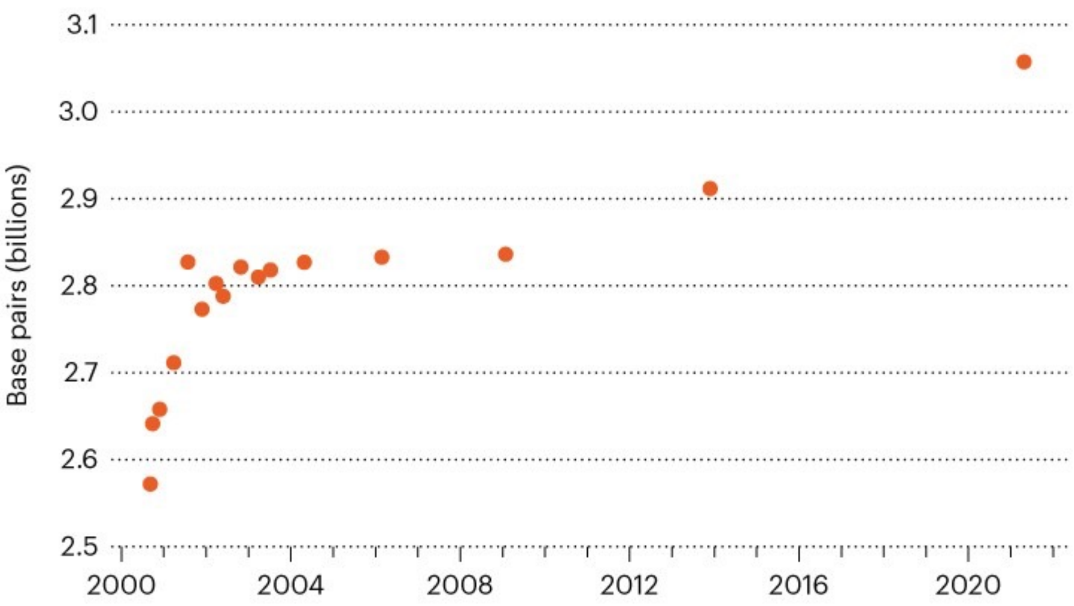
Session 8:
Large-scale genetic association studies - GWAS

The Human Genome Project (1990-2003) set out to sequence (“read”) every base pair in the human DNA

$2.7 billion

COMPLETING THE HUMAN GENOME

Researchers have been filling in incompletely sequenced parts of the human reference genome for 20 years, and have now almost finished it, with 3.05 billion DNA base pairs.



0.3% of sequence might still have errors. Includes X but not Y chromosome. Count excludes mitochondrial DNA.

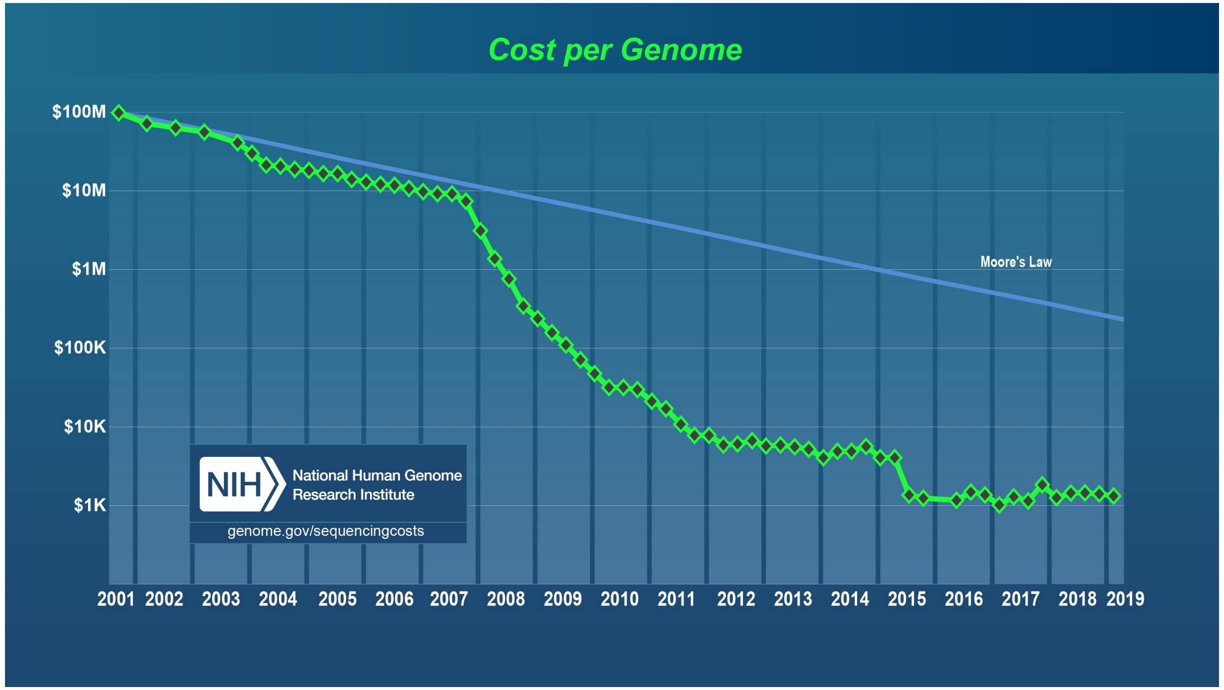
©nature

https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.full

Practical roadblocks to genome sequencing

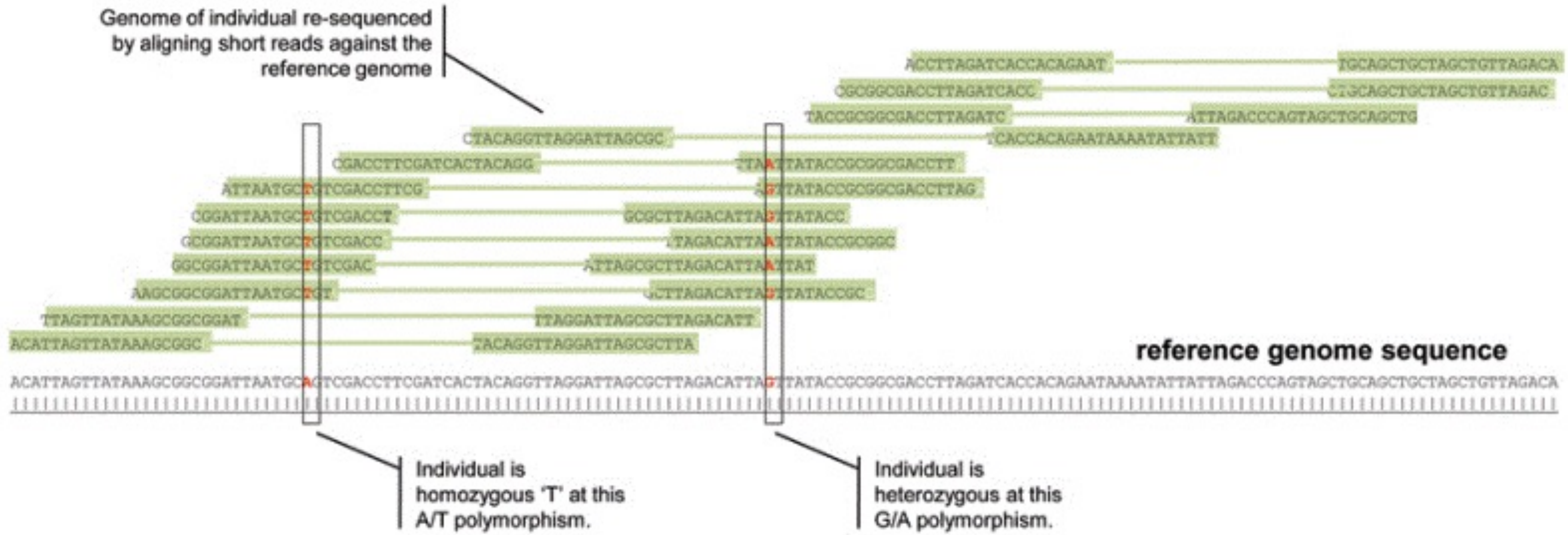
Sequencing cost per genome is currently ~$1,000

Sequencing one genome generates ~200 GB data



Sequencing alignment and depth

- Depth: The number of times a base-pair is sequenced

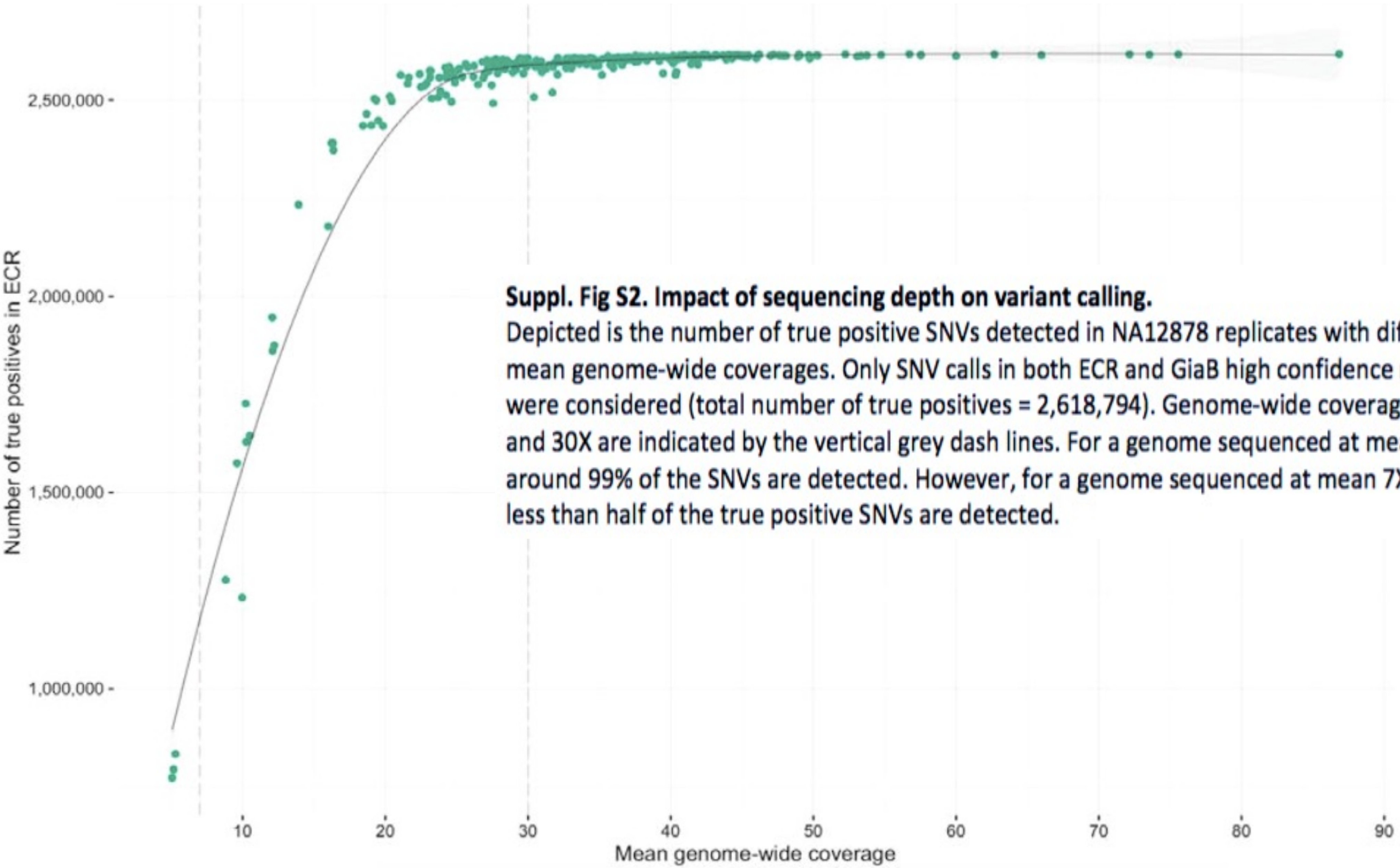


Genome of individual re-sequenced by aligning short reads against the reference genome

reference genome sequence

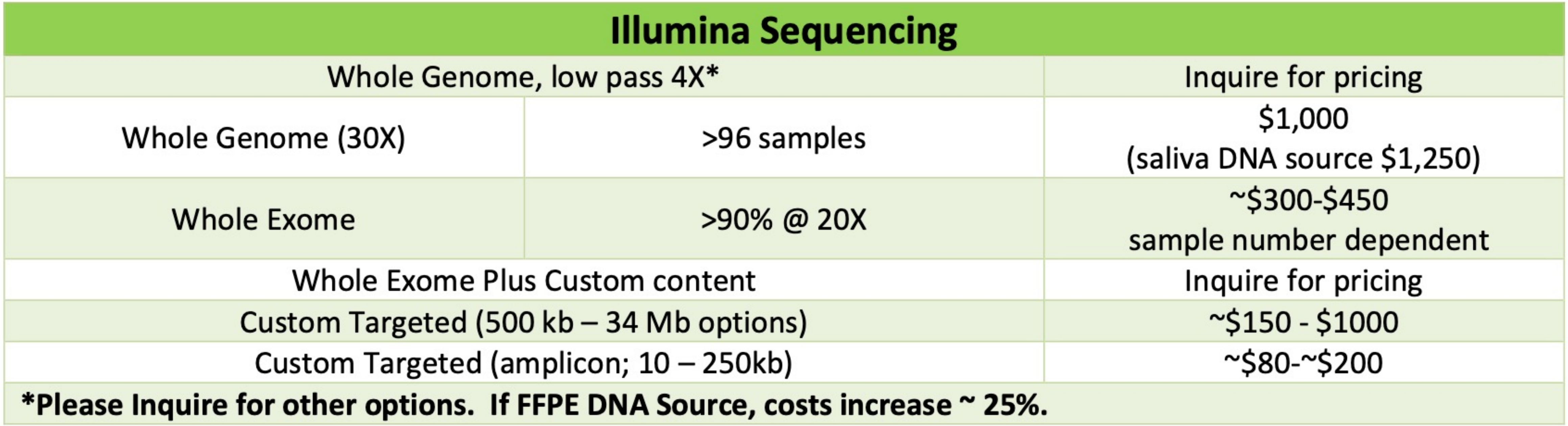
Individual is homozygous 'T' at this A/T polymorphism.

Individual is heterozygous at this G/A polymorphism.



Suppl. Fig S2. Impact of sequencing depth on variant calling.
Depicted is the number of true positive SNVs detected in NA12878 replicates with different mean genome-wide coverages. Only SNV calls in both ECR and GiaB high confidence regions were considered (total number of true positives = 2,618,794). Genome-wide coverages of 7X and 30X are indicated by the vertical grey dash lines. For a genome sequenced at mean 30X, around 99% of the SNVs are detected. However, for a genome sequenced at mean 7X coverage, less than half of the true positive SNVs are detected.

Pricing Sequencing (CIDR, March 2021)



Illumina Sequencing		
Whole Genome, low pass 4X*		Inquire for pricing
Whole Genome (30X)	>96 samples	$1,000 (saliva DNA source $1,250)
Whole Exome	>90% @ 20X	~$300-$450 sample number dependent
Whole Exome Plus Custom content		Inquire for pricing
Custom Targeted (500 kb – 34 Mb options)		~$150 - $1000
Custom Targeted (amplicon; 10 – 250kb)		~$80-~$200
***Please Inquire for other options. If FFPE DNA Source, costs increase ~ 25%.**		

https://www.cidr.jhmi.edu/services/pricing.pdf

Analysis of genetic association studies

1. Quality Control
   a. Sample level: Low call rate, heterozygosity, sex check, relatedness
   b. SNP level: Low call rate, minor allele frequency, HWE

2. Calculate PCs

3. Imputation
   1. Imputation Michigan Server (https://imputationserver.sph.umich.edu/index.html)

4. Analysis
   1. Model each SNP separately
   2. Linear/Logistic regression or general mixed models

$$Y = \alpha + \beta * SNP + X$$

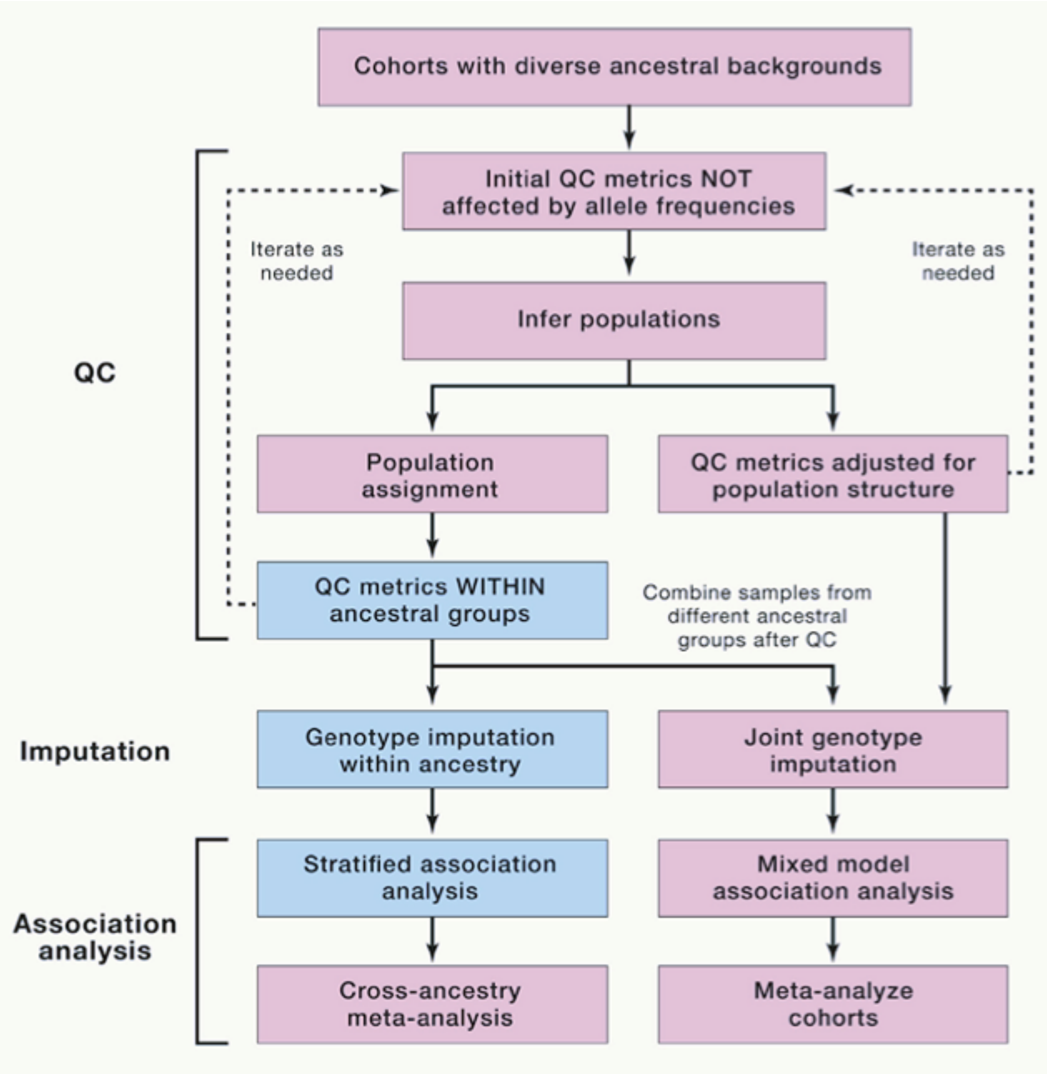
- β = SNP effect (log(OR) if logistic regression)
- X = additional covariates (e.g., sex, study, age, population stratification)

Peterson, Cell 2019

Leading Edge
Primer

Cell

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



QC
Imputation
Association analysis
Cohorts with diverse ancestral backgrounds
Initial QC metrics NOT affected by allele frequencies
Iterate as needed
Iterate as needed
Infer populations
Population assignment
QC metrics adjusted for population structure
QC metrics WITHIN ancestral groups
Combine samples from different ancestral groups after QC
Genotype imputation within ancestry
Joint genotype imputation
Stratified association analysis
Mixed model association analysis
Cross-ancestry meta-analysis
Meta-analyze cohorts

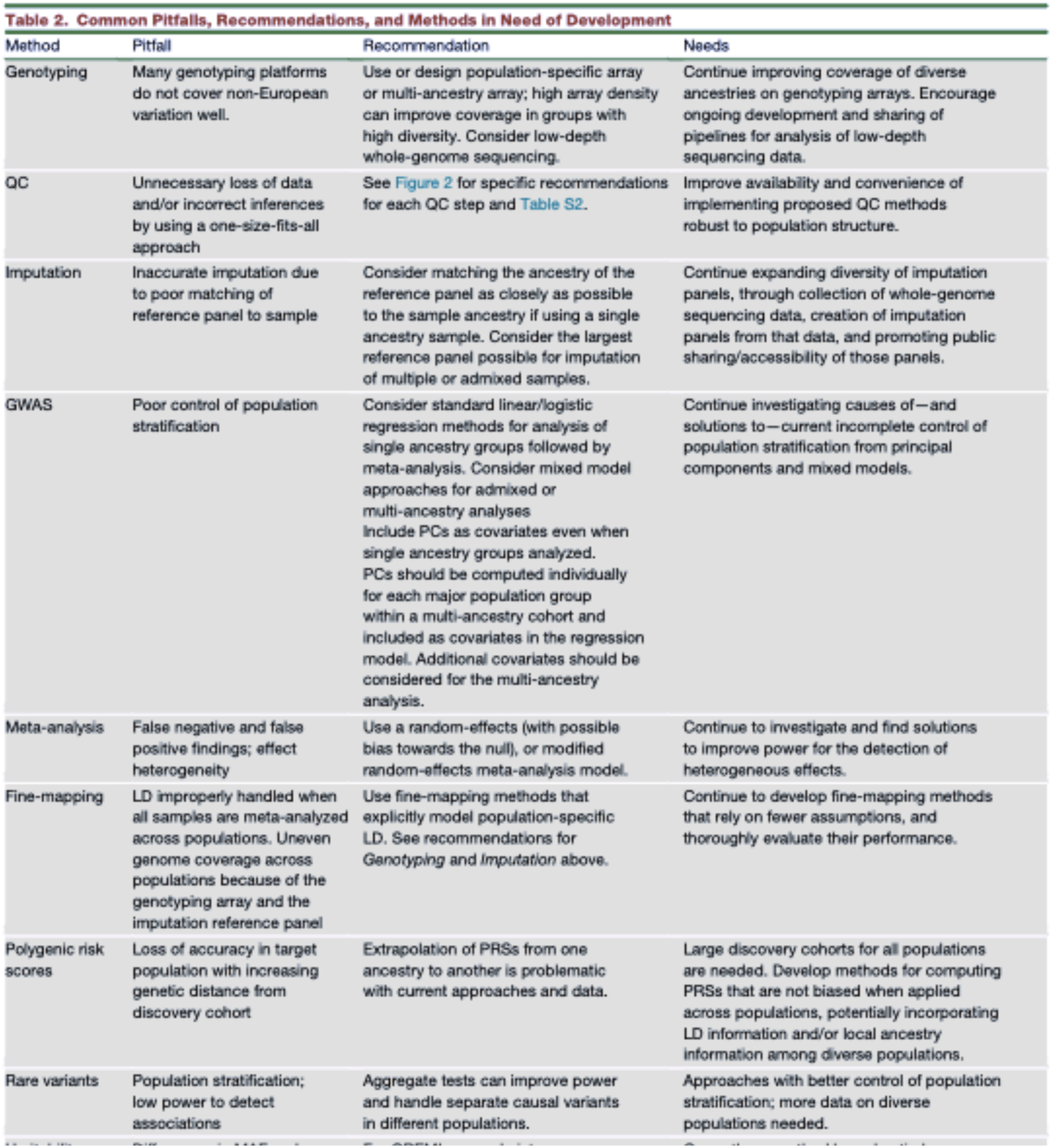
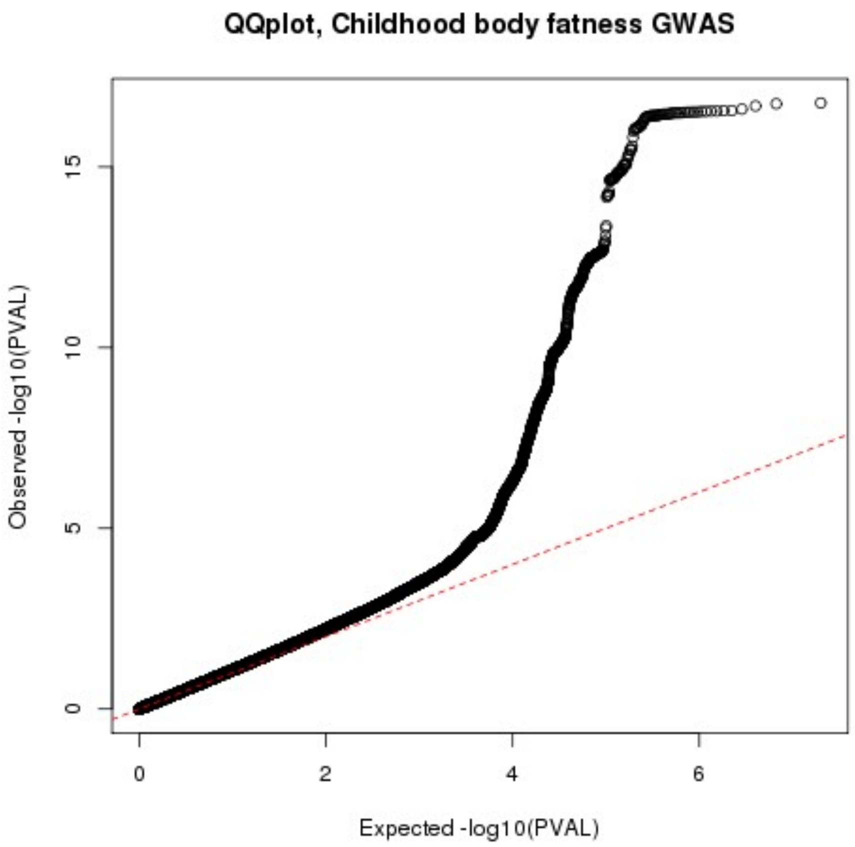


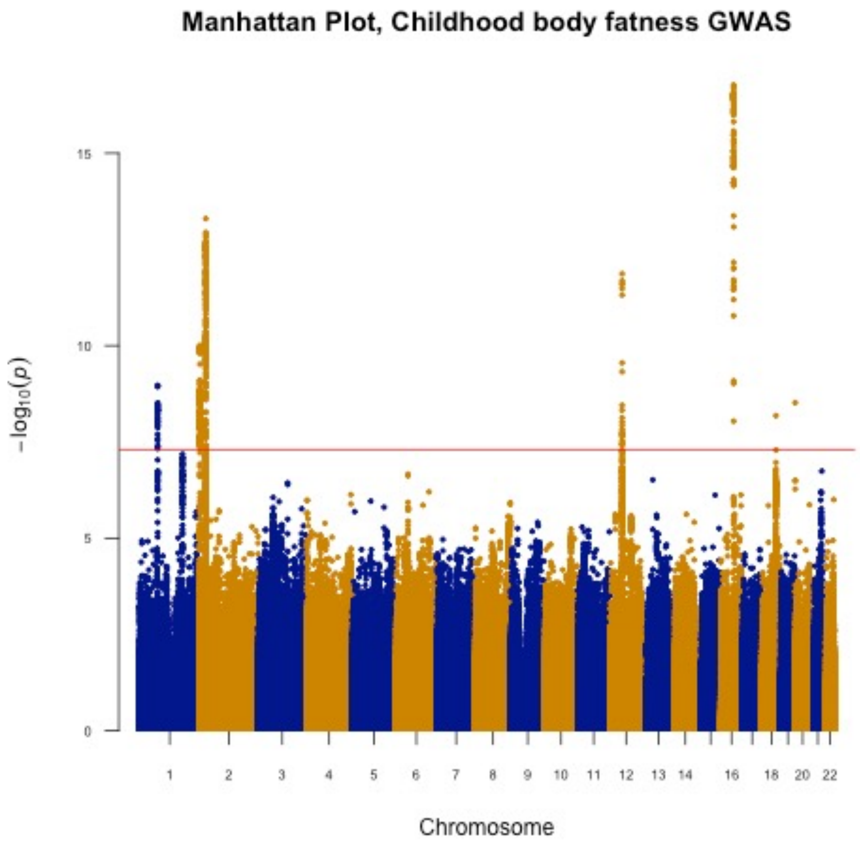
Table 2. Common Pitfalls, Recommendations, and Methods in Need of Development

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



QQplot, Childhood body fatness GWAS

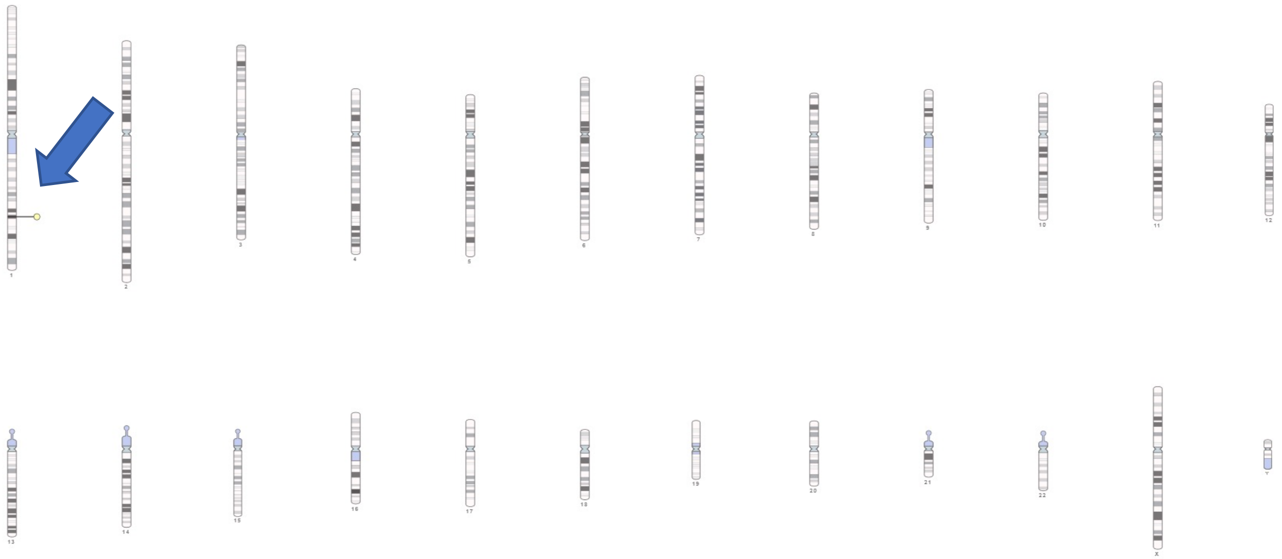


Manhattan Plot, Childhood body fatness GWAS

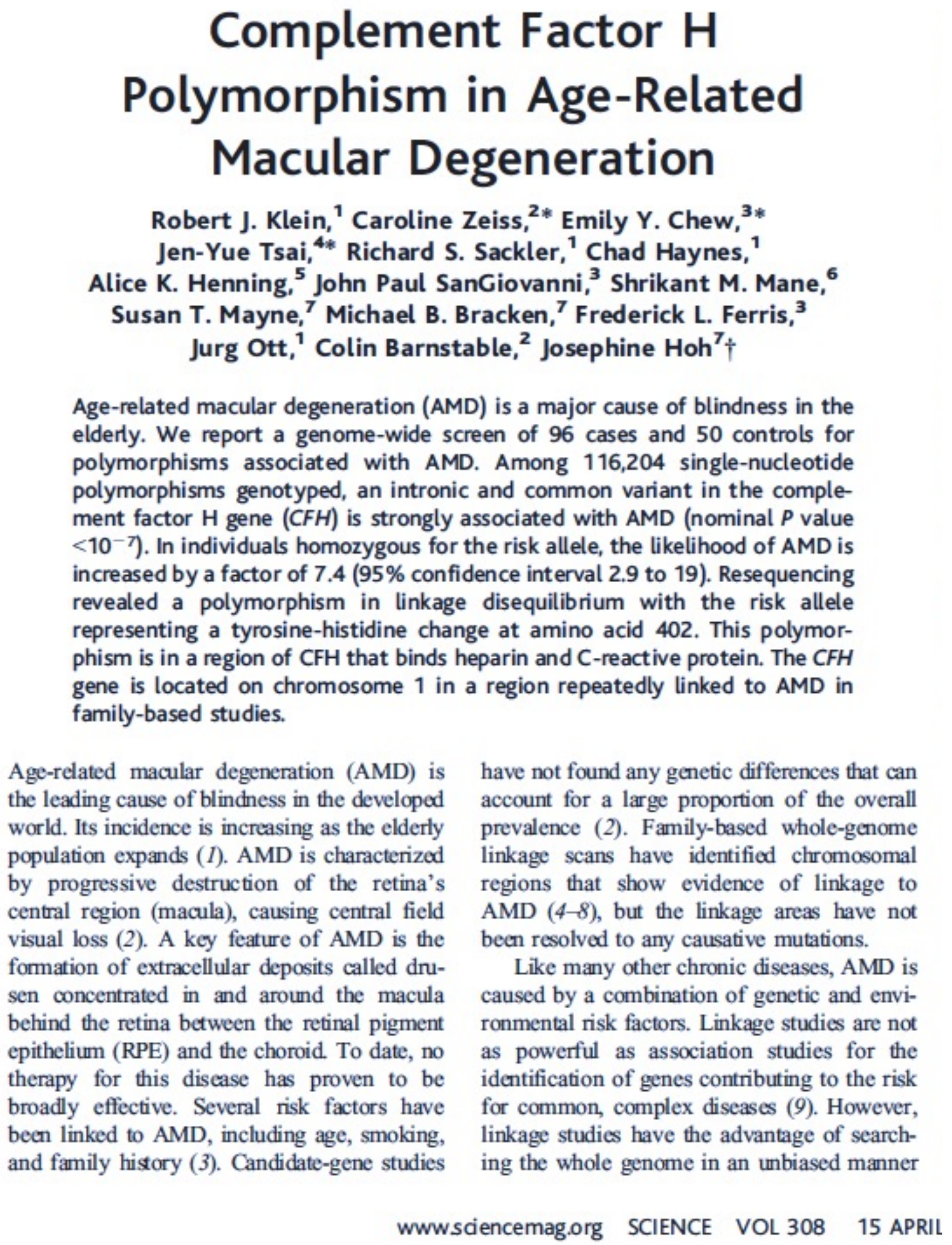
An association with p-value <5x10^{-8} is considered genome-wide significant

Warner, et al. Obesity 2021

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



Klein, Science 2005



Complement Factor H Polymorphism in Age-Related Macular Degeneration

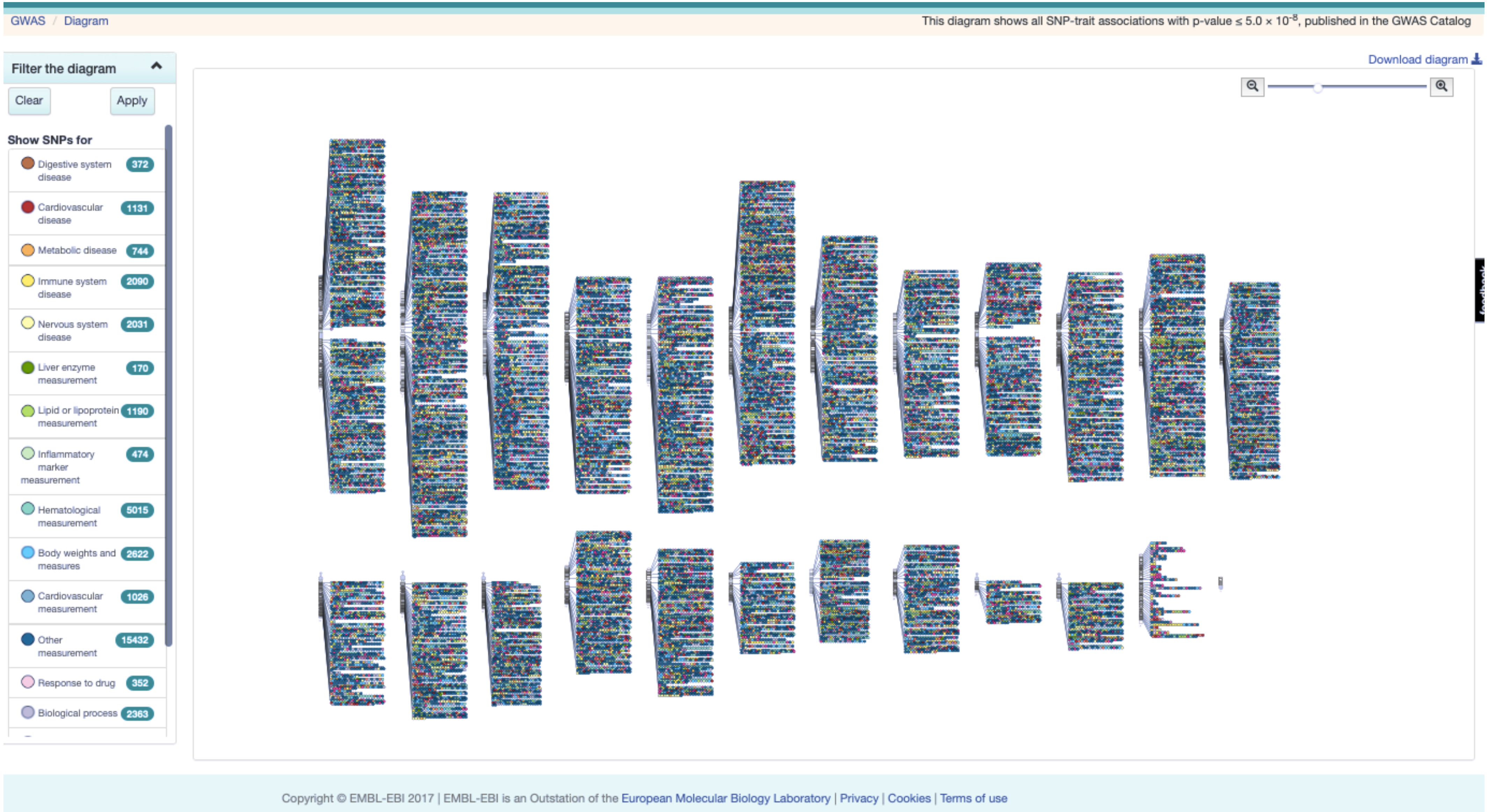
Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*]
Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1]
Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6]
Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3]
Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

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

April 2020



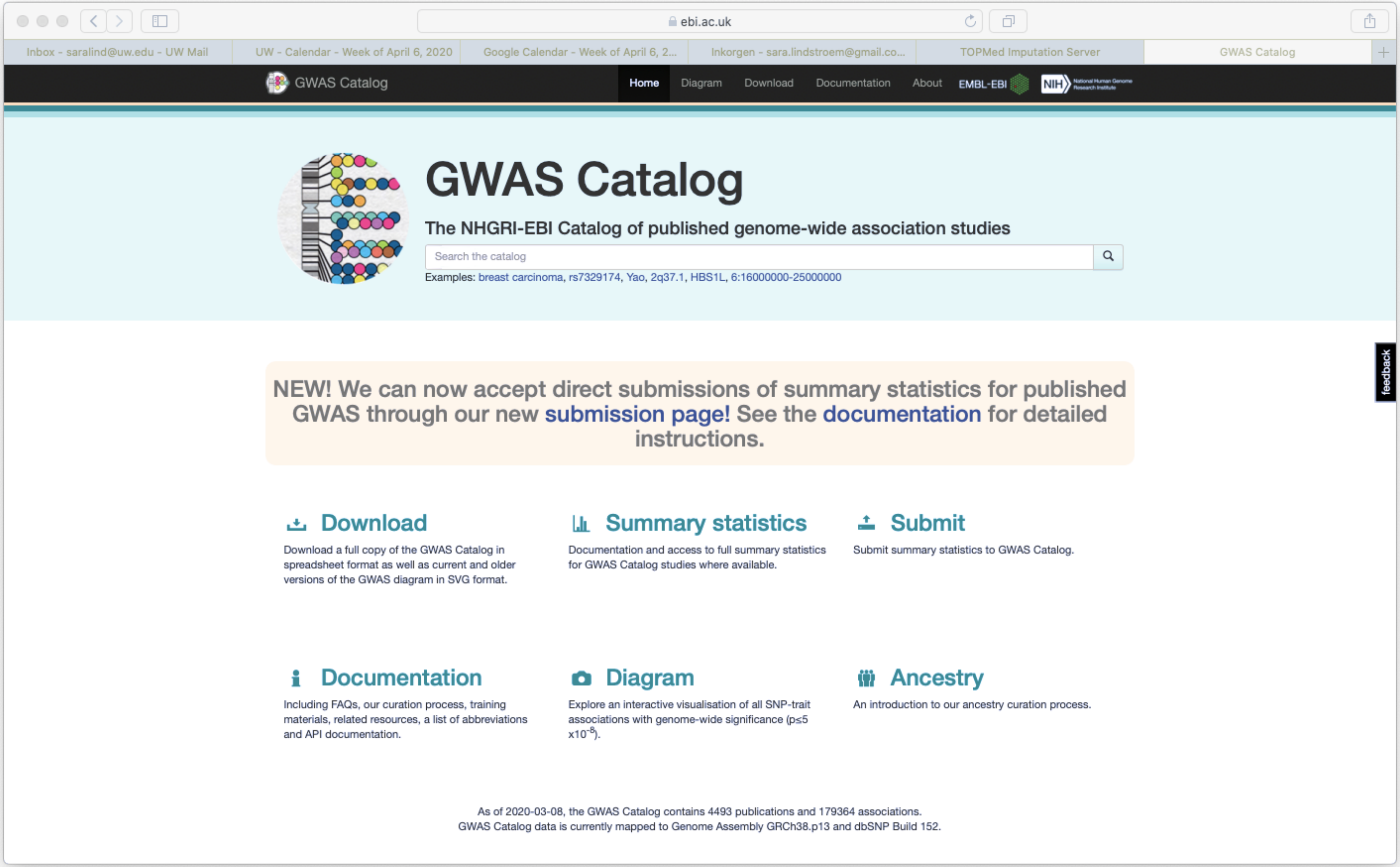
GWAS / Diagram

This diagram shows all SNP-trait associations with p-value ≤ 5.0×10^{-8}, published in the GWAS Catalog

Filter the diagram

Show SNPs for

● Digestive system disease	372
● Cardiovascular disease	1131
● Metabolic disease	744
● Immune system disease	2090
○ Nervous system disease	2031
● Liver enzyme measurement	170
● Lipid or lipoprotein measurement	1190
○ Inflammatory marker measurement	474
● Hematological measurement	5015
● Body weights and measures	2622
● Cardiovascular measurement	1026
● Other measurement	15432
○ Response to drug	352
● Biological process	2363


ebi.ac.uk
Inbox - saralind@uw.edu - UW Mail
UW - Calendar - Week of April 6, 2020
Google Calendar - Week of April 6, 2...
Inkorgen - sara.lindstroem@gmail.co...
TOPMed Imputation Server
GWAS Catalog
GWAS Catalog
Home
Diagram
Download
Documentation
About
EMBL-EBI
NIH
feedback

GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

NEW! We can now accept direct submissions of summary statistics for published GWAS through our new submission page! See the documentation for detailed instructions.

Download

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

Summary statistics

Documentation and access to full summary statistics for GWAS Catalog studies where available.

Submit

Submit summary statistics to GWAS Catalog.

Documentation

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

Diagram

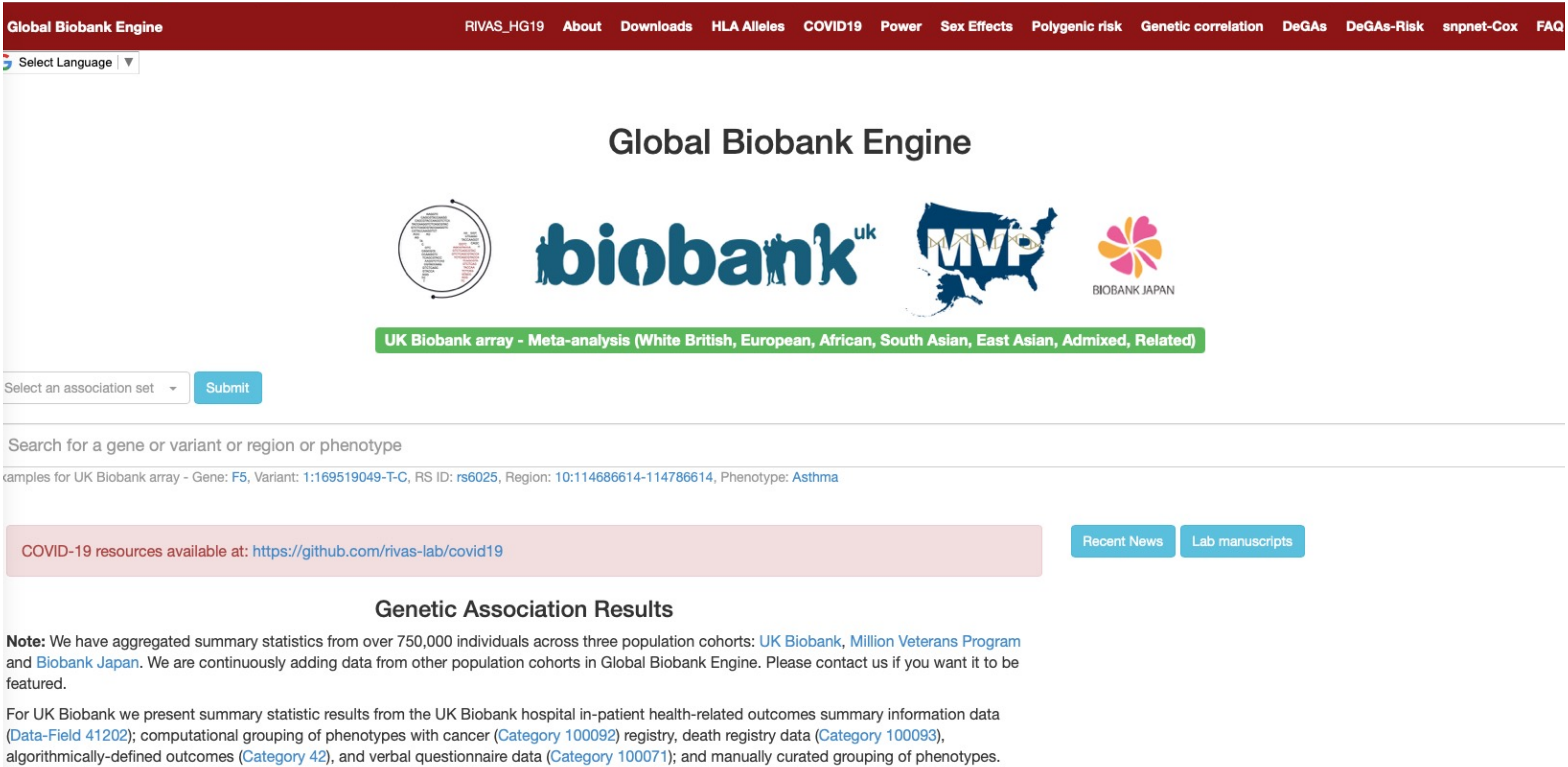
Explore an interactive visualisation of all SNP-trait associations with genome-wide significance ($p \leq 5 \times 10^{-8}$).

Ancestry

An introduction to our ancestry curation process.

As of 2020-03-08, the GWAS Catalog contains 4493 publications and 179364 associations.
GWAS Catalog data is currently mapped to Genome Assembly GRCh38.p13 and dbSNP Build 152.

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



Global Biobank Engine

UK Biobank array - Meta-analysis (White British, European, African, South Asian, East Asian, Admixed, Related)

xamples for UK Biobank array - Gene: F5, Variant: 1:169519049-T-C, RS ID: rs6025, Region: 10:114686614-114786614, Phenotype: Asthma

COVID-19 resources available at: https://github.com/rivas-lab/covid19

Genetic Association Results

Note: We have aggregated summary statistics from over 750,000 individuals across three population cohorts: UK Biobank, Million Veterans Program and Biobank Japan. We are continuously adding data from other population cohorts in Global Biobank Engine. Please contact us if you want it to be featured.

For UK Biobank we present summary statistic results from the UK Biobank hospital in-patient health-related outcomes summary information data (Data-Field 41202); computational grouping of phenotypes with cancer (Category 100092) registry, death registry data (Category 100093), algorithmically-defined outcomes (Category 42), and verbal questionnaire data (Category 100071); and manually curated grouping of phenotypes.

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

Some “classical” bias in the context of genetic epidemiology

- **Ascertainment bias**
    - Secondary phenotypes, e.g., Type 2 diabetes and BMI

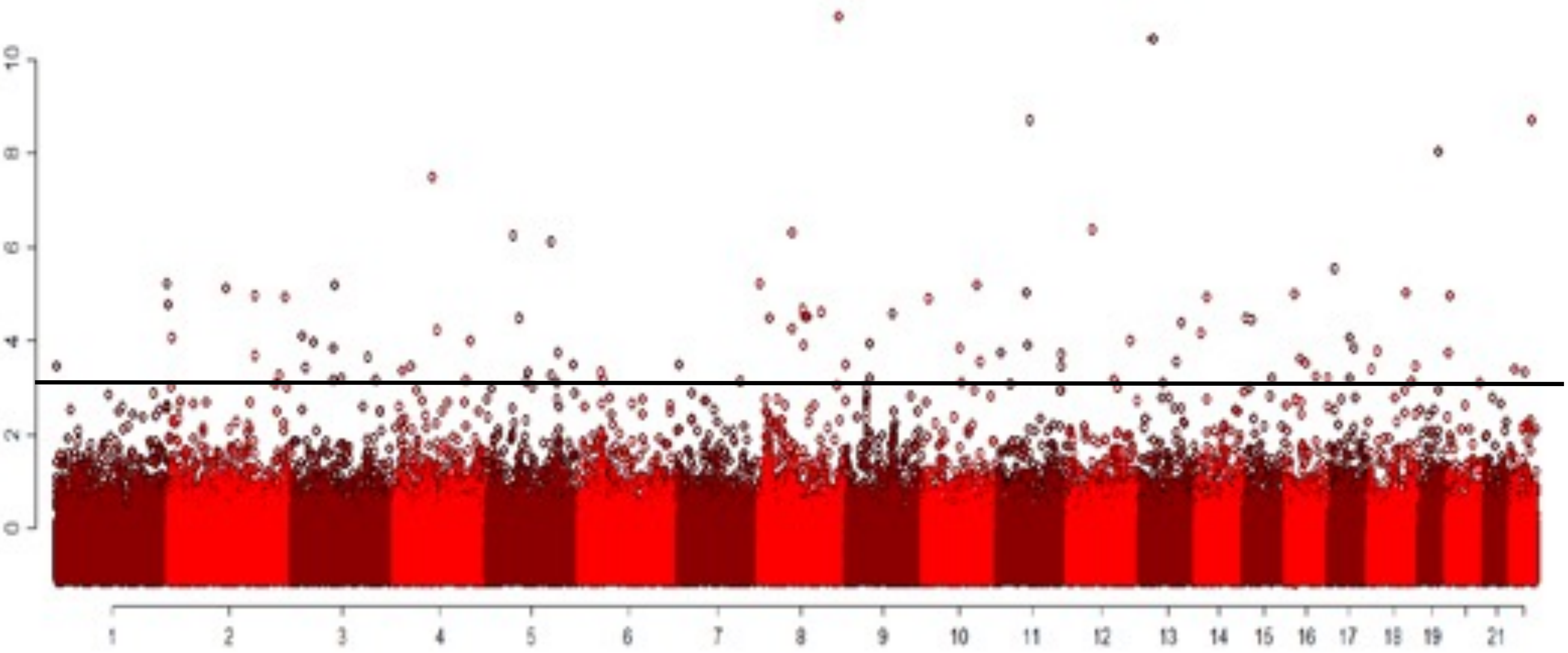
- **Survival bias**
    - When cases are recruited some time after they were diagnosed. Might lead to a milder form of disease. This is especially true for aggressive/fatal disease (e.g., pancreatic cancer, heart attack)

- **Diagnostic bias**
    - If the investigator determining the phenotype knows the genotype beforehand (e.g., if the radiologist knows that a potential pulmonary disease patient carries a high-risk genotype, she may look more carefully at the x-ray).

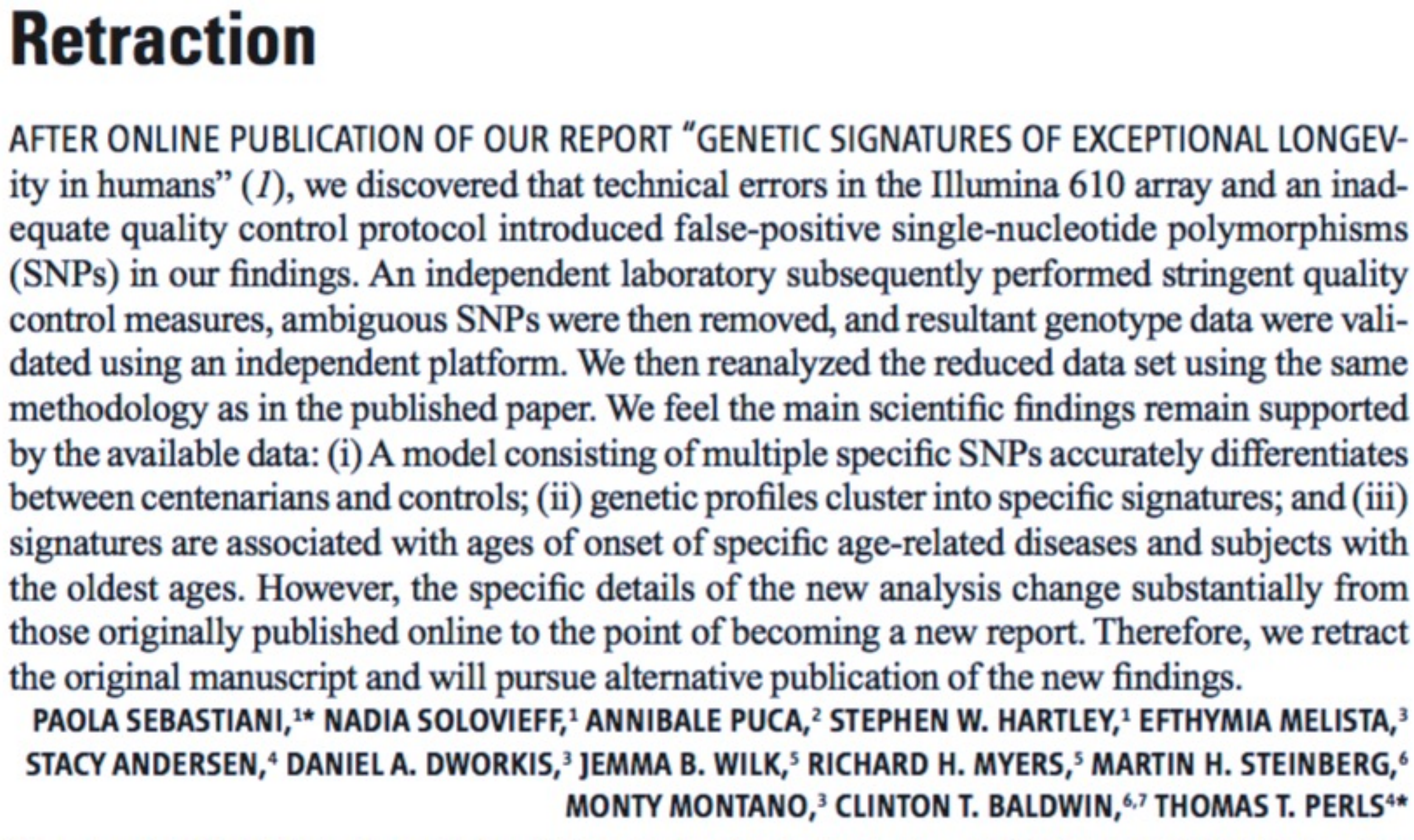
Differential genotyping error/missingness

- 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



Sebastiani, Science 2010



Retraction

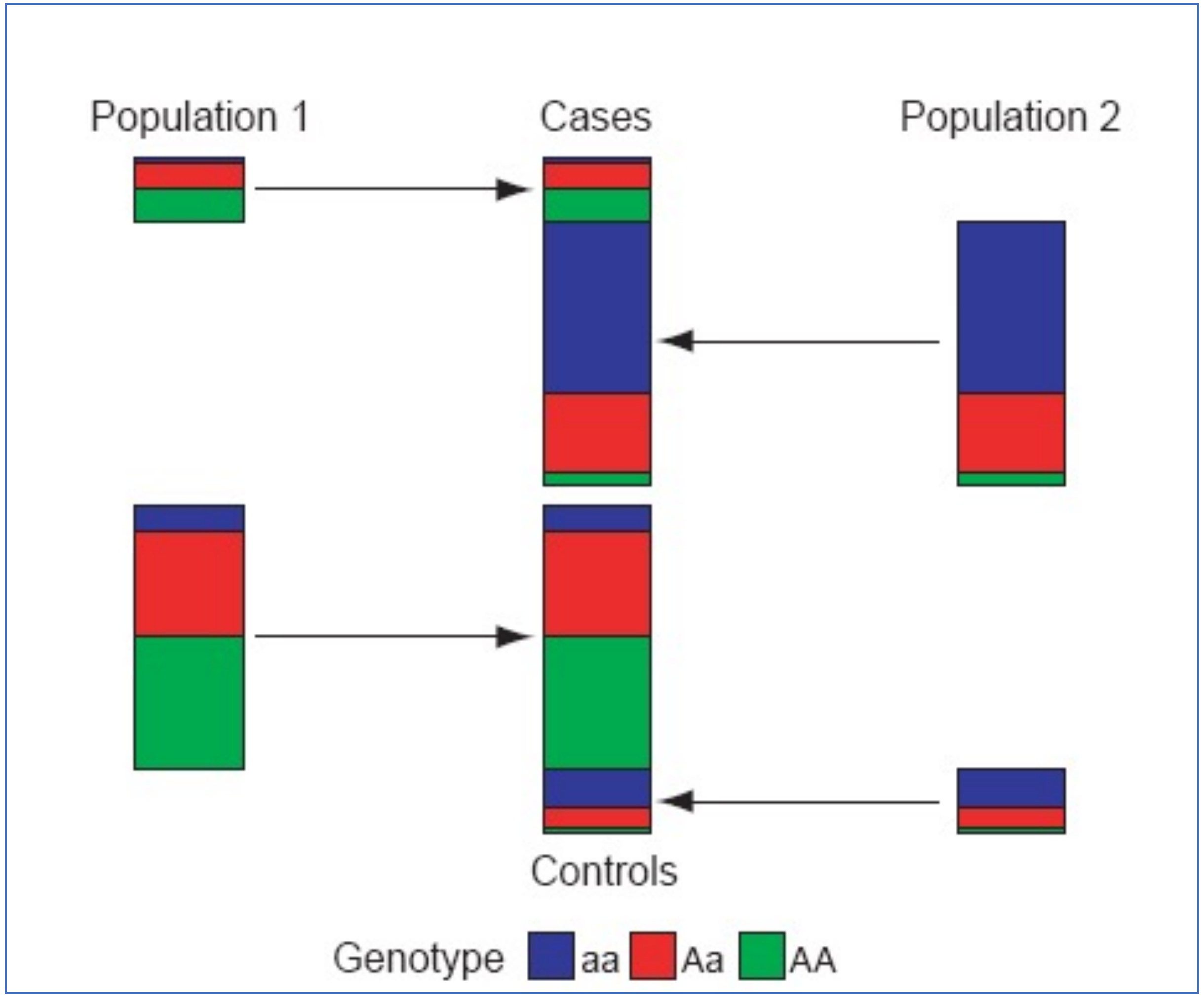
AFTER ONLINE PUBLICATION OF OUR REPORT “GENETIC SIGNATURES OF EXCEPTIONAL LONGEVity in humans” (*1*), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent 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]*

Population Stratification - Confounding by ancestry

- Group differences in allele frequencies AND outcome

- GWAS data pick up these differences! Use PCA to capture the information



Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

How to assess potential population stratification

- Most of the genetic markers in the genome (e.g. in a GWAS) are likely not associated with the disease

- The genomic control parameter (λ_{GC}) summarizes systematic inflation from a large number of association test results

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

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

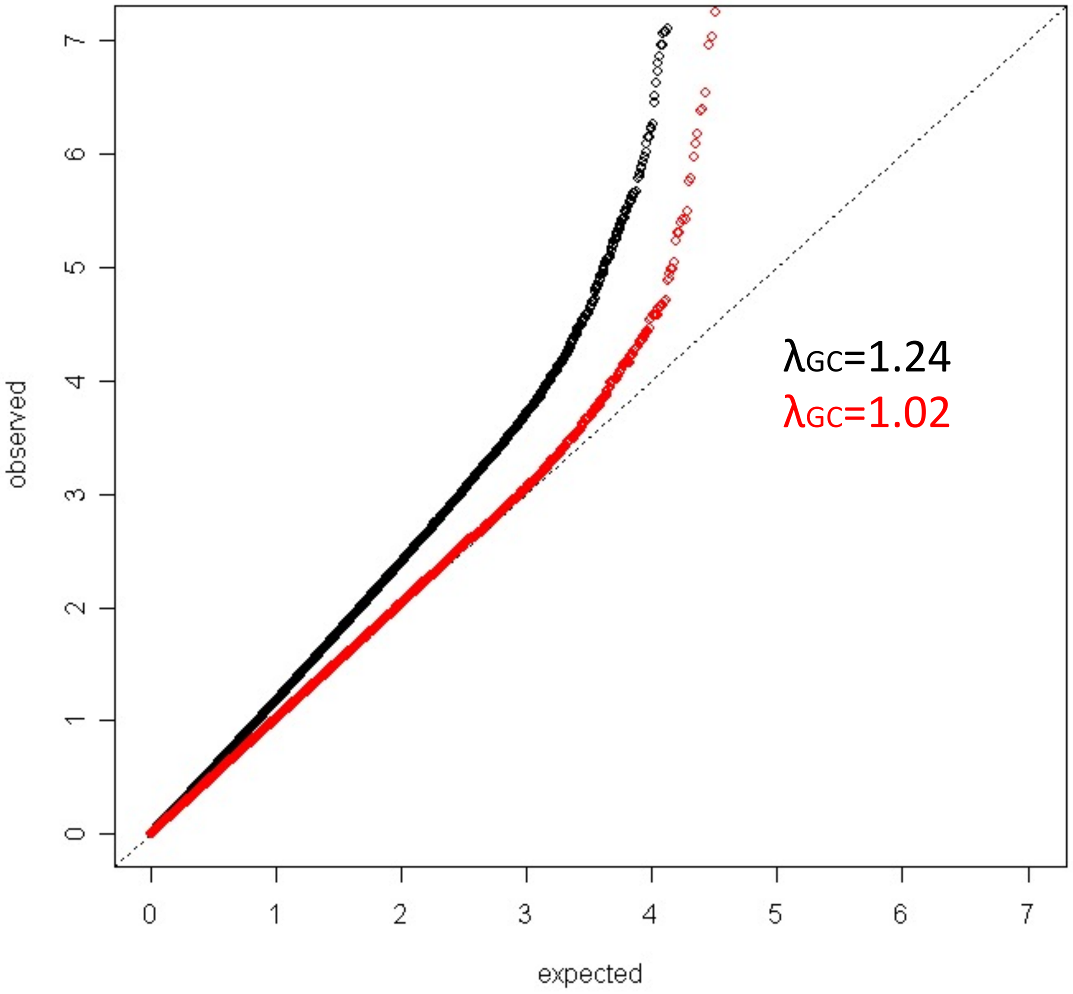
A few notes about λ_{GC}

- λ_{GC} should be close to 1 if no bias exists.
    - Rule of thumb: <1.05 is often ok, above 1.1 deserves attention (exception: when you have large sample sizes, we will come back to this)

- 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
    - λ_{GC} of 1.06 is a much bigger concern in studies with hundreds of samples compared to studies with thousands of samples

- A standard approach is to correct for inflation by dividing all test statistics by λ_{GC}
    - Drawback: Affects all SNPs, so SNPs that are not affected by bias are overpenalized and SNPs that are very affected by bias are underpenalized

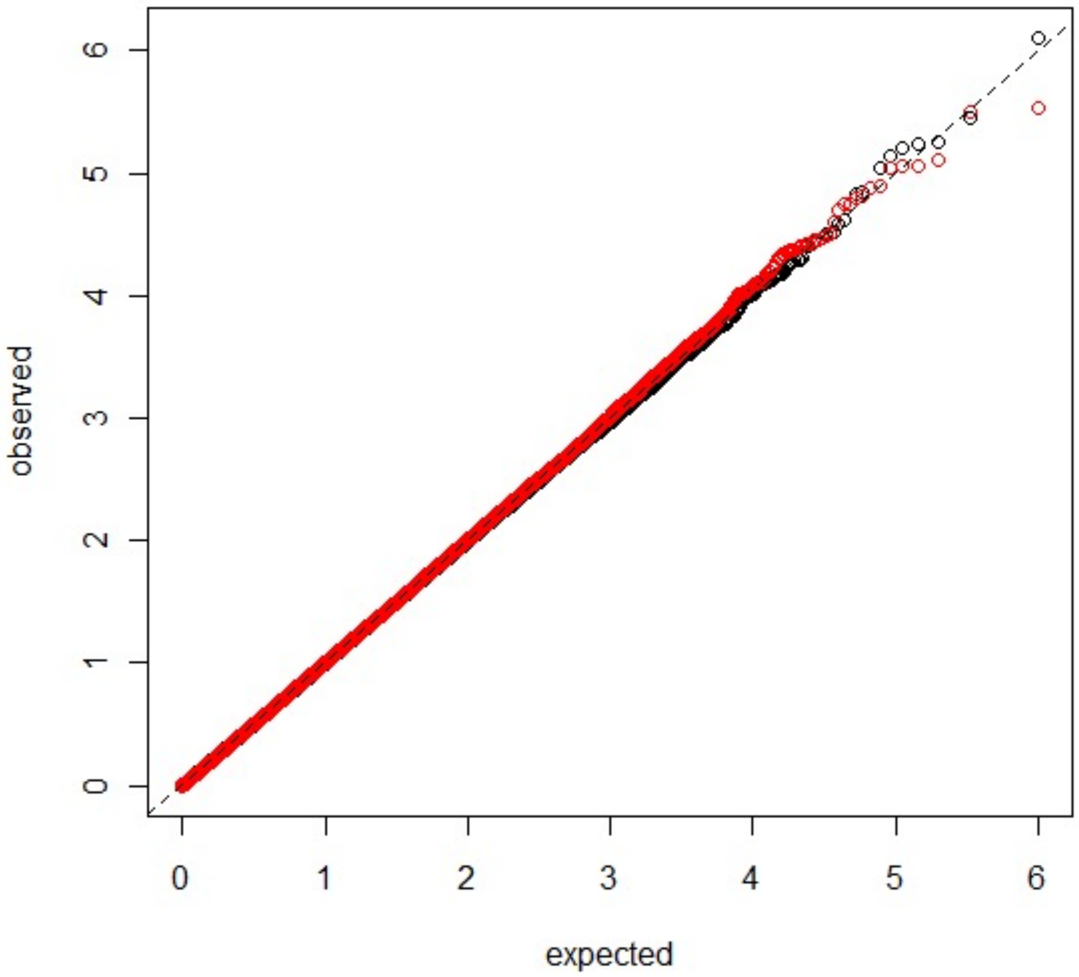
Hair Color in Nurses Health Study (n=2,287)



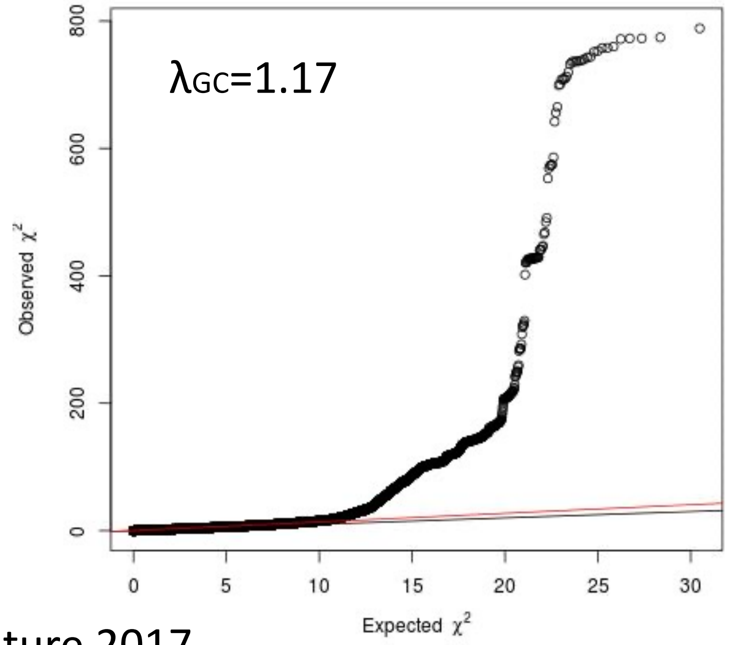
QQ plot for a GWAS of dark-light hair color in US European-ancestry subjects from the NHS. The black points are the p-values from the unadjusted tests. The red points are from **principal-component adjusted tests**.

Han 2008 Plos Genet

Breast Cancer GWAS



QQ plot for a GWAS of breast cancer in the same NHS samples (breast cancer risk does not correlate with European ancestry)



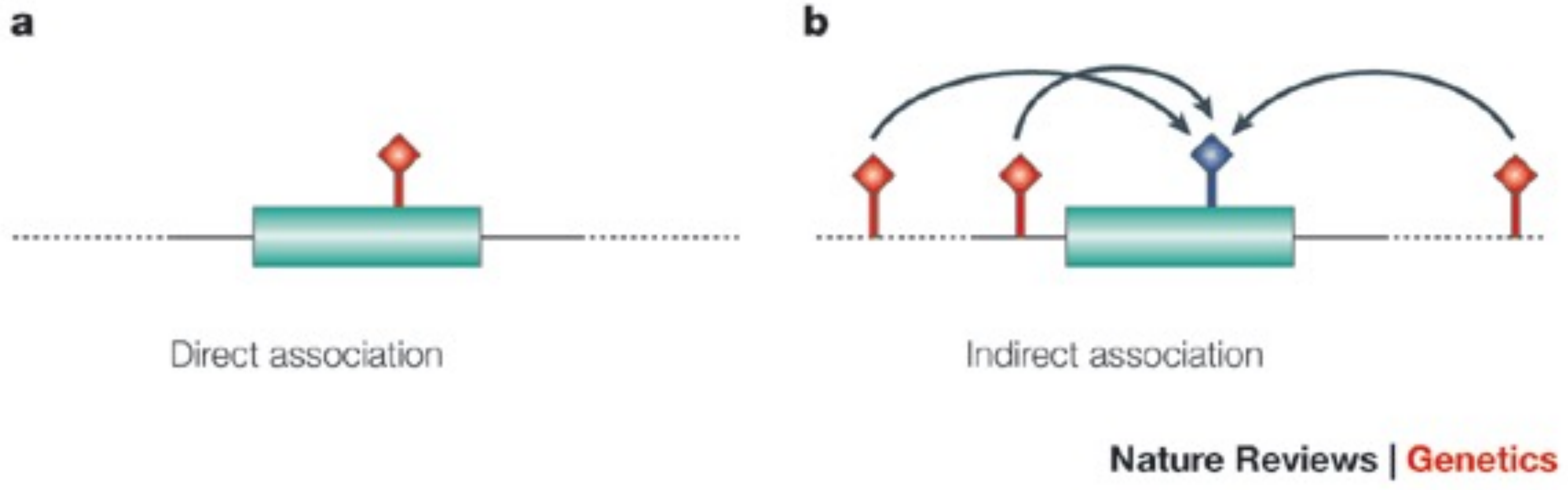
λ_{GC}=1.17

Michailidou, Nature 2017

A note about replication

- Want to see the signal in more than one population (e.g., longevity study)
- 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

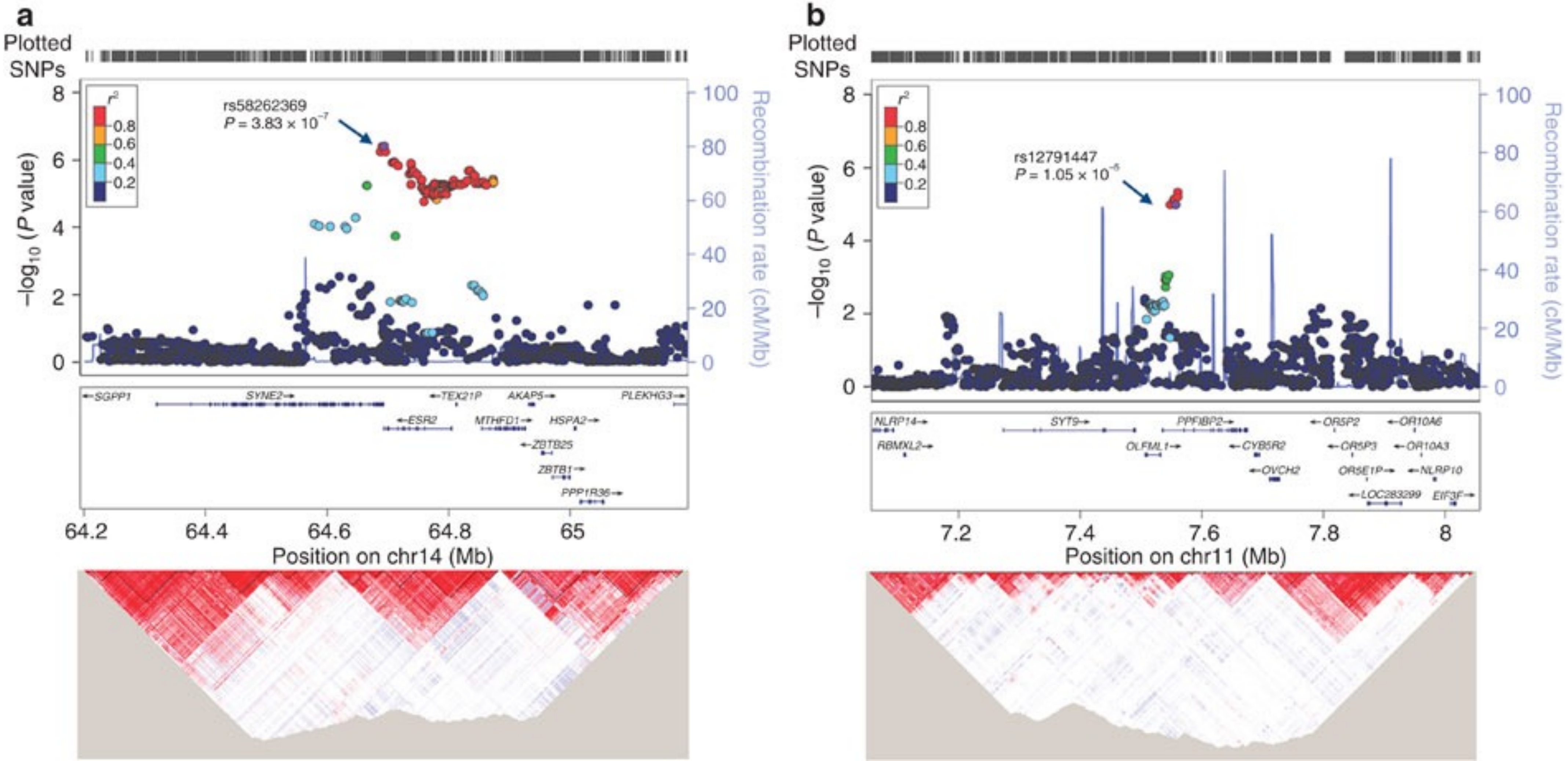
Follow up on GWAS hits: Fine-mapping



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

Hirschhorn & Daly. *Nature Reviews Genetics 2005*

Results from a prostate cancer GWAS



Wang. *Nature Comm, 2015*

Fine-mapping approaches

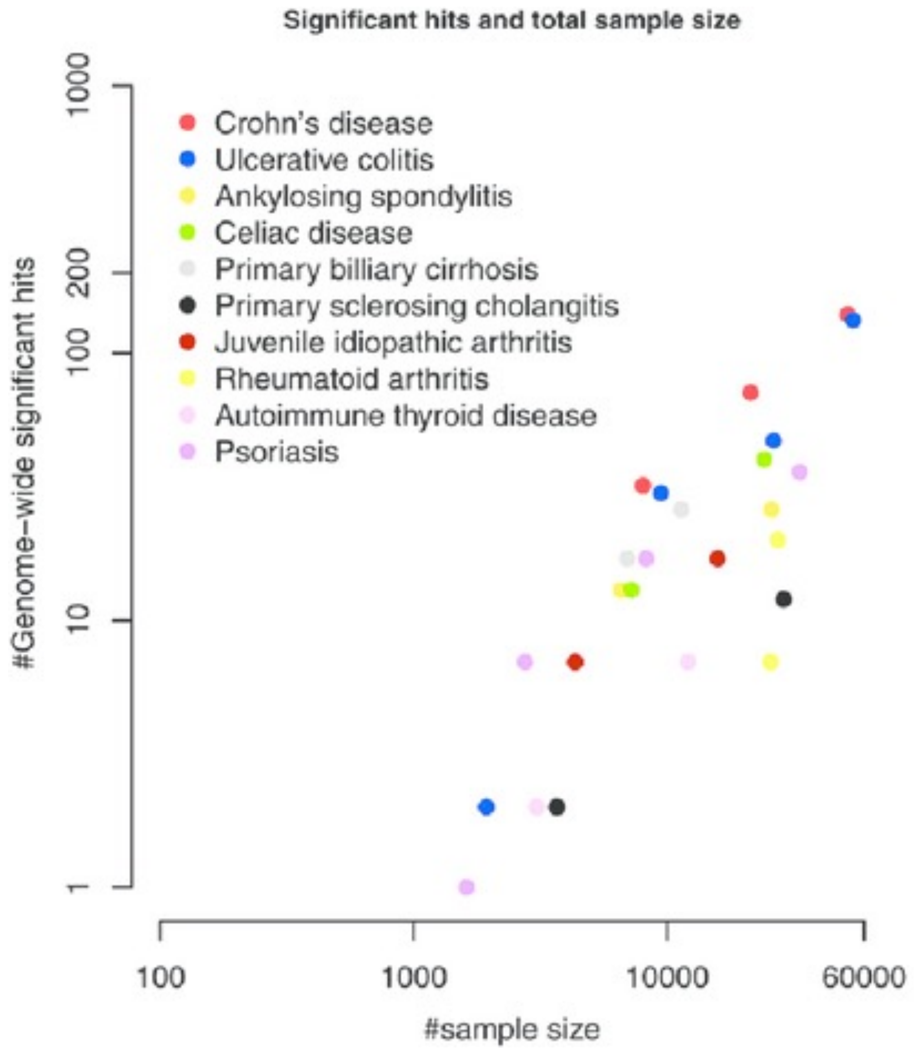
- Conditional regression analysis
    - Rerun analysis adjusting for the most significant SNP, see if any other SNP remains significant. Keep going until no more significant SNPs

- Calculate posterior probabilities for each SNP

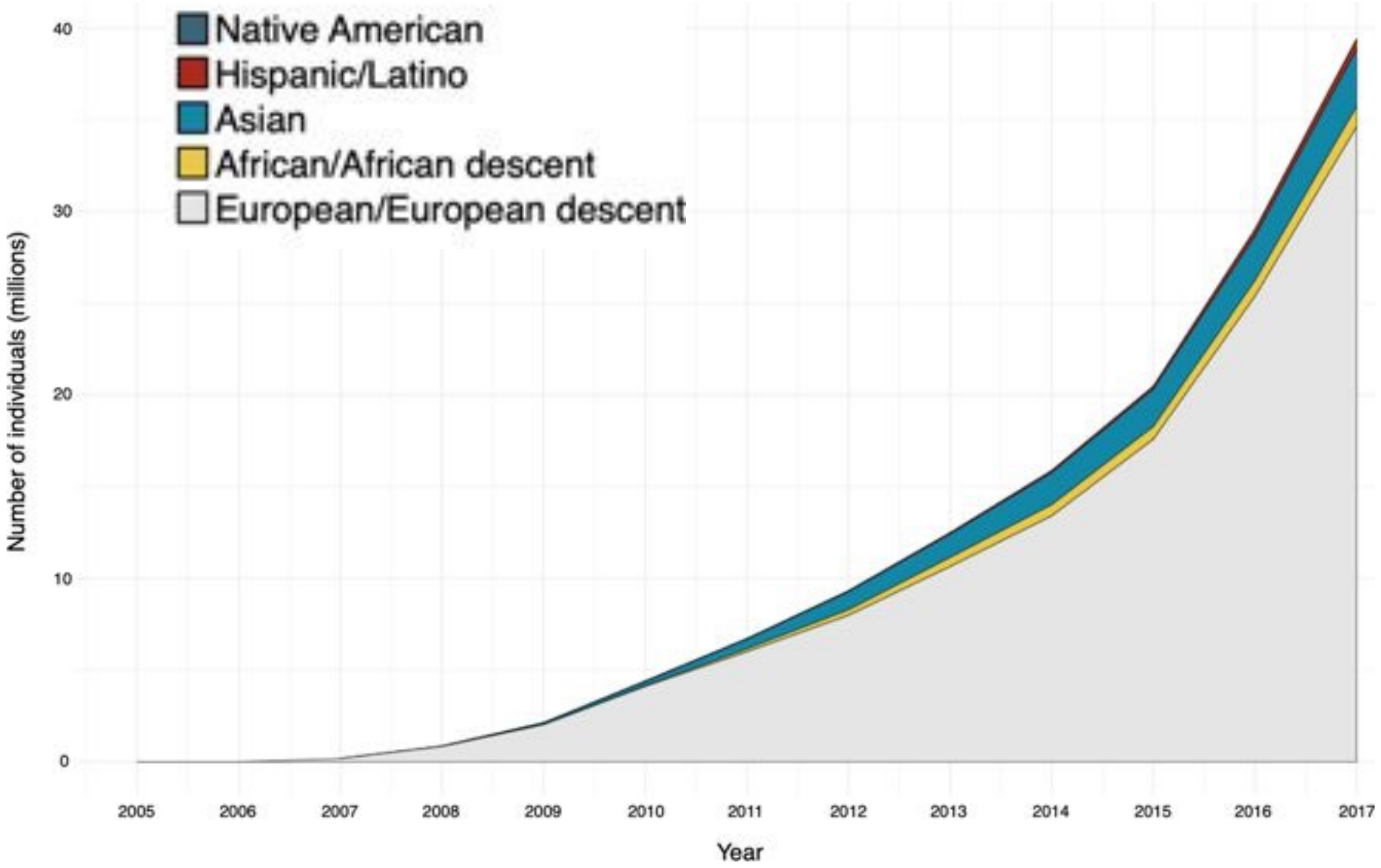
- Incorporate “functional” information to identify biological plausible SNPs

- Choose a set of “potentially causal variants” and take them forward for downstream analysis.

Sample Size is key to GWAS!



Chen, HMG 2014



Wojcik, Nature 2019

Meta-analysis

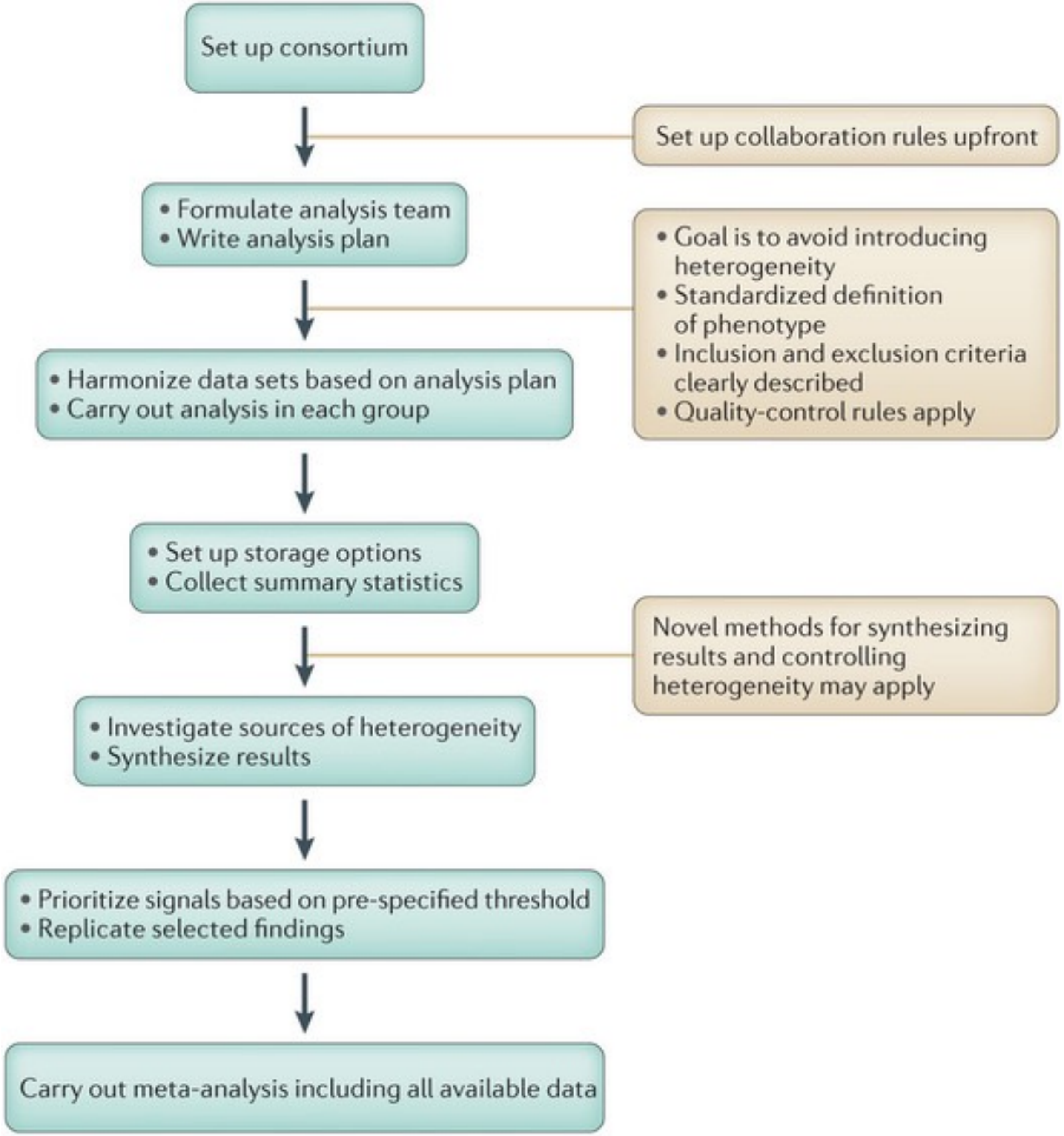
- Sample size is the key for a successful genetic association study

- International collaborations to pool data 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

de Bakker, Hum Mol Genetics 2008



Set up consortium

Set up collaboration rules upfront

- Formulate analysis team
- Write analysis plan

- Goal is to avoid introducing heterogeneity
- Standardized definition of phenotype
- Inclusion and exclusion criteria clearly described
- Quality-control rules apply

- Harmonize data sets based on analysis plan
- Carry out analysis in each group

- Set up storage options
- Collect summary statistics

Novel methods for synthesizing results and controlling heterogeneity may apply

- Investigate sources of heterogeneity
- Synthesize results

- Prioritize signals based on pre-specified threshold
- Replicate selected findings

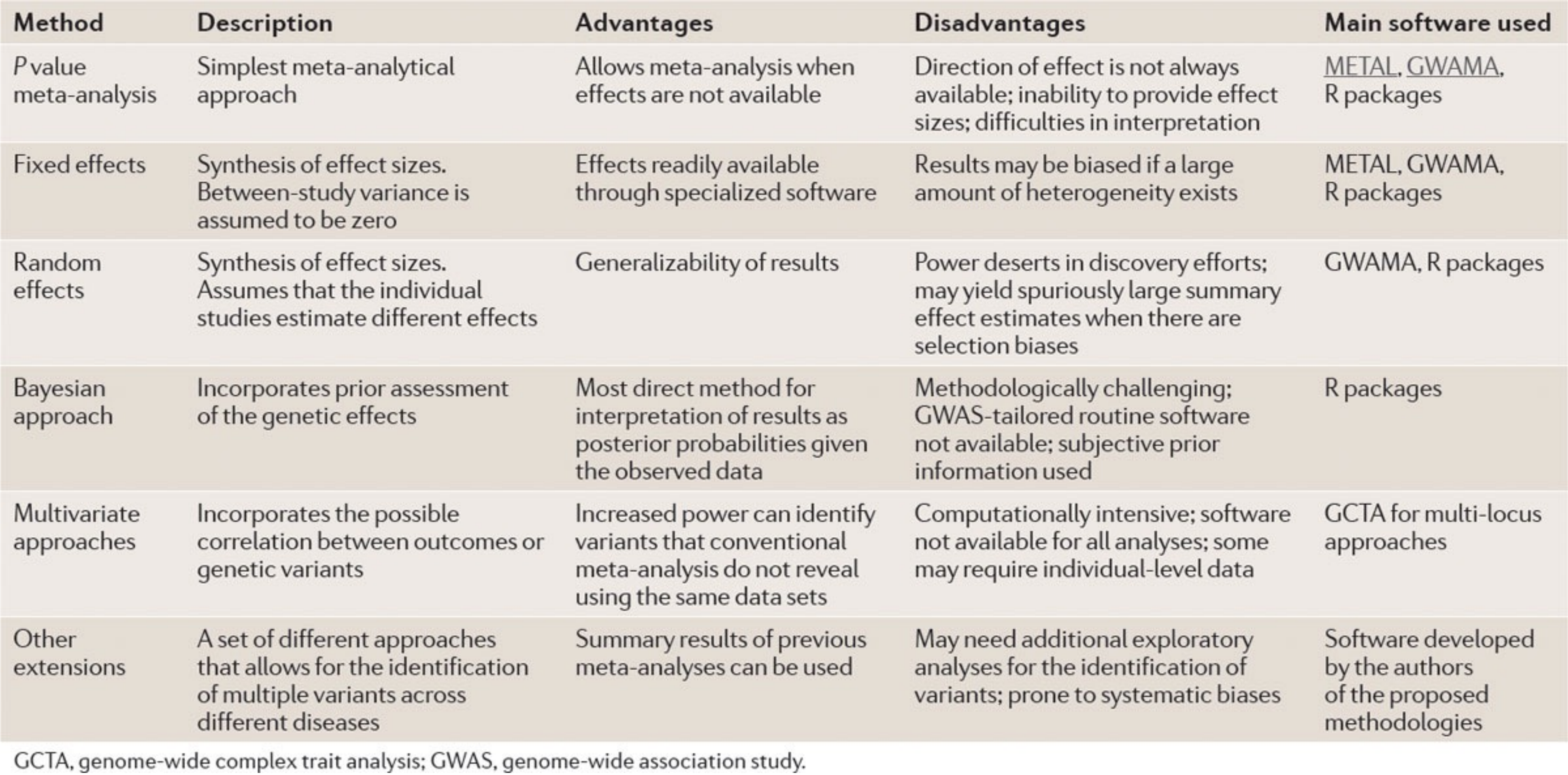
Carry out meta-analysis including all available data

Evangelou & Ioannidis, Nature Rev Genetics 2013

Nature Reviews | Genetics

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
P 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.

Evangelou & Ioannidis, Nature Rev Genetics 2013