array-specific group to determine sub-groups of European-ancestry, African-ancestry, and a group of samples that did not cluster with either the European- or African-ancestry individuals.

Note that in later text, the term "HIS" is used as a broad term representing samples of multiple ancestries that did not cluster with either European-ancestry or African-ancestry samples based on analysis of genetic markers (described below).

9.3.1 Principal component analysis (PCA)

To determine a homogenous group of European-ancestry samples, we performed PCA on the array- and region-specific groups described in **Section 9.2**. To perform PCA, we selected a high-quality set of SNPs from the data and merged it with HapMap 3(61) (HM3) in an identical manner to that described in **Section 9.1.3**.

As described above, PCs were calculated for the HM3 individuals and then the SiGN cases and controls were projected onto these PCs. Analyses were performed using EIGENSTRAT.

We calculated the average PC1 and PC2 values and standard deviations (s.d.) for the joint set of samples in CEU and TSI (i.e., the European-ancestry samples in HapMap 3). All SiGN samples within +/- 10 s.d. from the average PC1 and PC2 values of CEU + TSI were considered to be European-ancestry and are thus categorized as members of the "EUR" group.

After the EUR group was defined, hyperellipsoid clustering was performed in order to distinguish analysis groups corresponding to additional admixture populations (**Section 9.3.3**).

9.3.2 QC of self-reported Hispanic/Latino samples

In order to define the HIS analysis group, it was first necessary to create a combined dataset containing the CIDR-genotyped SiGN samples with HRS and HCHS/SOL. Initial project-specific QC of HCHS/SOL and HRS was previously performed and subject to the same protocol as SiGN individuals genotyped at CIDR (**Section 9.1.1**), with two exceptions.

(a) Illumina annotation for HRS was updated to version "H" from version "D", both genome build 37) before HRS cohort-specific data cleaning.

(b) For HRS and HCHS/SOL, Illumina misannotated chromosome information in pseudo-autosomal regions or the X-translocated region (**Section 6.2**) was corrected after genotype calling but before data cleaning.

The number of subjects (study participants, investigator controls, and HapMap controls) retained from each project who passed project-specific QC were as follows:

(a) SiGN: N = 11,187
(b) HRS: N = 12,595
(c) HCHS/SOL: N = 13,204

At this stage we removed 2 HRS Phase I – II participants who did not pass cross-phase QC during HRS Phase III cleaning.

Because each project was genotyped on a different array, we took the set of SNPs in common (consistency of rsID, chromosome, position, and alleles) across the 3 arrays on all unique subjects passing QC within each project.

Our combined dataset includes a total of 36,392 study participants and 592 HapMap subjects from the 3 projects across 2,302,224 variants. The variant QC recommendations from each of the 3 projects (**Section 9.1.1**) were carried forward into this combined dataset.

The remainder of this section describes further QC (**Supplementary Table 7**), beyond previously performed project-specific QC, to identify a set of higher quality variants on which to perform IBD estimation to obtain an unrelated set of individuals for PCA, necessary because of the use of multiple genotyping platforms and in particular due to cases genotyped separately from controls.

a. Identification of cross-study duplicates

Identify-by-descent (IBD) coefficients were estimated to identify cross-project duplicates across the 3 projects using the KING-robust procedure(63) as implemented in R using the package SNPRelate (function: *snpgdsIBDKING*).(58) The KING-robust procedure was chosen for its robustness to population structure and suitable for the dataset given the presence of self-reported white, black, Asian, and Hispanic/Latino individuals.

For each round of IBD, LD pruning was first performed to select a set of SNPs with pairwise $r^2 < 0.1$ in 10 Mb windows on the set of autosomal SNPs with missing call rate < 5% and minor allele frequency > 5%. A first round of IBD was performed on a selection of 158,037 autosomal LD pruned SNPs and resulted in identification of 289 pairs of cross-project duplicate subjects to use in duplicate sample discordance checking.

b. Cross-project duplicate sample discordance

The purpose of assessing cross-project duplicate sample discordance is to identify SNPs with artifactual differences across projects and/or arrays that may be assaying different variants which can lead to false positive associations in downstream association testing.

Cross-project duplicate subjects were identified by estimating initial IBD coefficients. All expected HapMap(64) and SiGN-HRS duplicates were observed, as were a few undocumented cross-study duplicates.

(a) A total of 79 HapMap and 30 study participants were genotyped in both SiGN and HRS Phases I – II
(b) 87 HapMap participants were genotyped in both HRS and HCHS/SOL
(c) 91 HapMap and 2 study participants were genotyped in both SiGN and HCHS/SOL studies.

The 2 study participants genotyped in SiGN and HCHS/SOL represent participants who enrolled in both studies, and the 30 study participants genotyped in SiGN and HRS represent investigator-provided controls who were intended duplicates to assess cross-project discordance.

Using the 109 cross-SiGN-HRS duplicate sample pairs, we tallied the number of discordant non-missing genotype calls and flagged any SNP with ≥ 1 discrepancies. This procedure was repeated across the 87 cross-HRS-HCHS/SOL duplicates and 93 cross-SiGN- HCHS/SOL duplicates. We identified 35,447 problematic SNPs, only 24% of which were also identified in project-specific QC.

Using the 109 cross-SiGN-HRS duplicate sample pairs, we also tallied the number of discordant missing calls among the 109 pairs. The probability of observing > "x" discordant missing calls out of a total of "N" duplicate sample pairs can be estimated using the binomial distribution. For the 109 pairs, we used a filtering threshold of 7 or more discordant missing calls where we would expect to fail at least 90% of SNPs with high genotyping error (> 0.01) but minimize failing of variants when the genotyping error is low (< 0.001).

This procedure was repeated across the 87 cross-HRS-HCHS/SOL duplicates and 93 cross-SiGN-HCHS/SOL duplicates, using a threshold of 5 or more discordant missing calls among the SiGN-

HCHS/SOL and HCHS/SOL-HRS pairs. We identified 8,121 problem SNPs; many were also flagged in the project-specific QC.

c. Identification of unrelated samples

After removing SNPs that failed any project-specific QC filter, map position duplicates, or failed for cross-project duplicate sample genotype or missingness discordance, we re-estimated IBD coefficients using this higher quality set of variants restricting to unduplicated study participants. The LD pruning step (**Section 9.3.2**) resulted in using 149,093 autosomal SNPs to estimate IBD coefficients, used to identify sets of unrelated individuals for downstream use including estimation of principal components of ancestry and association testing.

d. Estimating population structure

Several rounds of PCA were performed to identify population outliers and to compute PCs (sample eigenvectors) to use as covariates in downstream association testing to adjust for population stratification. PCA was performed as previously described,(65) but implemented in R (SNPRelate package(58)).

For each round of PCA, LD pruning to select a set of SNPs with pairwise $r^2 < 0.1$ in 10 Mb windows was performed on the set of autosomal SNPs with missing call rate < 5%, minor allele frequency > 5%, and after excluding the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions. In addition, for PCA rounds that included external HapMap datasets (not genotyped with study participants), SNPs with any genotype discordance for any HapMap individual genotyped both internally (with the study participants) and externally were excluded.

e. PCA (round 1)

We performed an initial round of PCA on a set of HapMap samples and unrelated study participants. Unrelated individuals were selected as one member per family with preference towards cases. Using previously-computed IBD coefficients, we defined families so that each family included all pairs of subjects with a kinship coefficient > 0.0625 (third degree relatives and higher), and an unrelated set included one person per family.

This analysis used 91,002 pruned SNPs on 33,843 unrelated study participants, 230 internal HapMap samples, and 1,201 external HapMap samples (**Supplementary Figure 12**). The HapMap3 population descriptions are provided in **Supplementary Table 10**.

The 33,843 unrelated study participants from this round are plotted by self-reported race/ethnicity in **Supplementary Figure 13**.

f. PCA (round 2)

We then repeated the PCA, restricting to only the 33,843 unrelated study participants from SiGN, HCHS/SOL, and HRS, using 147,177 pruned SNPs; this yielded a very similar clustering pattern to round 1. We examined the multidimensional PC space and verified that no patterns emerged due to case status or project (possible indications of genotyping artifacts due to study design). This set of PCs was used in determining genetic ancestry groups (**Supplementary Figure 14)**.

9.3.3 Hyperellipsoid clustering: defining the AFR and HIS analysis groups

We employed a hyperellipsoid clustering technique to determine genetic ancestry groups of study participants based on self-reported race/ethnicity as well as genetic markers.

This process yielded analysis groups of:

(a) admixed samples with African-ancestry, referred to henceforth as “AFR”, and

(b) admixed samples representing multiple ancestries, exclusive of the EUR and AFR samples, including samples self-identifying as Hispanic/Latino, and henceforth referred to as “HIS.”

These two groups, along with the EUR group described in **Section 9.3.1**, were used for ancestry-specific SNP QC and downstream association testing.

In order to identify more homogeneous sets of admixed individuals for association testing, a multi-dimensional algorithm was used. A hyperellipsoid clustering technique was implemented on the basis of genetic principal components within self-reported groups of non-Hispanic black, South Asian (SAS), and East Asian (EAS) participants on the combined dataset (M.A. Conomos, C.A. Laurie, et al., in preparation). This clustering technique yielded 3 hyperellipsoids (AFR, SAS, and EAS) corresponding to these self-reported groups.

Because the Hispanic/Latino population is a very diverse admixed group that does not form a well-defined hyperellipsoid, the HIS analysis group was defined as individuals who did not fall into the space of any other continental group (i.e., EUR, AFR, SAS, or EAS). The concordance with self-reported Hispanic/Latino ethnicity was very high (**Supplementary Figures 13** and **14**) and examination of plots of the PC space showed that the HIS group members appeared homogeneous and occupied a consistent PIC space as self-identified Hispanic/Latino individuals. The hyperellipsoid analysis resulted in identification of non-overlapping groups of participants and allowed for assignment of cases for which self-identification was unavailable.

Hyperellipsoids were defined using the minimum covariance determinant method(66) and resulted in the space of highest density of points for each self-identified group. AFR group members were defined as those lying within bounds of the hyperellipsoid informed by self-reported non-Hispanic black or African American participants. Two Asian groups were defined as members of two hyperellipsoids informed by self-identified Asian or Pacific Islander participants. Projected HapMap samples in this PC space corresponded to 2 distinct East and South Asian ancestry groups alongside the 2 groups of study participants.

Participants not lying within the EUR, AFR, EAS, or SAS spaces were tentatively defined as members of the HIS group, which would be subsequently thinned for outliers in additional rounds of PCA (**Section 9.4.2**).

a. Definition of parameters and hyperellipsoid computation

Hyperellipsoid limits were computed by using 4 or more PCs ($df = n_{PCs}$) of each set of participants with common self-identification to inform respective hyperellipsoid spaces. The PCs used were those computed in PCA on the unrelated set of study participants (**Section 9.3.2d**). The minimum covariance determinant (MCD) was computed over a fraction of each set of informing unrelated participants to define each hyperellipsoid space, allowing us to exclude the most outlying individuals (fraction $1 - \alpha$) when defining the hyperellipsoid space.

For each putative group, Mahalanobis distances were calculated on all unrelated and related study participants (PCs on relateds obtained using projection(67)) using the same number of PCs that were used for defining each space. In order to determine hyperellipsoid membership status, these distances, squared, on all participants were compared to a χ^2_{df} distribution. Hyperellipsoid group members were defined as those participants with squared distances which lie within 99% bounds ($p = 0.99$) of the χ^2_{df} distribution.

Due to the small sample sizes in the Asian-informing sets, α was chosen such that a minimum of 4 outlying participants could still be excluded in the MCD minimization step. Similarly, the use of additional PCs was necessary to better capture the SAS space.

b. Hyperellipsoid specifications

The African ancestry hyperellipsoid used 2,623 informing participants who self-identified as non-Hispanic black or African American and were not previously identified as an EUR group member (or relative). Four PCs were used (df = 4), the MCD was computed using α = 0.99, and group membership bounds were p = 0.99.

The EAS hyperellipsoid used 99 informing participants who self-identified as non-Hispanic Asian (and PC2 < -0.005) and were not previously identified as a EUR group member (or relative). Four PCs were used (df = 4), the MCD was computed using an α = 0.95, and group membership bounds were p = 0.99.

The SAS hyperellipsoid used 26 informing participants who self-identified as non-Hispanic Asian (and PC2 > -0.005) and were not previously identified as an EUR group member (or relative). Twelve PCs were used (df = 12), the MCD was computed using an α = 0.84, and group membership bounds were p = 0.99.

9.3.4 Confirming ancestry matching in European- and African-ancestry case-control groups

To confirm that our PCA-based approaches had successfully created ancestry-homogeneous groups of cases and controls, we performed two more rounds of PCA.

a. Checking for homogeneity within strata

The first round of PCA was within specific strata (comprised of cases and controls) to check that the samples were homogenous across PC1 and PC2 (**Supplemental Figure 15**). SNP selection was done as described in **Section 9.1.3** and principal components were calculated using SNPRelate.(58) Outliers were removed as necessary. We also plotted the top 10 PCs for each case group against its matched controls in parallel coordinates plots to check for stratification (**Supplemental Figure 16**).

b. Checking for homogeneity across all European-ancestry samples

The second round of PCA was performed after merging all EUR strata together. A high-quality set of SNPs was determined across all EUR samples as described in **Section 9.1.3** and PCs were calculated using SNPRelate.(58) This step was done to ensure that there was genetic homogeneity in terms of ancestry across the broader sample. This step was done across only the EUR strata as they contributed the bulk (> 90%) of all cases and controls in the SiGN discovery phase.

9.4 Quality control of study strata

Once we had determined study strata comprised of cases and controls that were matched on array and ancestry, we began a second round of sample- and SNP-level QC. The QC was performed on each stratum and, when necessary, was adapted to its ancestral composition. A full description of the applied filters and thresholds appears in **Supplementary Table 11**.

For sample QC in the EUR and AFR strata, missingness and inbreeding were calculated using PLINK.(60) PCA and relatedness were calculated using SNPRelate.(58)

For SNP QC in the EUR and AFR strata, frequency, Hardy-Weinberg Equilibrium, missingness, and differential missingness between cases and controls were calculated using PLINK.(60) Case–case and control–control comparisons were also done using logistic regression in PLINK.(60)

These QC steps are described below and include software-specific commands.

For the HIS stratum, the bulk of the sample QC was performed before the hyperellipsoid analysis and is described in **Section 9.3.2**. Additional QC is described below.

9.4.1 SNP QC: Cross-study genotyping concordance

Because the bulk of cases and controls had been separately genotyped on different platforms and at different times (a design that can increase the number of false-positive associations driven by genotyping batch), SNP QC was highly stringent.

To ensure comparability of case and control genotype data that was obtained in part across array platforms and core facilities, a number of samples from cohorts which have been genotyped using earlier smaller arrays were selected as quality control duplicate samples to be genotyped on the Illumina 5M arrays in CIDR core. These samples were:

(a) 30 INMA samples previously genotyped on the 1M array
(b) 30 HRS samples previously genotyped on the 2.5M array
(c) 28 GRAZ samples previously genotyped on the Illumina 610 array
(d) 30 OAI samples previously genotyped on the Illumina 2.5M array
(e) 25 LUND samples previously genotyped on the Omni Express 750k v1.0 array
(f) 5 LUND samples previously genotyped on the Omni Express 750k v1.1 array

The number of overlapping SNPs between the 5M and each of the smaller arrays used for comparison is provide in **Supplementary Table 12**. The overlapping SNPs were checked for genotyping discordance between duplicate samples. Any SNP that showed ≥ 1 discrepancy in any of the comparisons was removed from all study strata.

9.4.2 Sample QC: Missingness

Samples with missingness > 5% (--*missing*) were excluded from further analyses (EUR and AFR samples).

9.4.3 Sample QC: PCA

a. PCA in the EUR and AFR strata

We used a high-quality set of SNPs (see **Section 9.1.3**) and calculated PCA in each stratum and excluded outliers as needed; outliers included samples deviating substantially from the PC1 and PC2 values in the study stratum or, as was the case for the HABC cohort, the removal of a group that formed a separate population cluster on PCs 1 and 2 from the other samples in the cohort.

Using the set of SNPs used for PCA, we calculated inbreeding coefficients. Samples lying > 3 standard deviations from the mean of the distribution were removed from the data.

b. (Relatedness and) PCA in the HIS stratum

Controls with any previous report of stroke in HRS or HCHS/SOL were excluded at this time. IBD estimation showed that nearly all Hispanic/Latino stroke cases were unrelated through degree 3 relationships (first cousins). We identified 12,524 unrelated Hispanic/Latino individuals chosen as pairs of subjects that are not connected by a Degree 3 or closer relationship (kinship coefficient > 0.0625) and who were unrelated to members in either the EUR or AFR groups.

The KING-method(63) (function: *snpgdsIBDKING*) in SNPRelate,(58) which accounts for population structure, was used to calculate kinship coefficients.

PCA was performed on 12,524 unrelated Hispanic/Latino study participants, with and without HapMap ancestral populations. Four outlying participants with a high level of Asian ancestry were identified and excluded.

Additional PCA rounds were performed on the 12,520 remaining participants, with and without HapMap samples; no additional outlying participants were identified.

9.4.4 Sample QC: Relatedness (EUR and AFR strata)

Relatedness was also calculated using the same set of SNPs as was used for PCA. Pairs of samples with a kinship coefficient > 0.0625 (cousin-relationship or higher) were identified; if the pair of samples consisted of a case and a control, the case was kept and the control removed. Otherwise, the sample with the higher callrate was kept and the other was removed. If a sample appeared in more than one pair, it was preferentially removed so as to minimize sample loss.

For samples in EUR strata, the identity-by-descent function (PLINK method of moment (MoM) for IBD analysis, function: *snpgdsIBDMoM*) in SNPRelate(58) was used. For samples in AFR strata, the KING-method(63) (function: *snpgdsIBDKING*) in SNPRelate,(58) which accounts for population structure, was used to calculate kinship coefficients.

9.4.5 SNP QC: Frequency

Frequency was calculated in PLINK (*--freq*) for all SNPs in the EUR and AFR groups for three groups of samples (within a single stratum): cases, controls, and all samples. If a SNP's minor allele frequency was < 1% in any of the three groups, it was removed.

9.4.6 SNP QC: Hardy-Weinberg equilibrium and missingness

Missingness is often nonrandom and differences in missing call rate between cases and controls can lead to spurious findings in association tests.

a. EUR and AFR strata

An identical process as was used for the frequency QC step was used to check Hardy-Weinberg Equilibrium (HWE, *--hardy*) and missingness (*--missing*) using PLINK. Missingness and HWE p-values were calculated in cases only, controls only, and all samples. If missingness was > 1% or the HWE p-value was < 1 x 10^{-3} in any of the three groups, the SNP was removed from the analysis.

Differential missingness (*--test-missing*) was calculated. All SNPs with p < 1 x 10^{-3} from the Fisher's exact test were dropped.

b. HIS stratum

For each SNP, we performed a Fisher's exact test on the number of missing calls among the cases and controls (function fisher.test in R). Any SNP with p < 10^{-3} was flagged as failing QC. In order to keep high quality SNPs and bound missingness, we failed any SNP with a missing call rate of > 1% among either cases or controls.

Additionally, because the number of individuals was quite different across projects (there were 8.5 – 11 times as many HCHS/SOL samples as there were HIS samples in SiGN and HRS), a large difference in the number of missing genotype calls by study contributed to artifacts (visible in genotype cluster plots by study). Thus, we imposed study-specific filters failing any SNP with a missing call rate of > 0.5% among the 10,363 HCHS/SOL controls or > 1% among either the 1,214 HRS controls or the 942 SiGN cases.

9.4.7 SNP QC: Comparison to 1000 Genomes

This QC step was only performed on EUR strata.

The European-ancestry samples from the 1000 Genomes Phase I data(32) (1KGP-EUR) were coded as controls and the individual cohorts in a study stratum were coded as cases to make a pseudo case-control comparison. The 1KGP-EUR populations are:

(a) FIN: Finnish in Finland

(b) IBS: Spanish from the Iberian peninsula in Spain
(c) CEU: People of Northern and Western European ancestry living in Utah
(d) GBR: British living in Great Britain
(e) TSI: Tuscans in Italy

To run the association testing, we used FaST-LMM,(68) which uses a linear mixed model for association testing, including a genetic relationship matrix (GRM), which captures and corrects for population structure in the data. Aside from the GRM, no additional covariates were used in the regression analysis. SNPs with $p < 1 \times 10^{-3}$, indicating a large frequency difference between the SiGN cohorts and 1KGP-EUR were excluded from the analysis.

The comparison to 1KGP-AFR samples was not done because association testing between 1KGP-AFR and the SiGN AFR strata revealed strong population stratification between the sets of samples.

9.4.8 SNP QC: Case-case and control-control comparisons

This was the final SNP QC step before beginning initial genome-wide testing.

a. EUR and AFR strata

Wherever possible, case cohorts (or control cohorts) from the same study stratum were matched up, with one cohort coded as cases and the other as controls we performed association testing between the two sets of samples (adjusting for ten principal components). Because case cohorts were paired with other case cohorts, and controls with controls, the resulting association testing between the two cohorts should have a null distribution. SNPs indicating association between two cohorts ($p < 1 \times 10^{-3}$) were dropped from the analysis.

b. HIS stratum

In the HIS control cohorts, the pseudo-association tests were performed adjusting for age at baseline exam and the first 10 PCs. Any SNP with a likelihood-ratio test $p < 10^{-3}$ was flagged for failing QC.

A complete summary of all stages of variant QC in the Hispanic samples is presented in **Supplementary Table 13**.

9.4.9 Sample QC: Relatedness Check in all EUR strata

Once all of the EUR strata had been cleaned (sample and SNP QC), the full set of EUR samples (across all strata) were merged together across 180,085 SNPs.

From the merged data, we extracted high-quality SNPs (see **Section 9.1.3**) and calculated relatedness across the full set of EUR samples using the identity-by-descent function (*snpgdsIBDMoM*) in SNPRelate.(58) Sample pairs with a kinship parameter > 0.0625 (cousin-relationship or higher) were identified. In pairs composed of a case and a control, the control was removed from its respective stratum. Otherwise, for case-case and control-control pairs, the sample with higher genotype missingness was removed from its respective stratum. If a sample appeared in more than one pair, it was preferentially removed so as to minimize sample loss.

A pan-AFR relatedness check was not done. 10 pairs of related (kinship > 0.0625) remain in the analysis. Only 3 of the 10 pairs contain a case sample; the other 7 pairs are control-control pairs of individuals. Consequently (and also because the AFR strata contribute such a small fraction of the discovery sample), the related pairs likely have minimal to no impact on summary statistics in association testing.

After sample and SNP QC were complete (**Supplementary Tables 14 – 15**), the following number of cases and controls were available in each population group:

(a) EUR: 12,577 cases and 26,340 controls

(b) AFR: 1,353 cases and 2,383 controls
(c) HIS: 942 cases and 11,578 controls

9.5 Initial genome-wide association testing

a. EUR and AFR strata

Once all sample and SNP QC were complete, we calculated principal components for the QC-passing samples in each EUR and AFR stratum using a high-quality SNP set extracted from the cleaned data.

We ran an initial all-stroke (case/control) genome-wide association study (GWAS) in each of the 13 study strata, correcting for the top ten principal components. The purpose of the GWAS was to check the genomic inflation factor (lambda, λ). High genomic inflation (approximately λ > 1.05) indicated the need for additional QC due to potential population stratification and/or enrichment for false associations likely driven by study design; a lower genomic inflation (approximately λ < 1.05, consistent with previous ischemic stroke GWAS) indicated sufficient QC and that the data was ready for prephasing and imputation (**Supplementary Figure 16**).

After all QC was complete, the highest lambda across the EUR and AFR strata was 1.056 (in Group 2).

b. HIS stratum

We performed preliminary association testing of all stroke on the 942 SiGN cases, 1,214 HRS controls, and 10,363 HCHS/SOL controls on 1,801,834 QC-passing SNPs. In order to be consistent with covariates used in the analysis of the EUR and AFR strata in the meta-analysis, we did not adjust for age at exam in the analysis of HIS individuals as many of the publicly available datasets comprising the non-Hispanic/Latino strata did not include age at exam. However, we found that adjusting for 7 PCs and sex but not age in the Hispanic/Latino stratum yielded an inflated λ (1.110 versus 1.060) and a QQ plot with early departure from the expected distribution.

Consequently, we adopted an approach to reduce lambda for analysis without age adjustment by selecting a subset of controls for association testing. We reduced the number of HCHS/SOL controls to match the number of HRS controls, as balancing the two control sets might reduce the influence of artifacts specific to one control project. Further, the subset of HCHS/SOL controls could be selected to better match the age and sex distribution of the SiGN cases. If the age distributions were similar between cases and controls, age confounding would be less likely.

HRS and SiGN have roughly similar age at exam distributions, so all 1,214 HRS controls would be kept. However, the HCHS/SOL cohort enrolled younger participants relative to SiGN and HRS. As a result, we sampled at random within each of 16 age categories stratified by sex to achieve the same SiGN proportions, resulting in an HCHS/SOL subset of 1,214 controls (**Supplementary Table 16**).

We found that this approach, even without adjustment for age, yielded a reasonable λ of 1.029.

c. Additional PCA in the HIS stratum

An additional round of PCA was then performed on the 3,371 individuals and beginning with only the SNPs that passed all stages of QC in **Sections 9.3** and **9.4** (pruned to 143,314 SNPs). This final set of PCs computed on the 3,371 (**Supplementary Figure 17**) was used as covariates in further association testing regression models to adjust for population structure.

10. Prephasing

After the QC process was complete, all samples were prephased using SHAPEIT2.(69,70) Prephasing performs haplotype estimation for each sample. It improves downstream imputation accuracy and imputation runtime.

10.1 Prephasing in the EUR and AFR strata

Prephasing was completed on a per-chromosome basis to allow for parallelization. Consequently, per-chromosome missingness is crucial to being able to prephase accurately, and SHAPEIT2(69,70) performs a per-chromosome missingness check before beginning prephasing. A small number of samples failing the SHAPEIT2-based filter of < 10% missingness were dropped.

Prephasing was performed using genetic maps provided on the SHAPEIT2 website for data on hg19. The 1000 Genomes Project Phase I data (1,092 individuals from 4 continental populations)(32) was used as a reference panel for prephasing. The populations represented in 1000 Genomes Phase I are:

(a) FIN: Finnish in Finland (Continental group: Europe)
(b) IBS: Iberian population in Spain (Continental group: Europe)
(c) CEU: Utah residents of Northern and Western European ancestry (Continental group: Europe)
(d) GBR: British living in England and Scotland (Continental group: Europe)
(e) TSI: Tuscans in Italy (Continental group: Europe)
(f) YRI: Yoruba in Ibadan, Nigeria (Continental group: Africa)
(g) LWK: Luhya in Webuye, Kenya (Continental group: Africa)
(h) ASW: Americans of African ancestry in southwest USA (Continental group: Africa)
(i) JPT: Japanese in Tokyo (Continental group: East Asia)
(j) CHB: Chinese in Beijing (Continental group: East Asia)
(k) CHS: Southern Han Chinese (Continental group: East Asia)
(l) PUR: Puerto Ricans in Puerto Rico (Continental group: the Americas)
(m) MXL: Mexicans in Los Angeles (Continental group: the Americas)
(n) CLM: Colombians in Medellin, Colombia (Continental group: the Americas)

For prephasing on chromosome X, the *–chrX* option available in SHAPEIT2 was used. Prephasing was performed using four threads, so as to shorten computational time.

Due to large sample numbers in Group 2 (ESS, MUNICH, OXVASC, STGEORGE, KORA, WTCC) and Group 4 EUR samples (CIDR, HRS, OAI), these groups were split into smaller groups containing roughly the same proportions of cases and controls and chromosomes 1 – 6 were split into 5Mb windows to shorten runtime. The data were merged back together once prephasing was complete.

10.2 Prephasing in the HIS stratum

Before we began prephasing, we became aware of a new release of annotation for the Illumina HumanOmni5Exome-4v1 array. As a result, an additional 5,072 variants were flagged for removal prior to prephasing and imputation due to the updated annotation of chromosome, position, or a strand ambiguous SNP with updated reference strand between annotation versions “A” and “B”. Hence, the data for prephasing and imputation input consisted of 1,908,773 genotyped variants passing all stages of QC.

Prephasing of genotyped variants in the imputation basis was performed using SHAPEIT2.(69,70) All 15,056 study participants in the Hispanic/Latino composite analysis group or HCHS/SOL (including population outliers) who passed QC in the combined dataset were included in the prephasing step, even though only the 3,371 individuals described in **Section 9.5** would be included in association testing in **Section 12** because inclusion of relatives and the larger set of Hispanic/Latino individuals would improve the accuracy of the prephasing step.(71)

11. Imputation

After prephasing was complete, imputation reference panels based on next-generation sequencing (NGS) data were used to impute genotype dosages across the autosomal chromosomes and chromosome X.

11.1 Imputing the EUR strata

To impute the EUR samples, we first created a merged imputation panel using data from two next-generation sequencing projects.

The first project was the 1000 Genomes (Phase I) project (1KG)(32), consisting of 1,092 samples representing four different continental ancestries: Europe, East Asia, Africa, and the Americas. Samples were sequenced at 4x coverage outside the exome and at ~80x coverage over the exome.

The second project was the Genome of the Netherlands(72) (GoNL) project, consisting of medium-depth (~14x) coverage across 250 trios of Dutch ancestry. Data from the 499 unrelated individuals in GoNL was used for construction of the SiGN imputation panel.

The merged panel was constructed using the IMPUTE2(73–75) method for merging two reference panels, using the *–merge_ref_panels* option. Briefly, 1KG was used to impute 1KG-only variants into GoNL. Then, GoNL was used similarly to impute GoNL-variants into 1KG. For imputing the first panel into the second panel, and then imputing the second panel into the first, we imputed in 5Mb windows using a 250kb buffer region that was included in the merged reference panel produced. The effective sample size argument (*-Ne*) was set to 11418, as suggested by the IMPUTE2 best practices. The *k_hap* argument, which indicates to the software the number of haplotypes likely needed to be searched in order to perform accurate imputation at a given site, was set to 1,000 for 1KGP (roughly the number of EUR haplotypes in 1KGP) and 998 for GoNL (all of the haplotypes contained in GoNL). Finally, IMPUTE2 produced merged haplotypes based on 1KG and GoNL that can be used for downstream imputation.

Once the merged reference panel had been constructed, the stroke cases and controls were imputed using the 1KG+GoNL panel. With the exception of Group 2 (ESS, MUNICH, OXVASC, STGEORGE, KROA, WTCCC) and Group 4 (CIDR, HRS, OAI), the samples in all analysis groups were imputed together. Group 2 was split into three groups for imputation and Group 4 was split into six groups for imputation; the split groups were comprised of cases and controls held in approximately the same ratio.

Samples were imputed in 5Mb non-overlapping windows using a 250kb buffer. The effective sample size *(-Ne*) argument, was set to 11,418 per the IMPUTE2 best practices. The *–k_hap* argument was set to 1,000 and the *–use_prephased_g* argument was used, as the samples had been prephased.

11.2 Imputing the AFR strata

The AFR and HIS samples were imputed nearly identically to the European-ancestry samples. However, only the 1000 Genomes (Phase I)(32) samples were used as the reference panel for imputation; using the additional 499 GoNL samples in the reference panel would unlikely improve imputation substantially, as they only added more European-ancestry samples to the reference panel.

Samples were imputed in 5Mb non-overlapping windows using a 250kb buffer. The effective sample size (*-Ne*) argument, was set to 11,418 per the IMPUTE2 best practices. The *–k_hap* argument was set to 1,000 and the *–use_prephased_g* argument was used, as the samples had been prephased.

11.3 Imputing the HIS stratum

Imputation of all 15,056 prephased Hispanic/Latino participants was performed using IMPUTE2 v2.3.1(73–75) and the 1000 Genomes Phase 1(32) version 3 “nosing” reference panel released by the IMPUTE2 website. The “nosing” reference panel has filtered variants (pre-imputation) with only one copy of the minor allele across the 1,092 individuals in 1000 Genomes Phase 1. As is recommended for admixed study populations, the *k_hap* parameter was increased (from a default of 500 to 2,184, the number of 1000G haplotypes). The *-Ne* parameter was increased (to 20,000) and a buffer size of 500 kb (*buffer*) was used for the flanking regions of imputation segments over the default of 250 kb.

11.4 Post-imputation quality control

11.4.1 EUR and AFR strata

After imputation, 38.8 – 46.8M variants had been imputed in each of the analysis groups (the range in variants is primarily driven by the fact that the number of input SNPs varied across study strata and some strata contained information for chromosome X while others did not). Variants with an imputation info score < 0.5 or out of Hardy-Weinberg Equilibrium ($p < 10^{-6}$) were removed from the data, leaving 21.9 – 32.6M variants available for genome-wide association analysis (**Supplementary Table 17**).

11.4.2 HIS stratum

Prior to meta analysis variants were filtered if the imputation quality (info) score < 0.5 (IMPUTE2(73,74) and SNPTEST(75,76) imputation quality (info) scores) and expected effective heterozyosity count (effHC), calculated as $2 \times MAF \times (1-MAF) \times N \times oevar,$ < 20, where MAF is the minor allele frequency, N is minimum of the number of cases or controls, and oevar is the observed to expected variance ratio of the imputed allele dosages for a given SNP (oevar = 1 for genotyped variants). The effHC indicates the expected effective number of heterozygous individuals, and, for rare SNPs, estimates the minor allele count (since there are few or no homozygous minor individuals).

Note it was not necessary to apply the MAF < 0.01 filter (as applied in the EUR and AFR strata) as the filter on the expected effective number of heterozygotes was more stringent, given the sample size in the HIS stratum.

12. Stage I genome-wide association analysis

Genome-wide association testing was performed in all study strata for 15 different traits (**Supplementary Table 18**) for the CCS Causative (CCSc), CCS Phenotypic (CCsp) and TOAST subtyping methods. Please note that we use the terms “stage I” and “discovery” interchangeably in the following sections.

(a) IS: all ischemic stroke
(b) CCScCEmajor: cardioembolic stroke (CE)
(c) CCScLAA: large artery atherosclerosis (LAA)
(d) CCScSAO: small artery occlusion (SAO)
(e) CCScUNDETER: undetermined (UNDETER), includes cases with incomplete evaluation (CCScINCUNC) and multiple competing causes (CRYPTCE, cryptogenic and CE minor)
(f) CCScCRYPTCE: cryptogenic and cardioembolic minor
(g) CCScINCUNC: incomplete or unclassified
(h) CCSpCEmajinclCE: CE
(i) CCSpLAAmajincl: LAA
(j) CCSpSAOmajincl: SAO
(k) CCSpCryptoincl: Cryptogenic inclusive (undetermined)
(l) toastCE: CE
(m) toastLAA: LAA
(n) toastSAO: SAO
(o) toastUNDETER: undetermined

An “other” subtype is available in CCSc, CCSp, and TOAST. However, due to very small case counts (CCScOTHER = 594 cases, CCSpOTHER = 718 cases, toastOTHER = 373 cases) and consequently limited power (**Supplementary Figure 9**), these phenotypes were not considered in any GWAS.

12.1 Genome-wide association studies (GWAS) in individual strata

A GWAS was performed in each of the individual study strata for each of the 15 phenotypes over all of the SNPs available post imputation QC. If a single stratum contained < 40 cases for a given phenotype, a GWAS was not performed for that phenotype in that stratum (**Supplementary Tables 19 – 21**).

GWAS were performed using SNPTEST.(75,76) The *–frequentist 1* argument was used to test an additive model using logistic regression and the *–method expected* argument was used to test the imputed dosages produced from imputation. All GWAS included the top ten principal components and sex as covariates. The *–lower_sample_limit* argument was set to 10, because sample counts in some of the subtypes for some of the strata were small.

Chromosome X was not analyzed at the GWAS stage. Many of the control cohorts provided data only for the autosomal chromosomes and consequently, power for discovery across the X chromosome was substantially lowered.

12.2 Post-GWAS Quality control

After all GWAS within the individual study strata were complete, we performed additional QC. We generated a QQ plot across all SNPs for each GWAS in each stratum, and observed that lambda had increased from the genotype-only GWAS to the imputed-variant GWAS, potentially an artifact of the separate genotyping of cases and controls.

To identify the SNPs driving the increased genomic inflation, we used three different parameters to generate bins of SNPs and then calculated lambda across each of these bins. The parameters used to stratify SNPs were:

(a) Imputation quality (info score, provided by IMPUTE2)
(b) Frequency (calculated from genotype counts produced by SNPTEST)
(c) Expected effective heterozygosity count (effHC): effHC = 2 * MAF * (1-MAF) * n * oevar, where MAF is the minor allele frequency of the SNP, oevar is the observed over expected variance of the imputed dosages (equal to 1 for genotyped SNPs), and n is the minimum of (n_cases, n_controls).

Filtering SNPs based on imputation quality and effHC did not successfully decrease the genomic inflation factor. We observed that removing markers with a minor allele frequency < 1% removed the excess genomic inflation (**Supplementary Table 22**, **Supplementary Figure 19**), so all markers with MAF < 1% were consequently removed from the analysis.

13. Stage I meta-analysis

After quality control and imputation (with additional quality control) was complete, 16,851 cases and 31,259 controls (including additional cases and controls from VISP; see **Section 13.1**) with 22.0M – 31.9M SNPs were available for meta-analysis.

To combine summary-level results across the study strata, we used inverse variance-weighted fixed effects meta-analysis using MANTEL.(77)

13.1 Addition of VISP summary results

In addition to the study strata that had been constructed and cleaned (as described above), summary-level results from one additional cohort (VISP) were provided for the discovery phase of SiGN. Summary results were reported for both VISP Geneva (samples of European ancestry) and VISP Handls (samples of African ancestry). Subtypes were not available in VISP, so summary results were reported for the all stroke phenotype only. VISP underwent the same data quality control steps as the other study strata (next section, **Supplementary Figure 19**).

13.2 Additional data QC

Before running meta-analysis, the following filters were used to remove SNPs from analysis:

(a) Beta, standard error (SE), or p-value not available (NA)
(b) Beta > 100,000
(c) Comparison of frequency of imputed allele to frequency reported in 1000 Genomes(32) continental population (Europe for EUR stroke samples, Africa for AFR stroke samples, the Americas for HIS stroke samples), removing any SNP with a frequency difference > 30% between stroke and 1KG.

Once the data had been cleaned, data from each stratum was split into chunks of 125,000 SNPs for meta-analysis.

13.3 Fixed effects inverse variance-weighted meta-analysis

MANTEL(77) was used to perform all meta-analyses of IS and the various subtypes (**Supplementary Table 23**). A fixed effects beta, standard error, and p-value were calculated for each SNP. Lambda was calculated across all SNPs to check the overall behavior of the test statistic (**Supplementary Figure 7**). Manhattan and QQ plots were generated for all meta-analyses (**Supplementary Figure 2**).

13.4 Validation of discovery results

Once discovery results were finalized, all of the summary results (including all variants with frequency < 1% that were removed before discovery meta-analyses) from each of the study strata were sent to a separate analyst in the SiGN analysis group so that the results could be validated.

The second analyst independently carried out quality control on the individual strata using EasyQC.(78) All data was then meta-analyzed using METAL.(34) Results from the second external discovery meta-analyses were highly concordant with the initial discovery meta-analyses (**Supplementary Table 24**).

14. Stage II genome-wide analysis

14.1 Genetic and Phenotypic Correlation of CCS and TOAST

Because every existing CCS case had been included in the discovery (stage I) phase of SiGN, only TOAST-subtyped cases could be considered for replication (stage II). To check that use of TOAST cases in a discovery that was CCS-based was appropriate, we extracted the z-scores from each of the 15 stage I meta-analyses that we had performed and calculated the correlation, using all SNPs across the genome, between each pair of phenotypes (**Figure 1, Supplementary Figure 8, Supplementary Table 3**).

The z-score correlations across CCS Causative (CCSc), CCS Phenotypic (CCSp), and TOAST, within the subtypes, were:

(a) CE: 0.700 (CCSc – TOAST) and 0.698 (CCSp – TOAST)
(b) LAA: 0.678 (CCSc – TOAST) and 0.611 (CCSp – TOAST)
(c) SAO: 0.751 (CCSc – TOAST) and 0.734 (CCSp – TOAST)
(d) Undetermined: 0.620 (CCScUNDETER – TOAST), 0.390 (CCScINCUNC – TOAST), 0.533 (CCScCRYPTCE – TOAST), 0.429 (CCSpCryptoincl – TOAST)

The moderate to strong correlations indicated that it might be possible to replicate CCS-discovered loci using TOAST-subtyped cases.

We also examined the phenotypic correlations by looking at the phenotypic assignments for all of the cases, where, for each of the 15 phenotypes, a sample was annotated as "1" to indicate it was a case for a particular subtype (as defined by either CCSc, CCSp, or TOAST), and "0" otherwise. Within subtypes, the phenotypes were highly correlated (**Supplementary Figure 8**). Correlations across the different subtypes were weak and often inversely correlated, indicating that few cases were assigned to one subtype by the CCS Causative system and to a separate subtype by the CCS Phenotypic or TOAST system.

14.2 *In silico* lookups for stage II replication

Please note that we use the terms "stage II" and "replication" interchangeably in the following sections. All SNPs with $p < 1 \times 10^{-6}$ in any of the 15 discovery GWAS were selected for stage II follow-up. The SNPs selected for replication were not pruned based on linkage disequilibrium.

Stage II SNP lists were provided to each of the replication cohorts (**Supplementary Table 1**) for *in silico* lookup in pre-existing summary results from GWAS using TOAST-subtyped cases and controls. The SNPs were consolidated by subtype, so that e.g. all SNPs with $p < 1 \times 10^{-6}$ in the CCS Causative cardioembolic, CCS Phenotypic cardioembolic or TOAST cardioembolic GWAS were merged into a single list of cardioembolic SNPs to be looked up in the replication cohorts.

All stage II cohorts were checked for overlapping cases and controls included in SiGN; if stage II cohorts had included overlapping cases or controls, these samples were dropped from the replication group and summary results were recalculated. Summary results were extracted from the replication cohorts and reported back to SiGN.

For one replication cohort, SAHLSIS, imputed dosages for the replication SNPs were provided to the SiGN analysis group. The dosages were analyzed using logistic regression in PLINK,(60) correcting for sex and the first 5 principal components.

14.3 Joint analysis of stage I and stage II

We chose a multi-stage meta-analysis (in which stage I data undergoes goes genome-wide association testing, additional information is collected on potentially associated SNPs in a second stage, and then all data from stage I and stage II is jointly analyzed) because it is a statistically robust method that has improved power compared to separate analysis of discovery (stage I) and replication (stage II) data.(79) Before being merged into a joint analysis, the stage II data was checked for consistent SNP frequencies using the 1000 Genomes Phase I(32) continental populations as a reference. SNPs with a frequency difference > 30% compared to 1000 Genomes were removed. Additionally, SNP names and alleles were checked for consistency with the SiGN data. Replication SNPs were not cleaned further.

Once the stage II data was cleaned and formatted appropriately, the results from each replication cohort were added to the appropriate meta-analyses for the different traits analyzed in stage I (**Supplementary Figure 3, Supplementary Table 25**). For example, all of the SNPs selected for replication from one of the cardioembolic GWAS were first looked up in the cardioembolic GWAS of TOAST-subtyped cases (and matched controls) performed by the replication cohort. Then, the summary results for those SNPs were jointly meta-analyzed with the stage I results for CCS Causative cardioembolic, CCS Phenotypic cardioembolic, and TOAST cardioembolic.

Joint analysis was performed for all 15 traits examined in stage I using MANTEL.(77) SNPs exceeding $p < 1 \times 10^{-8}$, correcting for five traits tested in total, were considered to be genome-wide significant.

15. Genome-wide associated loci

After correcting for associating testing in five independent traits, all SNPs with a genome-wide p-value $< 1 \times 10^{-8}$ were considered to be genome-wide significant. We observed genome-wide significant p-values at four previously described ischemic stroke loci and at three novel loci (**Supplementary Table 26**). Additionally, we investigated the signal at four previously described risk loci for ischemic stroke that were not genome-wide significant in our study.

Genomic inflation (lambda) for all meta-analysis performed in the discovery and replication phases can be found in **Supplementary Table 27**.

15.1 Previously-described loci

The following four loci have been previously implicated in conferring risk for ischemic stroke or one of its subtypes and were also genome-wide significant in our study:

(a) *PITX2*, previously associated to cardioembolic (CE) stroke(80)
(b) *ZFHX3*, previously associated to CE(81)
(c) *HDAC9*, previously associated to large artery atherosclerosis (LAA)(82)
(d) 12q24.12, previously associated to all ischemic stroke(83)

PITX2, *ZFHX3*, and *HDAC9* are genome-wide significant in the subtypes for which they were previously implicated (**Supplementary Figures 19 – 23**). 12q24.12 was previously reported as associated with all ischemic stroke but in our study appeared associated primarily to the small artery occlusion (SAO) subtype (**Supplementary Figure 3c**).

Four additional loci have also been previously implicated as conferring risk to ischemic stroke or one its subtypes, but did not have a p-value $< 1 \times 10^{-6}$ after the discovery phase and were consequently not pushed forward to replication (**Supplementary Figure 25 – 28**). These loci are:

(a) *ABO*, previously implicated in all stroke (IS), CE, and LAA(84)
(b) *NINJ2*, previously implicated in IS(85,86)
(c) 6p21, previously implicated in LAA(87)
(d) *CDKN2B-AS1*, previously implicated in LAA(88)

and their odds ratios and p-values after the discovery phase are reported in **Table 2** of the main text.

NINJ2, discovered in a set of stroke cases and controls not used in the SiGN GWAS, shows no evidence for association and is likely a false-positive association (**Supplementary Table 2**).

ABO (previously associated to IS, CE, and LAA), 6p21 (previously associated to LAA), and *CDKN2B-AS1* (previously associated to LAA) show nominal evidence for association. To investigate how much of the signal is due to newly included cases and how much is due to previously analyzed samples, for each of these three loci, we removed the samples used to initially discover the gene and reran the meta-analysis (**Supplementary Table 2**). The overlapping cohorts were:

(a) *CDKN2B-AS1*: the ISGS cohort. These samples were included in Group 1 (EIR) and Group 4 (EUR, AFR, HIS). Only four ISGS samples were included in the HIS group and were not removed from the analysis, but are unlikely to have a substantial impact on the summary-level statistics.
(b) 6p21: the ASGC cohort (cases and controls in Group 10).
(c) *ABO*: the WTCCC cohort (cases and controls in Group 2).

For the subtype-specific analyses, we reran the meta-analyses that used the cases classified by the TOAST system, as this was the system originally used to discover these genes.

After removing the ASGC cohort, 6p21 showed no evidence for association (OR for the T allele = 1.04, p = 0.304, **Supplementary Table 2**) in the discovery meta-analysis of TOAST-determined LAA, indicating that the nominal evidence observed in the original GWAS was driven exclusively by the ASGC samples.

After removing the ISGS cohort from the discovery meta-analysis of IS, CE (TOAST), and LAA (TOAST), ABO remained nominally associated to all three phenotypes (OR for the C allele = 1.07, p = 2.5 x 10^{-4}; OR = 1.15, p_{TOAST} = 2.5 x 10^{-4}; OR = 1.10, p_{TOAST} = 0.007, **Supplementary Table 2**). After removing Group 2 from the discovery meta-analysis of LAA (TOAST), *CDKN2B-AS1* also showed nominal evidence for association (OR for the G allele = 1.09, p = 0.009). Future GWAS studies in larger samples will help determine the strength of these associations.

15.2 Testing the specificity of 12q24.12 to the SAO subtype

To formally test whether 12q24.12 was indeed specific to SAO rather than all ischemic stroke as first reported,(83) we performed a test of heterogeneity, comparing the odds ratios and p-values observed for the 12q24.12 SNP (rs10744777) between SAO and the other three subtypes (CE, LAA, and undetermined).

We used the CCS Phenotypic discovery data (where the variant was most significant) to obtain the odds ratios for each subtype (**Supplementary Figure 6**). If an individual had multiple classifications including SAO, we assigned that individual only to SAO. We then constructed a statistic,

$$S_{Diff} = \beta_{SAO} - \beta_{Others}$$

where β_{SAO} is the log odds ratio for SAO and β_{Others} is the log odds ratio for the union of LAA, CE, and undetermined. If an individual had multiple classifications including SAO, we assigned that individual only to SAO.

Obtaining the variance of S_{Diff} was complicated due to the fact that we used shared controls to obtain β_{SAO} and β_{Others}, which caused correlation between statistics. We wanted to calculate the correlation, but we could not use the correlation formula of Lin and Sullivan(89), because their formula is based on the null hypothesis of no association. In this particular situation, the null hypothesis is equal effect sizes between subtypes (that is, no heterogeneity of effect).

To empirically assess the correlation, we designed a permutation procedure that permutes individuals within cases only, effectively imposing the null hypothesis of equal effect sizes, while keeping the overall association intact. We performed 1,000 permutations to calculate the correlation between β_{SAO} and β_{Others}, and found $r_{SAO,Others}$ = -0.407 (a negative correlation was expected due to our within-case permutation scheme). Then, we calculated the variance of S_{Diff},

$$Var(S_{Diff}) = SE_{SAO}^2 + SE_{Others}^2 - 2 \times SE_{SAO} \times SE_{Others} \times r_{SAO,Others}$$

where SE_{SAO} and SE_{Others} refer to the standard errors of β_{SAO} and β_{Others} respectively. Using this variance, we were able to calculate the z-score and the corresponding p-value of our test.

Plots of the effects were generated using METASOFT(90) and ForestPMPlot(91) and are provided in **Supplementary Figure 6**.

15.3 Novel loci

In addition to the previously described loci, three additional loci were genome-wide significant after the combined meta-analysis of discovery and replication data. Rs12122341, near *TSPAN2*, was genome-wide significant in CCS Causative LAA discovery and TOAST LAA replication meta-analysis, as well as in the CCS Phenotypic LAA discovery and TOAST LAA meta-analysis. The forest plot and regional association plot for this locus appear in **Figure 2** of the main text.

Rs74475935, an intronic SNP in the ABCC1 gene, was genome-wide significant in the CCS Causative (Cryptogenic and CE minor) discovery and TOAST undetermined replication meta-analysis, and in the CCS Phenotypic cryptogenic and TOAST undetermined meta-analysis (**Supplementary Figure 5**).

This SNP is extremely rare in European samples (risk allele frequency ~0.1%) and low frequency in African-ancestry samples (risk allele frequency ~1.5%). Because of a small number (~5%) of cases in both the discovery and replication phase, future studies that interrogate more African-ancestry individuals are needed to determine the robustness of this association.

16. Online Sources

[1] clinicaltrials.gov
[2] http://hrsonline.isr.umich.edu/
[3] http://www.proyectoinma.org/en_index.html
[4] www.helmholtz-muenchen.de/en/kora-en
[5] http://www.oai.ucsf.edu/
[6] http://www.b58cgene.sgul.ac.uk/
[7] http://www.cls.ioe.ac.uk/studies.asp?section=000100020003
[8] www.gsf.de/kora/en/english.html
[9] https://ccs.mgh.harvard.edu
[10] https://ccs.mgh.harvard.edu/ccs_intro.php
[11] Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Builds: 136, 137, and 138). Available from: http://www.ncbi.nlm.nih.gov/SNP/

IV. Authors

Sara L. Pulit*‡, Patrick F. McArdle†*, Quenna Wong*, Rainer Malik*, Katrina Gwinn‡, Sefanja Achterberg, Ale Algra, Philippe Amouyel, Christopher D. Anderson, Donna K. Arnett†, Ethem Murat Arsava, John Attia, Hakan Ay‡, Traci M. Bartz, Thomas Battey, Oscar R. Benavente†, Steve Bevan, Alessandro Biffi, Joshua C. Bis, Susan H. Blanton, Giorgio B. Boncoraglio, Robert D. Brown Jr., Annette I. Burgess, Caty Carrera, Sherita N. Chapman Smith, Daniel I. Chasman, Ganesh Chauhan, Wei-Min Chen, Yu-Ching Cheng, Michael Chong, Lisa K. Cloonan, John W. Cole†, Ioana Cotlarciuc, Carlos Cruchaga, Elisa Cuadrado-Godia, Tushar Dave, Jesse Dawson, Stéphanie Debette, Hossein Delavaran, Cameron A. Dell, Martin Dichgans†, Kimberly F. Doheny, Chuanhui Dong, David J. Duggan, Gunnar Engström, Michele K. Evans, Xavier Estivill Pallejà, Jessica D. Faul, Israel Fernández-Cadenas, Myriam Fornage†, Philippe M. Frossard, Karen Furie, Dale M. Gamble, Christian Gieger, Anne-Katrin Giese, Eva Giralt-Steinhauer, Hector M. González, An Goris, Solveig Gretarsdottir, Raji P. Grewal†, Ulrike Grittner, Stefan Gustafsson, Buhm Han, Graeme J. Hankey, Laura Heitsch, Peter Higgins, Marc C. Hochberg, Elizabeth Holliday, Jemma C. Hopewell, Richard B. Horenstein, George Howard, M. Arfan Ikram, Andreea Ilinca, Erik Ingelsson, Marguerite R. Irvin, Rebecca D. Jackson, Christina Jern†, Jordi Jiménez Conde†, Julie A. Johnson†, Katarina Jood, Muhammad S. Kahn, Robert Kaplan, L. Jaap Kappelle, Sharon L.R. Kardia, Keith L. Keene, Brett M. Kissela, Dawn O. Kleindorfer, Simon Koblar, Daniel Labovitz, Lenore J. Launer, Cathy C. Laurie, Cecelia A. Laurie, Cue Hyunkyu Lee, Jin-Moo Lee†, Manuel Lehm, Robin Lemmens, Christopher Levi†, Didier Leys, Arne Lindgren†‡, W. T. Longstreth Jr., Jane Maguire, Ani Manichaikul, Hugh S. Markus†, Leslie A. McClure, Caitrin W. McDonough, Christa Meisinger, Olle Melander†, James F. Meschia†‡, Marina Mola-Caminal, Joan Montaner, Thomas H. Mosley, Martina Müller-Nurasyid, Mike A. Nalls, Jeffrey R. O'Connell, Martin O'Donnell, Ángel Ois, George J. Papanicolaou, Guillaume Paré, Leema Reddy Peddareddygari, Annie Pedersén, Joanna Pera, Annette Peters, Deborah Poole, Bruce M. Psaty, Raquel Rabionet, Miriam R. Raffeld, Kristiina Rannikmäe, Asif Rasheed, Petra Redfors, Alex P. Reiner, Kathryn Rexrode†, Marta Ribasés, Stephen S. Rich†, Wim Robberecht, Ana Rodriguez-Campello, Arndt Rolfs, Jaume Roquer, Lynda M. Rose, Daniel Rosenbaum, Natalia S. Rost, Peter M. Rothwell†, Tatjana Rundek†, Kathleen A. Ryan, Ralph L. Sacco†, Michèle M. Sale, Danish Saleheen, Veikko Salomaa, Cristina Sánchez-Mora, Carsten Oliver Schmidt, Helena Schmidt, Reinhold Schmidt†, Markus Schürks, Rodney Scott, Helen C. Segal, Stephan Seiler, Sudha Seshadri, Pankaj Sharma†, Alan R. Shuldiner, Brian Silver, Agnieszka Slowik†, Jennifer A. Smith, Martin Söderholm, Carolina Soriano, Mary J. Sparks, Tara Stanne, Kari Stefansson, O. Colin Stine, Konstantin Strauch, Jonathan Sturm, Cathie LM Sudlow†‡, Salman M. Tajuddin, Robert L. Talbert, Turgut Tatlisumak, Vincent Thijs†‡, Gudmar Thorleifsson, Unnur Thorsteindottir, Steffen Tiedt, Matthew Traylor, Stella Trompet, Valerie Valant, Melanie Waldenberger, Matthew Walters, Liyong Wang, Xin-Qun Wang, Sylvia Wassertheil-Smoller†, David R. Weir, Kerri L. Wiggins, Stephen R. Williams, Dorota Wloch-Kopec, Daniel Woo†‡, Rebecca Woodfield, Ona Wu, Huichun Xu, Alan B. Zonderman, Cervical Artery Dissection and Ischemic Stroke Patients (CADISP) study, Cohorts of Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium, Consortium of Minority Population genome-wide Association Studies of Stroke (COMPASS), METASTROKE consortium, Bradford B. Worrall*†‡, Paul I.W. de Bakker*†‡, Steven J. Kittner*†‡, Braxton D. Mitchell*†‡, Jonathan Rosand*†‡

†: principal investigators
‡: writing group
*: equal contributions

V. Author Affiliations

Sara L. Pulit, BA
Department of Medical Genetics, Institute for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands

Patrick F. McArdle, PhD
Department of Medicine and Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD

Quenna Wong, MS
Department of Biostatistics, University of Washington, Seattle, WA, USA

Rainer Malik, PhD
Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians University, Munich, Germany

Katrina Gwinn, MD
National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

Sefanja Achterberg, MD
Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

Ale Algra, MD
Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Philippe Amouyel, MD, PhD
INSERM U-1167, Lille, France
Institut Pasteur de Lille, Lille, France
Université Lille Nord de France, Lille, France
Lille University Hospital, Lille, France

Christopher D. Anderson, MD, MMSc
Department of Neurology and Center for Human Genetic Research
Massachusetts General Hospital, Boston, MA, USA
Program in Medical and Population Genetics
Broad Institute of MIT and Harvard, Cambridge, MA, USA
Harvard Medical School, Boston, MA, USA

Donna K. Arnett, PhD, MSPH
Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA

Ethem Murat Arsava, MD
AA Martinos Center for Biomedical Imaging, Department of Radiology,
Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

John Attia, MD, PhD, FRACP, FRCPC
School of Medicine and Public Health, University of Newcastle, NSW, Australia
Division of General Medicine, John Hunter Hospital, Newcastle, NSW, Australia

Hakan Ay, MD
AA Martinos Center for Biomedical Imaging, Department of Radiology,
Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
Stroke Service, Department of Neurology, Massachusetts General Hospital, Harvard
Medical School, Boston, MA, USA

Traci M. Bartz, MS
Department of Biostatistics, University of Washington, Seattle, WA, USA

Thomas Battey, BA
Department of Neurology and Center for Human Genetic Research, Massachusetts
General Hospital, Boston, MA, USA

Oscar R. Benavente, MD, FRCP
Department of Neurology, University of British Columbia, Vancouver, British
Columbia, Canada

Steve Bevan, PhD
Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

Alessandro Biffi, MD
Department of Neurology and Center for Human Genetic Research, Massachusetts
General Hospital, Boston, MA, USA

Joshua C. Bis, PhD
Cardiovascular Health Research Unit, Department of Medicine, University of
Washington, Seattle, WA, USA

Susan H. Blanton, PhD
John P. Hussman Institute for Human Genomics, Miller School of Medicine,
University of Miami, Miami, FL, USA

Giorgio B. Boncoraglio MD
Department of Cerebrovascular Diseases,
Fondazione IRCCS Istituto Neurologico Carlo Besta, Milano, Italy

Robert D. Brown, Jr., MD, MPH
Department of Neurology, Mayo Clinic, Rochester, MN, USA

Annette I. Burgess DPhil
Stroke Prevention Research Unit, Nuffield Department of Clinical Neurosciences,
University of Oxford, John Radcliffe Hospital, Oxford, UK

Caty Carrera MD MSc
Neurovascular Research Laboratory. Vall d'Hebron Institute of Research,
Vall d'Hebron Hospital, Universitat Autonoma Barcelona, Barcelona, Spain

Sherita N. Chapman Smith, MD
Department of Neurology, Virginia Commonwealth University, Richmond, VA, USA

Daniel I. Chasman, PhD
Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA
Harvard Medical School, Boston, MA, USA

Ganesh Chauhan, PhD
INSERM U897 Neuroepidemiology, Bordeaux, France
University of Bordeaux, Bordeaux, France

Wei-Min Chen, PhD
Center for Public Health Genomics; Department of Public Health Sciences; and
Department of Biochemistry and Molecular Genetics, University of Virginia,
Charlottesville, VA, USA

Yu-Ching Cheng, PhD
Department of Medicine, University of Maryland School of Medicine, Baltimore,
MD, USA

Michael Chong, MSc
Population Health Research Institute, McMaster University, DBCVS Research
Institute, Hamilton, Ontario, Canada

Lisa K. Cloonan, BA
Stroke Division, Department of Neurology, Massachusetts General Hospital, Harvard
Medical School, Boston, MA, USA

John W. Cole, MD, MS
Department of Neurology, University of Maryland School of Medicine and Veterans
Affairs Maryland Health Care System, Baltimore, MD, USA

Ioana Cotlarciuc, MSc, PhD
Institute of Cardiovascular Research, Royal Holloway University of London
(ICR2UL), Egham, UK

Carlos Cruchaga, PhD
Department of Psychiatry, Washington University School of Medicine, St. Louis,
MO, USA

Elisa Cuadrado-Godia, MD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques),
Universitat Autonoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Barcelona,
Spain

Tushar Dave, MS
Department of Medicine and Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

Jesse Dawson, MD, MBChB (hons), FRCP, BSc (hons)
Institute of Cardiovascular and Medical Sciences, College of Medical, Veterinary & Life Sciences, University of Glasgow, UK

Stéphanie Debette, MD, PhD
INSERM U897 Neuroepidemiology, Bordeaux, France
University of Bordeaux, Bordeaux, France
Department of Neurology, Bordeaux University Hospital, Bordeaux, France Boston University School of Medicine, Framingham Heart Study, Boston, MA, USA

Hossein Delavaran, MD
Department of Clinical Sciences Lund, Neurology, Lund University, Lund, Sweden
Department of Neurology and Rehabilitation Medicine, Neurology, Skåne University Hospital, Lund, Sweden

Cameron A. Dell, BS
Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA

Martin Dichgans, MD
Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians University, Munich, Germany
Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

Kimberly F. Doheny, PhD
Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD

Chuanhui Dong, PhD
Department of Neurology, Miller School of Medicine, University of Miami, Miami, FL, USA

David J. Duggan, PhD
Genetic Basis of Human Disease Division, Translational Genomics Research Institute (TGen), Phoenix, AZ, USA

Gunnar Engström, PhD
Cardio-vascular Epidemiology, Department of Clinical Sciences Malmö, Lund University, Skåne University Hospital Malmö, Sweden

Michele K. Evans, MD
Health Disparities Unit, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA

Xavier Estivill Pallejà, MD, PhD
Genomics and Disease group, Center for Genomic Regulation, Barcelona, Spain
Universitat Pompeu Fabra (UPF), Barcelona, Spain
Centro de Investigación Biomédica en Red (CIBERESP), Spain
IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

Jessica D. Faul, PhD, MPH
Survey Research Center, University of Michigan, Ann Arbor, MI, USA

Israel Fernández-Cadenas, PhD
Stroke Pharmacogenomics and Genetics, Fundació Docència i Recerca
MutuaTerrassa, Mutua de Terrassa Hospital,
Terrassa (Barcelona), Spain
Neurovascular Research Laboratory. Vall d'Hebron Institute of Research, Vall
d'Hebron Hospital, Universitat Autonoma Barcelona,
Barcelona, Spain

Myriam Fornage, PhD
Institute of Molecular Medicine, University of Texas Health Science Center at
Houston, Houston, TX, USA

Philippe M. Frossard, PhD, DSc
Center for Non-Communicable Diseases, Karachi, Pakistan
Nazarbayev University, Astana, Kazakhstan

Karen Furie, MD, MPH
Department of Neurology, Warren Alpert Medical School of Brown University,
Providence, RI, USA

Dale M. Gamble, MHSc, CCRP
Department of Neurology, Mayo Clinic, Jacksonville, FL, USA

Christian Gieger, PhD
Research unit of Molecular Epidemiology, Helmholtz Zentrum München - German
Research Center for Environmental Health, Neuherberg, Germany
Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center
for Environmental Health, Neuherberg, Germany.

Anne-Katrin Giese, MD
Albrecht-Kossel-Institute for Neuroregeneration, Medical University of Rostock,
Rostock, Germany
Eva Giralt-Steinhauer, MD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques),
Barcelona, Spain

Hector M. González, PhD
Department of Epidemiology & Biostatistics, Michigan State University, East
Lansing, MI, USA

An Goris, PhD
Department of Neurosciences, Laboratory for Neuroimmunology, KU Leuven-University of Leuven, Leuven, Belgium

Solveig Gretarsdottir, PhD
deCODE Genetics/Amgen, Reykjavik, Iceland

Raji P. Grewal, MD
Neuroscience Institute, Saint Francis Medical Center, School of Health and Medical Sciences, Seton Hall University, South Orange, New Jersey, USA

Ulrike Grittner, PhD
Department for Biostatistics and Clinical Epidemiology. Charité-University Medical Centre, Berlin, Germany

Stefan Gustafsson, PhD, MSc
Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.

Buhm Han, PhD
Asan Institute for Life Sciences, Asan Medical Center, Seoul 138-736, Republic of Korea
Department of Medicine, University of Ulsan College of Medicine, Seoul 138-736, Republic of Korea

Graeme J. Hankey, MD, FRACP, FRCP, FAHA
School of Medicine and Pharmacology, The University of Western Australia, Perth, Australia

Laura Heitsch, MD
Division of Emergency Medicine, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO, USA

Peter Higgins, MD
Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK

Marc C. Hochberg, MD
Division of Rheumatology and Clinical Immunology, Department of Medicine and Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD

Elizabeth Holliday, BSc(Hons), MSc, PhD
Public Health Research Program, Hunter Medical Research Institute, Newcastle, NSW, Australia

Jemma C. Hopewell, PhD
Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, UK

Richard B. Horenstein, MD, JD
Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD, USA
Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

George Howard, DrPH
Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA

M. Arfan Ikram, MD, PhD
Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands

Andreea Ilinca MD
Department of Clinical Sciences Lund, Neurology, Lund University, Lund, Sweden
Department of Internal Medicine, Neurology, Landskrona Hospital, Sweden

Erik Ingelsson, MD, PhD, FAHA
Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.
Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom.

Marguerite R. Irvin, PhD
Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA

Rebecca D. Jackson, MD
Division of Endocrinology, Diabetes and Metabolism, Department of Internal Medicine and the Center for Clinical and Translational Science, The Ohio State University, Columbus, OH.

Christina Jern, MD, PhD
Institute of Biomedicine, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Jordi Jiménez Conde, MD, PhD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques), Universitat Autonoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Barcelona, Spain

Julie A. Johnson, PharmD
Department of Pharmacotherapy and Translational Research and Center for Pharmacogenomics, College of Pharmacy, University of Florida, Gainesville FL, USA
Division of Cardiovascular Medicine, College of Medicine, University of Florida, Gainesville, FL, USA

Katarina Jood, MD, PhD
Institute of Neuroscience, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Muhammed S. Kahn, MSc
Institute of Cardiovascular Research, Royal Holloway University of London (ICR2UL), Egham, UK
St Peter's and Ashford Hospitals, UK

Robert Kaplan, PhD
Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

L. Jaap Kappelle, MD
Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

Sharon LR Kardia, PhD
Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

Keith L. Keene PhD
Department of Biology; Center for Health Disparities, East Carolina University, Greenville, NC, USA

Brett M. Kissela, MD, MS
University of Cincinnati College of Medicine, Cincinnati, OH, USA

Dawn O. Kleindorfer, MD, MS
University of Cincinnati College of Medicine, Cincinnati, OH, USA

Simon Koblar, BMBS, FRACP, PhD
School of Medicine, The Queen Elizabeth Hospital campus, Woodville South SA, Australia

Daniel Labovitz, MD, MS
Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

Lenore J. Launer, PhD
National Institute on Aging, National Institutes of Health, Bethesda, MD, USA

Cathy C. Laurie, PhD
Department of Biostatistics, University of Washington, Seattle, WA, USA

Cecelia A. Laurie, PhD
Department of Biostatistics, University of Washington, Seattle, WA, USA

Cue Hyunkyu Lee, MS
Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

Jin-Moo Lee, MD, PhD
Stroke Center, Department of Neurology, Washington University School of Medicine, St. Louis, MO, USA

Manuel Lehm
Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians University, Munich, Germany

Robin Lemmens, MD, PhD
KU Leuven - University of Leuven, Department of Neurosciences, Experimental Neurology and Leuven Research Institute for Neuroscience and Disease (LIND), Leuven, Belgium
VIB, Vesalius Research Center, Laboratory of Neurobiology, B-3000 Leuven, Belgium
University Hospitals Leuven, Department of Neurology, Leuven, Belgium

Christopher Levi, MBBS, BMed Sci, FRACP
John Hunter Hospital, Hunter Medical Research Institute and University of Newcastle, NSW, Australia

Didier Leys, MD, PhD
Université Lille Nord de France, Lille, France
Lille University Hospital, Lille, France
Department of Neurology, Equipe d'accueil 1046, Lille, France
INSERM U897, University of Bordeaux, Bordeaux, France

Arne Lindgren MD PhD
Department of Clinical Sciences Lund, Neurology, Lund University, Lund, Sweden
Department of Neurology and Rehabilitation Medicine, Neurology, Skåne University Hospital, Lund, Sweden

W. T. Longstreth Jr., MD
Department of Neurology, University of Washington, Seattle, WA
Department of Epidemiology, University of Washington, Seattle, WA

Jane Maguire, PhD, BNurs(Hons), BA, RN
School of Nursing and Midwifery, University of Newcastle, Callaghan, NSW, Australia

Ani Manichaikul, PhD
Center for Public Health Genomics, Biostatistics Section, Department for Public Health Sciences, University of Virginia, Charlottesville, VA, USA

Hugh S. Markus, DM
Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

Leslie A. McClure, PhD
Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA

Caitrin W. McDonough, PhD
Department of Pharmacotherapy and Translational Research and Center for Pharmacogenomics, College of Pharmacy, University of Florida, Gainesville FL, USA

Christa Meisinger, MD
Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

Olle Melander, MD, PhD
Lund University, Department of Clinical Sciences, Malmö University Hospital, Malmö, Sweden

James F. Meschia, MD
Department of Neurology, Mayo Clinic, Jacksonville, FL, USA

Marina Mola-Caminal, BSc
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques), Barcelona, Spain

Joan Montaner, MD, PhD
Neurovascular Research Laboratory, and Neurology Department
Vall d'Hebron Institute of Research (VHIR), Vall d'Hebron University Hospital, Autonomous University of Barcelona, Barcelona, Spain

Thomas H. Mosley, PhD
Department of Medicine (Geriatrics) and Neurology, University of Mississippi Medical Center, Jackson, MS, USA

Martina Müller-Nurasyid, PhD
Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
Department of Medicine I, Ludwig-Maximilians-University Munich, Munich, Germany
DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany

Mike A. Nalls, PhD
Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA

Jeffrey R. O'Connell, DPhil
Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD, USA

Martin O'Donnell, MD, PhD, MRCPI, BCh,BAO
DBCVS Research Institute, Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada

Ángel Ois, MD. PhD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques), Universitat Autonoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Barcelona, Spain

George J. Papanicolaou, PhD
Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, Bethesda, MD, USA

Guillaume Paré, MD, MSc, FRCPC
DBCVS Research Institute, Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada

Leema Reddy Peddareddygari, MBBS, MD
Neuroscience Institute, Saint Francis Medical Center, School of Health and Medical Sciences, Seton Hall University, South Orange, New Jersey, USA

Annie Pedersén, MD
Institute of Biomedicine, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Joanna Pera, MD, PhD
Department of Neurology, Jagiellonian University Medical College, Krakow, Poland

Annette Peters, PhD
Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany

Deborah Poole, HNC
Stroke Prevention Research Unit, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK

Bruce M. Psaty, MD, PhD
Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA
Department of Epidemiology, University of Washington, Seattle, WA, USA
Department of Health Services, University of Washington, Seattle, WA, USA
Group Health Research Institute, Group Health, Seattle, WA, USA

Raquel Rabionet, PhD
Genomics and Disease group, Center for Genomic Regulation, Barcelona, Spain
Universitat Pompeu Fabra (UPF), Barcelona, Spain
Centro de Investigación Biomédica en Red (CIBERESP), Spain
IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

Miriam R. Raffeld, BA
Department of Neurology and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA

Kristiina Rannikmäe, MD
Centre for Clinical Brain Sciences, University of Edinburgh, UK

Asif Rasheed, MBBS
Center for Non-Communicable Diseases, Karachi, Pakistan

Petra Redfors, MD, PhD
Institute of Neuroscience, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Alex P. Reiner, MD, MSc
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Kathryn Rexrode, MD, MPH
Brigham and Women's Hospital, Boston, MA, USA

Marta Ribasés, PhD, BSc
Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addictions, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain
Department of Psychiatry, Hospital Universitari Vall d'Hebron, Barcelona, Spain
Biomedical Network Research Centre on Mental Health (CIBERSAM), Barcelona, Spain

Stephen S. Rich, PhD
Center for Public Health Genomics, University of Virginia, Charlottesville, VA USA

Wim Robberecht, MD, PhD
KU Leuven - University of Leuven, Department of Neurosciences, Experimental Neurology and Leuven Research Institute for Neuroscience and Disease (LIND), Leuven, Belgium
VIB, Vesalius Research Center, Laboratory of Neurobiology, Leuven, Belgium
University Hospitals Leuven, Department of Neurology, Leuven, Belgium

Ana Rodríguez-Campello, MD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques), Universitat Autonoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Barcelona, Spain

Arndt Rolfs, MD
Albrecht-Kossel-Institute for Neuroregenation Medical Faculty, University of Rostock, Germany

Jaume Roquer, MD, PhD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques),
Universitat Autonoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Barcelona,
Spain

Lynda M. Rose, MS
Brigham and Women's Hospital, Boston, MA, USA

Daniel Rosenbaum, MD
State University of New York, Downstate, Brooklyn, NY, USA

Natalia S. Rost, MD, MPH
Stroke Division, Department of Neurology, Massachusetts General Hospital, Harvard
Medical School, Boston, MA 02114

Peter M. Rothwell, FMedSci
Stroke Prevention Research Unit, Nuffield Department of Clinical Neurosciences,
University of Oxford, John Radcliffe Hospital, Oxford, UK

Tatjana Rundek, MD, PhD, FANA
Department of Neurology, Miller School of Medicine, University of Miami, Miami,
FL, USA

Kathleen A. Ryan, MPH
Department of Medicine, University of Maryland School of Medicine, Baltimore,
MD, USA

Ralph L. Sacco, MD, MS, FAHA, FAAN, FANA
Department of Neurology, Miller School of Medicine, University of Miami, Miami,
FL, USA

Michèle M. Sale, PhD
Center for Public Health Genomics; Department of Public Health Sciences; and
Department of Biochemistry and Molecular Genetics, University of Virginia, VA,
USA

Danish Saleheen, MBBS, PhD
Department of Biostatistics and Epidemiology, University of Pennsylvania,
Philadelphia, PA, USA
Center for Non-Communicable Diseases, Karachi, Pakistan

Veikko Salomaa, MD, PhD
National Institute for Health and Welfare, Helsinki, Finland.

Cristina Sánchez-Mora, PhD, BSc
Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addictions, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain
Department of Psychiatry, Hospital Universitari Vall d'Hebron, Barcelona, Spain
Biomedical Network Research Centre on Mental Health (CIBERSAM), Barcelona, Spain

Carsten Oliver Schmidt, PD Dr.
Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

Helena Schmidt, MD, PhD
Institute of Molecular Biology and Biochemistry, Graz, Austria

Reinhold Schmidt, MD
Department of Neurology, Clinical Division of Neurogeriatrics, Medical University Graz, Graz, Austria

Markus Schürks, MD, MSc
Department of Neurology, University Hospital Essen, Essen, Germany

Rodney Scott, BSc(Hons), PhD, FRCPath, FHGSA, FFSc(RCPA)
School of Biomedical Sciences and Pharmacy, University of Newcastle, NSW, Australia

Helen C. Segal, PhD
Stroke Prevention Research Unit, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK

Stephan Seiler, MD
Department of Neurology, Clinical Division of Neurogeriatrics, Medical University Graz, Austria

Sudha Seshadri, MD
Department of Neurology, Boston University School of Medicine, Boston, MA, USA

Pankaj Sharma, MD, PhD, FRCP
Institute of Cardiovascular Research, Royal Holloway University of London (ICR2UL), Egham, UK
St Peter's and Ashford Hospitals, UK

Alan R. Shuldiner, MD
Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD, USA
Geriatric Research and Education Clinical Center, Veterans Administration Medical Center, Baltimore, MD, USA
Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

Brian Silver, MD
Department of Neurology, Alpert Medical School of Brown University, Providence, RI, USA

Agnieszka Slowik, MD, PhD
Department of Neurology,
Jagiellonian University Medical College, Krakow, Poland

Jennifer A. Smith, PhD, MPH
Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

Martin Söderholm, MD
Lund University, Department of Clinical Sciences, Malmö University Hospital, Malmö, Sweden

Carolina Soriano, Bsc, PhD
Department of Neurology, Neurovascular Research Group (NEUVAS)
IMIM-Hospital del Mar (Institut Hospital del Mar d'Investigacions Mèdiques), Barcelona, Spain

Mary J. Sparks, RN, BSN
Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA

Tara Stanne, PhD
Institute of Biomedicine, the Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Kari Stefansson, MD, PhD
deCODE Genetics/Amgen, Reykjavik, Iceland
Faculty of Medicine, University of Iceland, Reykjavik, Iceland

O. Colin Stine, PhD
Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD

Konstantin Strauch, PhD
Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany

Jonathan Sturm, MBChB, PhD, FRACP
Department of Neurosciences, Gosford Hospital, NSW, Australia

Cathie LM Sudlow, DPhil, FRCP(E)
Centre for Clinical Brain Sciences & Institute of Genomic and Molecular Medicine, University of Edinburgh, UK

Salman M. Tajuddin, MD, PhD, MPH
Laboratory of Epidemiology and Population Science, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA

Robert L. Talbert, PharmD
College of Pharmacy, University of Texas at Austin, Austin, Texas, USA

Turgut Tatlisumak, MD, PhD
Institute of Neuroscience and Physiology, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden
Department of Neurology, Sahlgrenska University Hospital, Gothenburg, Sweden
Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland

Vincent Thijs, MD, PhD
KU Leuven - University of Leuven, Department of Neurosciences, Experimental Neurology and Leuven Research Institute for Neuroscience and Disease (LIND), Leuven, Belgium
VIB, Vesalius Research Center, Laboratory of Neurobiology, Leuven, Belgium
University Hospitals Leuven, Department of Neurology, Leuven, Belgium
Gudmar Thorleifsson, PhD
deCODE Genetics/Amgen, Reykjavik, Iceland

Unnur Thorsteindottir, PhD
deCODE Genetics/Amgen, Reykjavik, Iceland
Faculty of Medicine, University of Iceland, Reykjavik, Iceland

Steffen Tiedt
Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians University, Munich, Germany

Matthew Traylor, PhD
Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

Stella Trompet, PhD
Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands
Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands

Valerie Valant, BA
University of Massachusetts Medical School, Worcester, MA, USA

Melanie Waldenberger, PhD
Research unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany

Matthew Walters, MD
Institute of Cardiovascular and Medical Sciences, University of Glasgow, UK

Liyong Wang, PhD
John P. Hussman Institute for Human Genomics, Miller School of Medicine,
University of Miami, Miami, FL, USA

Xin-Qun Wang, MS
Department of Public Health Sciences, University of Virginia, Charlottesville, VA

Sylvia Wassertheil-Smoller, PhD
Department of Epidemiology and Population Health, Albert Einstein College of
Medicine, Bronx, NY, USA

David R. Weir, PhD
Survey Research Center, University of Michigan, Ann Arbor, MI, USA

Kerri L. Wiggins, MS, RD
Cardiovascular Health Research Unit
Department of Medicine, University of Washington, Seattle, WA, USA

Stephen R. Williams PhD
Center for Public Health Genomics, Univeristy of Virginia, Charlottesville, VA, USA

Dorota Wloch-Kopec MD, PhD
Department of Neurology, Jagiellonian University Medical College, Krakow, Poland

Daniel Woo, MD, MS
University of Cincinnati College of Medicine, Cincinnati, OH, USA

Rebecca Woodfield, MBBChir, MRCP(UK)
Centre for Clinical Brain Sciences, University of Edinburgh, UK

Ona Wu, PhD
Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology,
Massachusetts General Hospital, Charlestown, MA, USA
Department of Radiology, Harvard Medical School, Boston, MA, USA

Huichun Xu, MD, PhD
Department of Medicine, University of Maryland School of Medicine, Baltimore,
MD, USA

Alan B. Zonderman, PhD
Laboratory of Personality and Cognition, National Institute on Aging, National
Institutes of Health, Baltimore, MD, USA

Bradford B. Worrall, MD, MSc
Departments of Neurology and Public Health Sciences, University of Virginia,
Charlottesville, VA, USA

Paul I.W. de Bakker, PhD
Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands

Steven J. Kittner, MD, MPH
Department of Neurology, University of Maryland School of Medicine and Veterans Affairs Maryland Health Care System, Baltimore, MD, USA

Braxton D. Mitchell, PhD, MPH
Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD, USA
Geriatric Research and Education Clinical Center, Veterans Administration Medical Center, Baltimore, MD, USA

Jonathan Rosand, MD, MSc
Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge MA, USA
Department of Neurology and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA
Department of Neurology, Harvard Medical School, Boston, MA, USA

VI. NINDS-SiGN Committee

Administrative:
Steven J. Kittner (chair)
Cameron A. Dell
Dale M. Gamble
Mary J. Sparks

Steering/PIs of Discovery Studies:
Donna K. Arnett
Oscar Benavente
John W. Cole
Martin Dichgans
Raji P. Grewal
Christina Jern
Jordi Jiménez Conde
Julie A. Johnson
Jin-Moo Lee
Christopher Levi
Arne Lindgren
Hugh S. Markus
Olle Melander
James F. Meschia
Kathryn Rexrode
Jonathan Rosand
Peter M. Rothwell
Tatjana Rundek
Ralph L. Sacco
Reinhold Schmidt
Pankaj Sharma
Agnieszka Slowik
Cathie LM Sudlow
Vincent Thijs
Sylvia Wasssertheil-Smoller
Daniel Woo
Bradford B. Worrall

PIs of Replication Studies:
Giorgio B. Boncoraglio
Daniel I. Chasman
Stéphanie Debette
Israel Fernández-Cadenas
Solveig Gretarsdottir
Peter Higgins
Jemma Hopewell
M. Arfan Ikram
Lenore J. Launer
Thomas H. Mosley
Guillaume Paré
Bruce M. Psaty
Alex P. Reiner
Arndt Rolfs
Michèle M. Sale
Danish Saleheen
Veikko Salomaa
Sudha Seshadri

PIs of Control Studies:
Rebecca D. Jackson
Martina Müller-Nurasyid
Mike A. Nalls
Marta Ribasés
David R. Weir

Data Management:
Patrick F. McArdle (chair)
Tushar Dave

Analysis:
Braxton D. Mitchell (chair)
Yu-Ching Cheng
Paul I.W. de Bakker
Myriam Fornage
Cathy C. Laurie
Ani Manichaikul
Jeffrey R. O'Connell
Sara L. Pulit
Stephen S. Rich
Quenna Wong
Huichun Xu

Phenotype:
James F. Meschia (co-chair)
Bradford B. Worrall (co-chair)
Hakan Ay
Robert D. Brown Jr.

Imaging:
Jonathan Rosand (chair)
Natalia S. Rost
Ona Wu

Publication:
Kathryn Rexrode (chair)
Tatjana Rundek (prior chair)
Agnieszka Slowik (prior co-chair)

Hakan Ay
Oscar R. Benavente
Steve Bevan
Katrina Gwinn
Steven J. Kittner
Jin-Moo Lee
Patrick F. McArdle
James F. Meschia
Braxton D. Mitchell
Jonathan Rosand
Sylvia Wasssertheil-Smoller
Daniel Woo
Bradford B. Worrall

Writing:
Jonathan Rosand (chair)
Braxton D. Mitchell (co-chair)
Hakan Ay

Paul I.W. de Bakker
Katrina Gwinn
Steven J. Kittner
Arne Lindgren
James F. Meschia
Sara L. Pulit
Cathie LM Sudlow
Vincent Thijs
Daniel Woo
Bradford B. Worrall

CIDR:
Kimberly F. Doheny

NINDS staff:
Roderick Corriveau
Katrina Gwinn

VII. Funding and Acknowledgements

The Stroke Genetics Network (SiGN) study was funded by a cooperative agreement grant from the National Institute of Neurological Disorders and Stroke (NINDS) U01 NS069208. Genotyping services were provided by the Johns Hopkins University Center for Inherited Disease Research (CIDR), which is fully funded through a federal contract from the National Institutes of Health (NIH) to the Johns Hopkins University (contract No.HHSN268200782096C). The Biostatistics Department Genetics Coordinating Center at the University of Washington (Seattle) provided more extensive quality control of the genotype data through a subcontract with CIDR. Additional support to the Administrative Core of SiGN was provided by the Dean's Office, University of Maryland School of Medicine.

Discovery Case-only & Case and Control Cohorts:

ASGC: Australian population control data were derived from the Hunter Community Study. We also thank the University of Newcastle for funding and the men and women of the Hunter region who participated in this study. This research was funded by grants from the Australian National and Medical Health Research Council (NHMRC Project Grant ID: 569257), the Australian National Heart Foundation (NHF Project Grant ID: G 04S 1623), the University of Newcastle, the Gladys M Brawn Fellowship scheme, and the Vincent Fairfax Family Foundation in Australia. Elizabeth G Holliday is supported by the Australian NHMRC Fellowship scheme.

BASICMAR: The Base de Datos de Ictus del Hospital del Mar (BASICMAR) Genetic Study was supported by the Ministerio de Sanidad y Consumo de España, Instituto de Salud Carlos III (ISC III) with the grants: Registro BASICMAR Funding for Research in Health (PI051737); GWA Study of Leukoaraiosis (GWALA) project from Fondos de Investigación Sanitaria ISC III (PI10/02064) and (PI12/01238); Agència de Gestió Ajuts Universitaris de Recerca (2014 SGR 1213) and Fondos European Regional Development Funding (FEDER/EDRF) Red de Investigación Cardiovascular (RD12/0042/0020). Additional support was provided by the Fundació la Marató TV3 with the grant GODS project. Genestroke Consortium (76/C/2011) Recercaixa'13 (JJ086116). Assistance with data cleaning was provided by the Research in Cardiovascular and Inflammatory Diseases Program of Institute Hospital del Mar of Medical Investigations, Hospital del Mar, and the Barcelona Biomedical Research Park.

BRAINS: The Bio-Repository of DNA in Stroke (BRAINS) was supported by the British Council (UKIERI), Henry Smith Charity, the UK Stroke Research Network and the Qatari National Research Fund. Prof Sharma was supported by a Department of Health (United Kingdom) Senior Fellowship.

EDINBURGH: The Edinburgh Stroke Study was supported by the Wellcome Trust and the Binks Trust. Sample processing occurred in the Genetics Core Laboratory of the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh, UK. Much of the neuroimaging occurred in the Scottish Funding Council Brain Imaging Research Centre (www.sbirc.ed.ac.uk), University of Edinburgh, a core area of the Wellcome Trust Clinical Research Facility and part of the Scottish Imaging

Network–A Platform for Scientific Excellence (SINAPSE) collaboration (www.sinapse.ac.uk), funded by the Scottish Funding Council and the Chief Scientist Office. Genotyping was performed at the Wellcome Trust Sanger Institute in the United Kingdom and funded by the Wellcome Trust as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z and WT084724MA).

GCNKSS: The Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS) was supported by the NIH (NS030678).

GEOS: The Genetics of Early Onset Stroke (GEOS) Study was supported by the NIH Genes, Environment, and Health Initiative (GEI) grant U01 HG004436, as part of the Gene Environment Association Studies (GENEVA) consortium under GEI, with additional support provided by the Mid-Atlantic Nutrition and Obesity Research Center (P30 DK072488) and the Office of Research and Development, Medical Research Service, and the Baltimore Geriatrics Research, Education, and Clinical Center of the Department of Veterans Affairs. Genotyping services were provided by the Johns Hopkins University CIDR, which is fully funded through a federal contract from the NIH to the Johns Hopkins University (contract No. HHSN268200782096C). Assistance with data cleaning was provided by the GENEVA Coordinating Center (U01 HG 004446; PI Bruce S Weir). Study recruitment and assembly of data sets were supported by a Cooperative Agreement with the Division of Adult and Community Health, Centers for Disease Control and by grants from the NINDS and the NIH Office of Research on Women's Health (R01 NS45012, U01 NS069208-01).

GRAZ: The Austrian Stroke Prevention Study was supported by the Austrian Science Fund (FWF) grant Nos. P20545-P05 and P13180 and I904-B13 (Era-Net). The Medical University of Graz supports the databases of the Graz Stroke Study and the Austrian Stroke Prevention Study.

ISGS and SWISS: The Ischemic Stroke Genetics Study (ISGS) was supported by the NINDS (R01 NS42733; PI Dr Meschia). The Sibling with Ischemic Stroke Study (SWISS) was supported by the NINDS (R01 NS39987; PI Dr Meschia). Both SWISS and ISGS received additional support, in part, from the Intramural Research Program of the National Institute on Aging (Z01 AG000954-06; PI Andrew Singleton). SWISS and ISGS used samples and clinical data from the NIH-NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds), human subject protocol Nos. 2003-081 and 2004-147. SWISS and ISGS used stroke-free participants from the Baltimore Longitudinal Study of Aging (BLSA) as controls with the permission of Dr Luigi Ferrucci. The inclusion of BLSA samples was supported, in part, by the Intramural Research Program of the National Institute on Aging (Z01 AG000015-50), human subject protocol No. 2003-078. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the NIH (http://biowulf.nih.gov). For SWISS and ISGS cases of African ancestry, a subset of the Healthy Aging in Neighborhoods of Diversity across the Life Span study (HANDLS) were used as stroke-free controls. HANDLS is funded by the National Institute of Aging (1Z01AG000513; PI Michele K. Evans).

KRAKOW: Phenotypic data and genetic specimens collection were funded by the grant from the Polish Ministry of Science and Higher Education for Leading National

Research Centers (KNOW) and by the grants from the Jagiellonian University Medical College in Krakow, Poland: K/ZDS/002848, K/ZDS/003844.

LEUVEN: The Leuven Stroke genetics study was supported by personal research funds from the Department of Neurology of the University Hospitals Leuven. Dr Thijs is supported by a Fundamental Clinical Research grant from FWO Flanders (Nos. 1800908N and 1800913N). Dr Lemmens is a Senior Clinical Investigator of FWO Flanders (FWO 1841913N) and is supported through Fonds Annie Planckaert-Dewaele. An Goris is supported by the Research Fund KU Leuven (OT/11/087) and Research Foundation Flanders (G073415N).

LUND: The Lund Stroke Register was supported by the Swedish Research Council (K2010-61X-20378-04-3), The Swedish Heart-Lung Foundation, Region Skåne, Skåne University Hospital, the Freemasons Lodge of Instruction EOS in Lund, King Gustaf V's and Queen Victoria's Foundation, Lund University, and the Swedish Stroke Association. Biobank services were provided by Region Skåne Competence Centre (RSKC Malmö), Skåne University Hospital, Malmö, Sweden, and Biobank, Labmedicin Skåne, University and Regional Laboratories Region Skåne, Sweden.

MALMÖ: The Malmö Diet and Cancer Study was supported by the Swedish Research Council (Vetenskapsrådet), Heart and Lung Foundation (Hjärt och Lungfonden), and Swedish Stroke Foundation (Strokeförbundet).

MCISS: The Middlesex County Ischemic Stroke Study (MCISS) was supported by intramural funding from the New Jersey Neuroscience Institute/JFK Medical Center, Edison, NJ, and The Neurogenetics Foundation, Cranbury, NJ. We acknowledge Dr Souvik Sen for his advice and encouragement in the initiation and design of this study.

MIAMISR and NOMAS: The Northern Manhattan Study (NOMAS) was supported by grants from the NINDS (R37 NS029993, R01 NS27517). The Cerebrovascular Biorepository at University of Miami/Jackson Memorial Hospital (The Miami Stroke Registry, Institutional Review Board No. 20070386) was supported by the Department of Neurology at University of Miami Miller School of Medicine and Evelyn McKnight Brain Institute. Biorepository and DNA extraction services were provided by the Hussmann Institute for Human Genomics at the Miller School of Medicine.

MGH-GASROS: The Massachusetts General Hospital Stroke Genetics Group was supported by the NIH Genes Affecting Stroke Risks and Outcomes Study (GASROS) grant K23 NS042720, the American Heart Association/Bugher Foundation Centers for Stroke Prevention Research 0775010N, and NINDS K23NS042695, K23 NS064052, the Deane Institute for Integrative Research in Atrial Fibrillation and Stroke, and by the Keane Stroke Genetics Fund. Genotyping services were provided by the Broad Institute Center for Genotyping and Analysis, supported by grant U54 RR020278 from the National Center for Research Resources.

MUNICH: The MUNICH study was supported by the Vascular Dementia Research Foundation, the Dr. Werner Jackstaedt-Stiftung, the FP7 EU project CVgenes@target

(261123), the German Federal Ministry of Education and Research (BMBF) in the context of the e:Med program (e:AtheroSysMed), and by grants from the Deutsche Forschungsgemeinschaft (SFB1123 B3/C1, Munich Cluster for Systems Neurology).

NHS: The Nurses' Health Study work on stroke is supported by grants from the NIH, including HL088521 and HL34594 from the National Heart, Lung, and Blood Institute, as well as grants from the National Cancer Institute funding the questionnaire follow-up and blood collection: CA87969 and CA49449.

OXVASC: The Oxford Vascular Study was supported by the Wellcome Trust, Wolfson Foundation, Stroke Association, Medical Research Council, Dunhill Medical Trust, NIH Research (NIHR), and NIHR Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford. Dr Rothwell is in receipt of Senior Investigator Awards from the Wellcome Trust and the NIHR.

REGARDS: The Reasons for Geographic and Racial Differences in Stroke (REGARDS) Study was supported by a cooperative agreement U01 NS041588 from the NINDS, NIH, and Department of Health and Human Service. A full list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org.

SAHLSIS: The Sahlgrenska Academy Study of Ischemic Stroke was supported by the Swedish Research Council (K2014-64X-14605-12-5), the Swedish Heart and Lung Foundation (20130315), the Swedish state/Sahlgrenska University Hospital (ALFGBG-429981), the Swedish Stroke Association, the Swedish Society of Medicine, and the Rune and Ulla Amlöv Foundation.

SPS3: The Secondary Prevention of Small Subcortical Strokes trial was funded by the US National Institute of Health and Neurological Disorders and Stroke grant No. U01NS38529-04A1 (principal investigator, Oscar R. Benavente; coprincipal investigator, Robert G. Hart). The SPS3 Genetic Substudy (SPS3-GENES) was funded by R01 NS073346 (coprincipal investigators, Julie A. Johnson, Oscar R. Benavente, and Alan R. Shuldiner) and U01 GM074492-05S109 (principal investigator, Julie A. Johnson).

ST. GEORGE'S: The principal funding for this study was provided by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z and WT084724MA). Collection of some of the St George's stroke cohort was supported by project grant support from the Stroke Association. Hugh Markus is supported by an NIHR Investigator award. Matthew Traylor is supported by project grant funding from the Stroke Association.(TSA 2013/01)

VISP: The GWAS component of the VISP study was supported by the United States National Human Genome Research Institute (NHGRI), Grant U01 HG005160 (PI Michèle Sale & Bradford Worrall), as part of the Genomics and Randomized Trials Network (GARNET). Genotyping services were provided by the Johns Hopkins University Center for Inherited Disease Research (CIDR), which is fully funded through a federal contract from the NIH to the Johns Hopkins University. Assistance with data cleaning was provided by the GARNET Coordinating Center (U01

HG005157; PI Bruce S Weir). Study recruitment and collection of datasets for the VISP clinical trial were supported by an investigator-initiated research grant (R01 NS34447; PI James Toole) from the United States Public Health Service, NINDS, Bethesda, Maryland. Control data for comparison with European ancestry VISP stroke cases were obtained through the database of genotypes and phenotypes (dbGAP) High Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation (phs000187.v1.p1; R01CA100264, 3P50CA093459, 5P50CA097007, 5R01ES011740, 5R01CA133996, HHSN268200782096C; PIs Christopher Amos, Qingyi Wei, Jeffrey E. Lee). For VISP stroke cases of African ancestry, a subset of the Healthy Aging in Neighborhoods of Diversity across the Life Span study (HANDLS) were used as stroke free controls. HANDLS is funded by the National Institute of Aging (1Z01AG000513; PI Michele K. Evans).

WHI-OS: The Women's Health Initiatives (WHI) program was funded by the National Heart, Lung, and Blood Institute, NIH, US Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118 to 32119, 32122, 42107-26, 42129-32, and 44221. The Hormones and Biomarkers Predicting Stroke (HaBPS) was supported by a grant from the National Institutes of Neurological Disorders and Stroke (R01NS042618).

WUSTL: Washington University St. Louis Stroke Study (WUSTL): The collection, extraction of DNA from blood, and storage of specimens were supported by 2 NINDS NIH grants (P50 NS055977 and R01 NS8541901). Basic demographic and clinical characterization of stroke phenotype was prospectively collected in the Cognitive Rehabilitation and Recovery Group (CRRG) registry. The Recovery Genomics after Ischemic Stroke (ReGenesIS) study was supported by a grant from the Barnes-Jewish Hospital Foundation.

Control-only Cohorts:

ADHD: Financial support was received from "Fundació La Marató de TV3" (ref. 092330/31), Instituto de "Salud Carlos III-FIS", grants PI11/00571, PI11/01629, PI12/01139 and PI14/01700, cofinanced by the European Regional Development Fund (ERDF*),* "Agència de Gestió d'Ajuts Universitaris i de Recerca-AGAUR, Generalitat de Catalunya" (2014SGR1357) and "Departament de Salut", Government of Catalonia, Spain. Genotyping was performed at the Spanish National Genotyping Centre (CEGEN-Barcelona). Marta Ribasés is a recipient of a Miguel de Servet contract from the "Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación", Spain. Authors wish to thank all participants who kindly participated in this research.

Health ABC: The Health Aging and Body Composition Study was supported by NIA contracts N01AG62101, N01AG62103, and N01AG62106 and, in part, by the NIA Intramural Research Program. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C.

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov).

HCHS/SOL: The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), San Diego State University (N01-HC65237), and University of Washington (HHSN268201300005C). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Center on Minority Health and Health Disparities, the National Institute of Deafness and Other Communications Disorders, the National Institute of Dental and Craniofacial Research, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements.

The authors thank the staff and participants of HCHS/SOL for their important contributions. A complete list of staff and investigators has been provided by Sorlie P., et al. in Ann Epidemiol. 2010 Aug;20: 642-649 and is also available on the study website http://www.cscc.unc.edu/hchs/.

HRS: HRS is supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded as a separate award from the National Institute on Aging (RC2 AG036495). Genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington. HRS genotype data have been deposited in the NIH GWAS repository (dbGaP). Researchers wishing to use the HRS genetic data must first apply to dbGaP for access. The process to request access to any dbGaP study is done via the dbGaP authorized access system. Researchers who wish to obtain HRS phenotype measures that are not in dbGaP must submit a data access use agreement to HRS. For further information, contact hrsquestions@umich.edu. Relevant websites describing HRS genotype and phenotype data are: www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000428.v1.p1 and http://hrsonline.isr.umich.edu.

INMA: This study was funded by grants from Instituto de Salud Carlos III (CB06/02/0041, G03/176, FIS PI041436, PI081151, PI041705, PI061756, PI091958, and PS09/00432, FIS-FEDER 03/1615, 04/1509, 04/1112, 04/1931 , 05/1079, 05/1052, 06/1213, 07/0314, 09/02647, 11/01007, 11/02591, 11/02038, 13/1944, 13/2032 and CP11/0178), Spanish Ministry of Science and Innovation (SAF2008-00357), European Commission (ENGAGE project and grant agreement HEALTH-F4-2007-201413, HEALTH.2010.2.4.5-1, FP7-ENV-2011 cod 282957), Fundació La Marató de TV3, Generalitat de Catalunya-CIRIT 1999SGR 00241 and Conselleria de Sanitat Generalitat Valenciana. Part of the DNA extractions and genotyping was performed at the Spanish National Genotyping Centre (CEGEN-Barcelona). The authors are grateful to Silvia Fochs, Anna Sànchez, Maribel López, Nuria Pey, Muriel Ferrer, Amparo Quiles, Sandra Pérez, Gemma León, Elena Romero, Maria Andreu, Nati Galiana, Maria Dolores Climent, Amparo Cases and Cristina Capo for their assistance in contacting the families and administering the questionnaires. The authors would particularly like to thank all the participants for their generous collaboration. A full roster of the INMA Project Investigators can be found at http://www.proyectoinma.org/presentacion-inma/listado-investigadores/en_listado-investigadores.html.

KORA: The KORA research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München - German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

OAI: The OAI is a public–private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Genotyping support was provided by grant RC2-AR-058950 from NIAMS/NIH. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health.

WTCCC2: *Wellcome Trust Case-Control Consortium 2 (WTCCC2)* was principally funded by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z and WT084724MA). The Stroke Association provided additional support for collection of some of the St George's, London cases. The Oxford cases were collected as part of the Oxford Vascular Study which is funded by the MRC, Stroke Association, Dunhill Medical Trust, National Institute of Health Research (NIHR) and the NIHR Biomedical Research Centre, Oxford. The Edinburgh Stroke Study was supported by the Wellcome Trust (clinician scientist award to C Sudlow), and the Binks Trust. Sample processing occurred in the Genetics Core Laboratory of the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh. Much of the neuroimaging occurred in the Scottish Funding Council Brain Imaging Research Centre (www.sbirc.ed.ac.uk), Division of Clinical Neurosciences, University of Edinburgh, a core area of the Wellcome Trust Clinical Research Facility and part of the SINAPSE (Scottish Imaging Network—A Platform for Scientific Excellence) collaboration (www.sinapse.ac.uk), funded by the Scottish Funding Council and the Chief Scientist Office. Collection of the Munich cases and data analysis was supported by the Vascular Dementia Research Foundation. M Farrall and A Helgadottir acknowledge support from the BHF Centre of Research Excellence in Oxford and the Wellcome Trust core award (090532/Z/09/Z). *Barcelona* The Neurovascular Research Laboratory takes part in the International Stroke Genetics Consortium (ISGC), the Spanish Stroke Genetics Consortium (www.genestroke.com), and the Cooperative Neurovascular Research RENEVAS (RD06/0026/0010). This study was funded by a grant of the Spanish government (PI10/01212.). The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreements #201024 and #202213 (European Stroke Network). *Belgium Stroke Study (BSS)* was supported by Erasme Funds. *Edinburgh Stroke Study (ESS)* (which contributed discovery cases as part of WTCCC2 and additional replication cases) was supported as described above. *Lothian Birth Cohort 1936* was supported in part by Research into Aging, Help the Aged (Sidney De Haan Award and The Disconnected Mind Major Gift Campaign), MRC, and UK Biotechnology and Biological Sciences Research Council (BBSRC). Lothian Birth Cohort 1936 was also supported by a programme grant from Research Into Ageing and continues with programme grants from Help the Aged/Research Into Ageing (Disconnected Mind). The work was undertaken by The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698). Funding from the BBSRC, Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council (ESRC), and MRC is gratefully acknowledged. Genotyping of the LBC1936 was funded by the BBSRC.

Replication Cohorts:

AGES: The Age, Gene/Environment Susceptibility study was funded by the National Institute on Aging (NIA)(N01-AG-12100), Hjartavernd (the Icelandic Heart Association), and the Althingi (the Icelandic Parliament), with contributions from the Intramural Research Programs at the NIA and at the National Heart, Lung, and Blood Institute (Z01 HL004607-08 CE).

ARIC: The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

CADISP: The Cervical Artery Dissection and Ischemic Stroke Patients (CADISP) study has been supported by Inserm, Lille 2 University, Institut Pasteur de Lille and Lille University Hospital and received funding from the ERDF (FEDER funds) and Région Nord-Pas de Calais in the frame of Contrat de Projets Etat-Region 2007-2013 Région Nord-Pas-de-Calais - Grant N°09120030, Centre National de Genotypage, Emil Aaltonen Foundation, Paavo Ilmari Ahvenainen Foundation, Helsinki University Central Hospital Research Fund, Helsinki University Medical Foundation, Päivikki and Sakari Sohlberg Foundation, Aarne Koskelo Foundation, Maire Taponen Foundation, Aarne and Aili Turunen Foundation, Lilly Foundation, Alfred Kordelin Foundation, Finnish Medical Foundation, Orion Farmos Research Foundation, Maud Kuistila Foundation, the Finnish Brain Foundation, Biomedicum Helsinki Foundation, Projet Hospitalier de Recherche Clinique Régional, Fondation de France, Génopôle de Lille, Adrinord, Basel Stroke-Funds, Käthe-Zingg-Schwichtenberg-Fonds of the Swiss Academy of Medical Sciences, Swiss Heart Foundation. Stéphanie Debette is a recipient of a "Chaire d'Excellence Junior" grant from the Agence Nationale de la Recherche (ANR) and is supported by a grant from the Fondation Leducq. Vincent

Thijs is supported by a Fundamental Clinical Research Fellowship from FWO Flanders.

CHS: The Cardiovascular Health Study (CHS) research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

deCODE: deCODE Genetics Work performed at deCODE was funded, in part, through a grant from the European Community's Seventh Framework Programme (FP7/2007-2013), the ENGAGE project grant agreement HEALTH-F4-2007 to 201413. The deCODE CAD/MI Study was sponsored by NIH grant, National Heart, Lung, and Blood Institute R01HL089650-02.

FHS: The work of the Framingham Heart Study was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). A portion of this research utilized the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center. This study was also supported by grants from the National Institute of Neurological Disorders and Stroke (R01 NS17950), the National Heart, Lung and Blood Institute (R01 HL093029) and the National Institute of Aging (P30 AG10129, R01s AG033193, AG08122, AG16495).

FINRISK: Vaikko Salomaa was supported by the Academy of Finland (grant #139635) and the Finnish Foundation for Cardiovascular Research.

Glasgow ImmunoChip: The work was supported by NHS Greater Glasgow Endowment funds.

HPS: *Heart Protection Study (HPS)* (ISRCTN48489393) was supported by the UK Medical Research Council (MRC), British Heart Foundation, Merck and Co (manufacturers of simvastatin), and Roche Vitamins Ltd (manufacturers of vitamins). Genotyping was supported by a grant to Oxford University and CNG from Merck and Co. Jemma C Hopewell acknowledges support from the British Heart Foundation (FS/14/55/30806).

HVH: The research of the Heart and Vascular Health Studies has been funded in part by National Institute of Health grants HL40628, HL043201, HL053375, HL060739, HL68639, HL068986, HL073410, HL74745, HL085251, HL095080.

INTERSTROKE: We would like to acknowledge the Canadian Institutes of Health Research, Heart and Stroke Foundation of Canada, Canadian Stroke Network, Pfizer Cardiovascular Award, Merck, AstraZeneca, and Boehringer Ingelheim.

MESA: The Multi-Ethnic Study of Atherosclerosis was supported by the following grants: N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, and N01-HC-95169

MILANO: *Milano- Besta Stroke Register* Collection and genotyping of the Milan cases within CEDIR were supported by Annual Research Funding of the Italian Ministry of Health (Grant Numbers: RC 2007/LR6, RC 2008/LR6; RC 2009/LR8; RC 2010/LR8). FP6 LSHM-CT-2007-037273 for the PROCARDIS control samples.

PROSPER: The Prospective Study on Pravastatin in the Elderly at Risk (PROSPER) was supported by an investigator initiated grant obtained from Bristol-Myers Squibb. Prof. Dr. J. W. Jukema is an Established Clinical Investigator of the Netherlands Heart Foundation (grant 2001 D 032). Support for genotyping was provided by the seventh framework program of the European commission (grant 223004) and by the Netherlands Genomics Initiative (Netherlands Consortium for Healthy Aging grant 050-060-810).

RACE: We are thankful to the RACE study participants. Fieldwork in RACE was funded by the R-21 grant provided by the NINDS and the Fogarty International Center (1R21NS064908-01) and educational grants available to Dr. Saleheen at the Center for Non-Communicable Diseases, Pakistan. We would also like to acknowledge the contributions made by Professor John Danesh, Dr. Ayeesha Kamal and Professor Panos Deloukas.

Rotterdam: The generation and management of GWAS genotype data for the Rotterdam Study is supported by the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) project nr. 050-060-810. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. M. Arfan Ikram is supported by an NWO Veni grant (916.13.054).

SHIP: The Study of Health in Pomerania (SHIP) is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the

Federal State of Mecklenburg-West Pomerania, and the network ‘Greifswald Approach to Individualized Medicine (GANI_MED)’ funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data have been supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg- West Pomerania. The University of Greifswald is a member of the ‘Center of Knowledge Interchange’ program of the Siemens AG and the Caché Campus program of the InterSystems GmbH.

SIFAP: The sifap study (Stroke In Young Fabry Patients, http://www.sifap.eu; ClinicalTrials.gov: NCT00414583) has been supported partially by an unrestricted scientific grant from Shire Human Genetic Therapies. Funding for genotyping and analysis of samples were supported by the National Institutes of Health Genes, Environment and Health Initiative (GEI) Grant U01 HG004436, as part of the GENEVA consortium.

SIGNET: The Sea Islands Genetics Network (SIGNET) was supported by R01 DK084350 (MM Sale) from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and consists of data from the REasons for Geographic And Racial Differences in Stroke (REGARDS) cohort, (U01 NS041588; G Howard), Project SuGAR (Sea Islands Genetic African American Registry) (W.M. Keck Foundation; WT Garvey), a South Carolina Center of Biomedical Research Excellence (COBRE) in Oral Health Project P20 RR017696 (PI: Kirkwood; Sub-award: JK Fernandes), and the Systemic Lupus Erythematosus in Gullah Health (SLEIGH) study (PI: GS Gilkeson; K23 AR052364, DL Kamen; UL1 RR029882, KT Brady). Only data from the SIGNET-REGARDS sub-study were included in the current analyses. REGARDS is supported by a cooperative agreement U01 NS041588 from the National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health (NIH), Department of Health and Human Service. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDDK, NINDS or the NIH. Representatives of the funding agencies have been involved in the review of the manuscript but not directly involved in the collection, management, analysis or interpretation of the data. The authors thank the other investigators, the staff, and the participants of the REGARDS study for their valuable contributions. A full list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org

Three-City: The Three-City Study is conducted under a partnership agreement among the Institut National de la Santé et de la Recherche Médicale (INSERM), the Victor Segalen–Bordeaux II University, and Sanofi-Aventis. The Fondation pour la Recherche Médicale funded the preparation and initiation of the study. The 3C Study is also supported by the Caisse Nationale Maladie des Travailleurs Salariés, Direction Générale de la Santé, Mutuelle Générale de l’Education Nationale (MGEN), Institut de la Longévité, Conseils Régionaux of Aquitaine and Bourgogne, Fondation de France, and Ministry of Research–INSERM Programme “Cohortes et collections de données biologiques.” Christophe Tzourio has received investigator-initiated research funding from the French National Research Agency (ANR). Stéphanie Debette is a recipient of a “Chaire d’Excellence Junior” grant from the Agence Nationale de la Recherche (ANR) and is supported by a grant from the Fondation Leducq.

TWINGENE: This work was supported by grants from the Ministry for Higher Education, the Swedish Research Council (M-2005-1112 and 2009-2298), GenomEUtwin (EU/QLRT-2001-01254; QLG2-CT-2002-01254), NIH grant DK U01-066134, The Swedish Foundation for Strategic Research (SSF; ICA08-0047).

ULSAM: This work was supported by grants from Uppsala University, Swedish Research Council (2012-1397), Swedish Heart-Lung Foundation (20140422), Knut och Alice Wallenberg Foundation, European Research Council, Swedish Diabetes Foundation (2013-024).

Utrecht ImmunoChip/PROMISe: S.Achterberg was supported by in part by a grant from the Netherlands Heart Foundation, (grant no 2005B031) and a grant from the Dutch Brain Foundation (project 2008(1).10).

VHIR-FMT-Barcelona: The Barcelona GWAs Study was supported by the Genetic contribution to functional Outcome and Disability after Stroke (GODS) project, Fundació la Marató de TV3 and by the Miguel Servet grant (Pharmastroke project: CP12/03298). Neurovascular research Laboratory takes part in the INVICTUS network. I. F-C. is supported by the Miguel Servet programme (CP12/03298), Instituto de Salud Carlos III.

WGHS: The WGHS is supported by HL043851 and HL080467 from the National Heart, Lung, and Blood Institute and CA047988 from the National Cancer Institute with collaborative scientific support and funding for genotyping provided by Amgen. Additional funding for verification of cardiovascular endpoints was provided under ARRA by HL099355.

WHI-HT: WHI Funding support for WHI-GARNET was provided through the NHGRI GARNET (Grant Number U01 HG005152). Assistance with phenotype harmonisation and genotype cleaning, as well as with general study coordination, was provided by the GARNET Coordinating Center (U01 HG005157). Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH Genes, Environment, and Health Initiative (GEI; U01 HG004424).

1. Ay H, Furie KL, Singhal A, Smith WS, Sorensen AG, Koroshetz WJ. An evidence-based causative classification system for acute ischemic stroke. Ann Neurol. 2005;58(5):688–97.

2. Adams H, Bendixen B, Kappelle L, Biller J, Love B, Gordon D, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 1993;24(1):35–41.

3. Hatano S. Experience from a multicentre stroke register: a preliminary report. Bull World Health Organ. 1976;54(5):541–53.

4. Yadav S, Schanz R, Maheshwari A, Khan MS, Slark J, de Silva R, et al. Bio-Repository of DNA in stroke (BRAINS): a study protocol. BMC Med Genet. 2011;12:34.

5. Cotlarciuc I, Khan MS, Maheshwari A, Yadav S, Khan FY, Al-Hail H, et al. Bio-repository of DNA in stroke: a study protocol of three ancestral populations. J R Soc Med Cardiovasc Dis. 2012;1(10).

6. Kasner SE. Clinical interpretation and use of stroke scales. Lancet Neurol. 2006;5(7):603–12.

7. Johnson CJ, Kittner SJ, McCarter RJ, Sloan MA, Stern BJ, Buchholz D, et al. Interrater reliability of an etiologic classification of ischemic stroke. Stroke. 1995;26(1):46–51.

8. Kittner SJ, Stern BJ, Wozniak M, Buchholz DW, Earley CJ, Feeser BR, et al. Cerebral infarction in young adults: the Baltimore-Washington Cooperative Young Stroke Study. Neurology. 1998;50(4):890–4.

9. Foulkes MA, Wolf PA, Price TR, Mohr JP, Hier DB. The Stroke Data Bank: design, methods, and baseline characteristics. Stroke. 1988;19(5):547–54.

10. Meschia JF, Brown RD, Brott TG, Chukwudelunzu FE, Hardy J, Rich SS. The Siblings With Ischemic Stroke Study (SWISS) protocol. BMC Med Genet. 2002;3:1.

11. Eliasziw M, Smith RF, Singh N, Holdsworth DW, Fox AJ, Barnett HJ. Further comments on the measurement of carotid stenosis from angiograms. North American Symptomatic Carotid Endarterectomy Trial (NASCET) Group. Stroke. 1994;25(12):2445–9.

12. Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. Classification and natural history of clinically identifiable subtypes of cerebral infarction. Ann Intern Med. 1991;115(SUPPL.3):89.

13. Romano JG, Arauz A, Koch S, Dong C, Marquez JM, Artigas C, et al. Disparities in stroke type and vascular risk factors between 2 Hispanic populations in Miami and Mexico city. J Stroke Cerebrovasc Dis [Internet].

2013 Aug [cited 2015 Apr 25];22(6):828–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22749627

14. White H, Boden-Albala B, Wang C, Elkind MS V, Rundek T, Wright CB, et al. Ischemic stroke subtype incidence among whites, blacks, and Hispanics: The northern Manhattan study. Circulation. 2005;111(10):1327–31.

15. Howard VJ, Kleindorfer DO, Judd SE, McClure LA, Safford MM, Rhodes JD, et al. Disparities in stroke incidence contributing to disparities in stroke mortality. Ann Neurol. 2011;69(4):619–27.

16. Jood K, Ladenvall C, Rosengren A, Blomstrand C, Jern C. Family history in ischemic stroke before 70 years of age: The Sahlgrenska academy study on ischemic stroke. Stroke. 2005;36(7):1383–7.

17. Olsson S, Holmegaard L, Jood K, Sjögren M, Engström G, Lövkvist H, et al. Genetic variation within the interleukin-1 gene cluster and ischemic stroke. Stroke. 2012;43(9):2278–82.

18. Benavente OR, White CL, Pearce L, Pergola P, Roldan A, Benavente MF, et al. The Secondary Prevention of Small Subcortical Strokes (SPS3) study. Int J Stroke. 2011;6(2):164–75.

19. Meschia JF, Brott TG, Chukwudelunzu FE, Hardy J, Brown RD, Meissner I, et al. Verifying the stroke-free phenotype by structured telephone interview. Stroke. 2000;31(5):1076–80.

20. Toole JF, Malinow MR, Chambless LE, Spence JD, Pettigrew LC, Howard VJ, et al. Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial. JAMA J Am Med Assoc. 2004;291(5):565–75.

21. McEvoy M, Smith W, D'Este C, Duke J, Peel R, Schofield P, et al. Cohort profile: The Hunter Community Study. Int J Epidemiol [Internet]. 2010 Dec [cited 2015 Mar 5];39(6):1452–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20056765

22. Schmidt R, Lechner H, Fazekas F, Niederkorn K, Reinhart B, Grieshofer P, et al. Assessment of cerebrovascular risk profiles in healthy persons: definition of research goals and the Austrian Stroke Prevention Study (ASPS). Neuroepidemiology. 1994;13(6):308–13.

23. Schmidt R, Fazekas F, Kapeller P, Schmidt H, Hartung HP. MRI white matter hyperintensities: three-year follow-up of the Austrian Stroke Prevention Study. Neurology. 1999;53(1):132–9.

24. LaVange LM, Kalsbeek WD, Sorlie PD, Avilés-Santa LM, Kaplan RC, Barnhart J, et al. Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. Ann Epidemiol. 2010;20(8):642–9.

25. Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, Kaplan RC, Daviglus ML, Giachello AL, et al. Design and Implementation of the Hispanic Community Health Study/Study of Latinos. Ann Epidemiol. 2010;20(8):629–41.

26. Juster FT, Suzman R. An overview of the health and retirement study. J Hum Resour. 1995;30:S7–56.

27. Lemmens R, Abboud S, Robberecht W, Vanhees L, Pandolfo M, Thijs V, et al. Variant on 9p21 strongly associates with coronary heart disease, but lacks association with common stroke. Eur J Hum Genet. 2009;17(10):1287–93.

28. Li C, Liu Z, Wang LE, Gershenwald JE, Lee JE, Prieto VG, et al. Haplotype and genotypes of the VDR gene and cutaneous melanoma risk in non-Hispanic whites in Texas: A case-control study. Int J Cancer. 2008;122(9):2077–84.

29. Li C, Zhao H, Hu Z, Liu Z, Wang L-E, Gershenwald JE, et al. Genetic variants and haplotypes of the caspase-8 and caspase-10 genes contribute to susceptibility to cutaneous melanoma. Hum Mutat [Internet]. 2008 Dec [cited 2015 May 20];29(12):1443–51. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2937220&tool=pmcentrez&rendertype=abstract

30. Evans MK, Lepkowski JM, Powe NR, LaVeist T, Kuczmarski MF, Zonderman AB. Healthy aging in neighborhoods of diversity across the life span (HANDLS): overcoming barriers to implementing a longitudinal, epidemiologic, urban study of health, race, and socioeconomic status. Ethn Dis [Internet]. 2010 Jan [cited 2015 May 20];20(3):267–75. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3040595&tool=pmcentrez&rendertype=abstract

31. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium design of prospective meta-analyses of genome-wide association studies from 5 Cohorts. Circulation: Cardiovascular Genetics. 2009. p. 73–80.

32. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature [Internet]. Nature Publishing Group; 2012 Nov 1 [cited 2014 Jan 21];491(7422):56–65. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23128226&retmode=ref&cmd=prlinks

33. The International HapMap Consortium., International HapMap C. A haplotype map of the human genome. Nature [Internet]. 2005 Oct 27 [cited 2014 Jan 20];437(7063):1299–320. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16255080

34. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.

35. Klungel OH, Heckbert SR, Longstreth WT, Furberg CD, Kaplan RC, Smith NL, et al. Antihypertensive drug therapies and the risk of ischemic stroke. Arch Intern Med. 2001;161(1):37–43.

36. Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Raghunathan TE, Weiss NS, et al. The Risk of Myocardial Infarction Associated With Antihypertensive Drug Therapies. JAMA: The Journal of the American Medical Association. 1995. p. 620–5.

37. Psaty BM, Heckbert SR, Atkins D, Lemaitre R, Koepsell TD, Wahl PW, et al. The risk of myocardial infarction associated with the combined use of estrogens and progestins in postmenopausal women. Arch Intern Med. 1994;154(12):1333–9.

38. Price TR, Psaty B, O'Leary D, Burke G, Gardin J. Assessment of cerebrovascular disease in the Cardiovascular Health Study. Ann Epidemiol. 1993;3(5):504–7.

39. O'Donnell M, Xavier D, Diener C, Sacco R, Lisheng L, Zhang H, et al. Rationale and design of INTERSTROKE: a global case-control study of risk factors for stroke. Neuroepidemiology [Internet]. 2010 Jan [cited 2015 May 6];35(1):36–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20389123

40. Howard VJ, Cushman M, Pulley L, Gomez CR, Go RC, Prineas RJ, et al. The reasons for geographic and racial differences in stroke study: Objectives and design. Neuroepidemiology. 2005;25(3):135–43.

41. Rolfs A, Fazekas F, Grittner U, Dichgans M, Martus P, Holzhausen M, et al. Acute cerebrovascular disease in the young: The stroke in young fabry patients study. Stroke. 2013;44(2):340–9.

42. Rolfs A, Martus P, Heuschmann PU, Grittner U, Holzhausen M, Tatlisumak T, et al. Protocol and methodology of the stroke in young fabry patients (sifap1) study: A prospective multicenter european study of 5,024 young stroke patients aged 18-55 years. Cerebrovasc Dis. 2011;31(3):253–62.

43. Achterberg S, Kappelle LJ, Algra a. Prognostic modelling in ischaemic stroke study, additional value of genetic characteristics: Rationale and design. Eur Neurol. 2008;59(5):243–52.

44. Achterberg S, Kappelle LJ, de Bakker PIW, Traylor M, Algra A. No Additional Prognostic Value of Genetic Information in the Prediction of Vascular Events after Cerebral Ischemia of Arterial Origin: The PROMISe Study. PLoS One [Internet]. 2015;10(4):e0119203. Available from: http://dx.plos.org/10.1371/journal.pone.0119203

45. Simons PCG, Algra A, Van De Laak MF, Grobbee DE, Van Der Graaf Y. Second manifestations of ARTerial disease (SMART) study: Rationale and design. Eur J Epidemiol. 1999;15(9):773–81.

46. Trynka G, Hunt K, Bockett N, Romanos J, Mistry V, Szperl a, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet [Internet]. Nature Publishing Group; 2011;43(12):1193–201. Available from: http://discovery.ucl.ac.uk/1337303/

47. Ridker PM, Chasman DI, Zee RYL, Parker A, Rose L, Cook NR, et al. Rationale, design, and methodology of the Women's Genome Health Study: A genome-wide association study of more than 25 000 initially healthy American women. Clin Chem. 2008;54(2):249–55.

48. Ridker PM, Cook NR, Lee IM, Gordon D, Gaziano JM, Manson JE, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. N Engl J Med [Internet]. 2005;352(13):1293–304. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15753114

49. Ay H, Benner T, Arsava EM, Furie KL, Singhal AB, Jensen MB, et al. A computerized algorithm for etiologic classification of ischemic stroke: The causative classification of stroke system. Stroke. 2007;38(11):2979–84.

50. Arsava EM, Ballabio E, Benner T, Cole JW, Delgado-Martinez MP, Dichgans M, et al. The causative classification of stroke system: An international reliability and optimization study. Neurology. 2010;75(14):1277–84.

51. McArdle PF, Kittner SJ, Ay H, Brown RD, Meschia JF, Rundek T, et al. Agreement between TOAST and CCS ischemic stroke classification: the NINDS SiGN study. Neurology [Internet]. 2014 Oct 28 [cited 2015 May 7];83(18):1653–60. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4223086&tool=pmcentrez&rendertype=abstract

52. Ay H, Arsava EM, Andsberg G, Benner T, Brown RD, Chapman SN, et al. Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network. Stroke [Internet]. 2014;45(12):3589–96. Available from: http://stroke.ahajournals.org/cgi/doi/10.1161/STROKEAHA.114.007362

53. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. 2006;16(9):1136–48.

54. Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, et al. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Genet. 2010;19(7):1263–75.

55. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007;23(6):657–63.

56. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nature Genetics. 2012. p. 642–50.

57. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics [Internet]. 2012;28(24):3329–31. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3519456&tool=pmcentrez&rendertype=abstract

58. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics [Internet]. 2012 Dec 15 [cited 2014 Jan 21];28(24):3326–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3519454&tool=pmcentrez&rendertype=abstract

59. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010;34(6):591–602.

60. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

61. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature [Internet]. 2010 Sep 2 [cited 2014 Jan 21];467(7311):52–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3173859&tool=pmcentrez&rendertype=abstract

62. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.

63. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73.

64. The International HapMap Consortium. The International HapMap Project. Nature [Internet]. 2003 Dec [cited 2014 Jan 21];426:789–96. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3584704&tool=pmcentrez&rendertype=abstract

65. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):2074–93.

66. Rousseeuw PJ, van Driessen K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics [Internet]. 1999;41(3):212–23. Available from: http://www.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670\npapers2://publication/doi/10.2307/1270566?ref=search-gateway:f11cc048465292d7e0104d547b154f19

67. Zhu X, Li S, Cooper RS, Elston RC. A Unified Association Analysis Approach for Family and Unrelated Samples Correcting for Stratification. Am J Hum Genet. 2008;82(2):352–65.

68. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods [Internet]. 2011 Jan [cited 2014 Jan 21];8(10):833–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21892150

69. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods [Internet]. Nature Publishing Group; 2013;10(1):5–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23269371

70. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature Methods. 2011. p. 179–81.

71. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. 2014;10(4).

72. Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, de Bakker PI, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet [Internet]. 2014 Jun 29 [cited 2014 Jun 30];46(8):818–25. Available from: http://www.nature.com/doifinder/10.1038/ng.3021

73. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6).

74. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet [Internet]. Nature Publishing Group; 2012;44(8):955–9. Available from: http://dx.doi.org/10.1038/ng.2354

75. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499–511.

76. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906–13.

77. De Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 2008;17(R2):R122–8.

78. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. Nat Protoc [Internet]. 2014;9(5):1192–212. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24762786

79. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38(2):209–13.

80. Gretarsdottir S, Thorleifsson G, Manolescu A, Styrkarsdottir U, Helgadottir A, Gschwendtner A, et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. Ann Neurol. 2008;64(4):402–9.

81. Gudbjartsson DF, Holm H, Gretarsdottir S, Thorleifsson G, Walters GB, Thorgeirsson G, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. Nat Genet. 2009;41(8):876–8.

82. Bellenguez C, Bevan S, Gschwendtner A, Spencer CCA, Burgess AI, Pirinen M, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. Nat Genet [Internet]. 2012 Mar [cited 2014 Jan 21];44(3):328–33. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3303115&tool=pmcentrez&rendertype=abstract

83. Kilarski LL, Achterberg S, Devan WJ, Traylor M, Malik R, Lindgren A, et al. Meta-analysis in more than 17,900 cases of ischemic stroke reveals a novel association at 12q24.12. Neurology [Internet]. 2014;83(8):678–85. Available from: http://www.neurology.org.ezp-prod1.hul.harvard.edu/content/early/2014/07/16/WNL.0000000000000707

84. Williams FMK, Carter AM, Hysi PG, Surdulescu G, Hodgkiss D, Soranzo N, et al. Ischemic stroke is associated with the ABO locus: the EuroCLOT study. Ann Neurol [Internet]. 2013 Jan [cited 2014 Jan 21];73(1):16–31. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3582024&tool=pmcentrez&rendertype=abstract

85. Kim J, Chae YK. Genomewide association studies of stroke. N Engl J Med [Internet]. 2009 Aug 13 [cited 2014 Jan 21];361(7):722; author reply 722. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19675338

86. Bis JC, DeStefano A, Liu X, Brody JA, Choi SH, Verhaaren BFJ, et al. Associations of NINJ2 sequence variants with incident ischemic stroke in the Cohorts for Heart and Aging in Genomic Epidemiology (CHARGE) consortium. PLoS One. 2014;9(6).

87. Holliday EG, Maguire JM, Evans T-J, Koblar S a, Jannes J, Sturm JW, et al. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. Nat Genet [Internet]. 2012 Oct [cited 2014 Feb 20];44(10):1147–51. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3651583&tool=pmcentrez&rendertype=abstract

88. Matarin M, Brown WM, Singleton A, Hardy J a., Meschia JF. Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. Stroke. 2008;39(5):1586–9.

89. Lin DY, Sullivan PF. Meta-Analysis of Genome-wide Association Studies with Overlapping Subjects. Am J Hum Genet. 2009;85(6):862–72.

90. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet [Internet]. The American Society of Human Genetics; 2011 May 13 [cited 2014 Jun 4];88(5):586–98. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146723&tool=pmcentrez&rendertype=abstract

91. Han B, Eskin E. Interpreting meta-analyses of genome-wide association studies. PLoS Genet [Internet]. 2012 Jan [cited 2014 Jun 10];8(3):e1002555. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3291559&tool=pmcentrez&rendertype=abstract