

SCHOOL OF PUBLIC HEALTH

EPIDEMIOLOGY

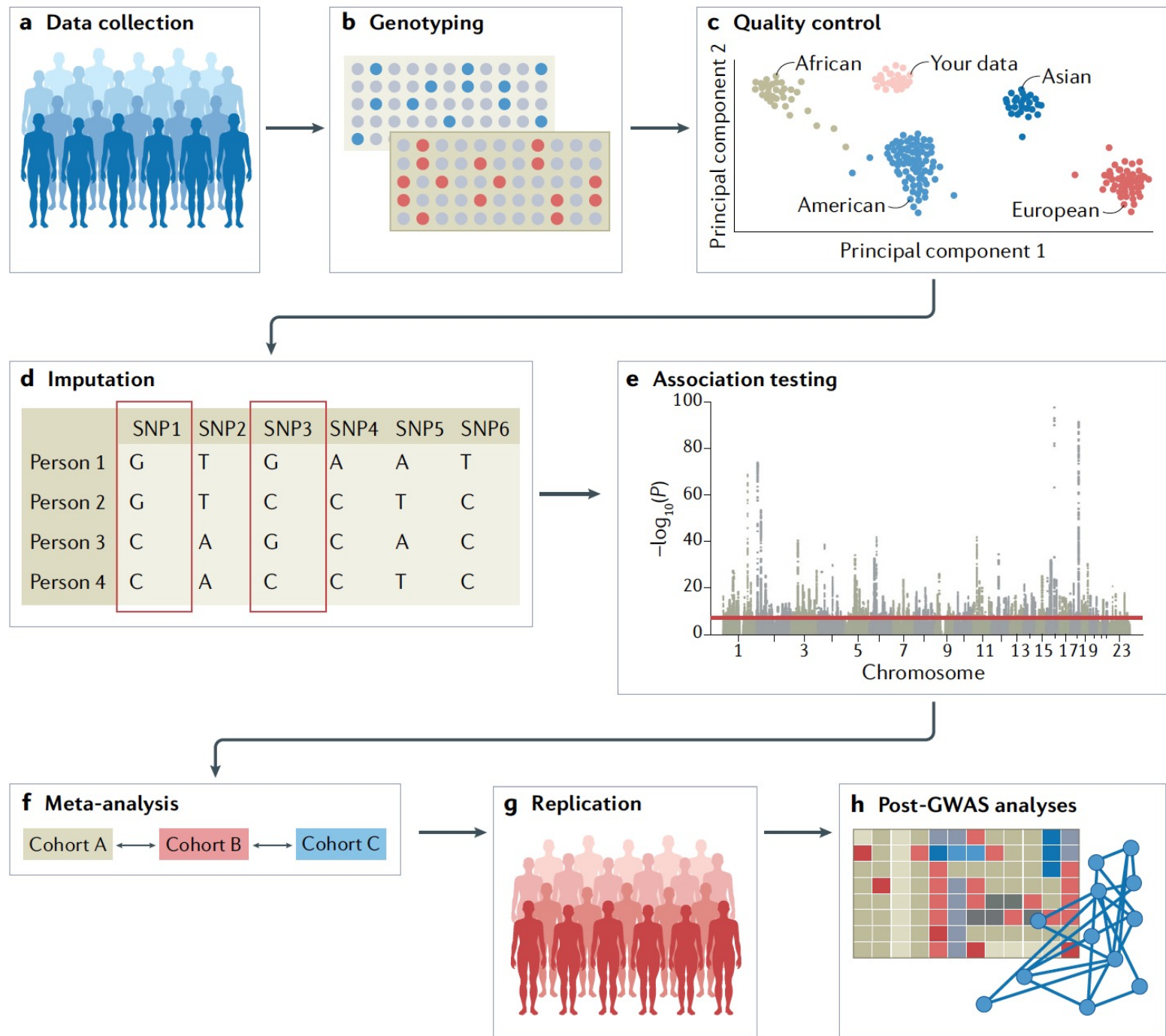
UNIVERSITY *of* WASHINGTON

Session 8:

Large-scale genetic association studies - GWAS



Workflow for GWAS



Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations

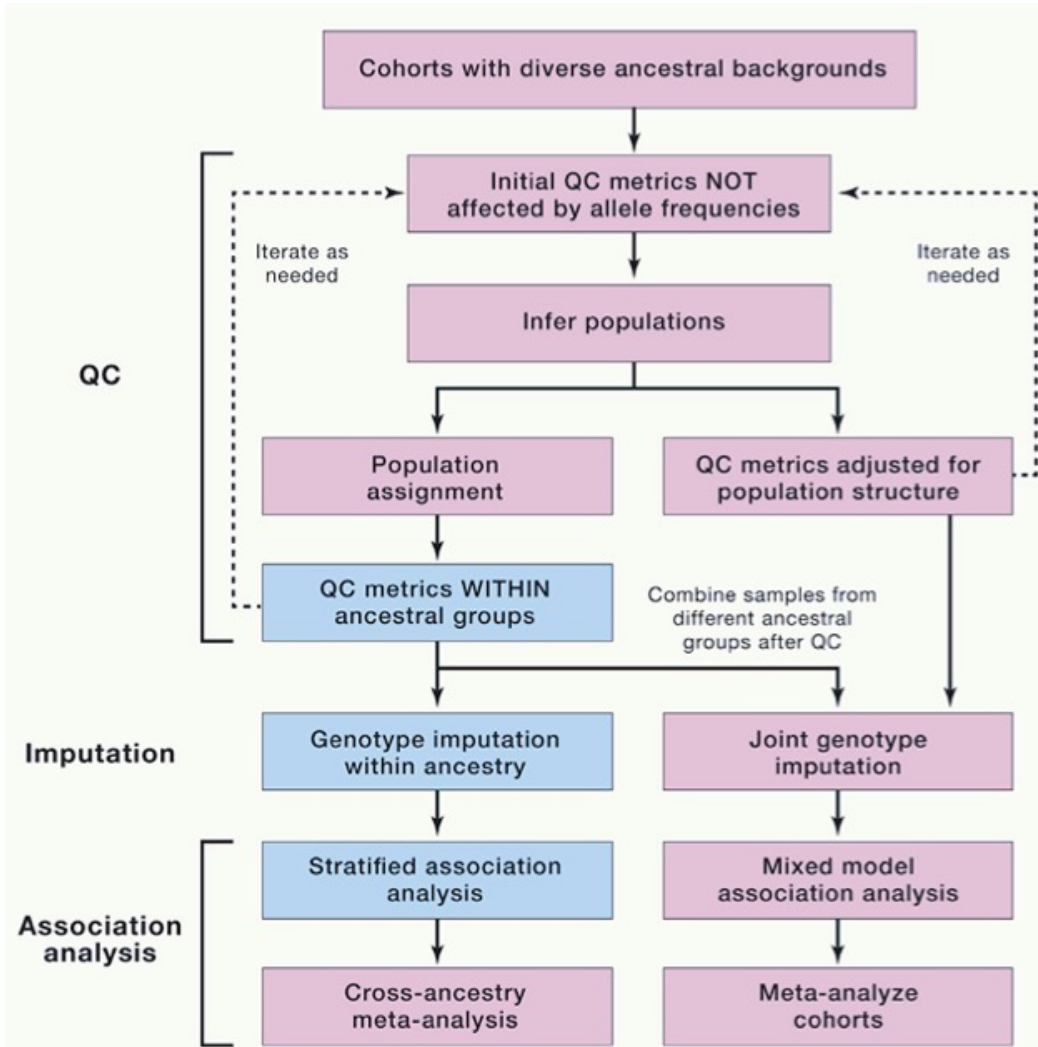
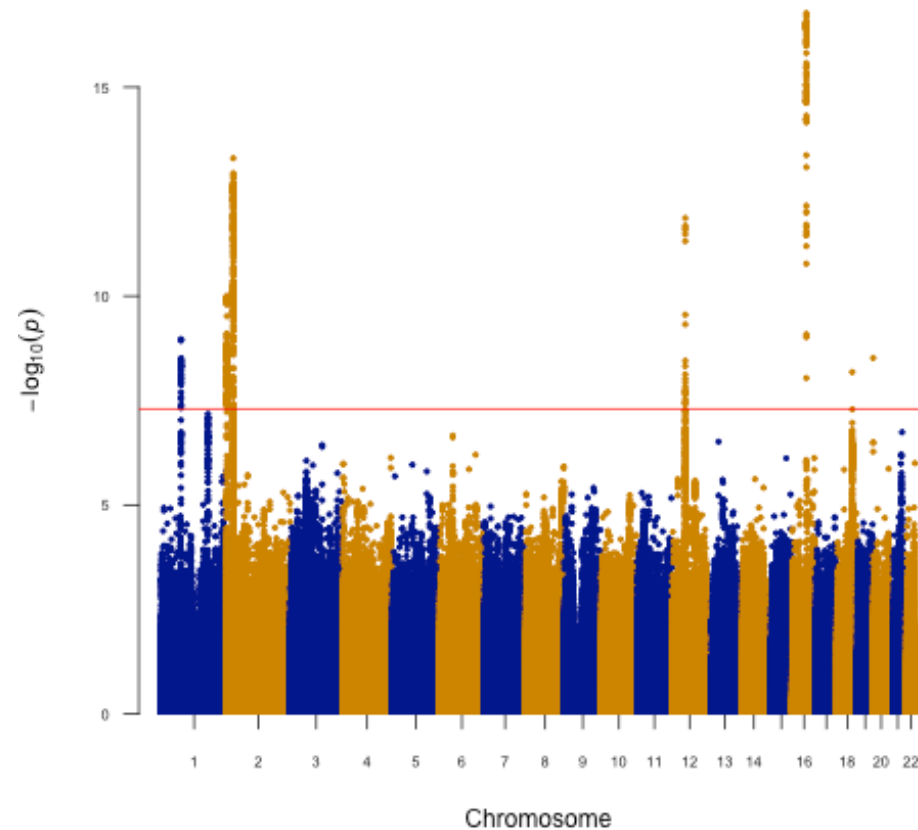
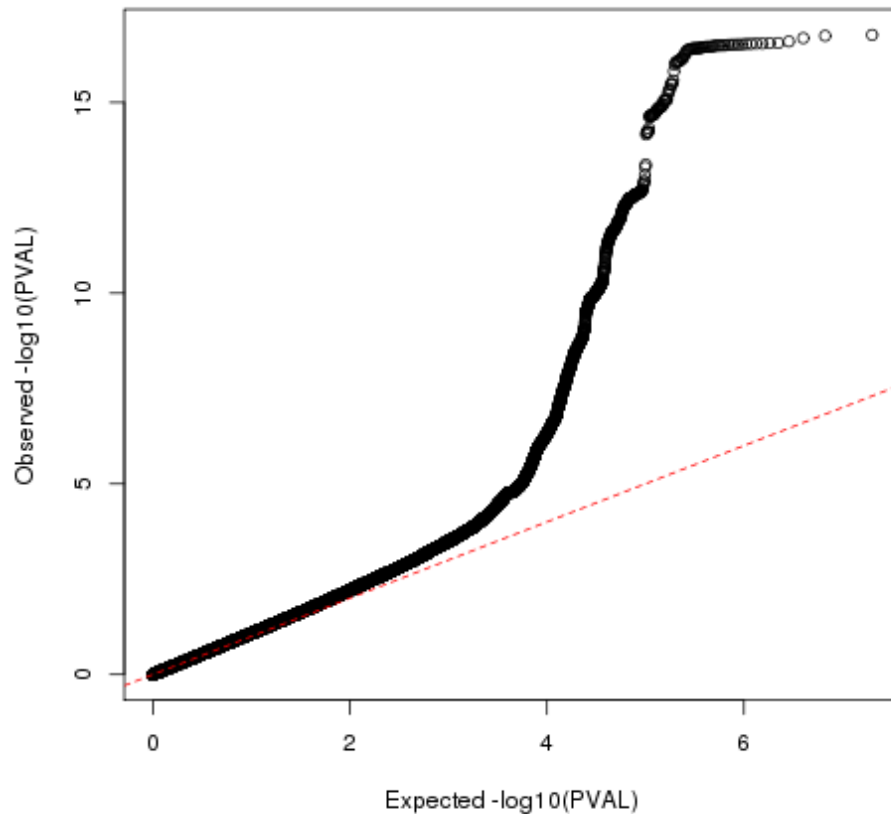


Table 2. Common Pitfalls, Recommendations, and Methods in Need of Development

Method	Pitfall	Recommendation	Needs
Genotyping	Many genotyping platforms do not cover non-European variation well.	Use or design population-specific array or multi-ancestry array; high array density can improve coverage in groups with high diversity. Consider low-depth whole-genome sequencing.	Continue improving coverage of diverse ancestries on genotyping arrays. Encourage ongoing development and sharing of pipelines for analysis of low-depth sequencing data.
QC	Unnecessary loss of data and/or incorrect inferences by using a one-size-fits-all approach	See Figure 2 for specific recommendations for each QC step and Table S2.	Improve availability and convenience of implementing proposed QC methods robust to population structure.
Imputation	Inaccurate imputation due to poor matching of reference panel to sample	Consider matching the ancestry of the reference panel as closely as possible to the sample ancestry if using a single ancestry sample. Consider the largest reference panel possible for imputation of multiple or admixed samples.	Continue expanding diversity of imputation panels, through collection of whole-genome sequencing data, creation of imputation panels from that data, and promoting public sharing/accessibility of those panels.
GWAS	Poor control of population stratification	Consider standard linear/logistic regression methods for analysis of single ancestry groups followed by meta-analysis. Consider mixed model approaches for admixed or multi-ancestry analyses. Include PCs as covariates even when single ancestry groups analyzed. PCs should be computed individually for each major population group within a multi-ancestry cohort and included as covariates in the regression model. Additional covariates should be considered for the multi-ancestry analysis.	Continue investigating causes of—and solutions to—current incomplete control of population stratification from principal components and mixed models.
Meta-analysis	False negative and false positive findings; effect heterogeneity	Use a random-effects (with possible bias towards the null), or modified random-effects meta-analysis model.	Continue to investigate and find solutions to improve power for the detection of heterogeneous effects.
Fine-mapping	LD improperly handled when all samples are meta-analyzed across populations. Uneven genome coverage across populations because of the genotyping array and the imputation reference panel	Use fine-mapping methods that explicitly model population-specific LD. See recommendations for Genotyping and Imputation above.	Continue to develop fine-mapping methods that rely on fewer assumptions, and thoroughly evaluate their performance.
Polygenic risk scores	Loss of accuracy in target population with increasing genetic distance from discovery cohort	Extrapolation of PRSs from one ancestry to another is problematic with current approaches and data.	Large discovery cohorts for all populations are needed. Develop methods for computing PRSs that are not biased when applied across populations, potentially incorporating LD information and/or local ancestry information among diverse populations.
Rare variants	Population stratification; low power to detect associations	Aggregate tests can improve power and handle separate causal variants in different populations.	Approaches with better control of population stratification; more data on diverse populations needed.

Presentation of results from large-scale genetic association studies



An association with p-value $<5 \times 10^{-8}$ is considered genome-wide significant

The first GWAS was published in December 2005 (96 cases and 50 controls)

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,¹ Caroline Zeiss,^{2*} Emily Y. Chew,^{3*} Jen-Yue Tsai,^{4*} Richard S. Sackler,¹ Chad Haynes,¹ Alice K. Henning,⁵ John Paul SanGiovanni,³ Shrikant M. Mane,⁶ Susan T. Mayne,⁷ Michael B. Bracken,⁷ Frederick L. Ferris,³ Jurg Ott,¹ Colin Barnstable,² Josephine Hoh^{7†}

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal P value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of *CFH* that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

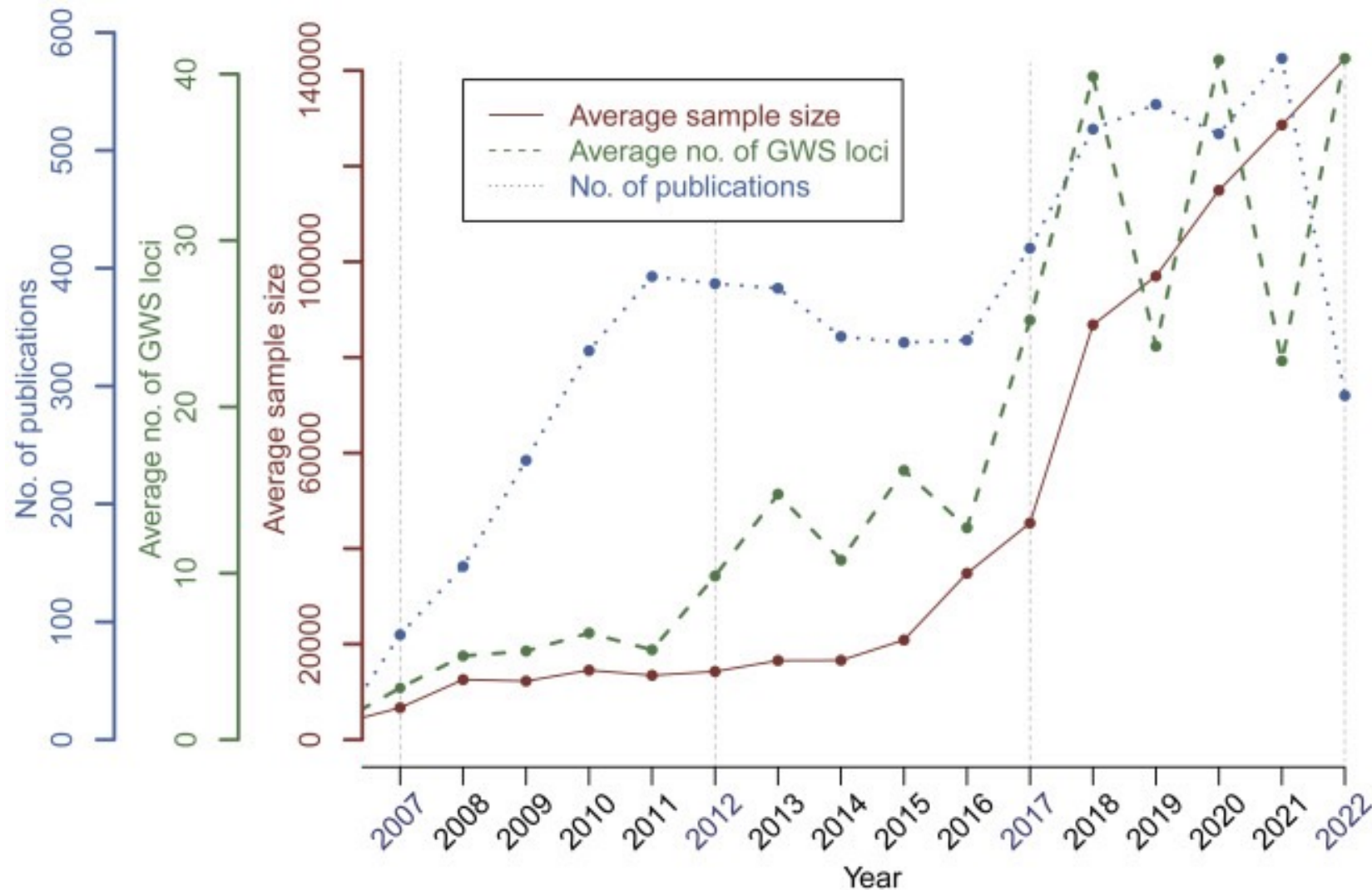
Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world. Its incidence is increasing as the elderly population expands (1). AMD is characterized by progressive destruction of the retina's central region (macula), causing central field visual loss (2). A key feature of AMD is the formation of extracellular deposits called drusen concentrated in and around the macula behind the retina between the retinal pigment epithelium (RPE) and the choroid. To date, no therapy for this disease has proven to be broadly effective. Several risk factors have been linked to AMD, including age, smoking, and family history (3). Candidate-gene studies

have not found any genetic differences that can account for a large proportion of the overall prevalence (2). Family-based whole-genome linkage scans have identified chromosomal regions that show evidence of linkage to AMD (4–8), but the linkage areas have not been resolved to any causative mutations.

Like many other chronic diseases, AMD is caused by a combination of genetic and environmental risk factors. Linkage studies are not as powerful as association studies for the identification of genes contributing to the risk for common, complex diseases (9). However, linkage studies have the advantage of searching the whole genome in an unbiased manner

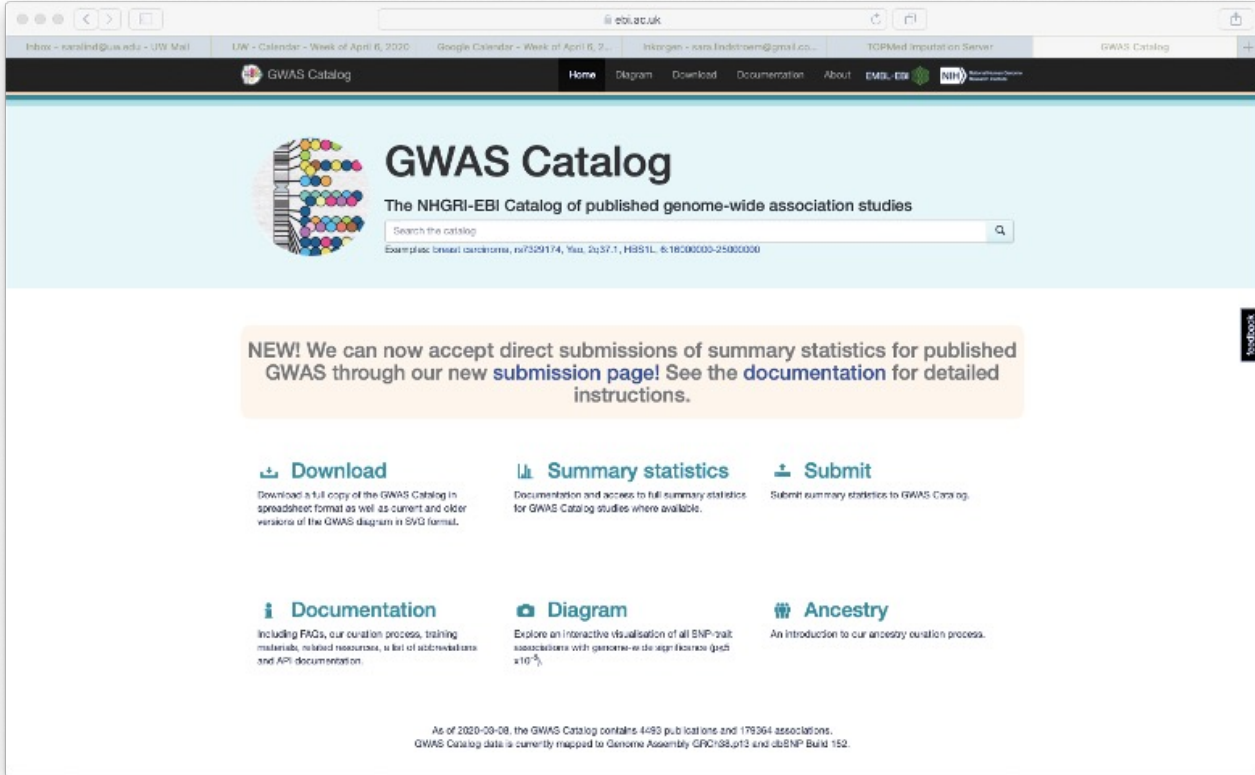
Klein, Science 2005

GWAS sample sizes over the years

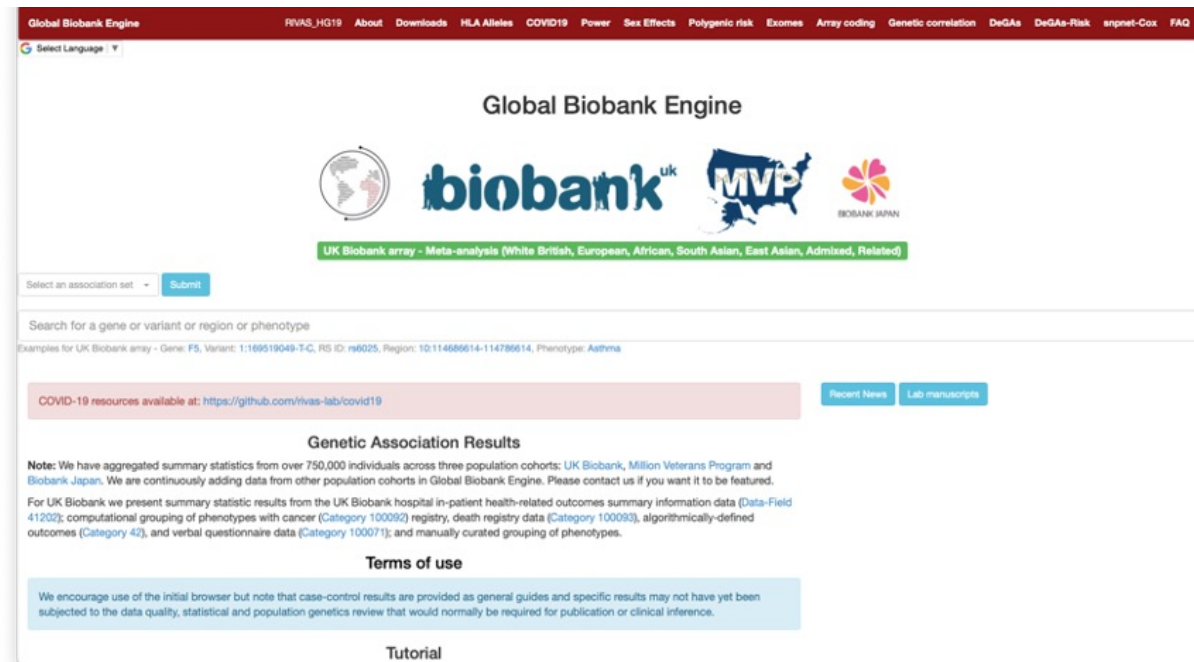


Based on 5,771 GWAS that shared summary stats in GWAS Catalog by Nov 8, 2022

Abdellaoui, AJHG 2023



<https://www.ebi.ac.uk/gwas/>



<https://biobankengine.stanford.edu>

Breakout Activity

- > Explore the NHGRI-EBI GWAS catalog: <https://www.ebi.ac.uk/gwas/home>. This website will introduce you to existing GWAS on many different phenotypes.
- > Using the GWAS catalog, determine what SNP rs6025 has been associated with in previous studies.
- > Explore the Global Biobank Engine (<https://biobankengine.stanford.edu>), which has collated GWAS results on a wide range of phenotypes based on large biobanks (UK Biobank, Biobank Japan, Million Veterans Program). Using this resource, what associations do you see with rs6025?

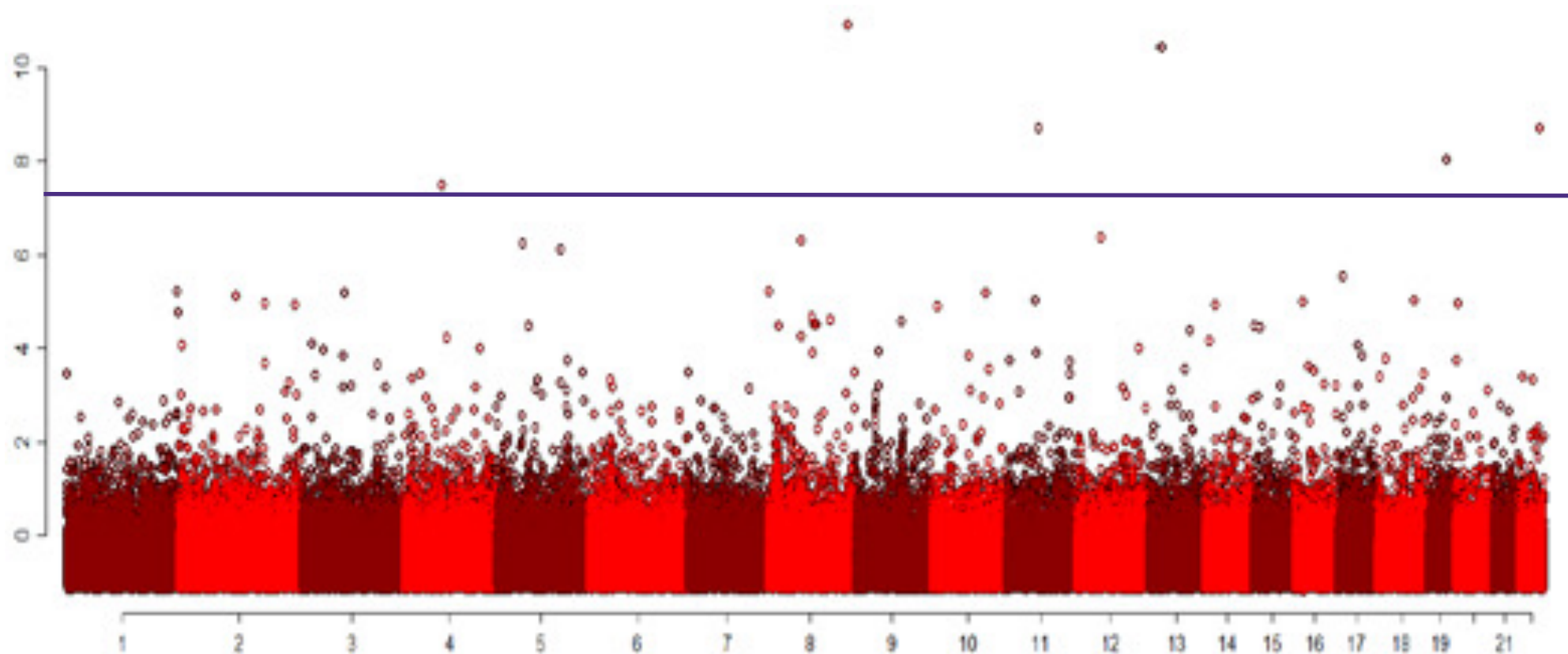
Practical issues in GWAS and other large-scale association studies

- > Bias
- > Differential genotyping error/missingness
- > Population Stratification
- > Replication
- > Follow up of identified signals: fine-mapping
- > Meta-analysis of GWAS

Differential genotyping error/missingness

- > Systematic differences in how case and control samples were collected, handled, or genotyped can lead to spurious associations
 - DNA was collected from blood samples for cases and from cheek swabs for controls
 - Case samples have been sitting in the freezer for 15 years, control samples are new
 - Cases and controls were genotyped in different genotyping labs or by different platforms

Genetic signatures of exceptional longevity in humans



Anything odd about this plot compared to other Manhattan plots we've seen?

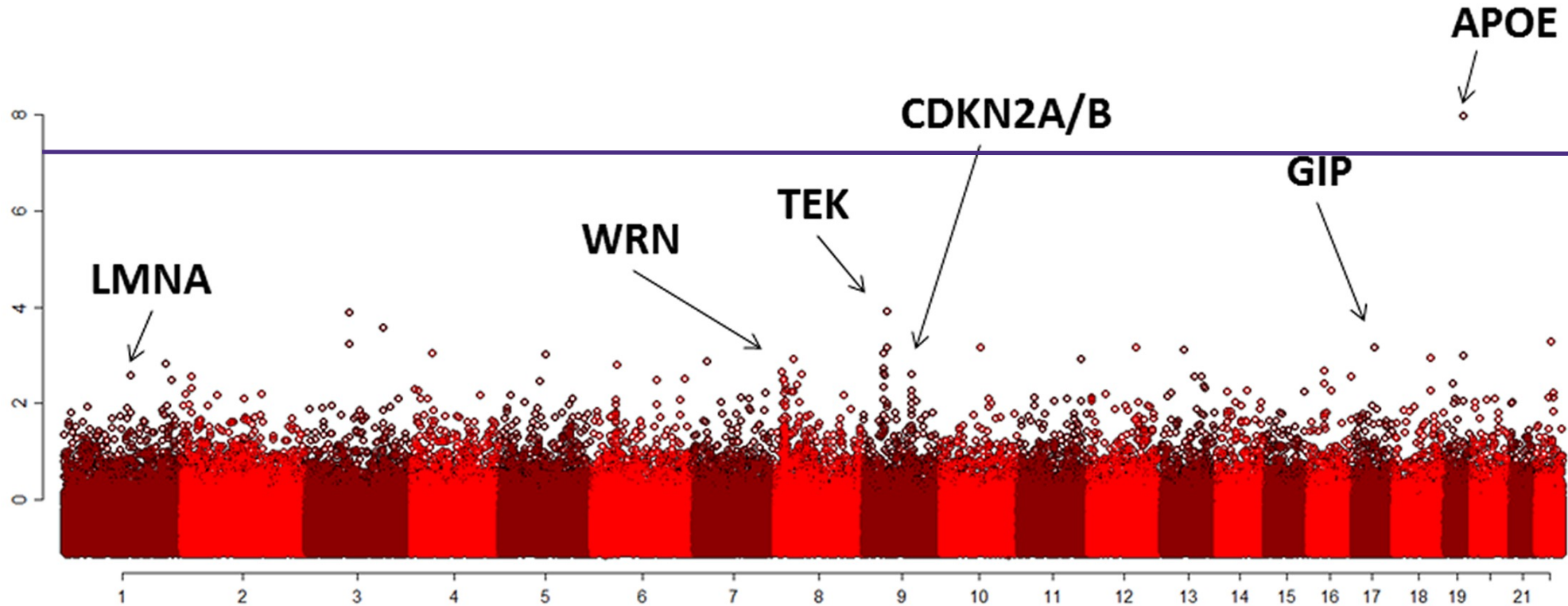
Retraction

AFTER ONLINE PUBLICATION OF OUR REPORT “GENETIC SIGNATURES OF EXCEPTIONAL LONGEVITY IN HUMANS” (1), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent laboratory subsequently performed stringent quality control measures, ambiguous SNPs were then removed, and resultant genotype data were validated using an independent platform. We then reanalyzed the reduced data set using the same methodology as in the published paper. We feel the main scientific findings remain supported by the available data: (i) A model consisting of multiple specific SNPs accurately differentiates between centenarians and controls; (ii) genetic profiles cluster into specific signatures; and (iii) signatures are associated with ages of onset of specific age-related diseases and subjects with the oldest ages. However, the specific details of the new analysis change substantially from those originally published online to the point of becoming a new report. Therefore, we retract the original manuscript and will pursue alternative publication of the new findings.

PAOLA SEBASTIANI,^{1*} NADIA SOLOVIEFF,¹ ANNIBALE PUCA,² STEPHEN W. HARTLEY,¹ EFTHYMIA MELISTA,³
STACY ANDERSEN,⁴ DANIEL A. DWORKIS,³ JEMMA B. WILK,⁵ RICHARD H. MYERS,⁵ MARTIN H. STEINBERG,⁶
MONTY MONTANO,³ CLINTON T. BALDWIN,^{6,7} THOMAS T. PERLS^{4*}

Genetic signatures of exceptional longevity in humans

Published version, post retraction



How to assess population stratification (and other sources of inflation) in your GWAS

- > Most of the genetic markers in the genome (e.g., in a GWAS) are likely not associated with the trait of interest
- > The genomic control parameter (λ_{GC}) summarizes systematic inflation in your data and is based on a large number of association tests

$$\lambda_{GC} = \frac{\textit{The median of the observed } \chi^2 \textit{ statistics}}{\textit{The median of the } \chi^2 \textit{ statistics under the NULL}}$$

For a 1 d.f. χ^2 test, the denominator is 0.455

A few notes about λ_{GC}

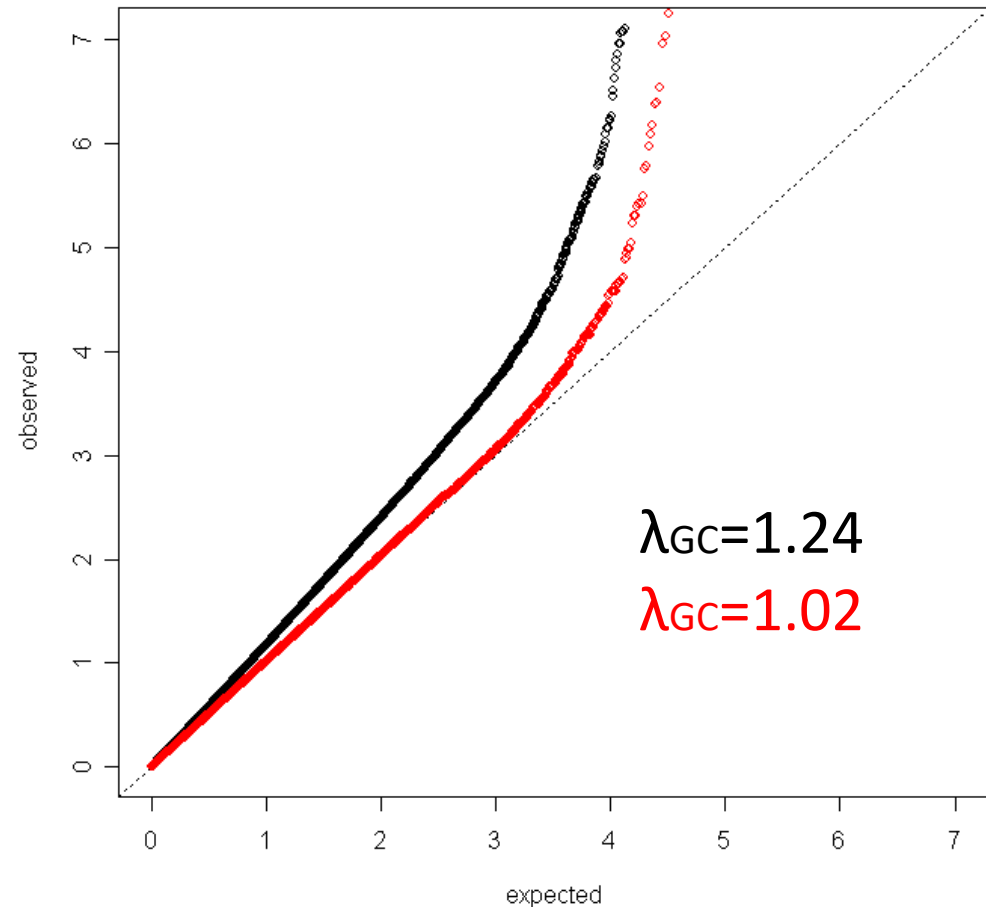
- > λ_{GC} should be close to 1 if no bias exists.
 - Rule of thumb for a small-to-moderate GWAS: <1.05 is often ok, above 1.1 deserves attention
- > However, λ_{GC} scales with sample size
 - Under a polygenic model, many SNPs with small effect sizes will be detected with very large sample size -> expect λ_{GC} to increase
 - This can be accounted for by scaling inflation to an equivalent study of 1,000 cases and 1,000 controls (λ_{1000})

$$\lambda_{1000} = 1 + \frac{500 (\lambda - 1)}{\left(\frac{1}{\#cases} + \frac{1}{\#controls}\right)^{-1}}$$

- > There are methods (e.g., LD score regression) that allows you to assess if the inflation in test statistics is due to a **true polygenic signal or due to bias.**

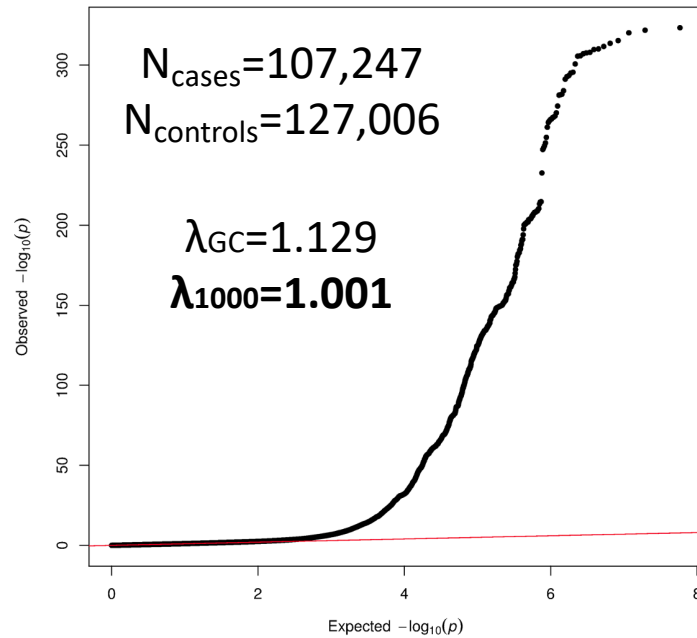
Inflation with and without adjustment for ancestry

- > QQ plot for a GWAS of dark-light hair color in European ancestry women from the Nurses Health Study (N=2,287). The black points are the test statistics from the unadjusted tests. The red points are from PC adjusted tests.

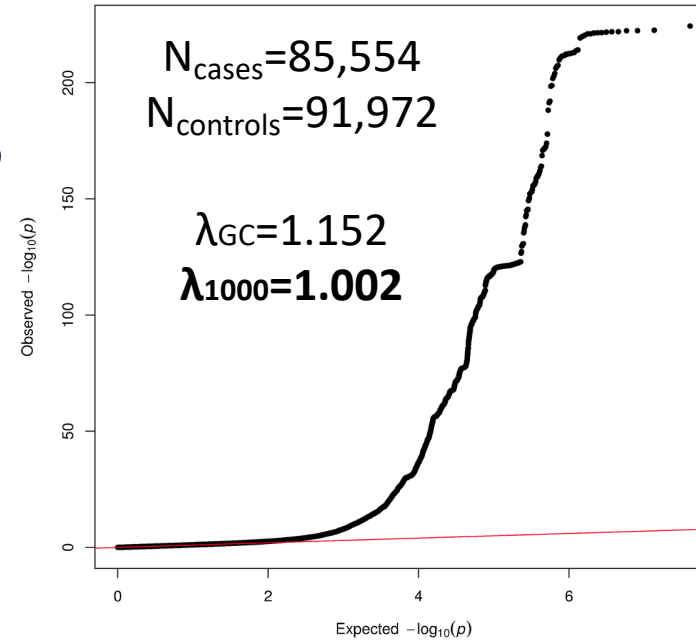


Accounting for inflation in a large prostate cancer GWAS

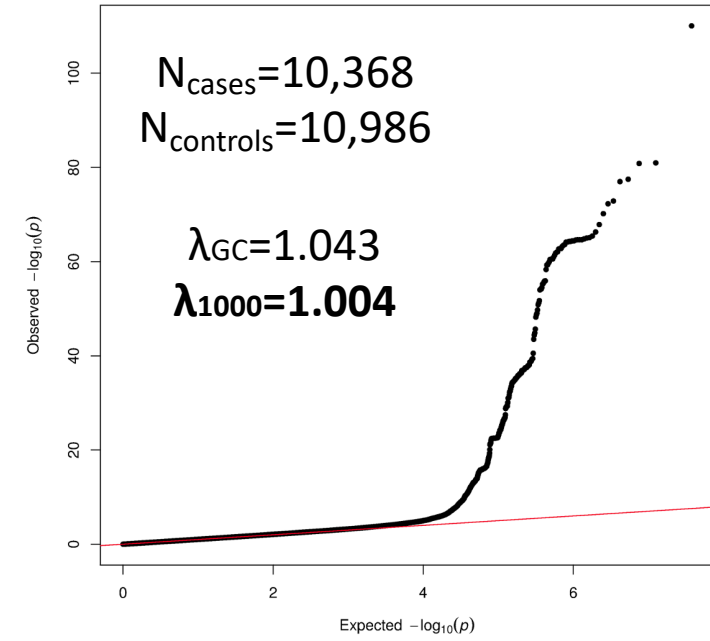
Multi-ancestry



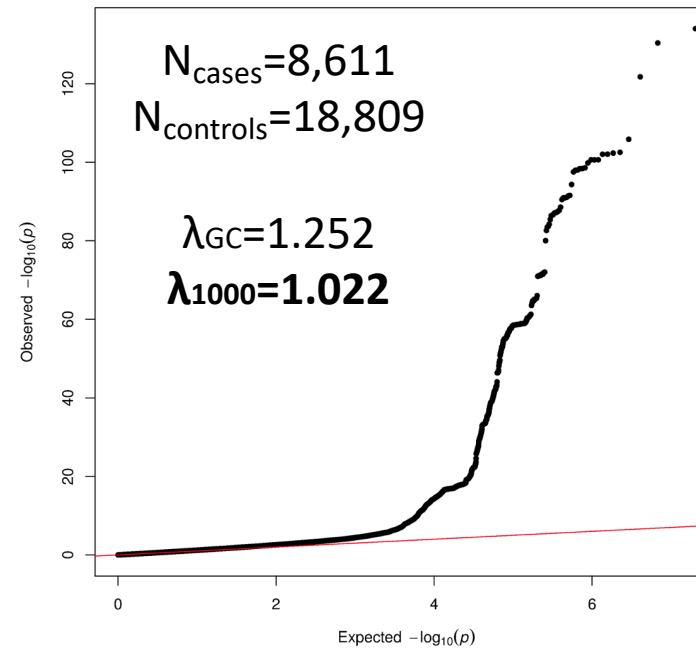
European



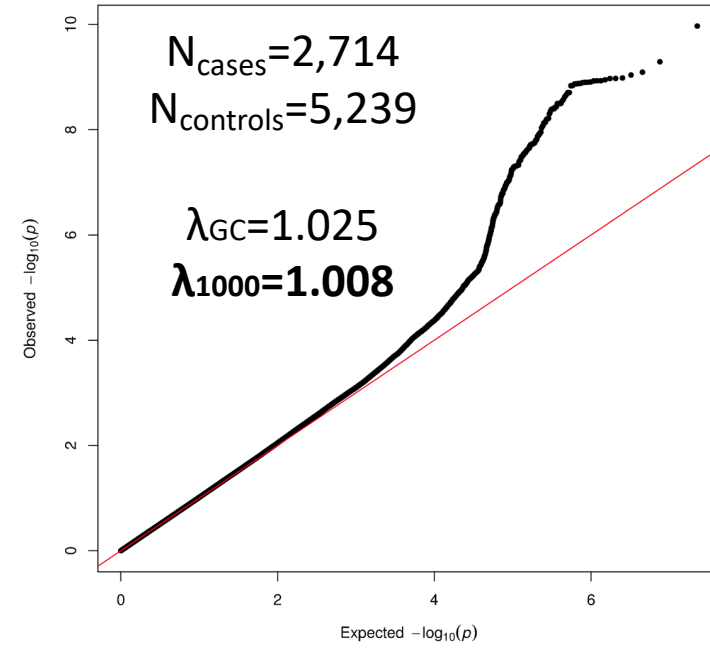
African



East Asian

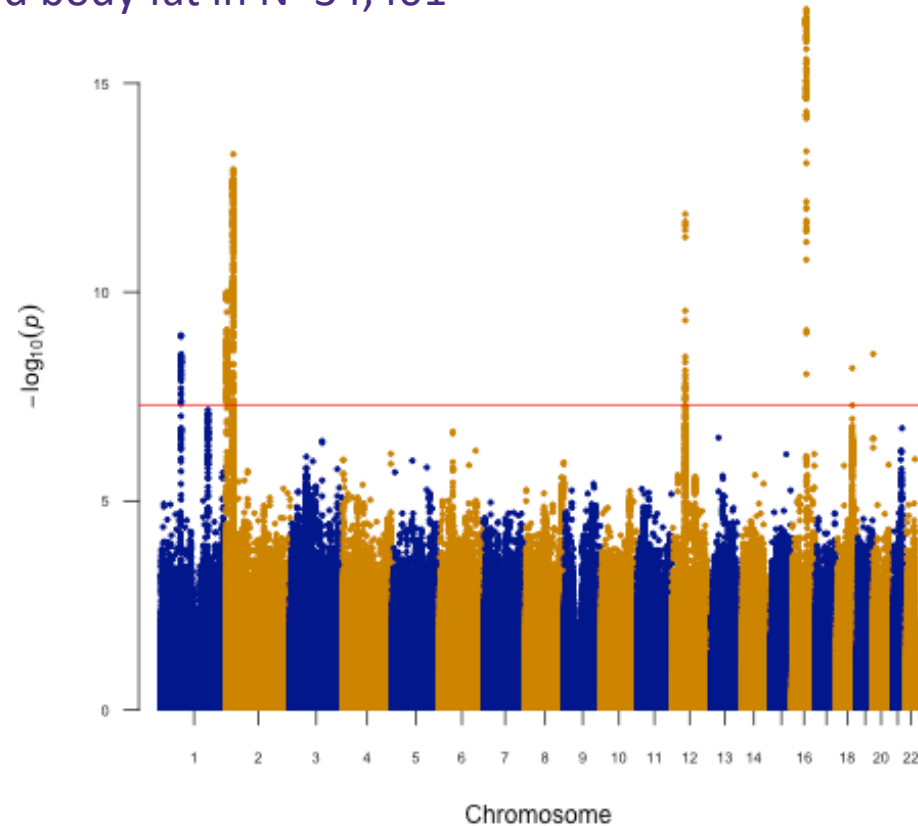
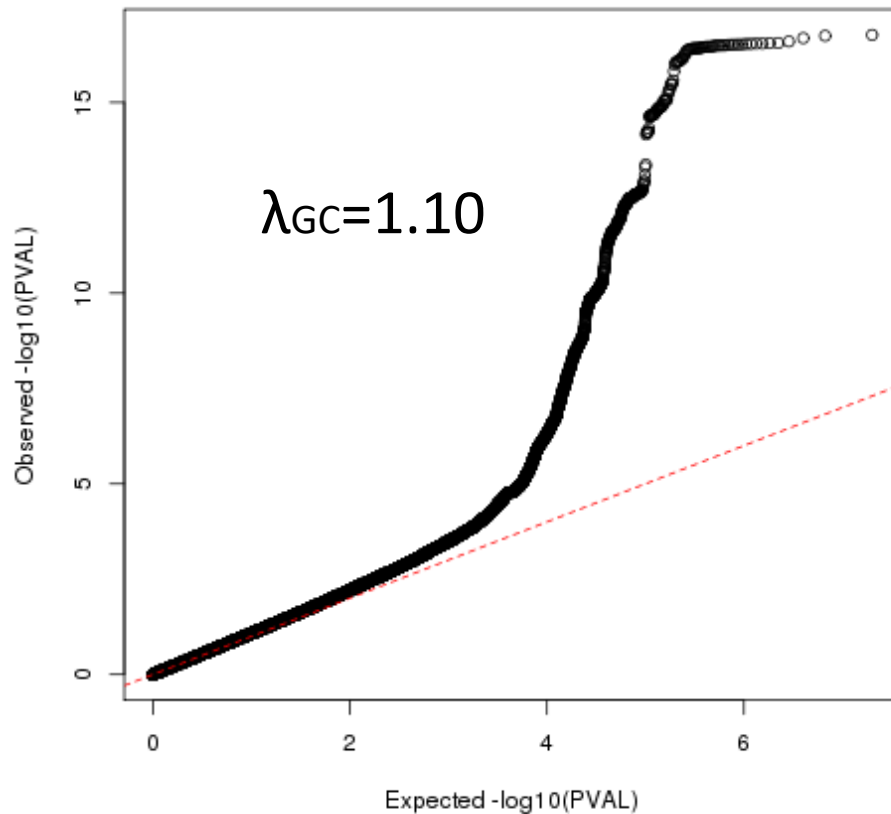


Hispanic



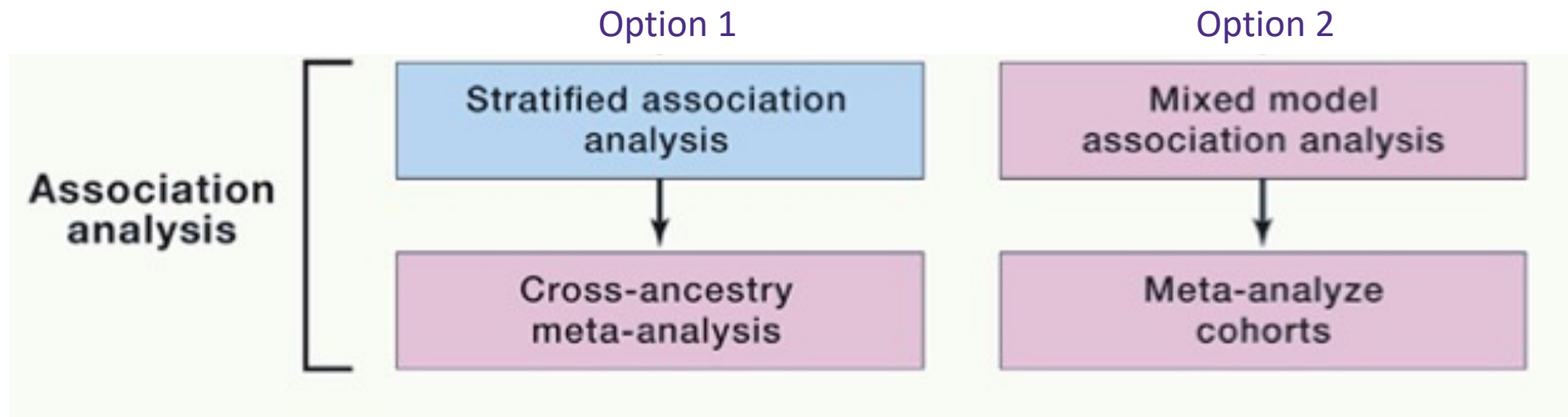
Assessing whether “inflation” could reflect true polygenic signal

GWAS on childhood body fat in N=34,401



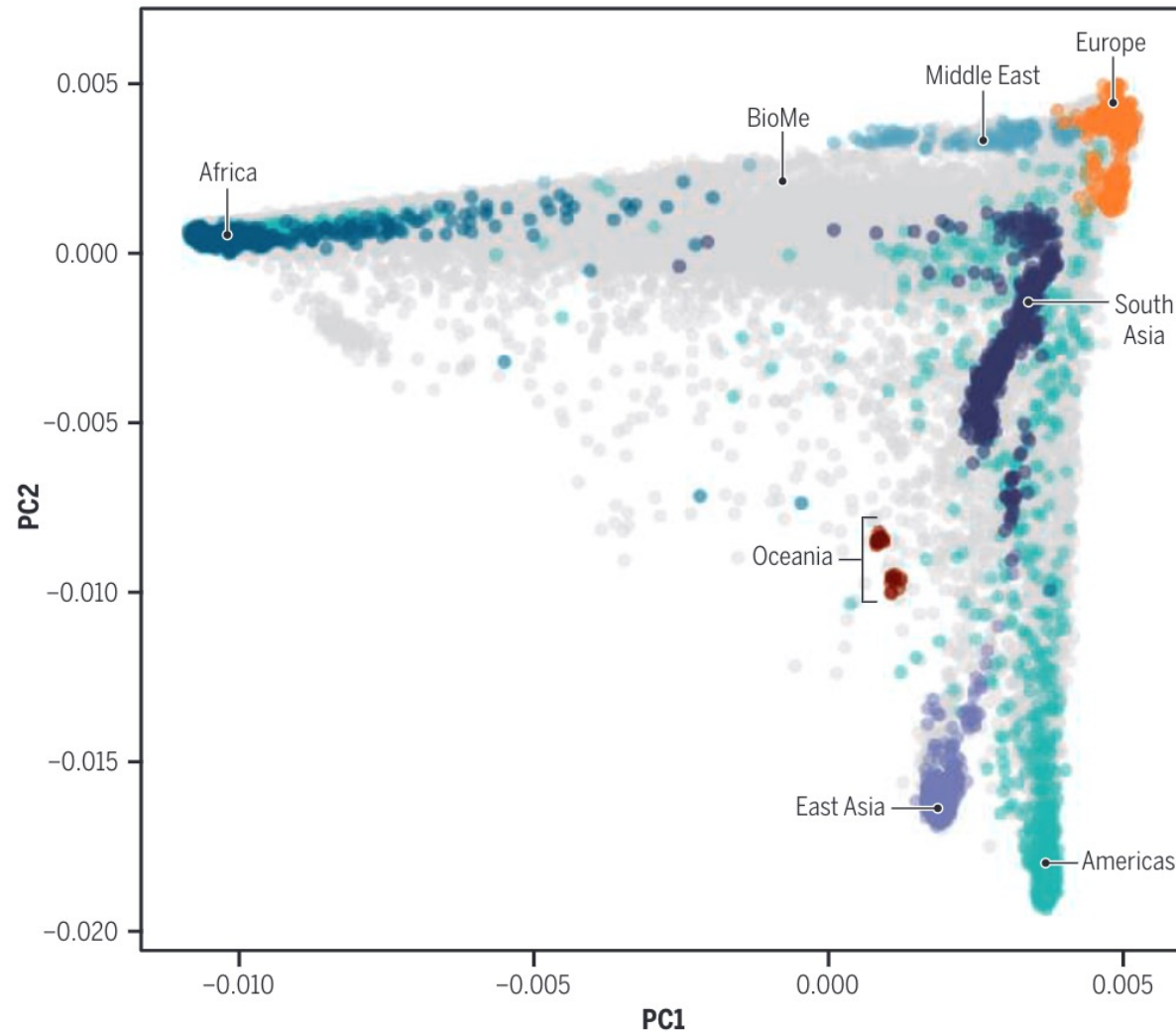
LD Score intercept=1.02 suggests this likely reflects a true polygenic signal

How can we conduct GWAS across diverse populations?



Peterson, Cell 2019

Challenges with Option 1: The continuous category-free nature of genetic variation



GRAPHIC: K. FRANKLIN/SCIENCE BASED ON (12)

Colored dots (n = 4149) are reference individuals representing ancestry from seven regions projected onto the first two PCs of genetic similarity. Gray dots (n = 31,705) are participants from BioMe, a diverse biobank based in New York City.

Clearly delineated continental ancestry categories (dots in color) are really a by-product of sampling strategy. They are not reflective of the diversity in this real-world dataset, which is made evident by the continuous sea of gray.

Option 2: Mixed model association analysis

- > Model any sample structure as a random effect in a mixed model
- > More sensitive to **cryptic relatedness** and **complex population structure** not easily captured by PCA
- > Historically not used due to computational limitation (especially for large datasets)
- > Software: BOLT-LMM, SAIGE, GENESIS (R Package) and many others...
- > Relies on building a genetic relatedness matrix (GRM)

Generalized linear mixed models (GLMM) in GWAS

$$\mathbf{y} = \mathbf{x}_{snp}\beta_{snp} + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{g} + \mathbf{e} \quad [1]$$

$$\mathbf{g} \sim N(0, \boldsymbol{\pi}\sigma_g^2) \quad \mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$$

- y is the phenotypes of interest
 - x_{snp} is the variant of interest with its effect β_{snp} ;
 - \mathbf{X}_c is any fixed covariates (e.g., sex, age) with their corresponding coefficients $\boldsymbol{\beta}_c$;
 - \mathbf{g} is the total genetic effects
 - \mathbf{e} is an error term
-
- $\boldsymbol{\pi}$ is the SNP-derived genetic relationship matrix (GRM)
 - σ_g^2 is the additive genetic variance tagged by SNPs (unknown)

The genetic relationship matrix (GRM)

- > The genetic relationship π_{jk} between two individuals j and k can be estimated by the following equation:

$$\pi_{jk} = \sum_i \frac{(g_{ij} - 2p_i)(g_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

g_{ij} is the number of copies of reference allele for SNP i in individual j

p_i is the frequency of the reference allele for SNP i

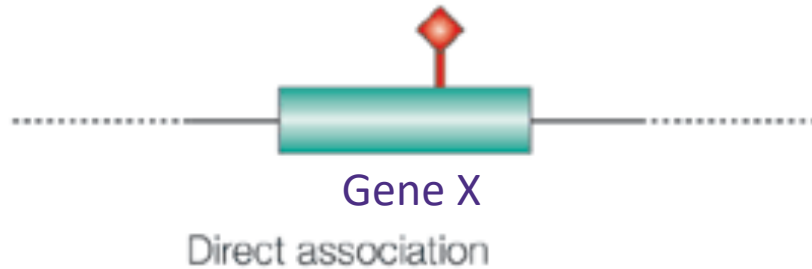
g_{ik} is the number of copies of reference allele for SNP i in individual k

Replication of GWAS findings

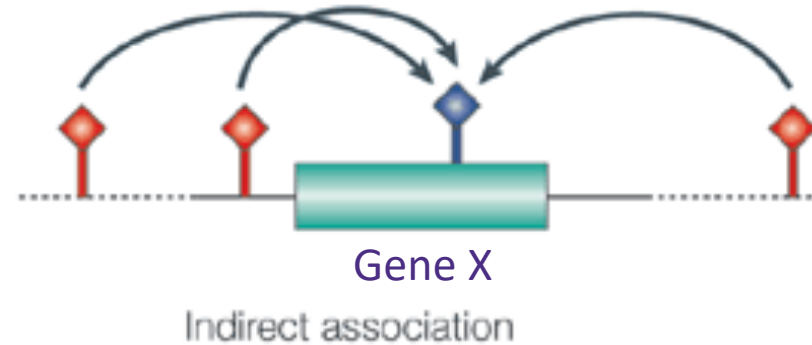
- > Want to see the signal in more than one population
- > Originally, replication was a way to maintain sample size while reducing costs
 - Stage 1: many SNPs in few samples
 - Stage 2: few SNPs (selected from stage 1) in many samples
- > It has been shown that it is more powerful to combine data up-front instead of subsequent replication (or “look-ups”)
 - Politics will play a role

Follow up on GWAS hits: Fine-mapping

a Candidate approach



b Genome-wide approach



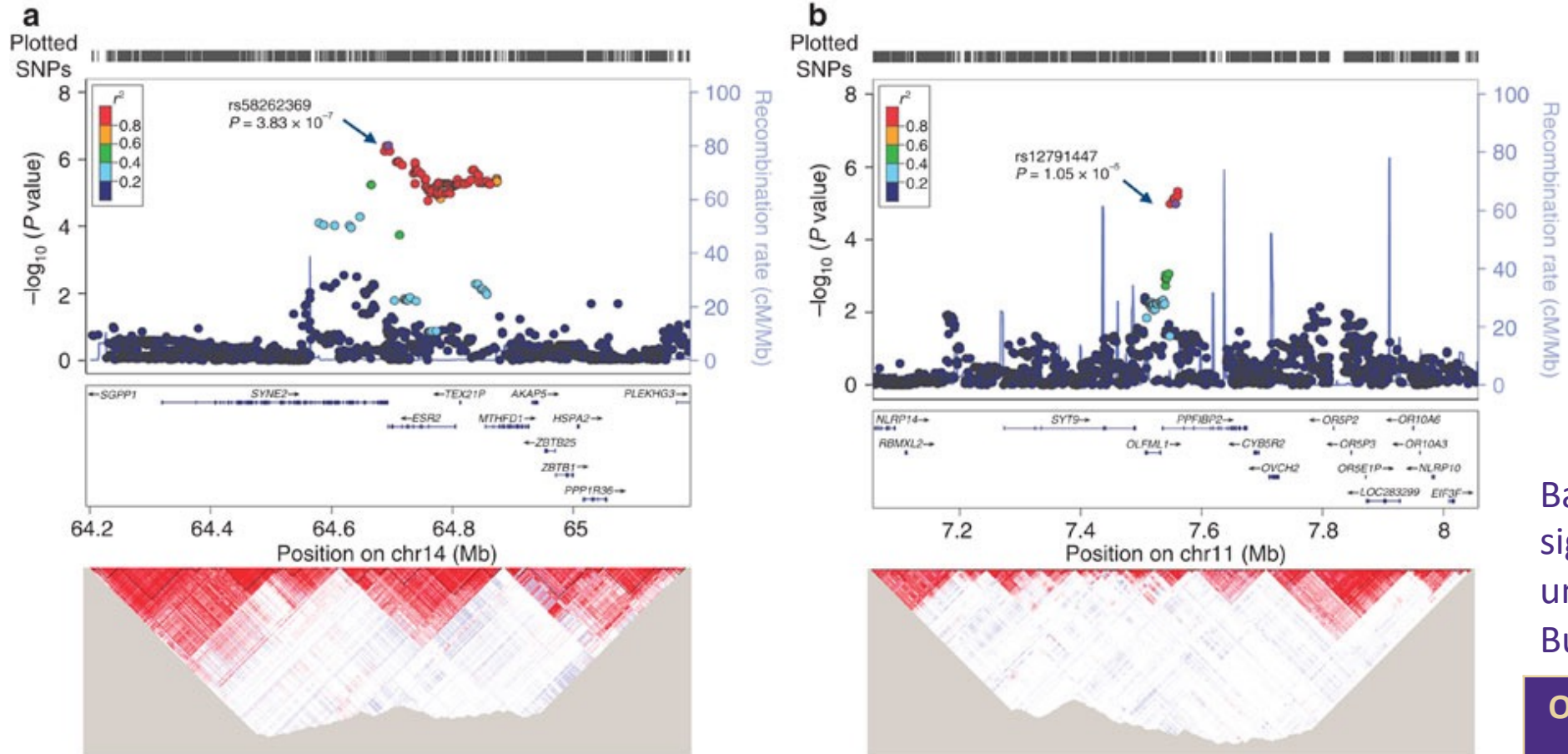
Genotyped SNPs

Indirectly tested SNP based on LD

Nature Reviews | **Genetics**

LD complicates things: Which SNP(s) is the causal SNP?

Results from a prostate cancer GWAS



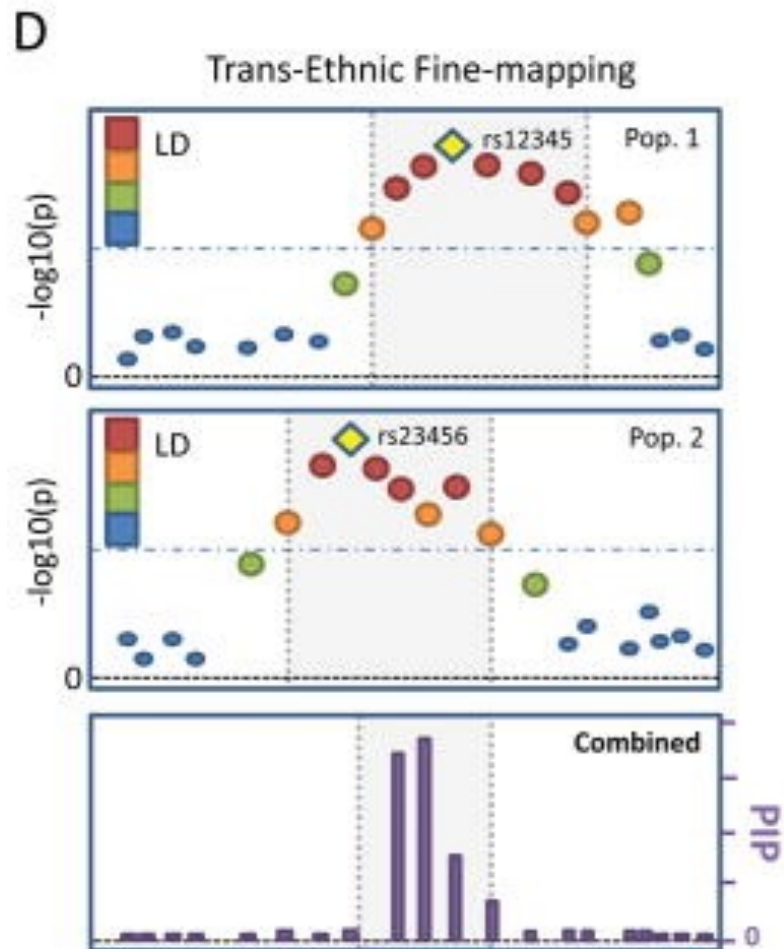
Based on simulations, lead signal in a region is typically unlikely to be causal (van de Bunt et al., PLoS Genet 2015)

OR	RAF	Probability lead SNP is causal
1.5	50%	79%
1.1	5%	2.4%

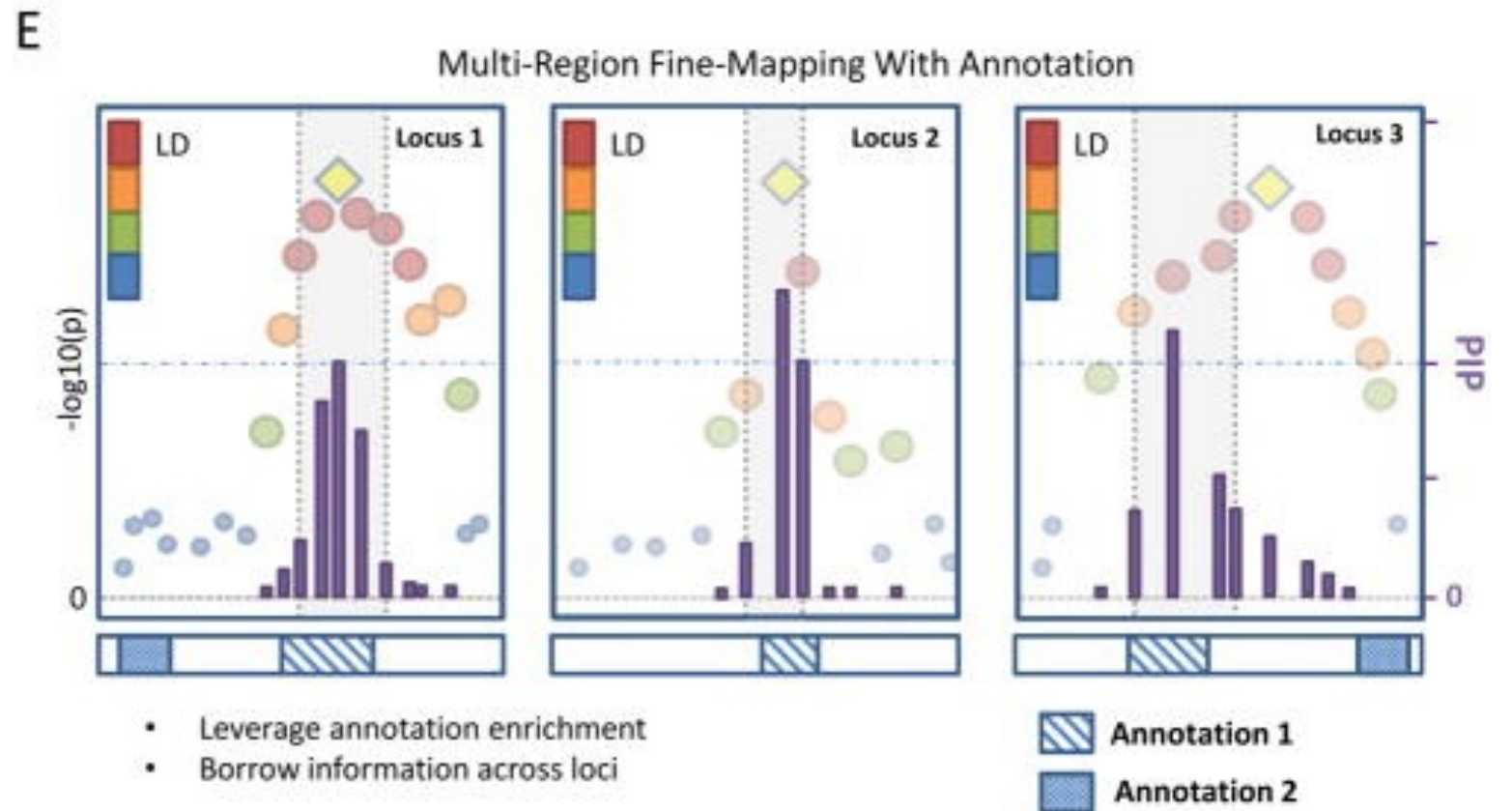
Fine-mapping approaches

- > Conditional forward stepwise regression analysis
 - Rerun analysis adjusting for the most significant SNP, see if any other SNP remains significant. Keep going until no more significant SNPs
- > Calculate posterior probabilities for each SNP
- > Incorporate “functional” information to identify biological plausible SNPs
- > Choose a set of “potentially causal variants” and take them forward for downstream analysis.

Fine-mapping approaches



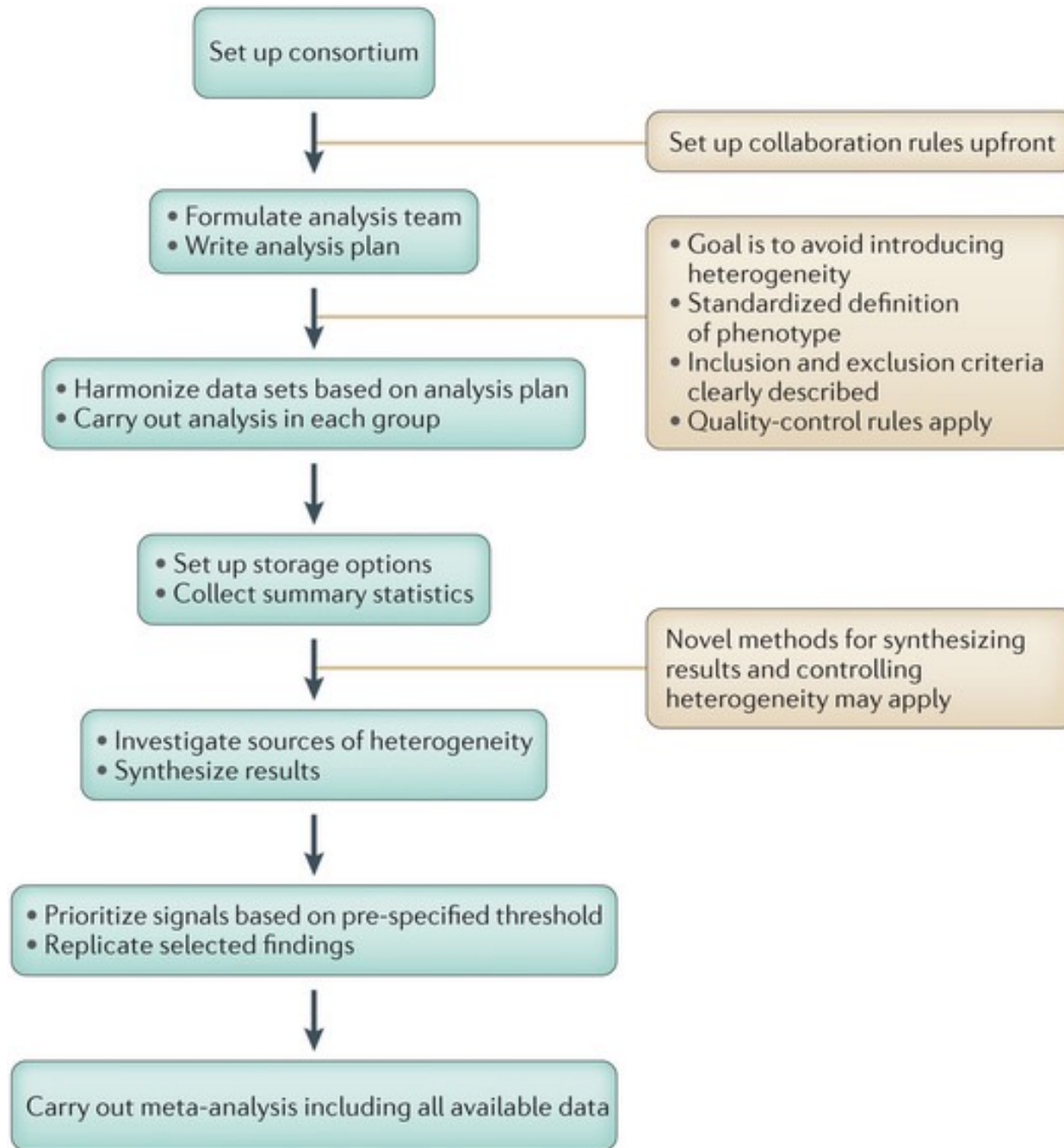
- Leverage Ethnic Differences in LD at a given locus



- Leverage annotation enrichment
- Borrow information across loci

Meta-analysis

- > Sample size is the key for a successful genetic association study
- > International collaborations to summary statistics from multiple GWAS are common
- > Issues with sharing individual-level data
 - Ethical approvals, IRBs, large files, ownership of the data...



Evangelou & Ioannidis, Nature Rev Genetics 2013

Meta-analysis in practice

- > Common protocol
 - Imputation reference panel
 - Association analysis (test for the same thing across studies)
- > QC of summary stats
 - Are the alleles the expected?
 - Are the minor allele frequencies the expected?
 - Are beta estimates/standard errors reasonable?
 - QQ-plots, Manhattan plots
 - Note: “Clean data” is most often not cleaned.

Method	Description	Advantages	Disadvantages	Main software used
<i>P</i> value meta-analysis	Simplest meta-analytical approach	Allows meta-analysis when effects are not available	Direction of effect is not always available; inability to provide effect sizes; difficulties in interpretation	METAL , GWAMA , R packages
Fixed effects	Synthesis of effect sizes. Between-study variance is assumed to be zero	Effects readily available through specialized software	Results may be biased if a large amount of heterogeneity exists	METAL, GWAMA, R packages
Random effects	Synthesis of effect sizes. Assumes that the individual studies estimate different effects	Generalizability of results	Power deserts in discovery efforts; may yield spuriously large summary effect estimates when there are selection biases	GWAMA, R packages
Bayesian approach	Incorporates prior assessment of the genetic effects	Most direct method for interpretation of results as posterior probabilities given the observed data	Methodologically challenging; GWAS-tailored routine software not available; subjective prior information used	R packages
Multivariate approaches	Incorporates the possible correlation between outcomes or genetic variants	Increased power can identify variants that conventional meta-analysis do not reveal using the same data sets	Computationally intensive; software not available for all analyses; some may require individual-level data	GCTA for multi-locus approaches
Other extensions	A set of different approaches that allows for the identification of multiple variants across different diseases	Summary results of previous meta-analyses can be used	May need additional exploratory analyses for the identification of variants; prone to systematic biases	Software developed by the authors of the proposed methodologies

GCTA, genome-wide complex trait analysis; GWAS, genome-wide association study.