

Combining cases and publicly-available controls

for discovery of common disease loci through GWAS

Sara L. Pulit

Department of Genetics, University Medical Center Utrecht

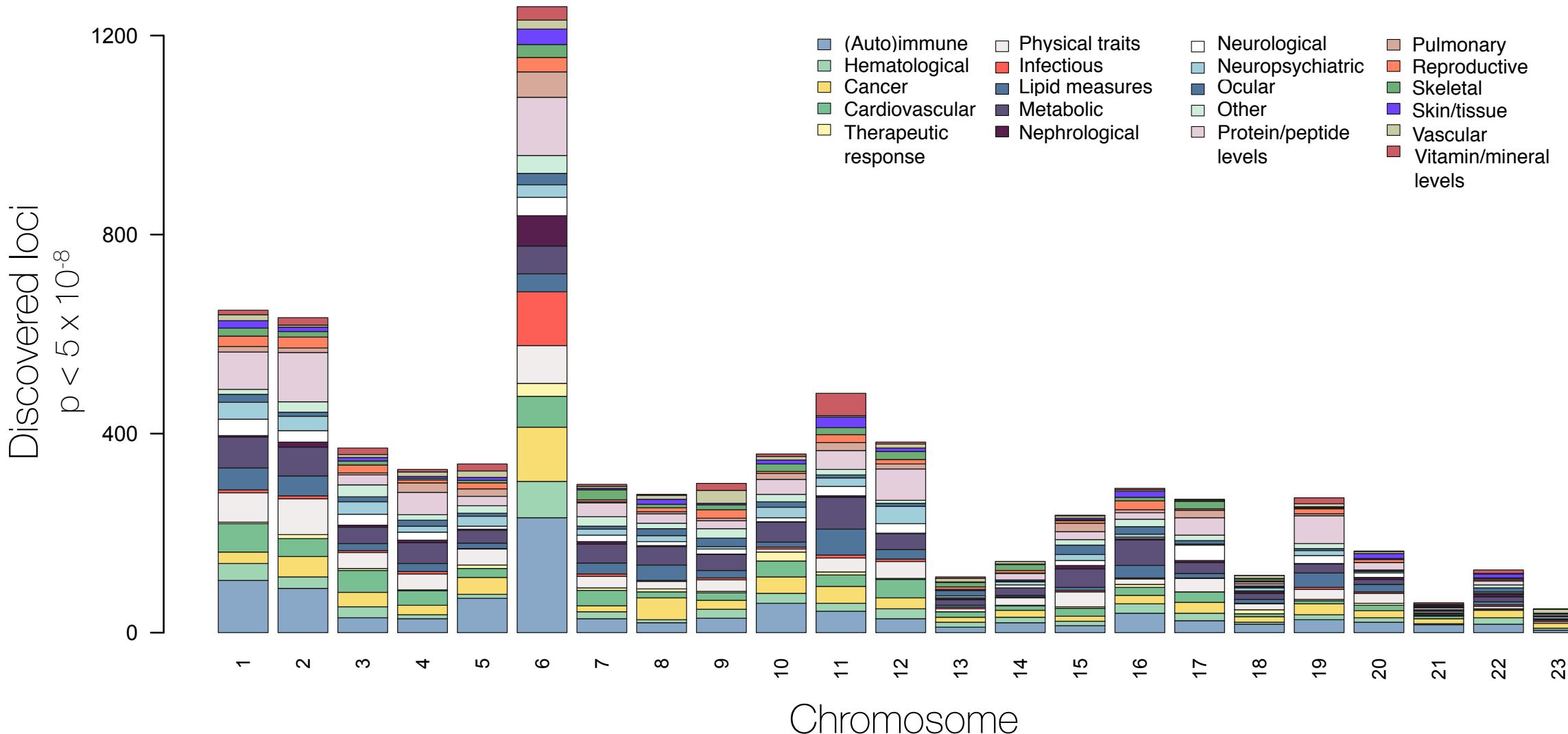


s.l.pulit@umcutrecht.nl



@saralpulit

Thousands of SNP-trait associations identified



Massive sample collections

UK Biobank



N = 500,000
European ancestry

China Kadoorie Biobank



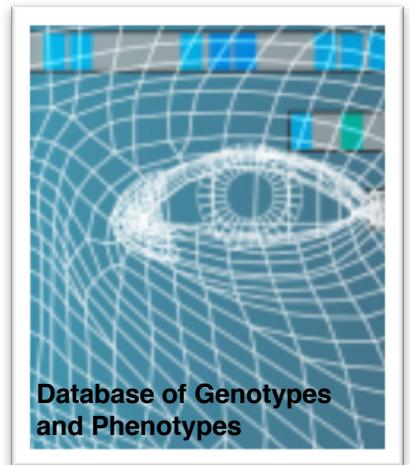
N = 510,000
East Asian ancestry

Human Genome Diversity Project



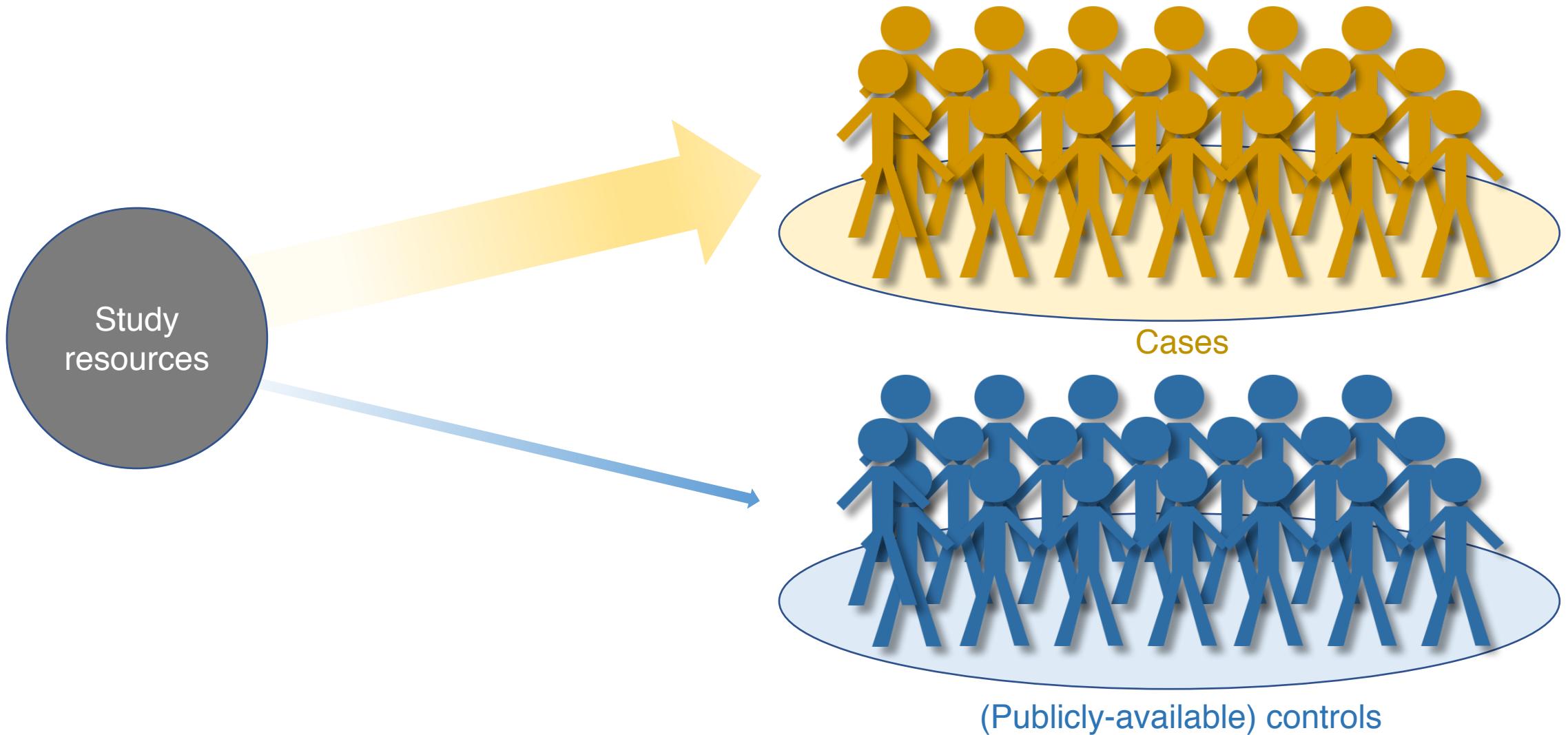
N = 1,043
Multi-ancestry

dbGaP

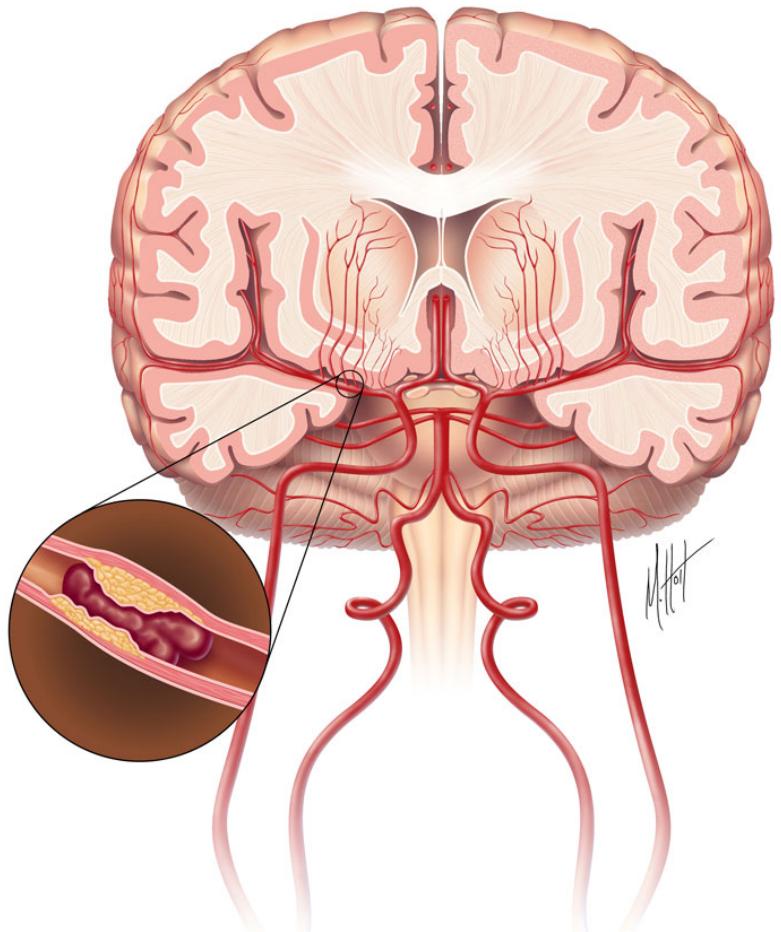


N > 500,000
Multi-ancestry

Investing resources in cases only

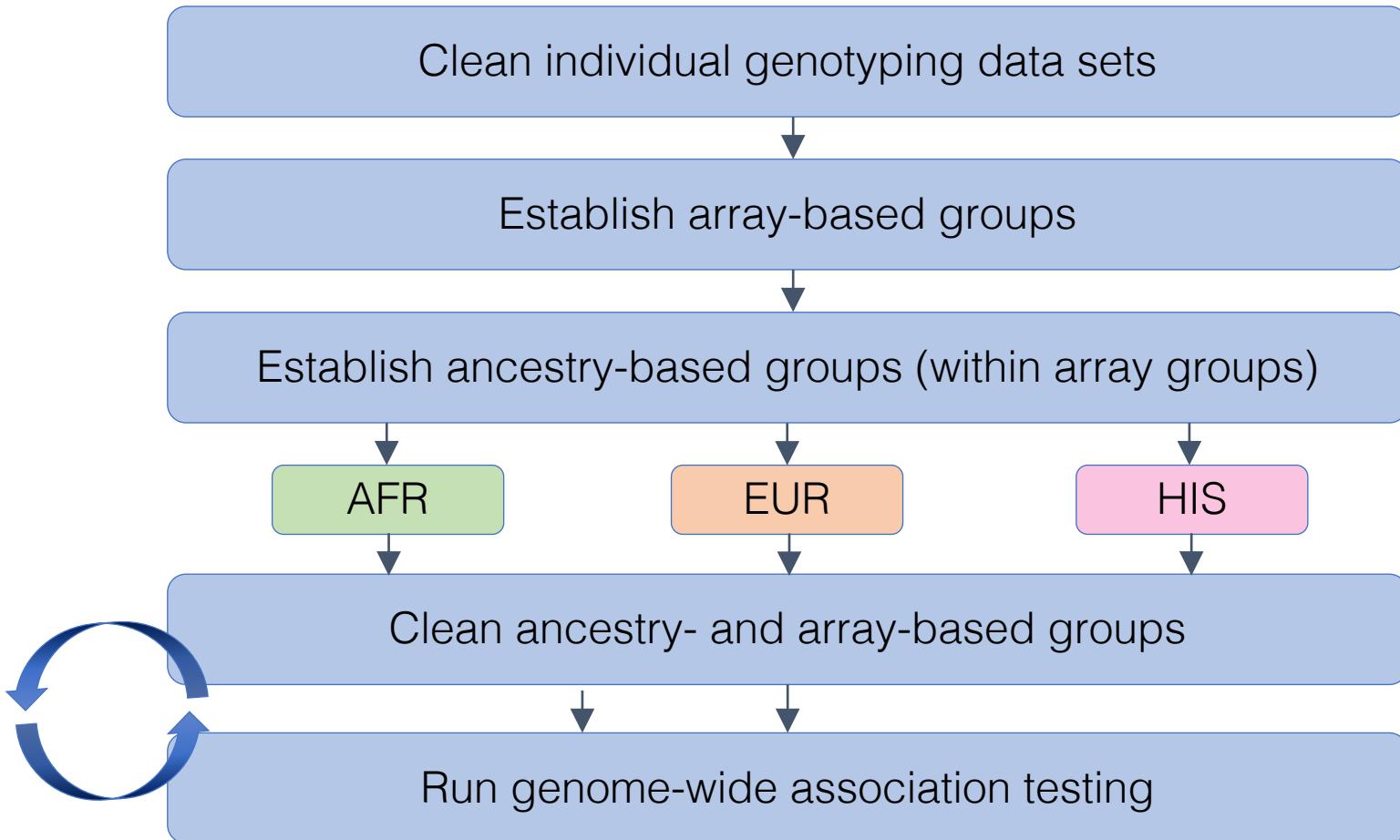


The Stroke Genetics Network (SiGN)



- GWAS of ischemic stroke
- 16,000 cases and 30,000 controls
 - > 28,000 controls drawn from publicly-available datasets

SiGN workflow at a glance



1. Know your data
2. Know your phenotype
3. Know your limitations

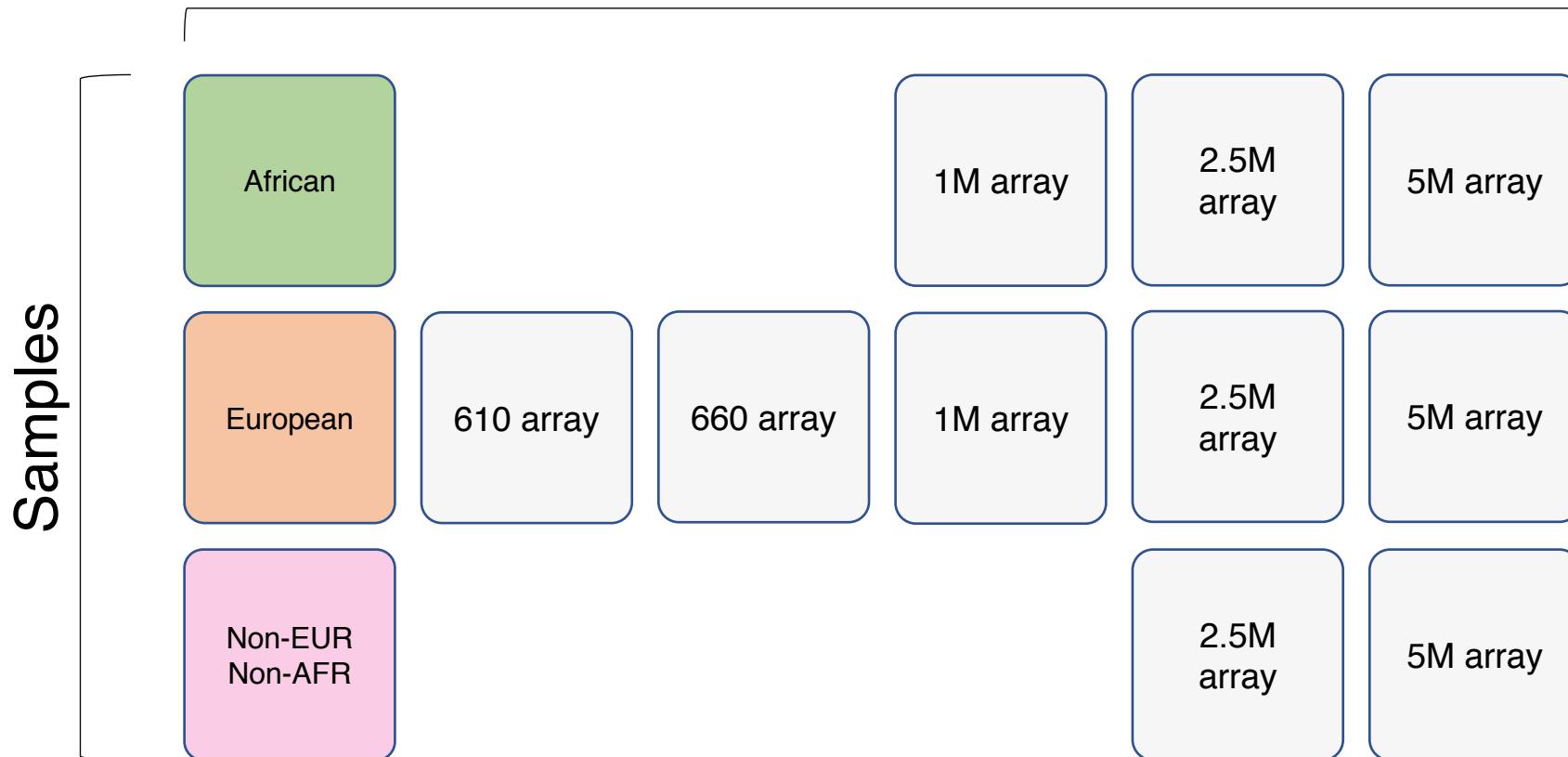
1. Know your data

2. Know your phenotype

3. Know your limitations

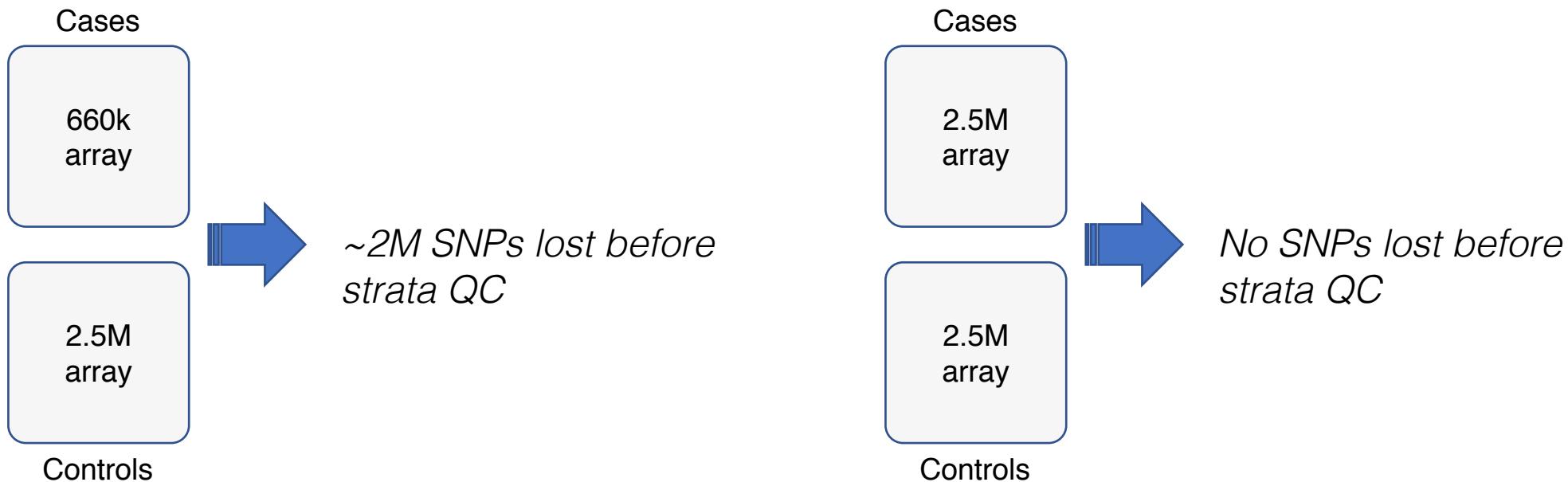
Know your data

Genotyping arrays (Illumina)



Creating case-control strata

- Match samples by genotyping array
 - Maximize the total number of SNPs available for genotype imputation

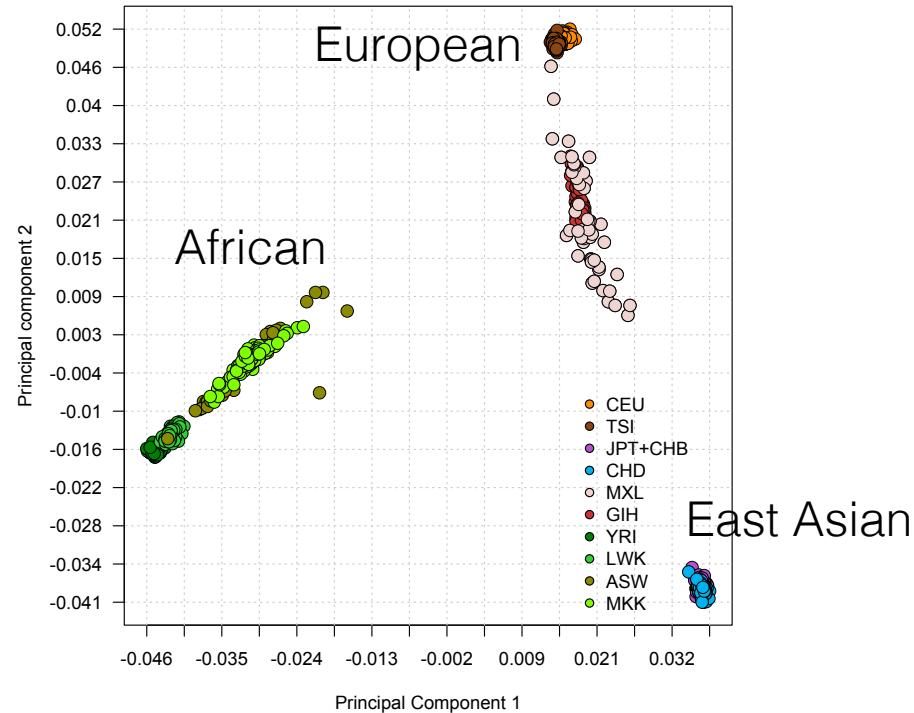


Creating case-control groups

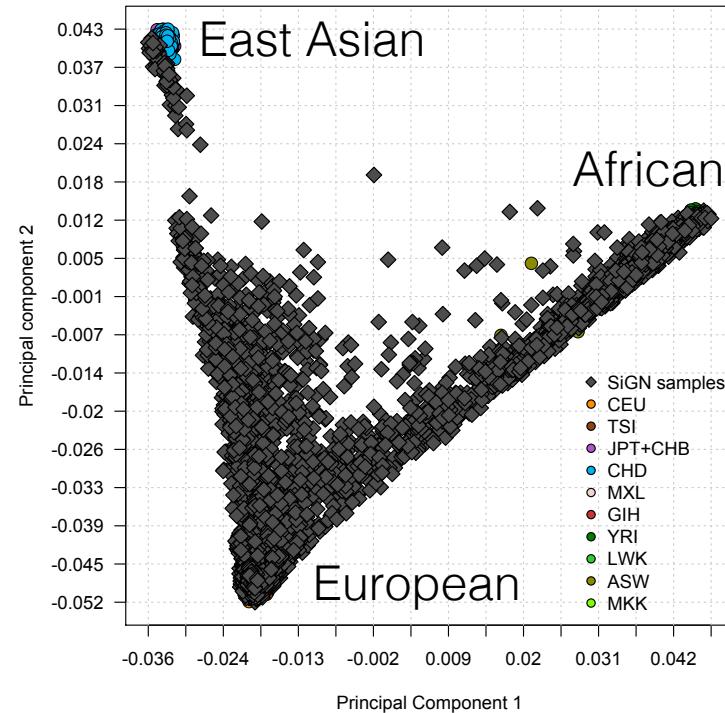
- Match samples by genotyping array
 - Maximize the total number of SNPs available for genotype imputation
- When possible, maintain a 1:2 case-control ratio within groups
- When possible, merge cohorts collected in the same geographic region
 - Maximize chance of ancestry-matched samples
- Identify ancestry-matched cases and controls

PCA-based approaches for matching samples by ancestry

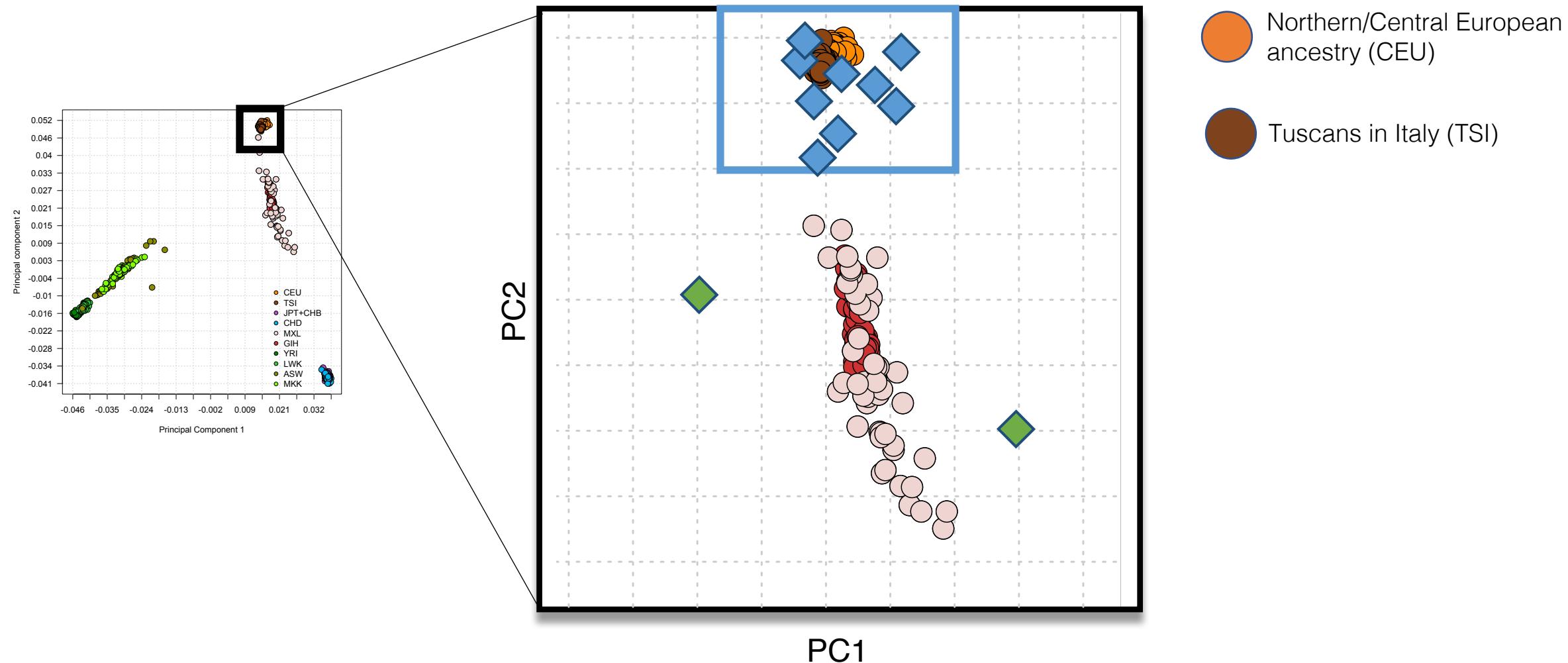
Global populations (HapMap 3)



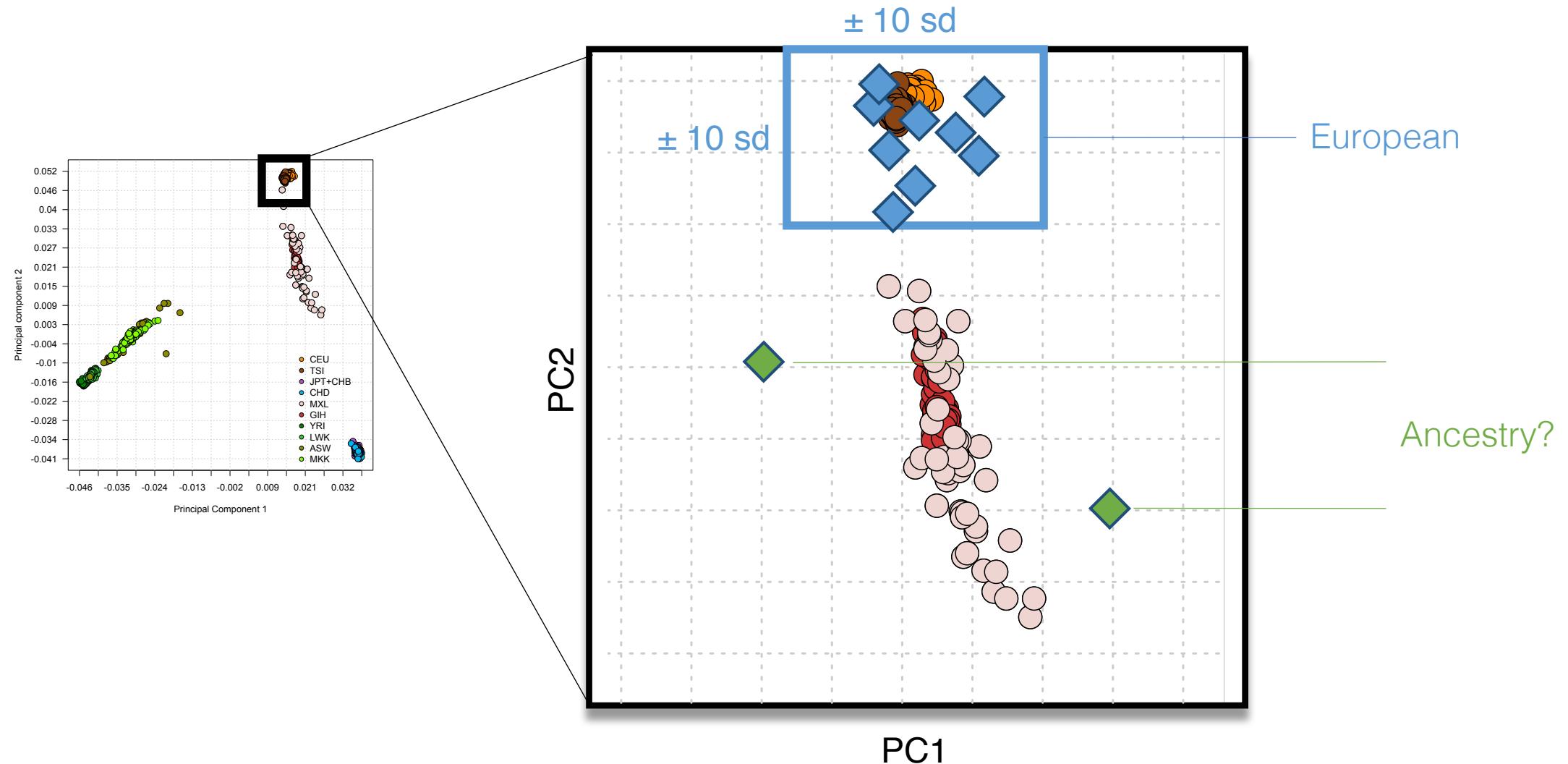
SiGN data



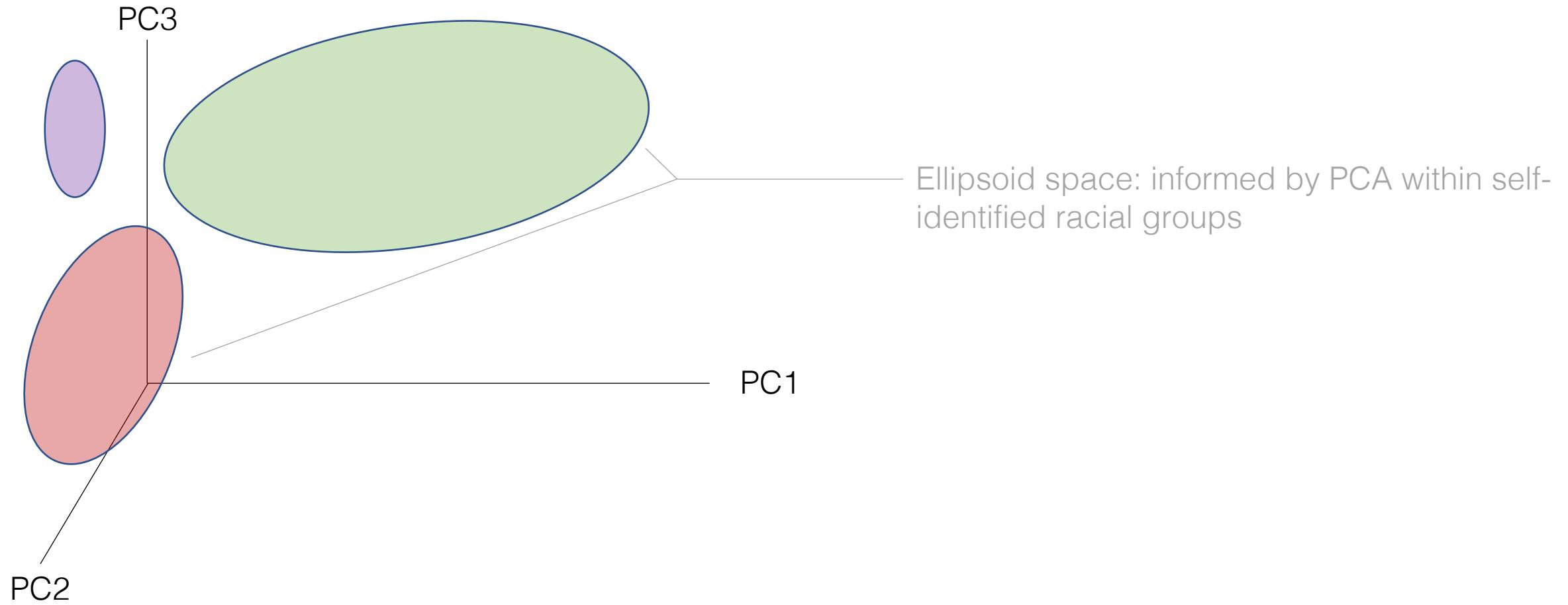
Defining European-ancestry groups with PCA



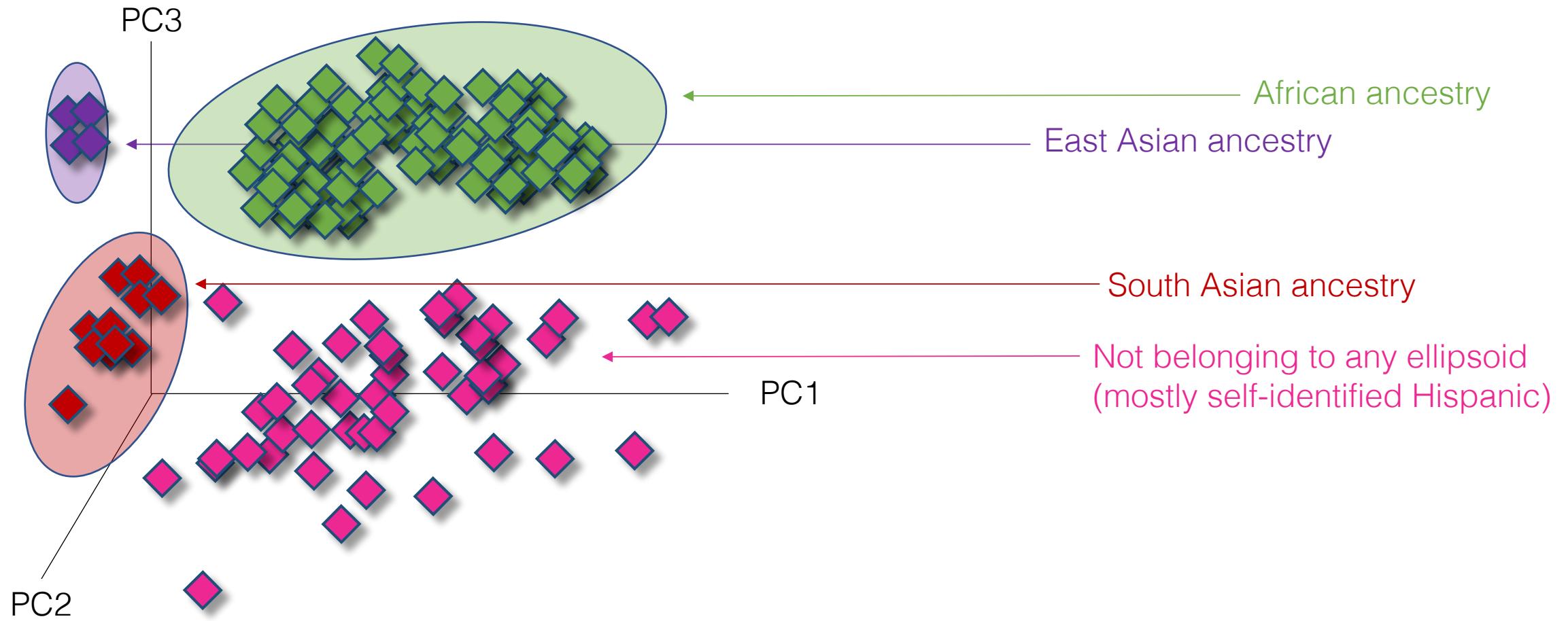
Defining European-ancestry groups with PCA



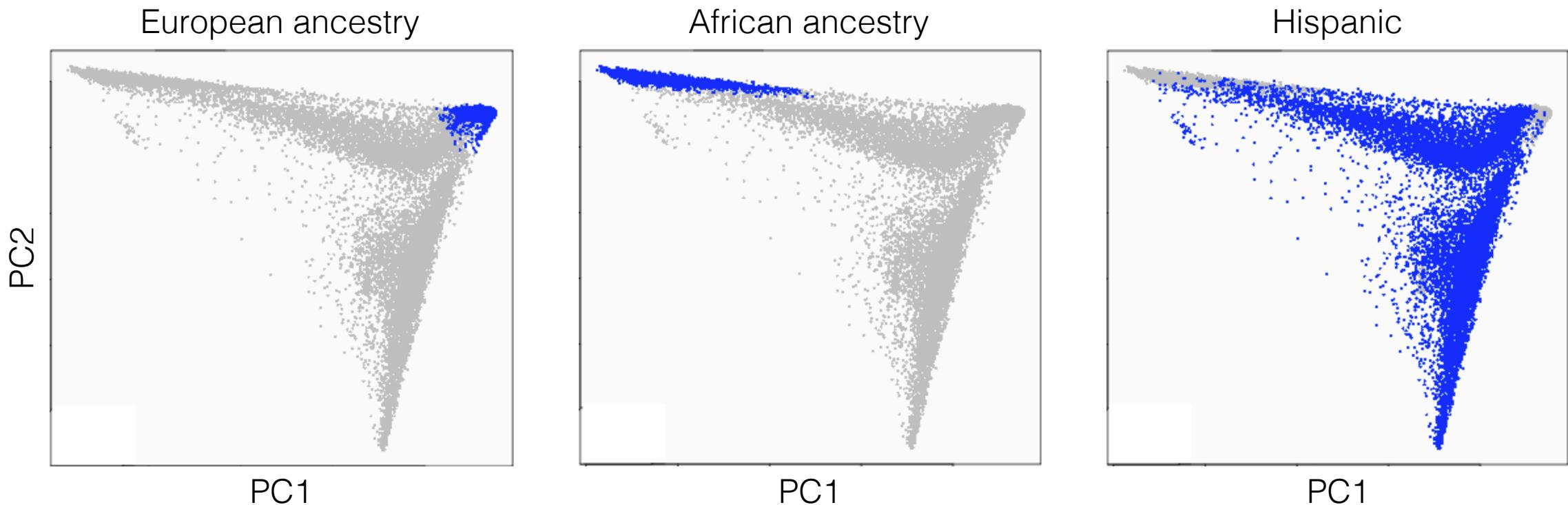
Defining additional ancestral groups with hyperellipsoid analysis



Defining additional ancestral groups with hyperellipsoid analysis



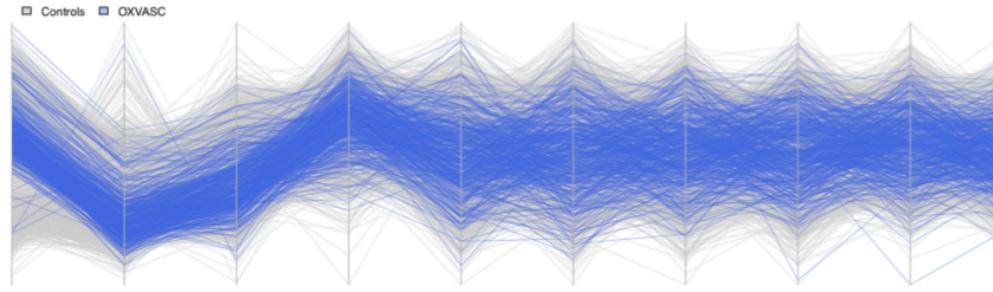
Ancestral groups in two-dimensional PCA space



Case-control matching across 10 PCs

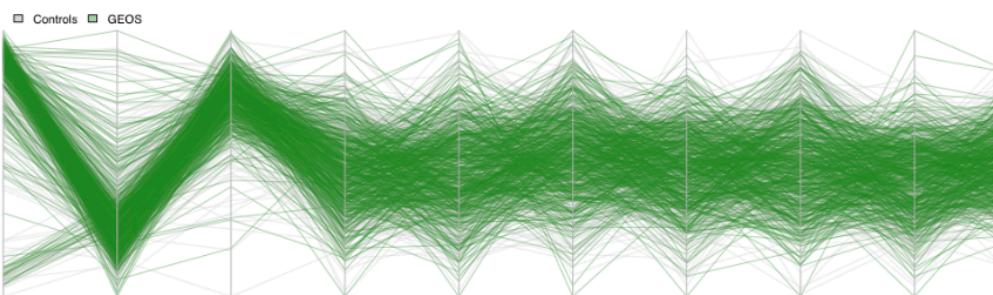
Case —————

Control —————



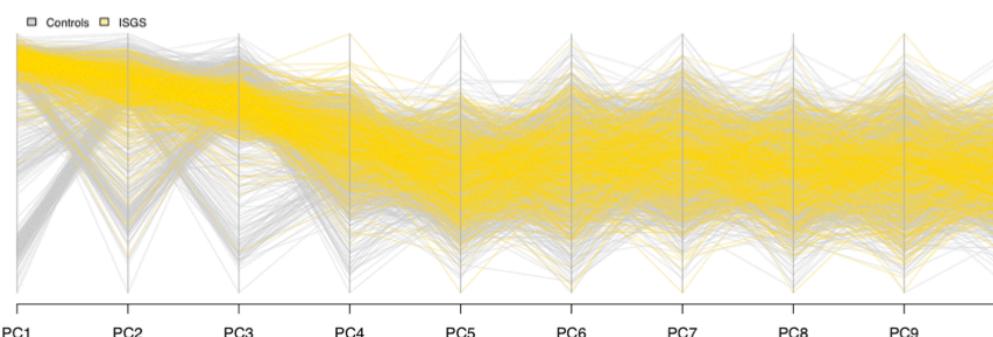
Case —————

Control —————



Case —————

Control —————



Know your data

Know your phenotype

Know your limitations

Stroke: what should we expect?

Risk Variants for Atrial Fibrillation on Chromosome 4q25 Associate with Ischemic Stroke

Solveig Gretarsdottir, PhD,¹ Gudmar Thorleifsson, PhD,¹ Andrei Manolescu, PhD,¹ Unnur Stykarsdottir, PhD,¹ Anna Helgadottir, MD,¹ Andreas Gschwendtner, MD,² Konstantinos Kostulas, MD, PhD,³ Gregor Kuhlenbäumer, MD,^{4,5} Steve Bevan, PhD, Thorbjorg Jonsdottir, BSc,¹ Hjordis Bjarnason, BSc,¹ Jona Sæmundsdottir, BSc,¹ Stefan Palsson, David O. Arnar, MD, PhD,⁷ Hilma Holm, MD,¹ Gudmundur Thorgerisson, MD, PhD, Einar Mar Valdimarsson, MD,² Sigurlaug Sveinbjörnsdóttir, MD,¹ Christian Gieger, PhD, Klaus Berger, MD,¹⁹ H-Erich Wichmann, MD,¹⁴ Jan Hillert, MD,¹ Hugh Markus, PhD, Jeffrey Robert Gulcher, MD, PhD,¹ E. Bernd Ringelstein, MD,¹⁸ Augustine Kong, PhD, Martin Dichgans, MD,² Daniel Pannar Gudbjartsson, PhD,¹ Unnur Thorsteinsdottir, PhD, Kari Stefansson, MD, PhD,^{1,4}

A sequence variant in *ZFHX3* on 16q22 associates with atrial fibrillation and ischemic stroke

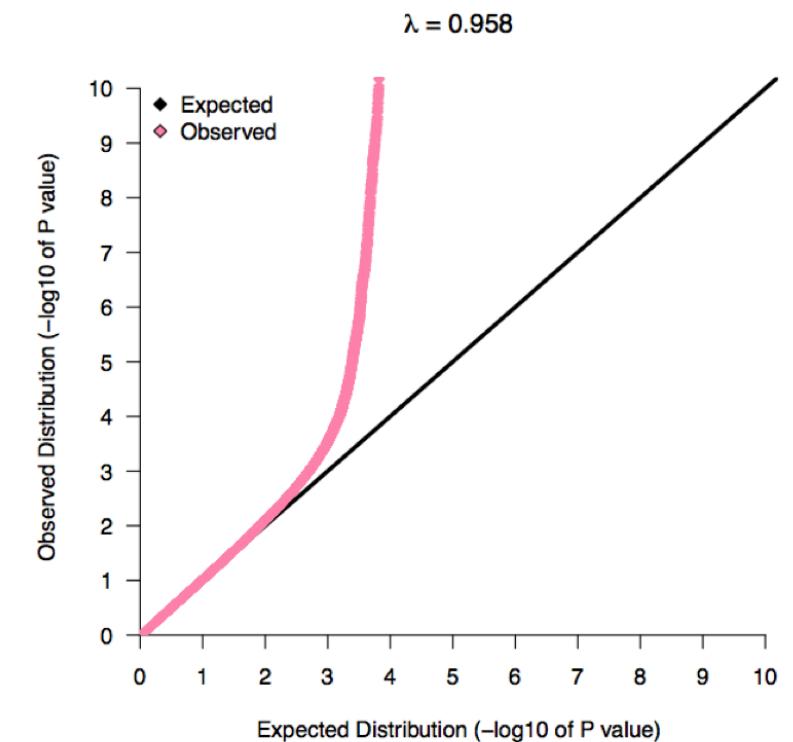
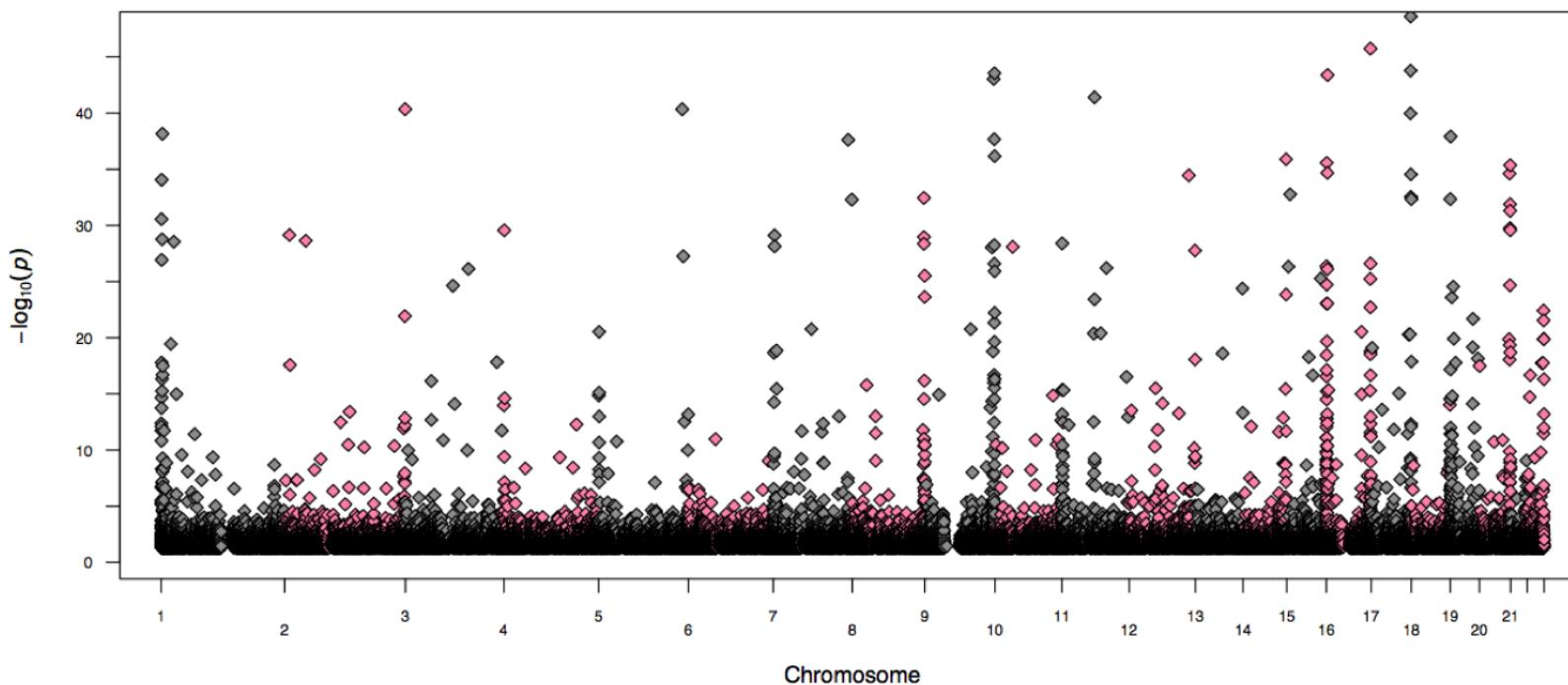
Daniel F Gudbjartsson^{1,22}, Hilma Holm^{1,2,22}, Solveig Gretarsdottir¹, Gudmar Thorleifsson¹, G Bragi Walters¹, Gudmundur Thorgerisson^{3,4}, Jeffrey Gulcher¹, Ellisiv B Mathiesen^{5,6}, Inger Njolstad⁷, Audhild Nyren^{7,8}, Tom Wilsgaard⁷, Erin M Hald⁹, Kristian Hveem¹⁰, Camilla Stoltzenberg¹¹, Gayle Kucera¹², Tanya Stubblefield¹², Shannon Carter¹², Dan Roden¹², Maggie C Y Ng¹³, Larry Baum¹³, Wing Yee So¹³, Ka Sing Wong¹³, Juliana C N Chan¹³, Christian Gieger¹⁴, H-Erich Wichmann¹⁴, Andreas Gschwendtner¹⁵, Martin Dichgans¹⁵, Gregor Kuhlenbäumer¹⁶, Klaus Berger¹⁷, E Bernd Ringelstein¹⁸, Steve Bevan¹⁹, Hugh S Markus¹⁹, Konstantinos Kostulas²⁰, Jan Hillert²⁰, Sigurlaug Sveinbjörnsdóttir²¹, Einar M Valdimarsson²¹, Maja-Lisa Lochen^{7,9}, Ronald C W Ma¹³, Dawood Darbar¹², Augustine Kong¹, David O Arnar^{3,4}, Unnur Thorsteinsdottir^{1,4} & Kari Stefansson^{1,4}

Genome-wide association study identifies a variant in *HDAC9* associated with large vessel ischemic stroke

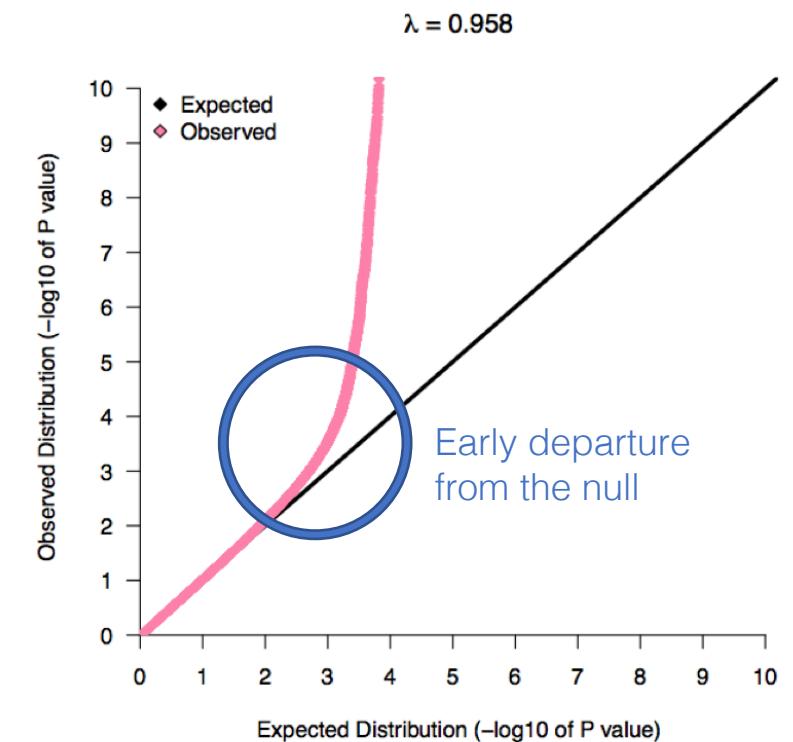
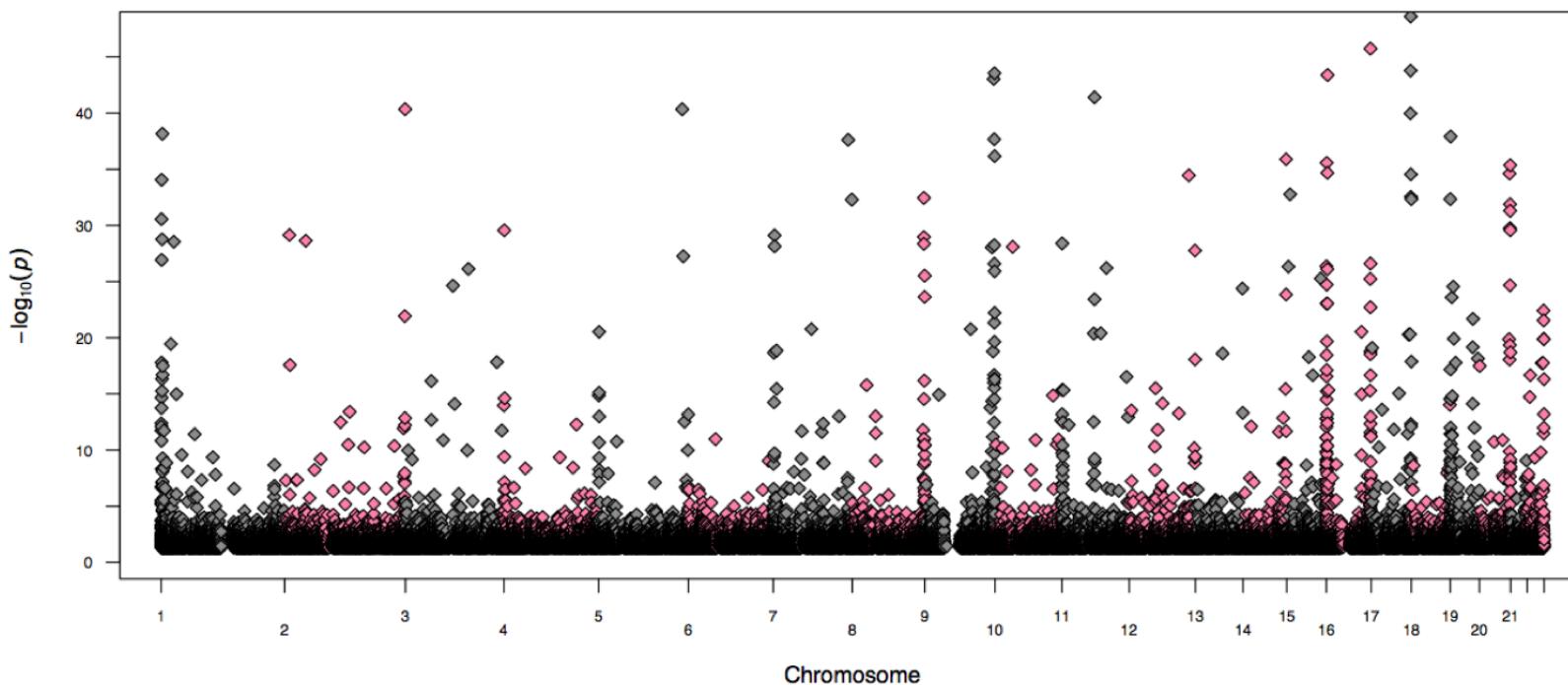
The International Stroke Genetics Consortium (ISGC)¹ & the Wellcome Trust Case Control Consortium 2 (WTCCC2)¹

- GWAS have revealed a modest number of hits
 - True positives
- Common, complex disease
 - Likely underpinned by modestly penetrant SNPs
- Associated SNPs are typically associated to a subtype
 - Reduce sample size (and power)

This is not what we expect



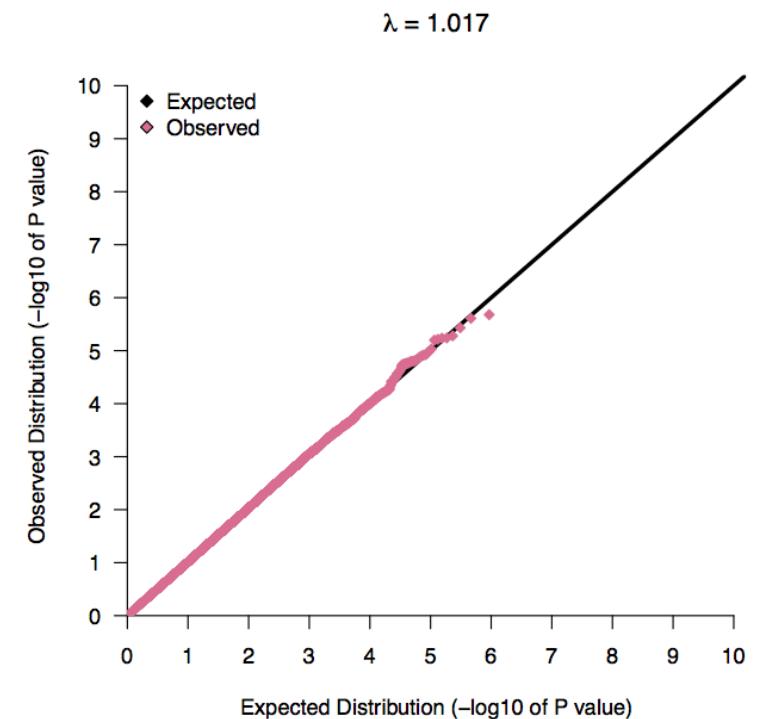
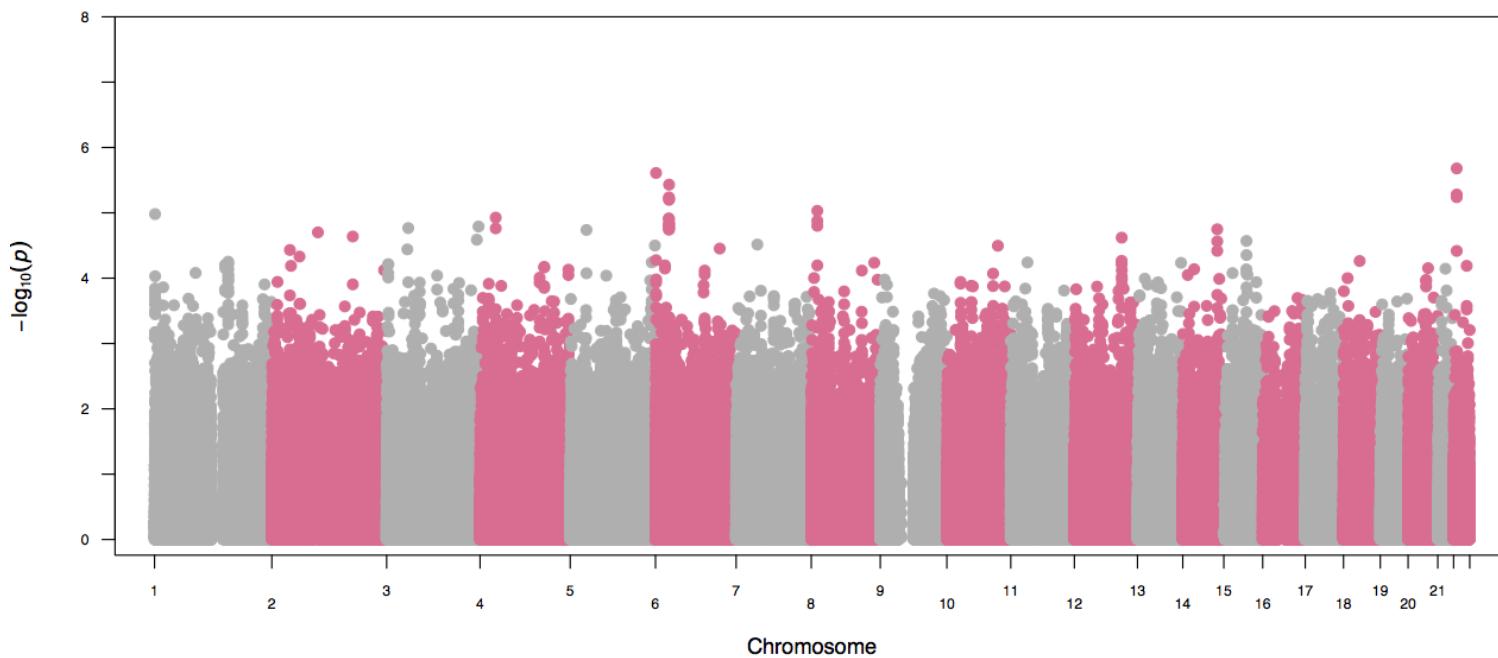
This is not what we expect

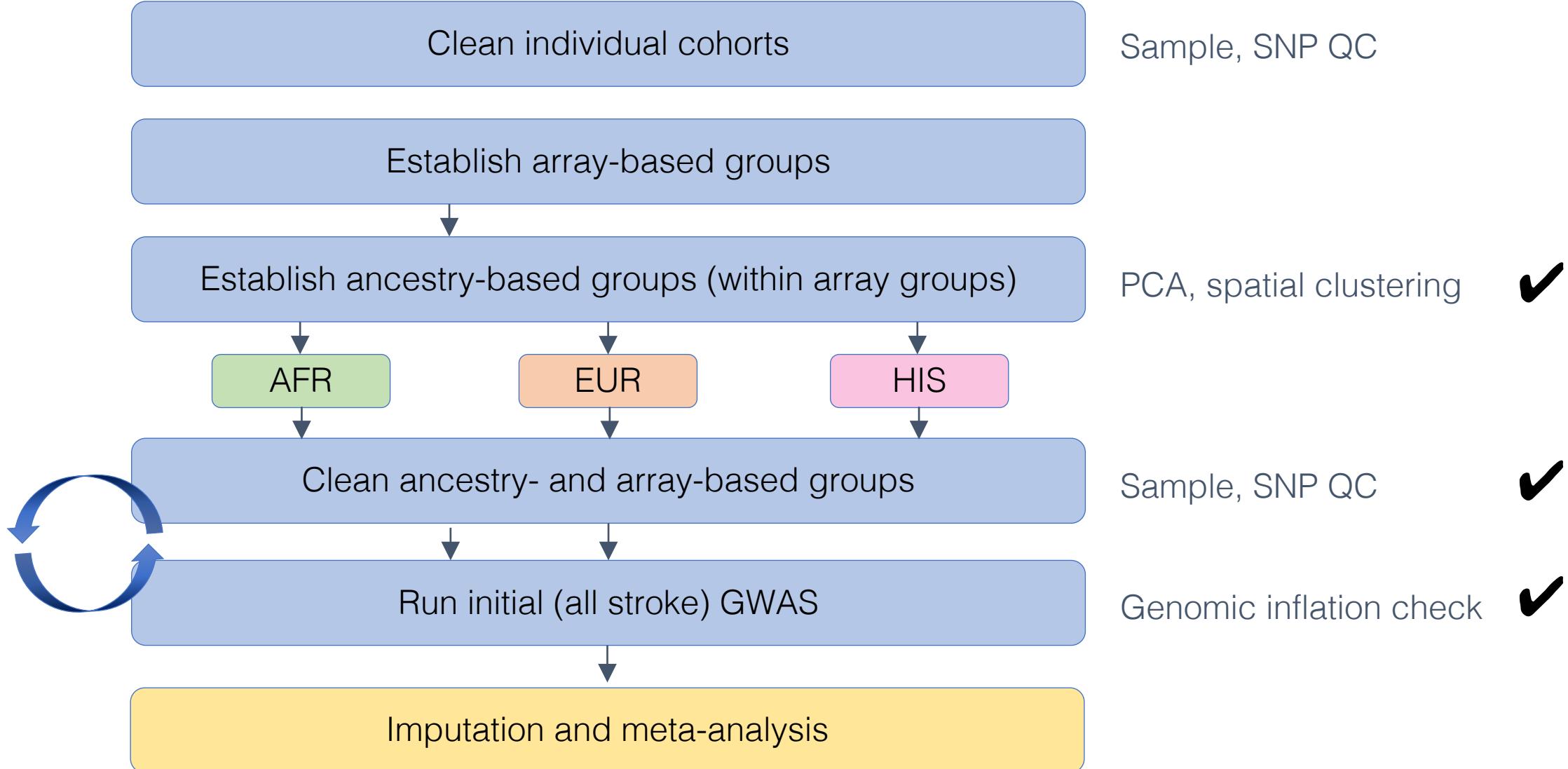


SNP quality control \longleftrightarrow association testing

Quality control step	SiGN exclusion threshold
Missingness	> 1%
Minor allele frequency	< 1%
Hardy-Weinberg equilibrium	$p < 10^{-3}$ (controls only)
Frequency differences between control groups	$p < 10^{-3}$
Frequency differences between controls and 1000 Genomes	$p < 10^{-3}$ (European controls only)
Differential missingness	$p < 10^{-3}$
Mismatching genotypes across arrays	≥ 1 mismatch (for samples genotyped on >1 array)

This is what we expect





1. Know your data

2. Know your phenotype

3. Know your limitations

Quality control has its limitations

Case-control group array	Sample ancestry	All ischemic stroke Genomic inflation (λ)		
		Genotyped SNPs (post QC)	Imputed SNPs	Imputed SNPs (post QC)
610k	European	1.020	1.181	1.026
660k	European	1.056	1.127	1.045
1M	African	1.017	1.086	1.027
1M	European	1.007	1.123	1.014

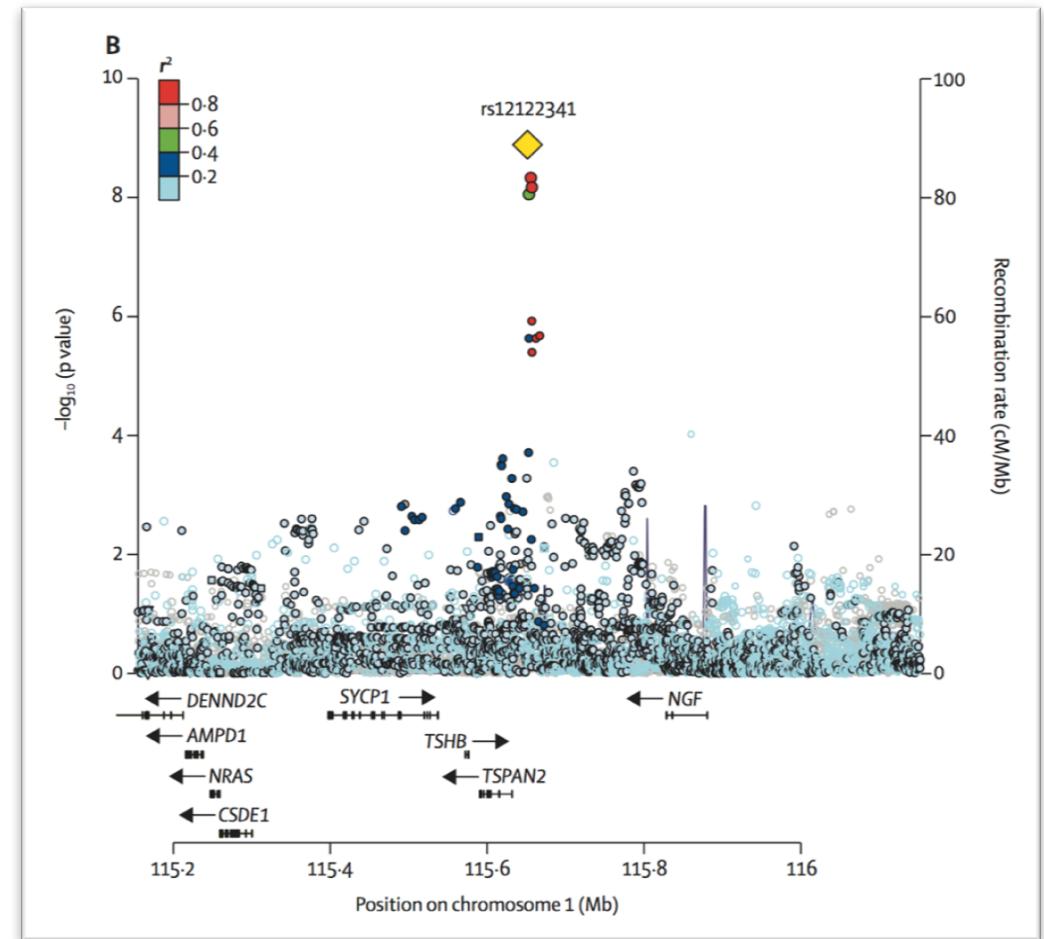
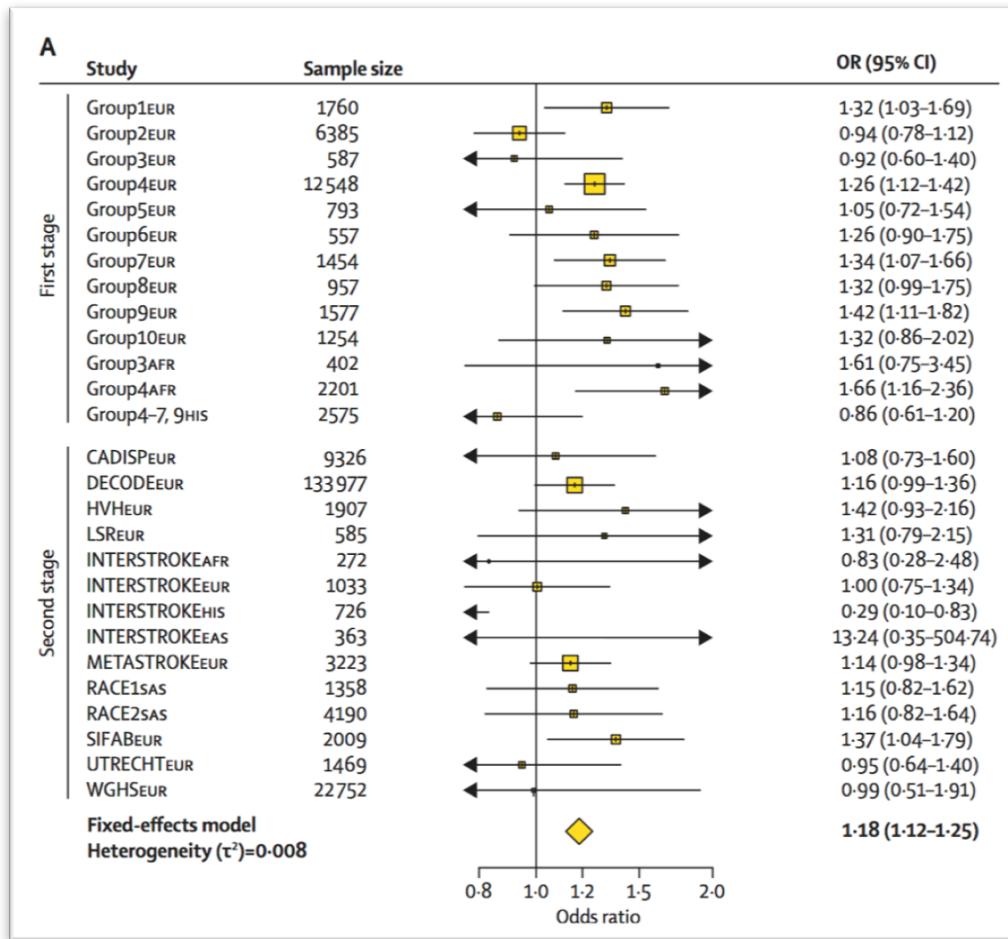
Quality control has its limitations

Sample array	Sample ancestry	All ischemic stroke Genomic inflation (λ)		
		Genotyped SNPs (post QC)	Imputed SNPs	Imputed SNPs (post QC)
610k	European	1.020	1.181	1.026
660k	European	1.056	1.127	1.045
1M	African	1.017	1.086	1.027
1M	European	1.007	1.123	1.014



Drop SNPs with frequency < 1%

An association to large vessel stroke



Conclusions

- Characteristics of your study design can inform analytic path
- Can ancestry match samples using PCA-based methods
- Can use (highly) stringent SNP quality control to reduce false positives
- Be aware of potential stratification even after QC is complete

Acknowledgements

The NINDS Stroke Genetics Network (SiGN) and the Internal Stroke Genetics Consortium (ISGC)

SiGN Analysis Group

Brackie Mitchell	Tushar Dave	Huichun Xu	Paul de Bakker
Steven Kittner	Quenna Wong	Mary J Sparks	
Patrick McArdle	Cathie Laurie	Sara Pulit	

SiGN Study Design

Brackie Mitchell	Hakan Ay	Arne Lindgren	Vincent Thijs
Steven Kittner	Paul de Bakker	Sara Pulit	Dan Woo
Jonathan Rosand	Katrina Gwinn	Cathie Sudlow	Brad Worrall

SiGN/ISGC Writing Committee

Jonathan Rosand	Paul de Bakker	Arne Lindgren	Cathie Sudlow
Brackie Mitchell	Katrina Gwinn	James Meschia	Vincent Thijs
Hakan Ay	Steven Kittner	Sara Pulit	Daniel Woo
			Brad Worrall

SiGN Steering Committee

Donna Arnett	James Meschia
Oscar Benavente	Kathryn Rexrode
John Cole	Jonathan Rosand
Martin Dichgans	Peter Rothwell
Raji Grewal	Tatjana Rundek
Christina Jern	Ralph Sacco
Jordi Conde	Reinhold Schmidt
Julia Johnson	Pankaj Sharma
Steven Kittner	Agnieszka Slowik
Jin-Moo Lee	Cathie Sudlow
Christopher Levi	Vincent Thijs
Arne Lindgren	Sylvia Wassertheil-Smoller
Hugh Markus	
Olle Melander	Daniel Woo
	Brad Worrall