Combining cases and publicly-available controls

for discovery of common disease loci through GWAS

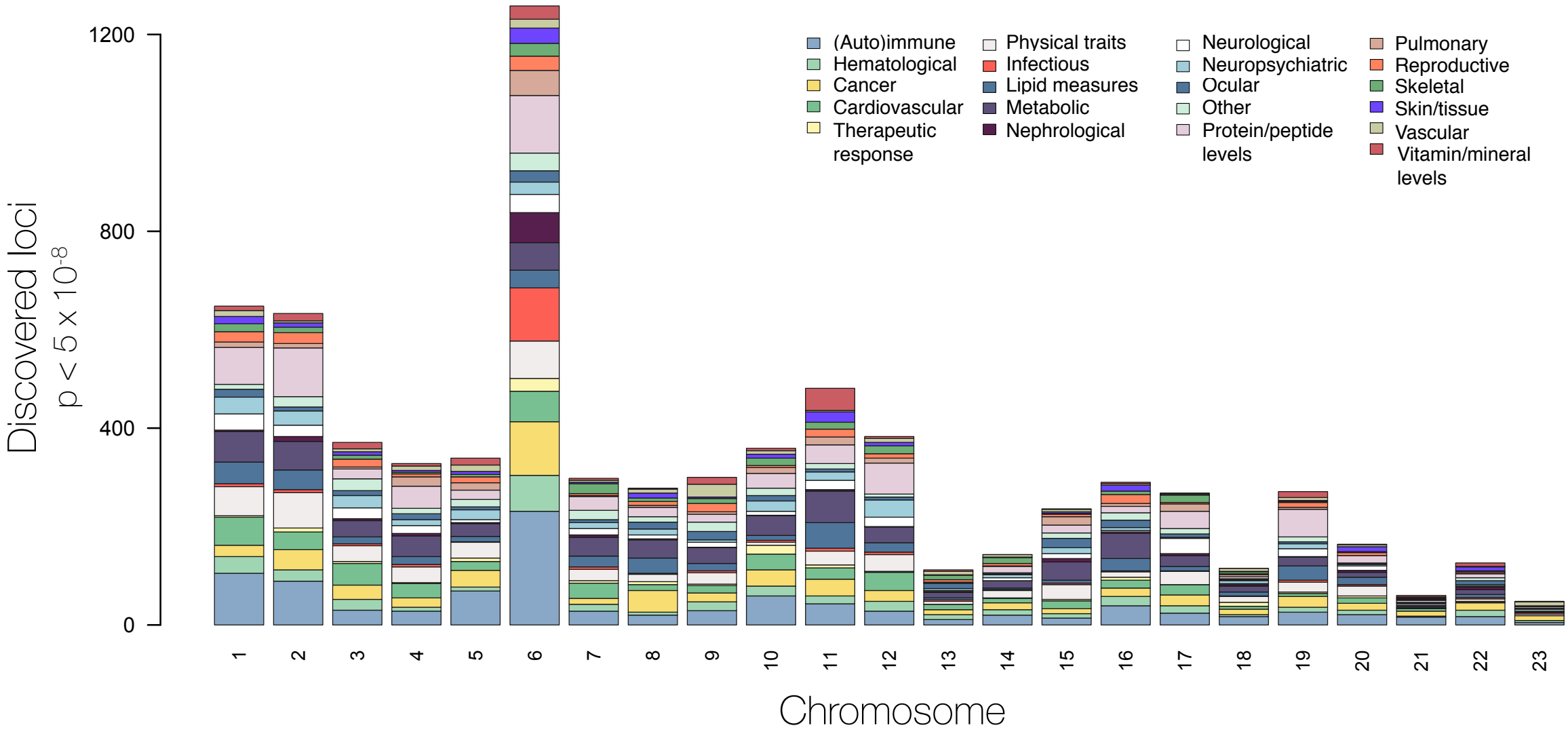
Sara L. Pulit

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

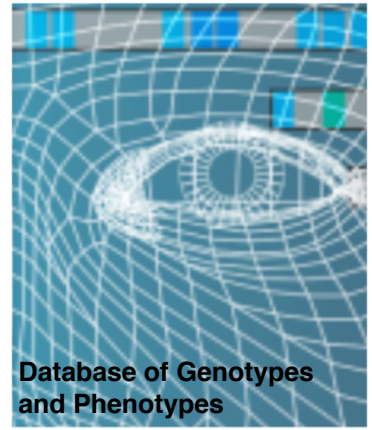
China Kadoorie Biobank

N = 510,000
East Asian ancestry

Human Genome Diversity Project

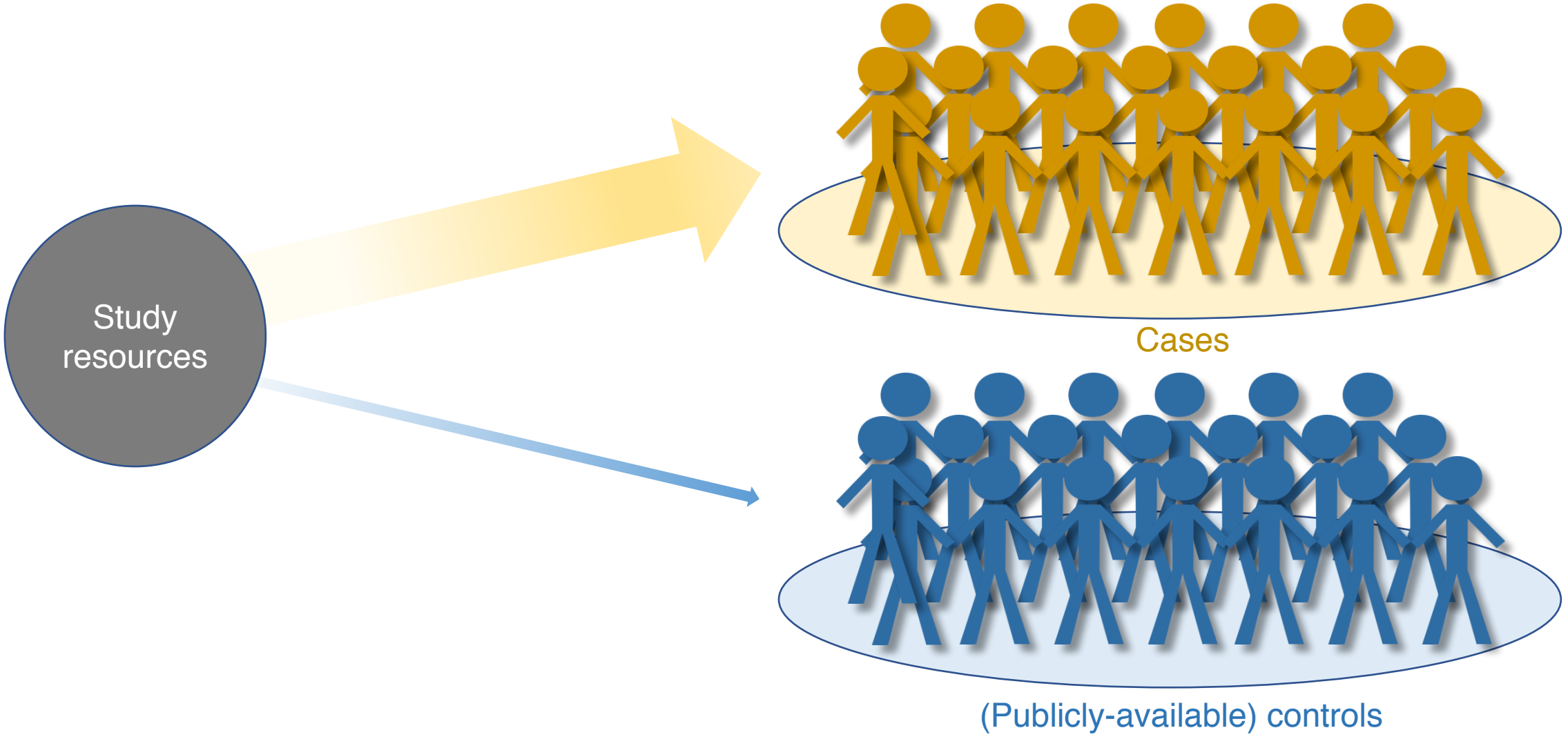
N = 1,043
Multi-ancestry

dbGaP

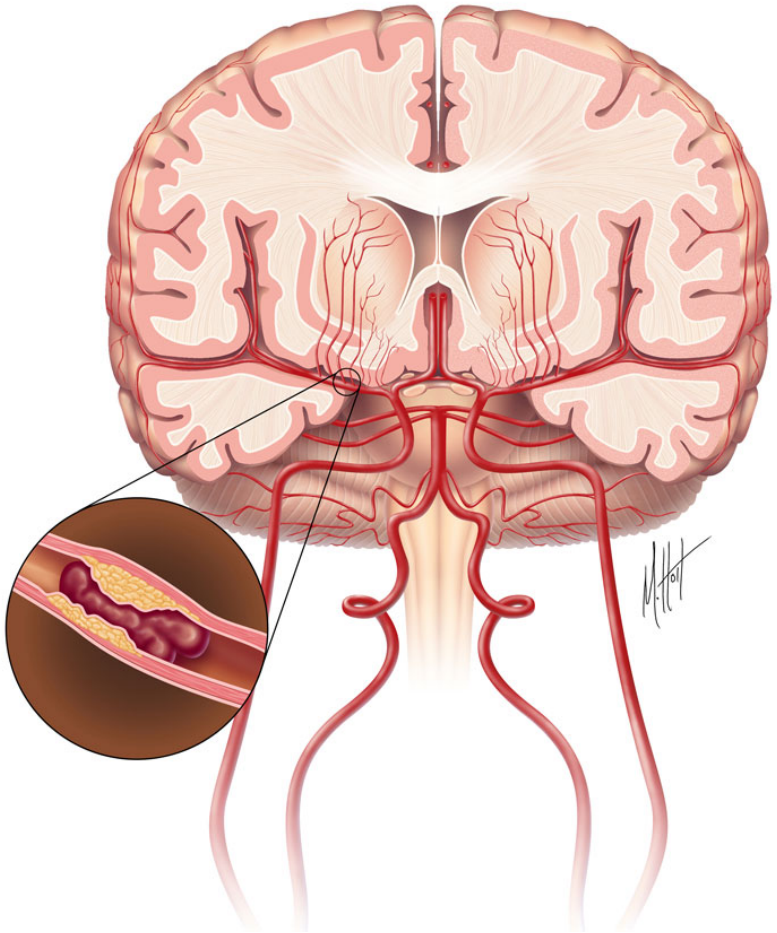


N > 500,000
Multi-ancestry

Investing resources in cases only

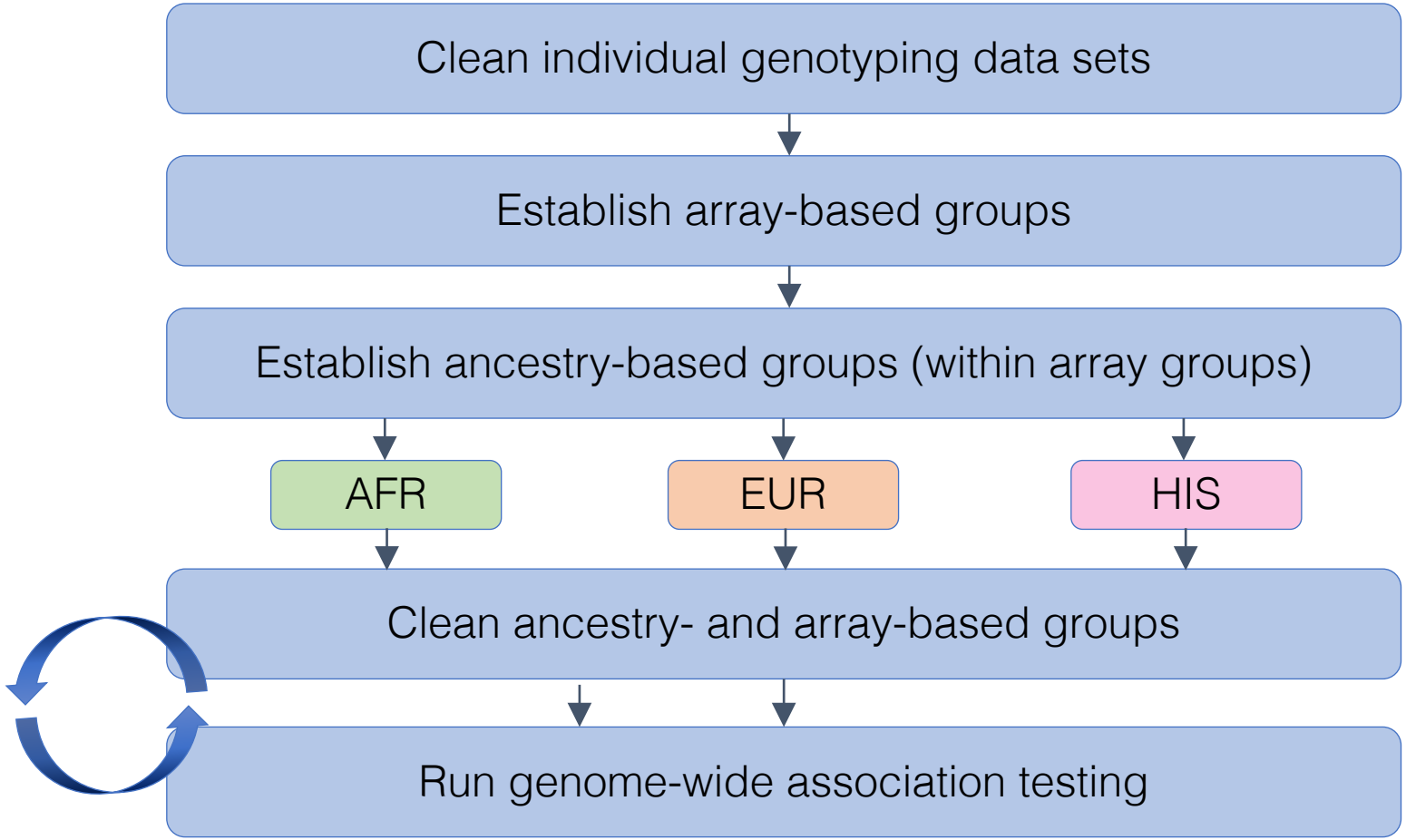


The Stroke Genetics Network (SiGN)



- 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

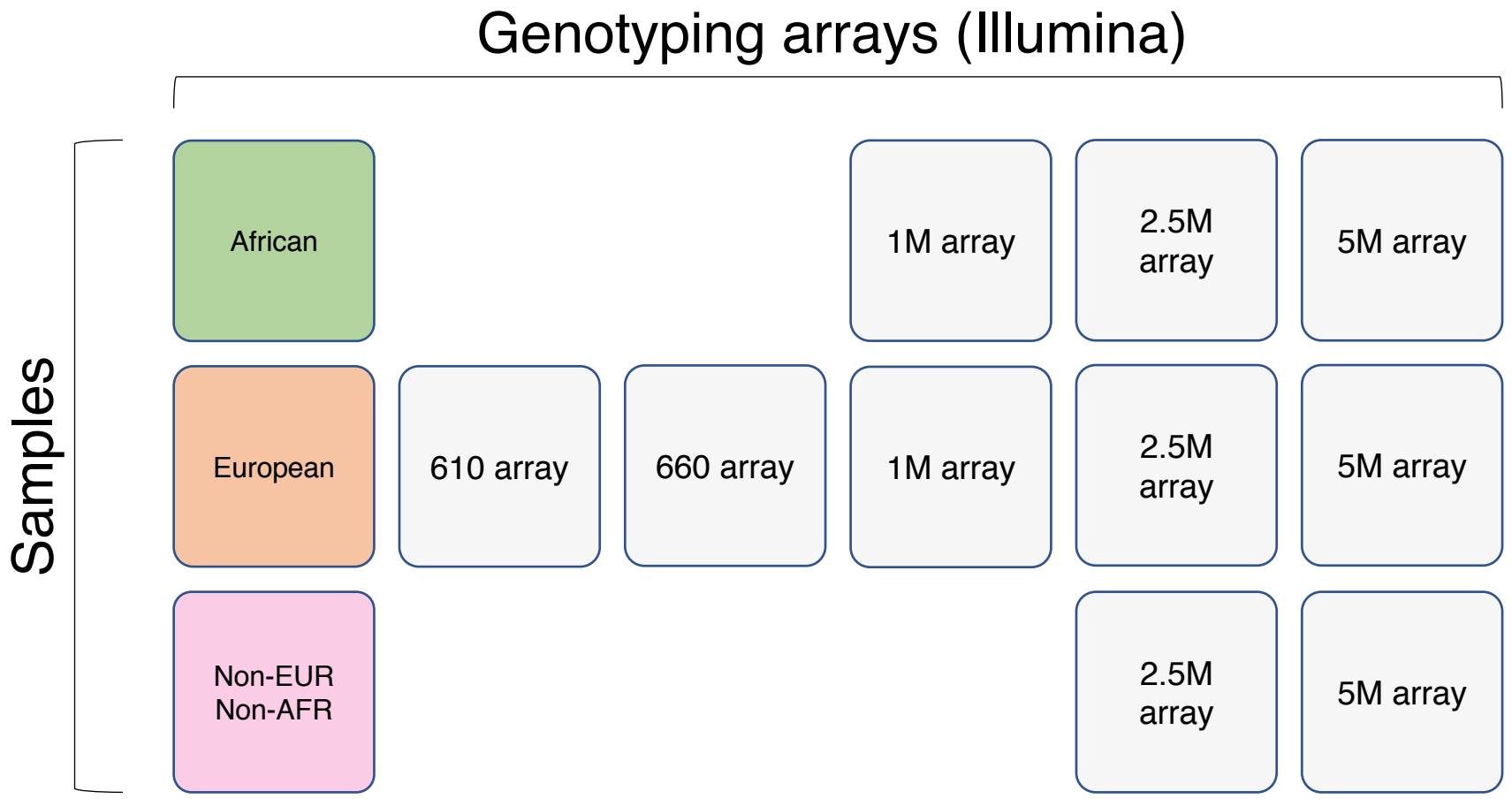
3. Know your limitations

1. Know your data

2. Know your phenotype

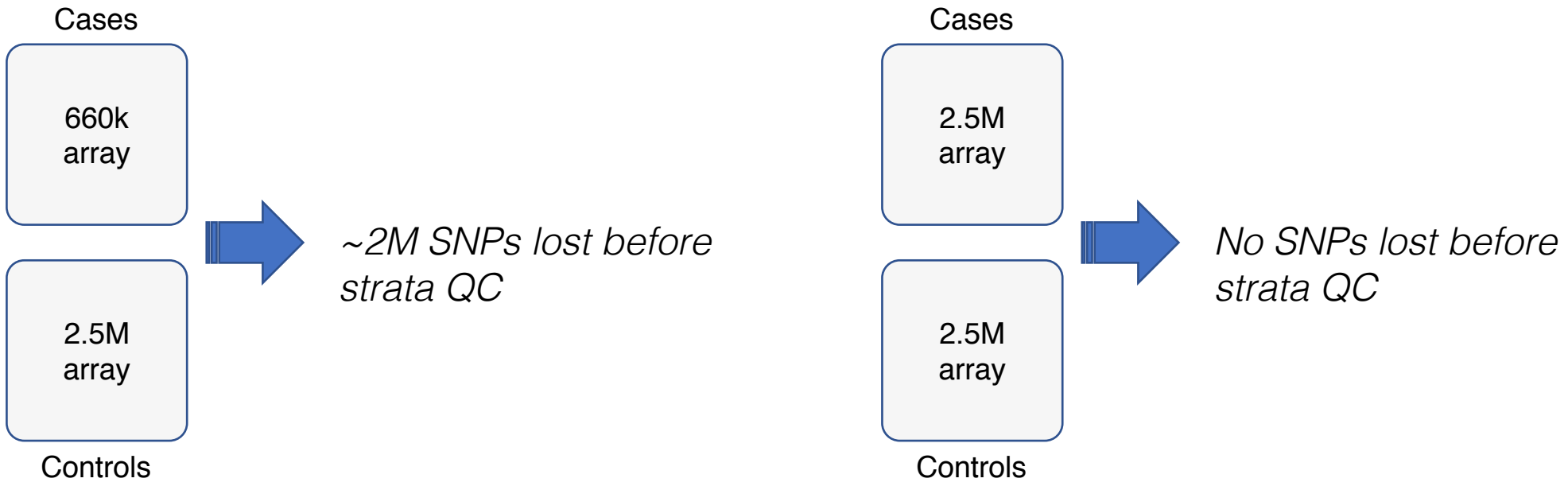
3. Know your limitations

Know your data



Creating case-control strata

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

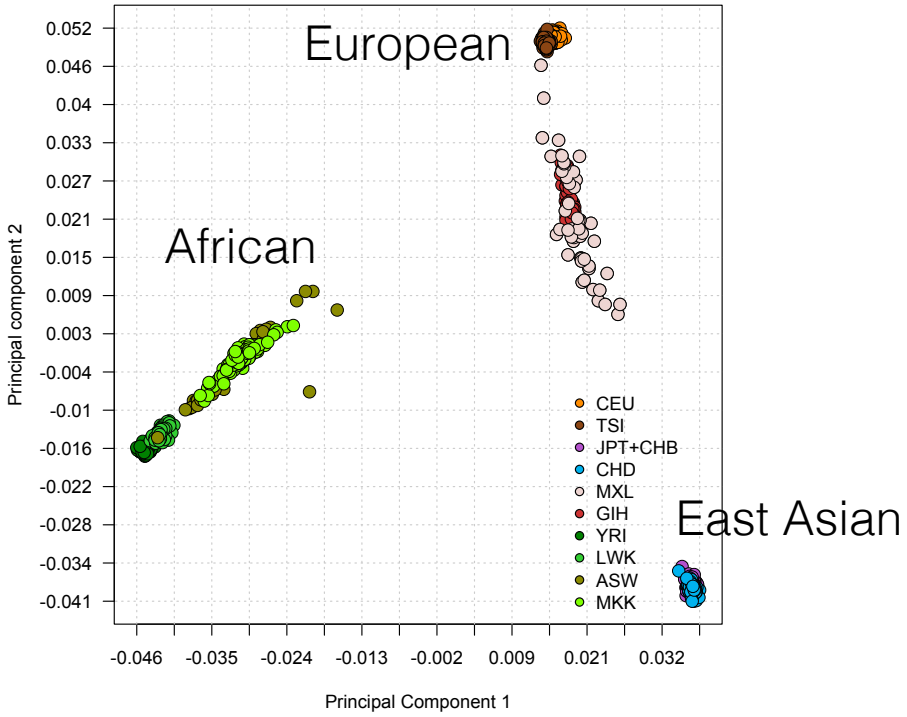


Creating case-control groups

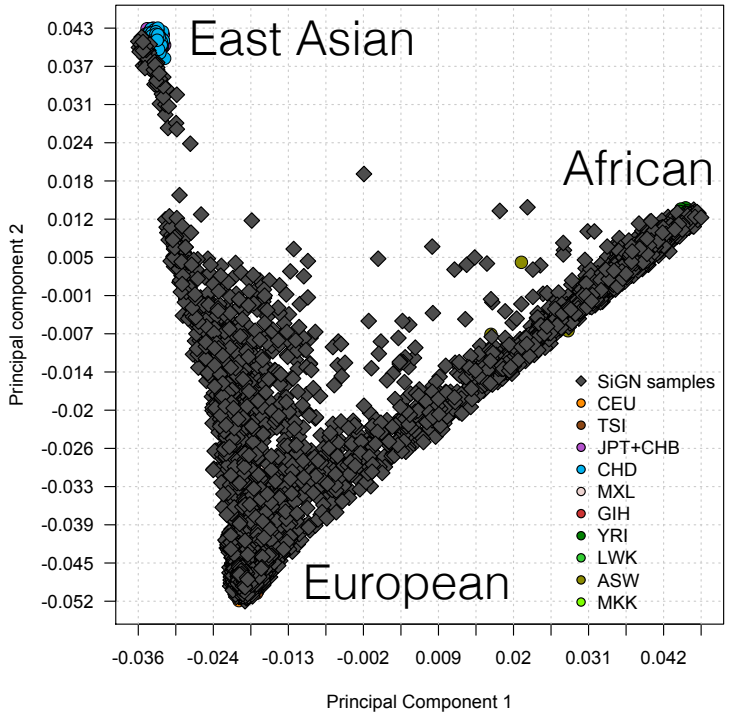
- Match samples by genotyping array
  - Maximize the total number of SNPs available for genotype imputation
- When possible, maintain a 1:2 case-control ratio within groups
- When possible, merge cohorts collected in the same geographic region
  - Maximize chance of ancestry-matched samples
- Identify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

PCA-based approaches for matching samples by ancestry

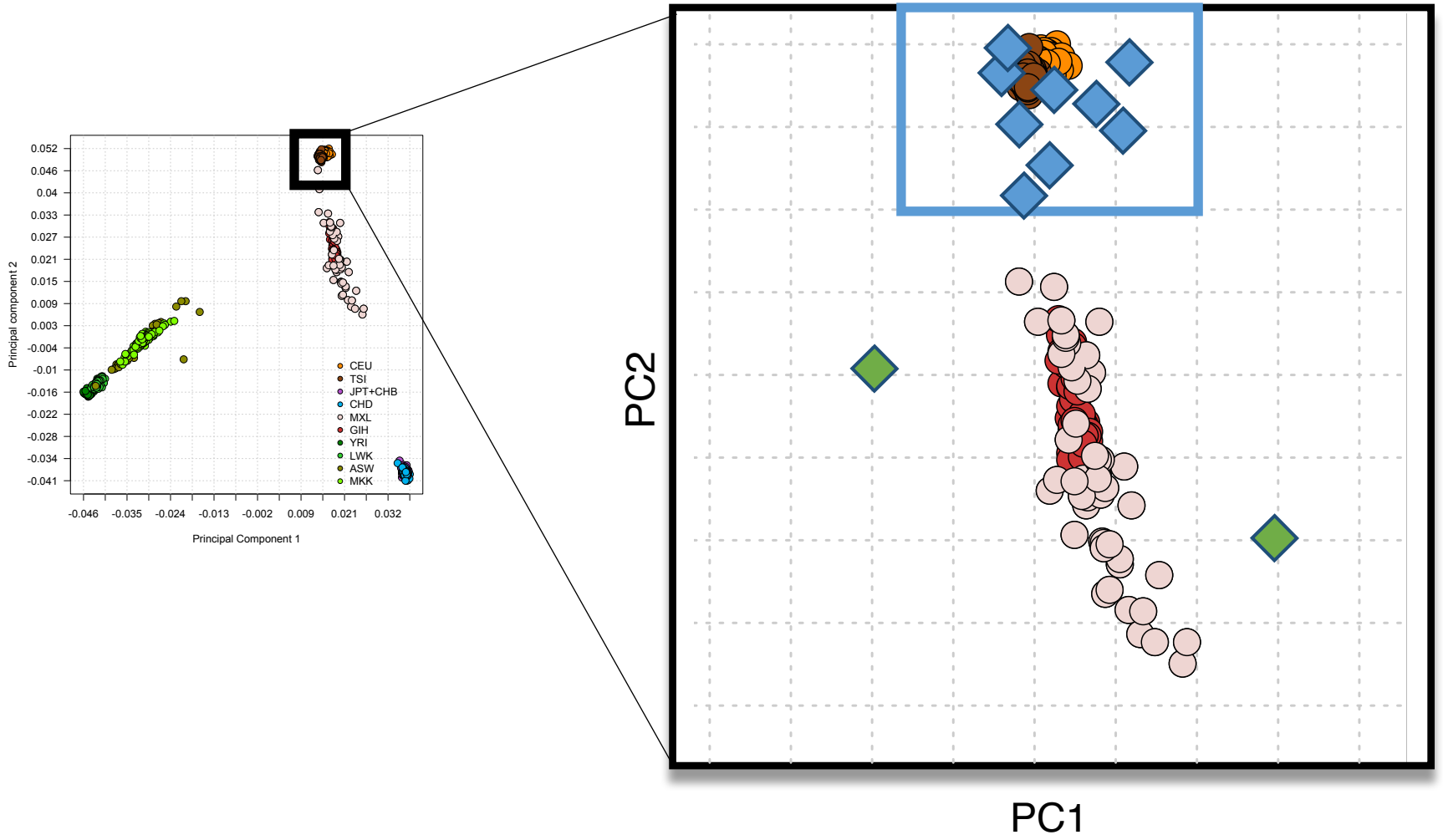
Global populations (HapMap 3)



SiGN data



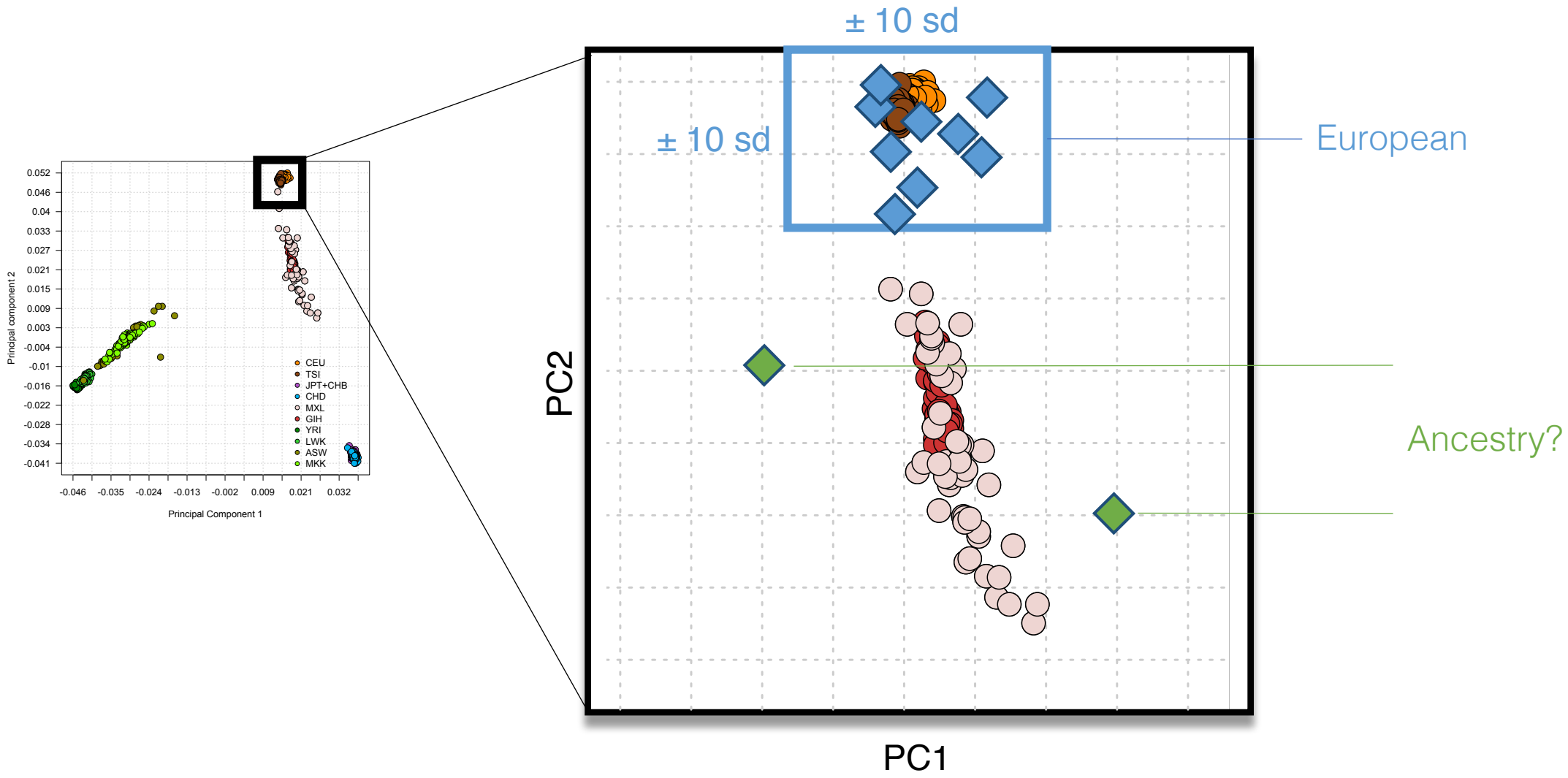
Defining European-ancestry groups with PCA



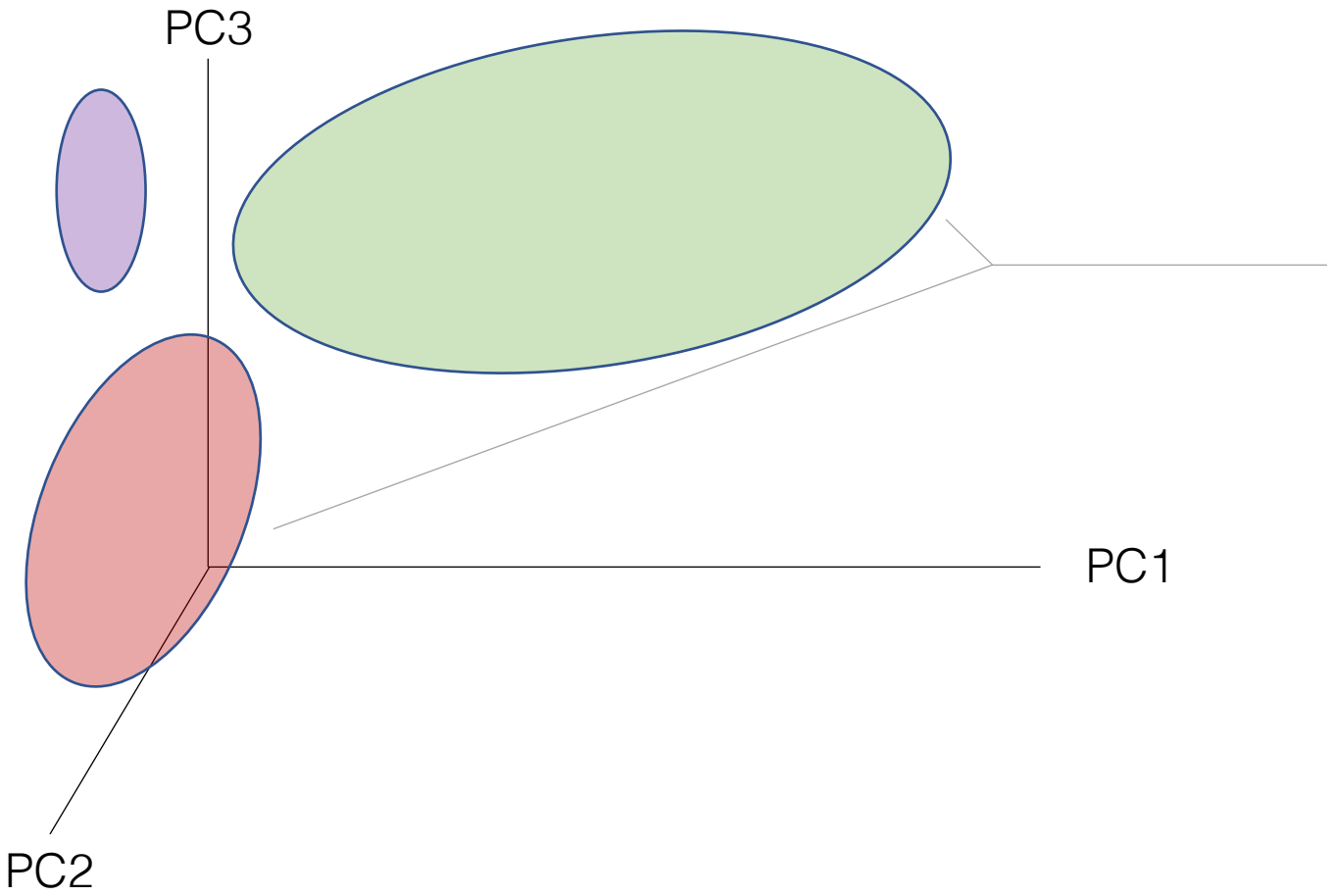
Northern/Central European ancestry (CEU)

Tuscans in Italy (TSI)

Defining European-ancestry groups with PCA

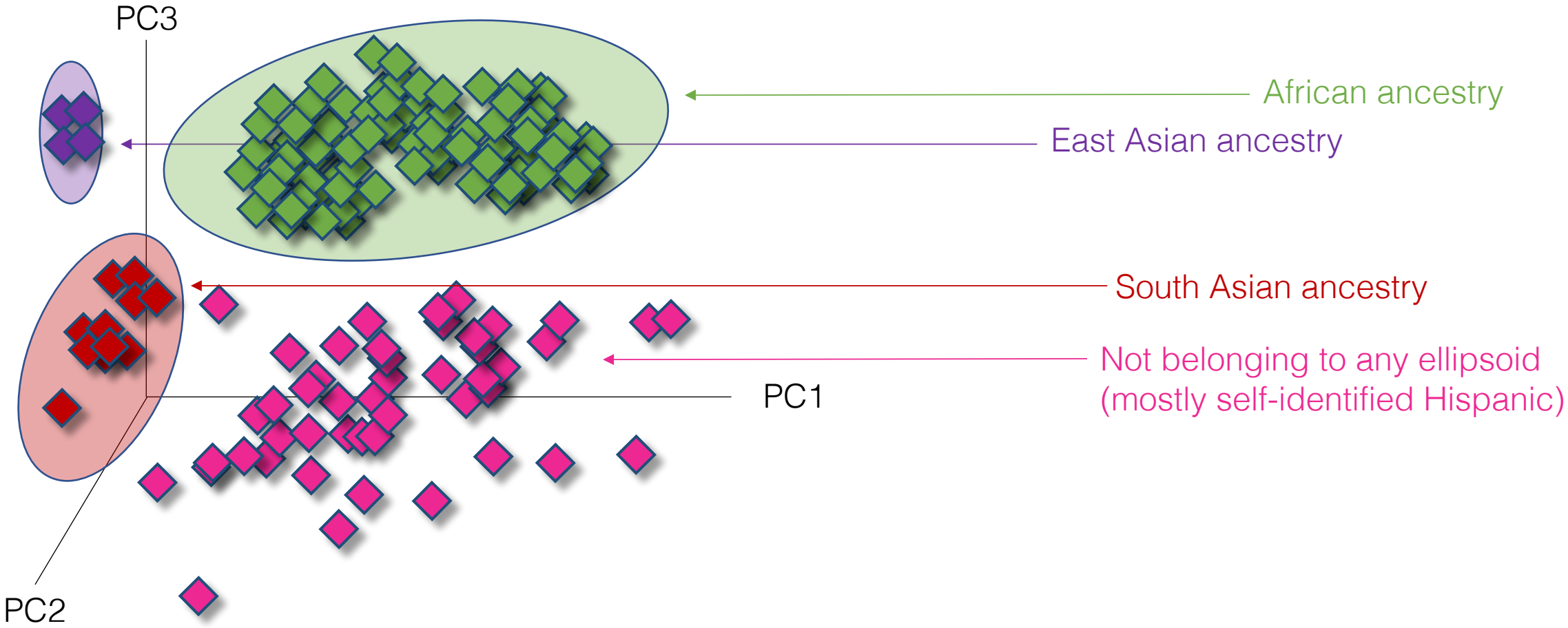


Defining additional ancestral groups with hyperellipsoid analysis

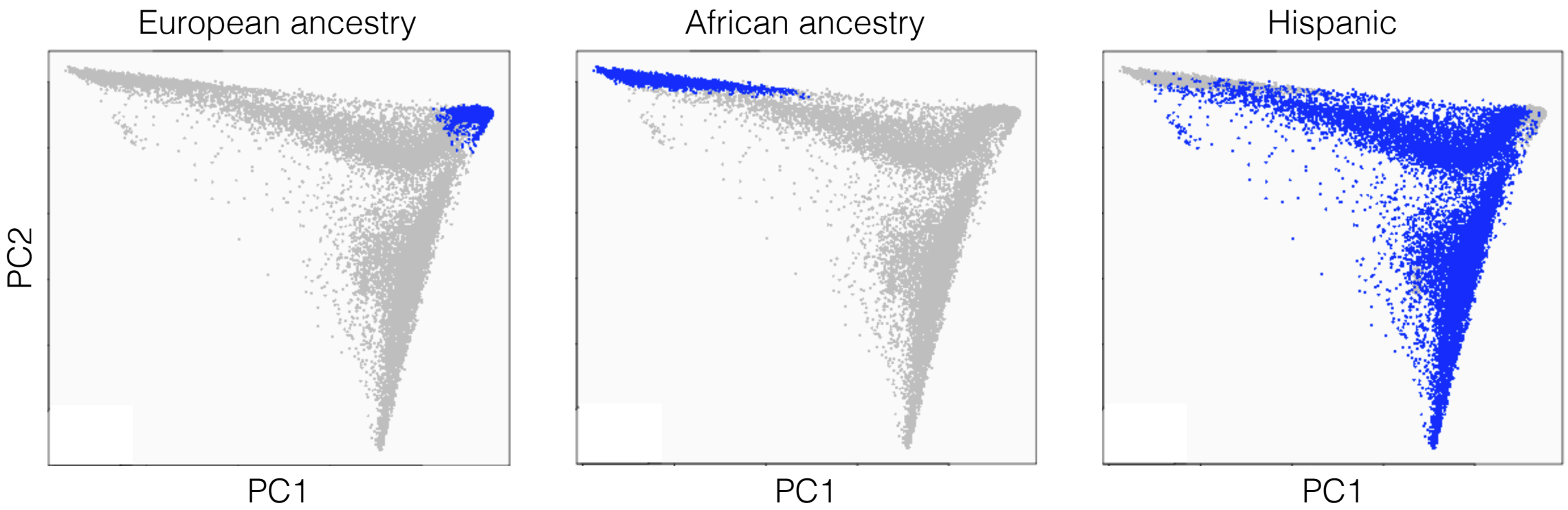


Ellipsoid space: informed by PCA within self-identified racial groups

Defining additional ancestral groups with hyperellipsoid analysis



Ancestral groups in two-dimensional PCA space



Case-control matching across 10 PCs

Case

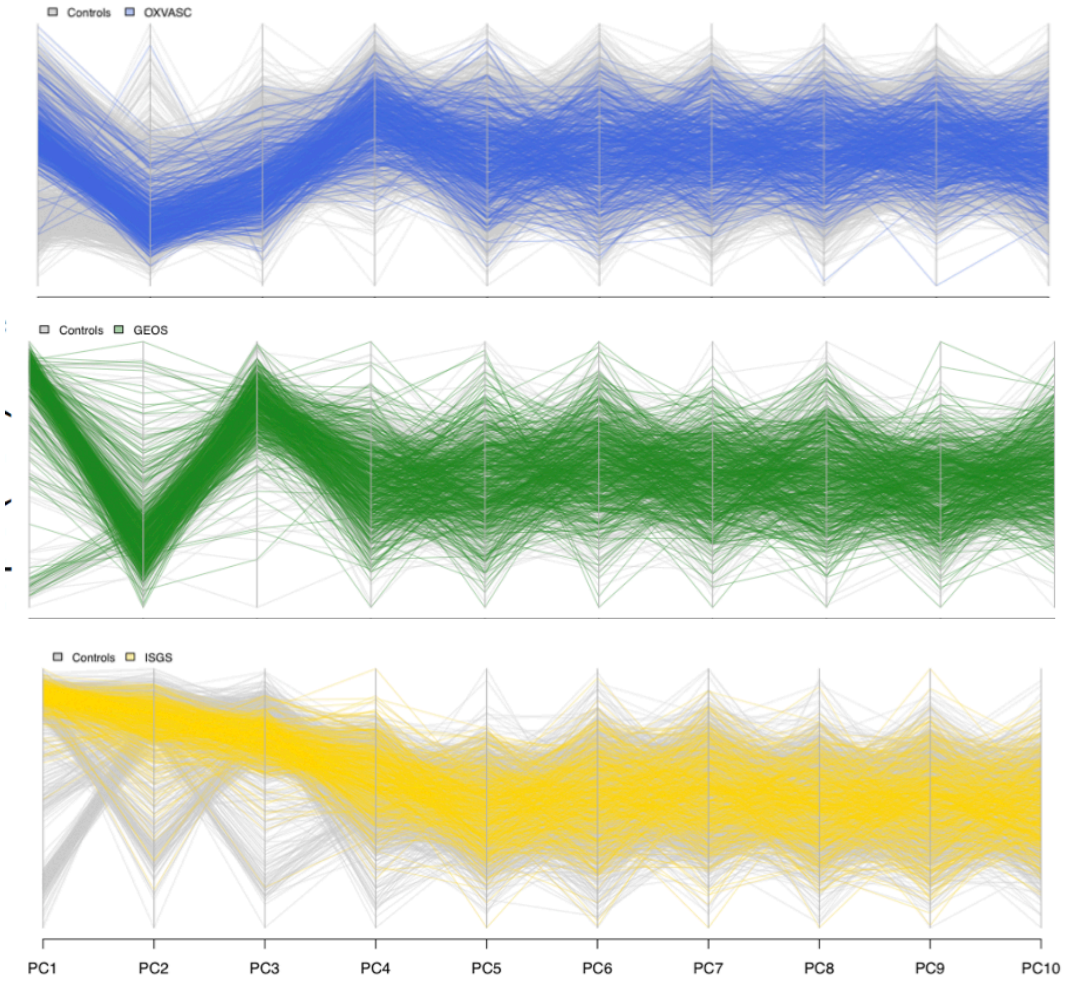
Control

Case

Control

Case

Control

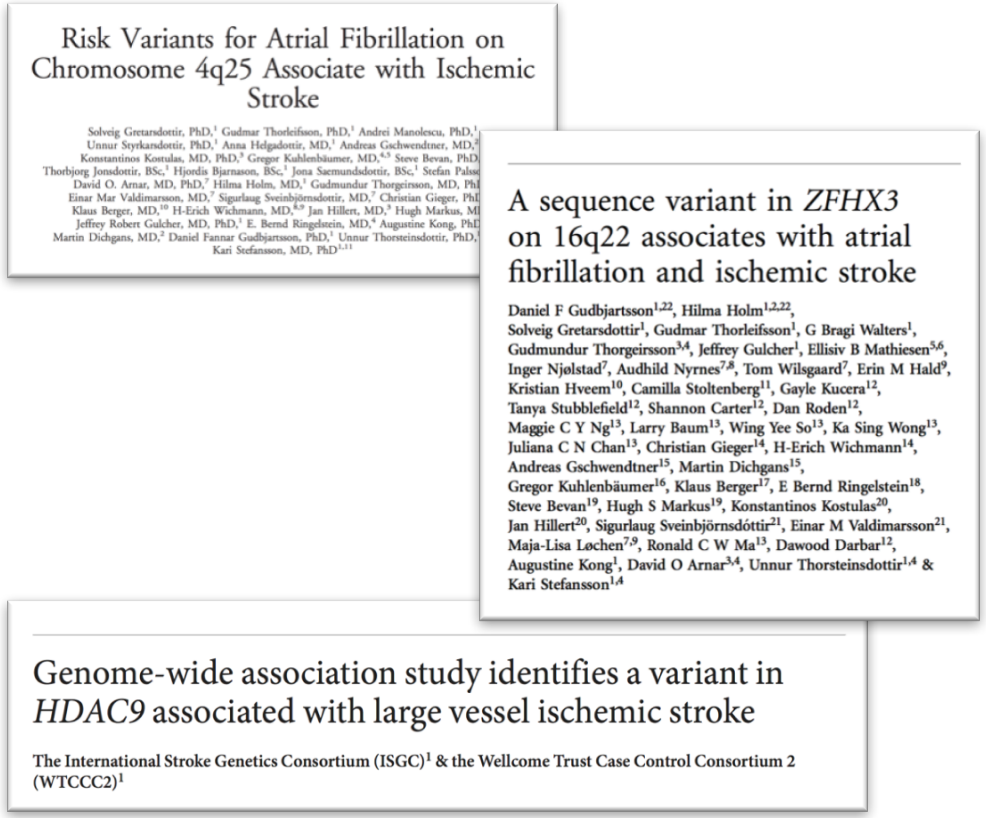


Know your data

Know your phenotype

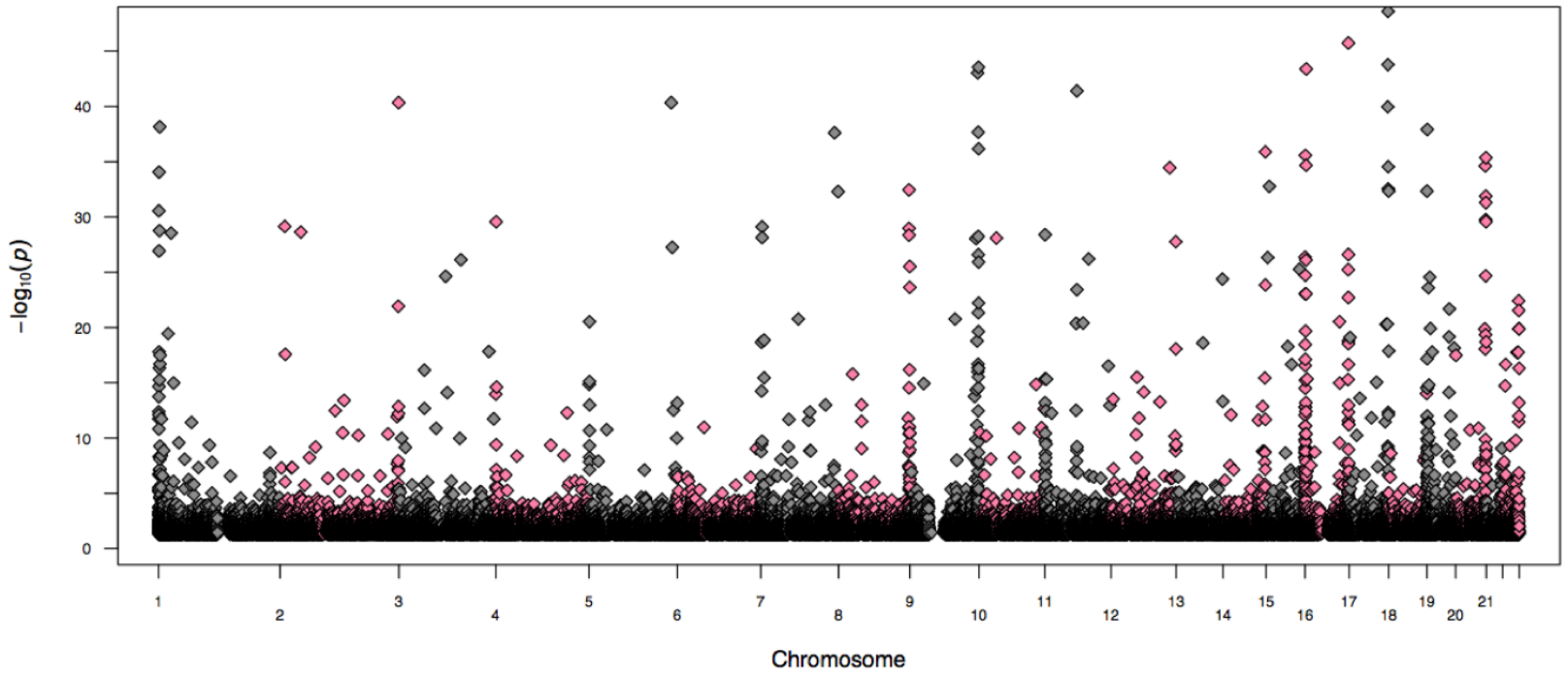
Know your limitations

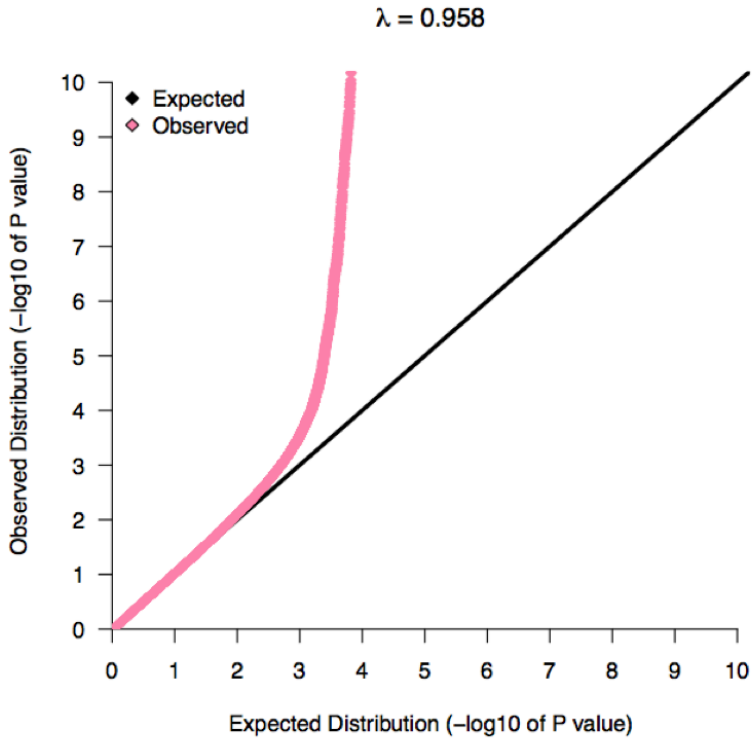
Stroke: what should we expect?



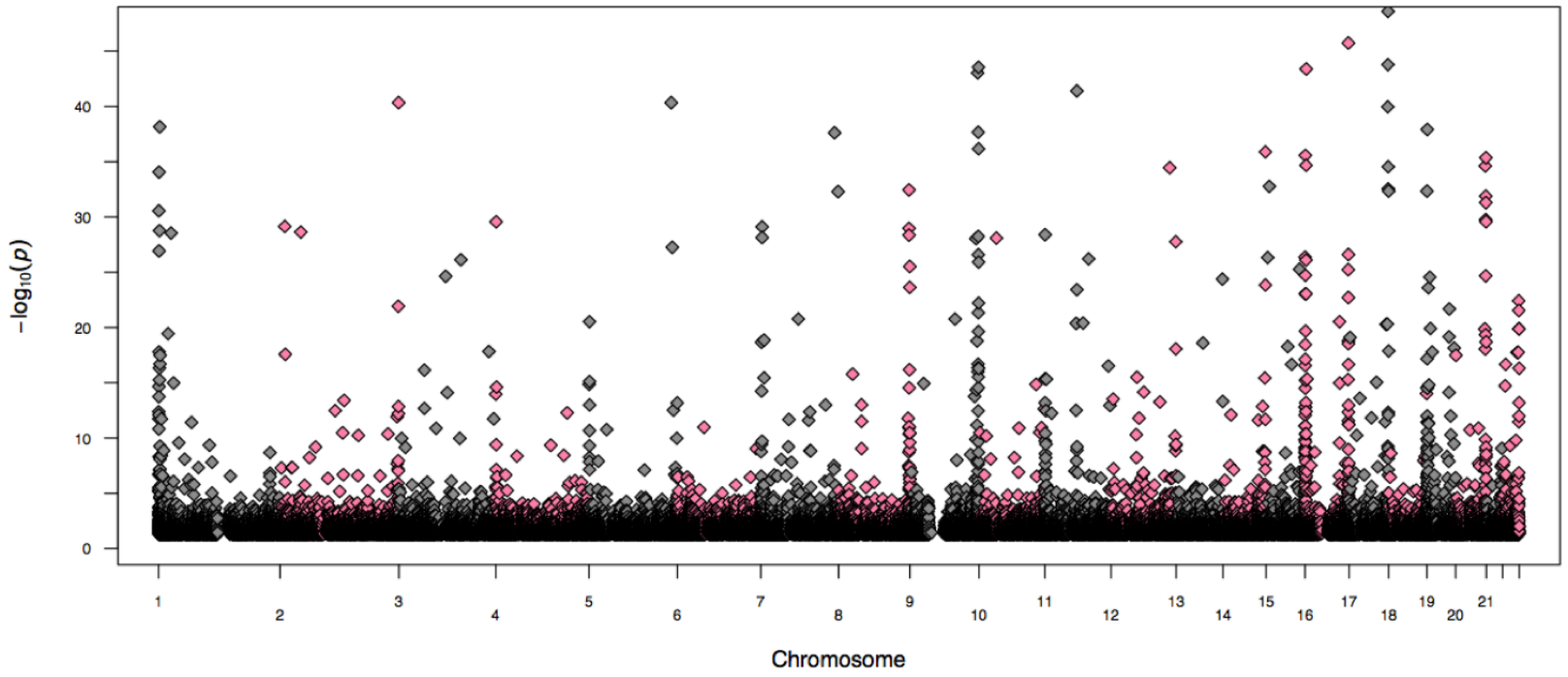
- GWAS have revealed a modest number of hits
  - True positives
- Common, complex disease
  - Likely underpinned by modestly penetrant SNPs
- Associated SNPs are typically associated to a subtype
  - Reduce sample size (and power)

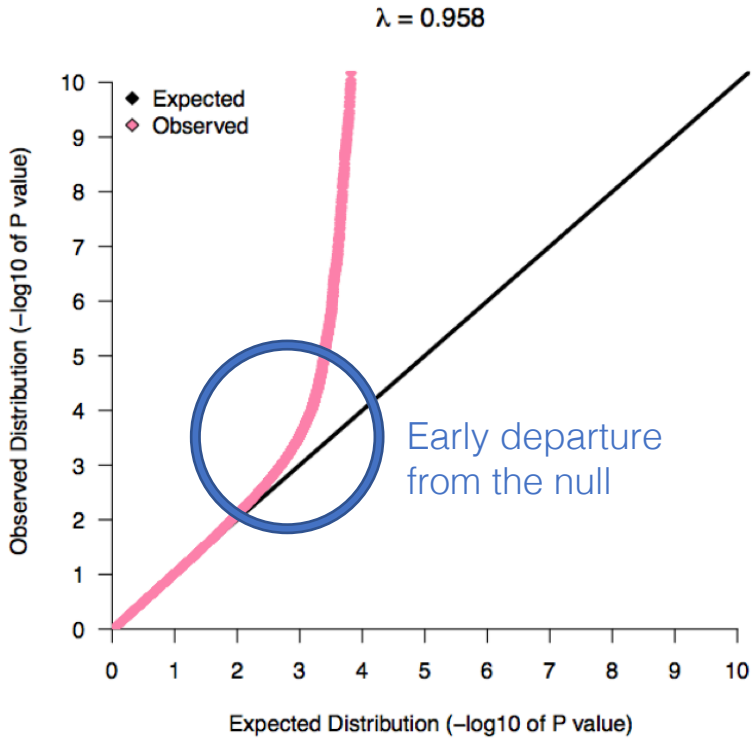
This is not what we expect





This is not what we expect

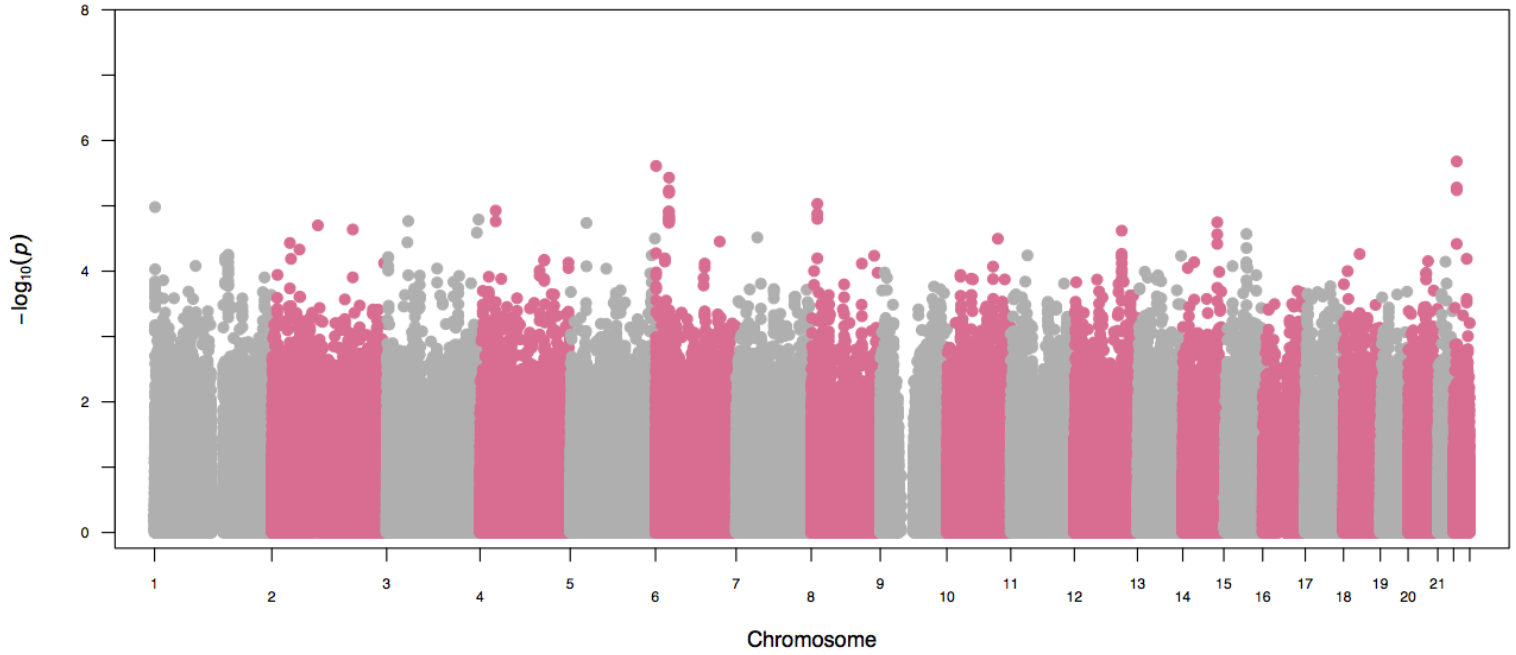


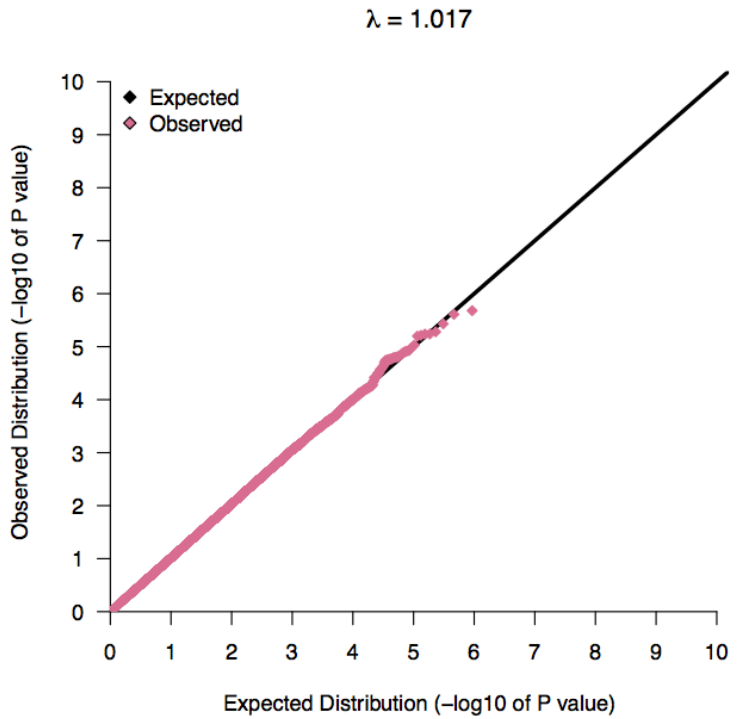


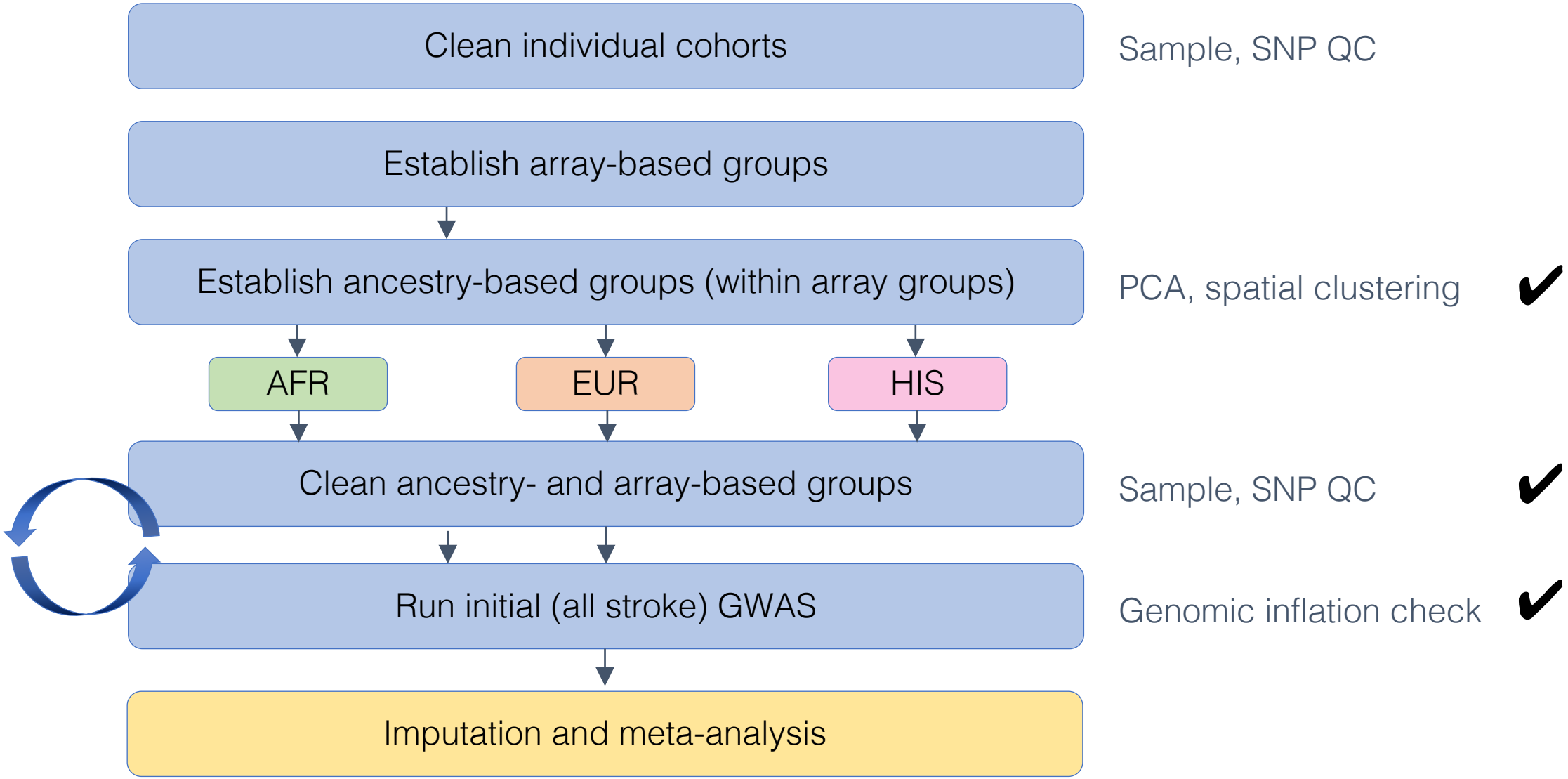
SNP quality control ⇌ association testing

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






Clean individual cohorts
Sample, SNP QC
Establish array-based groups
Establish ancestry-based groups (within array groups)
PCA, spatial clustering
AFR
EUR
HIS
Clean ancestry- and array-based groups
Sample, SNP QC
Run initial (all stroke) GWAS
Genomic inflation check
Imputation and meta-analysis

1. Know your data

2. Know your phenotype

3. Know your limitations

Quality control has its limitations

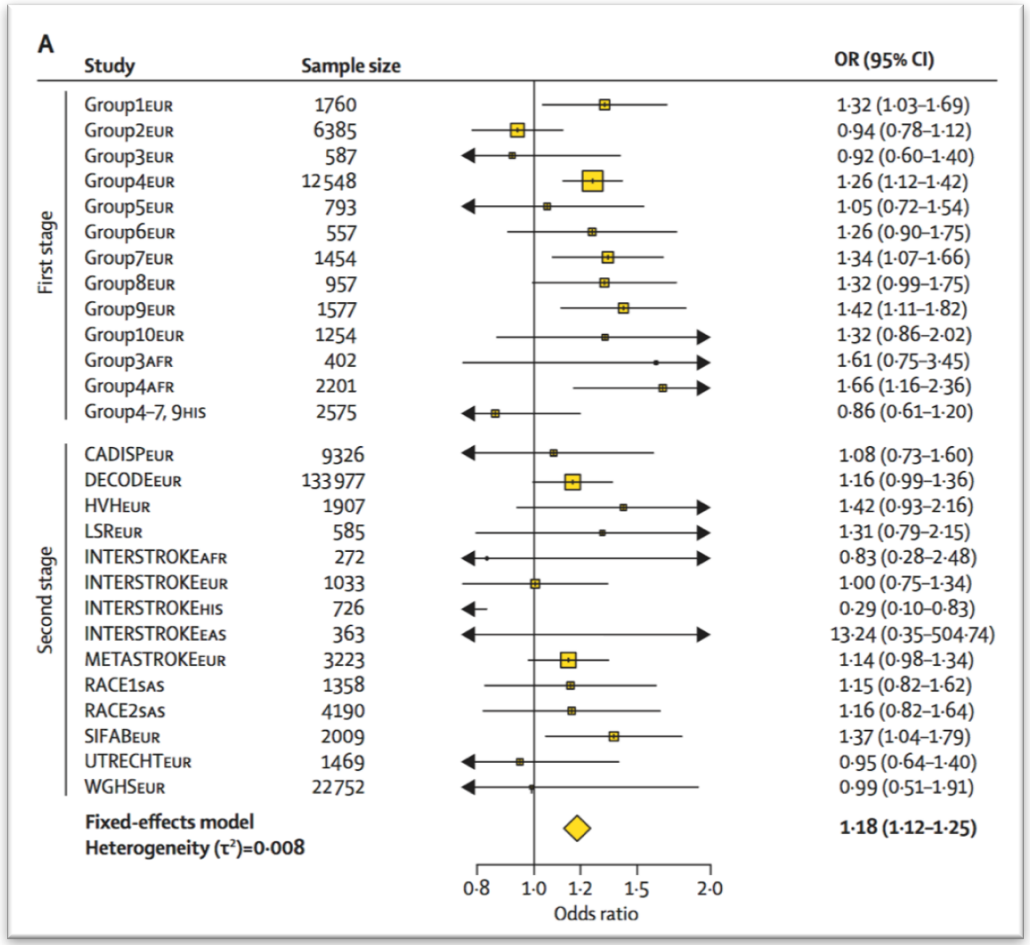
Case-control group array	Sample ancestry	All ischemic stroke Genomic inflation (λ)		
		Genotyped SNPs (post QC)	Imputed SNPs	Imputed SNPs (post QC)
610k	European	1.020	1.181	1.026
660k	European	1.056	1.127	1.045
1M	African	1.017	1.086	1.027
1M	European	1.007	1.123	1.014

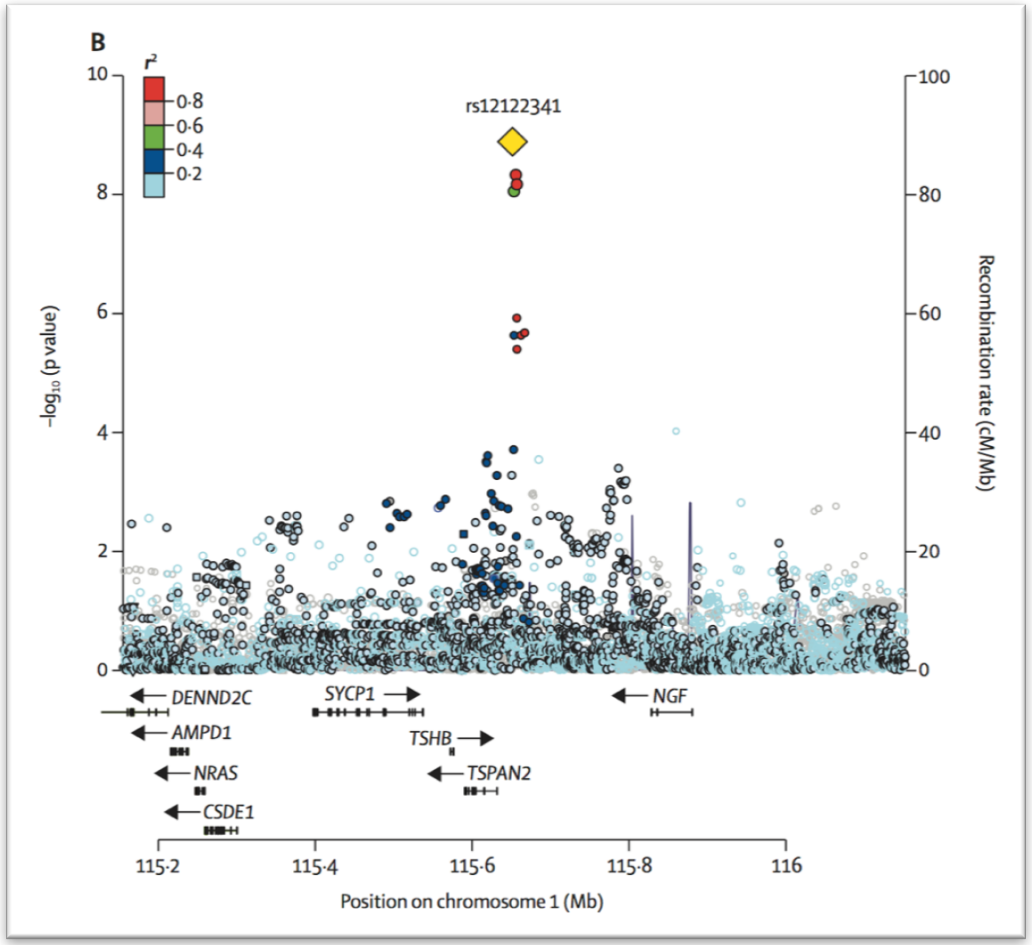
Quality control has its limitations

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

SiGN Study Design

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

SiGN/ISGC Writing Committee

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

SiGN Steering Committee

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