

# Genome-wide association studies (GWAS): benefits and limitations

Julian Hecker

Instructor in Medicine, Harvard Medical School

Channing Division of Network Medicine, Brigham and Women's Hospital

[rejhe@channing.harvard.edu](mailto:rejhe@channing.harvard.edu)

# Genome-wide association studies (GWAS)

- Testing single nucleotide variants/polymorphisms (SNVs/SNPs) for association with complex trait/disease
  - $\geq 1$  million SNPs tested (whole-genome sequencing, imputed chip data)
  - significance level at  $5 * 10^{-8}$  (multiple testing)
- First GWAS in 2005 (age-related macular degeneration)
- Polygenic architecture, small effect sizes, large sample sizes required (meta-analyses)
- Observed 'gap' between amount of GWAS findings and heritability
- Goal: identify disease mechanisms, drug targets, personalized medicine

## REVIEW

### Five Years of GWAS Discovery

Peter M. Visscher,<sup>1,2,\*</sup> Matthew A. Brown,<sup>1</sup> Mark I. McCarthy,<sup>3,4</sup> and Jian Yang<sup>5</sup>

AJHG 2012

## REVIEW

### 10 Years of GWAS Discovery: Biology, Function, and Translation

Peter M. Visscher,<sup>1,2,\*</sup> Naomi R. Wray,<sup>1,2</sup> Qian Zhang,<sup>1</sup> Pamela Sklar,<sup>3</sup> Mark I. McCarthy,<sup>4,5,6</sup> Matthew A. Brown,<sup>7</sup> and Jian Yang<sup>1,2</sup>

AJHG 2017

## REVIEW

### 15 years of GWAS discovery: Realizing the promise

Abdel Abdellaoui,<sup>1,\*</sup> Loic Yengo,<sup>2</sup> Karin J.H. Verweij,<sup>3</sup> and Peter M. Visscher<sup>2</sup>

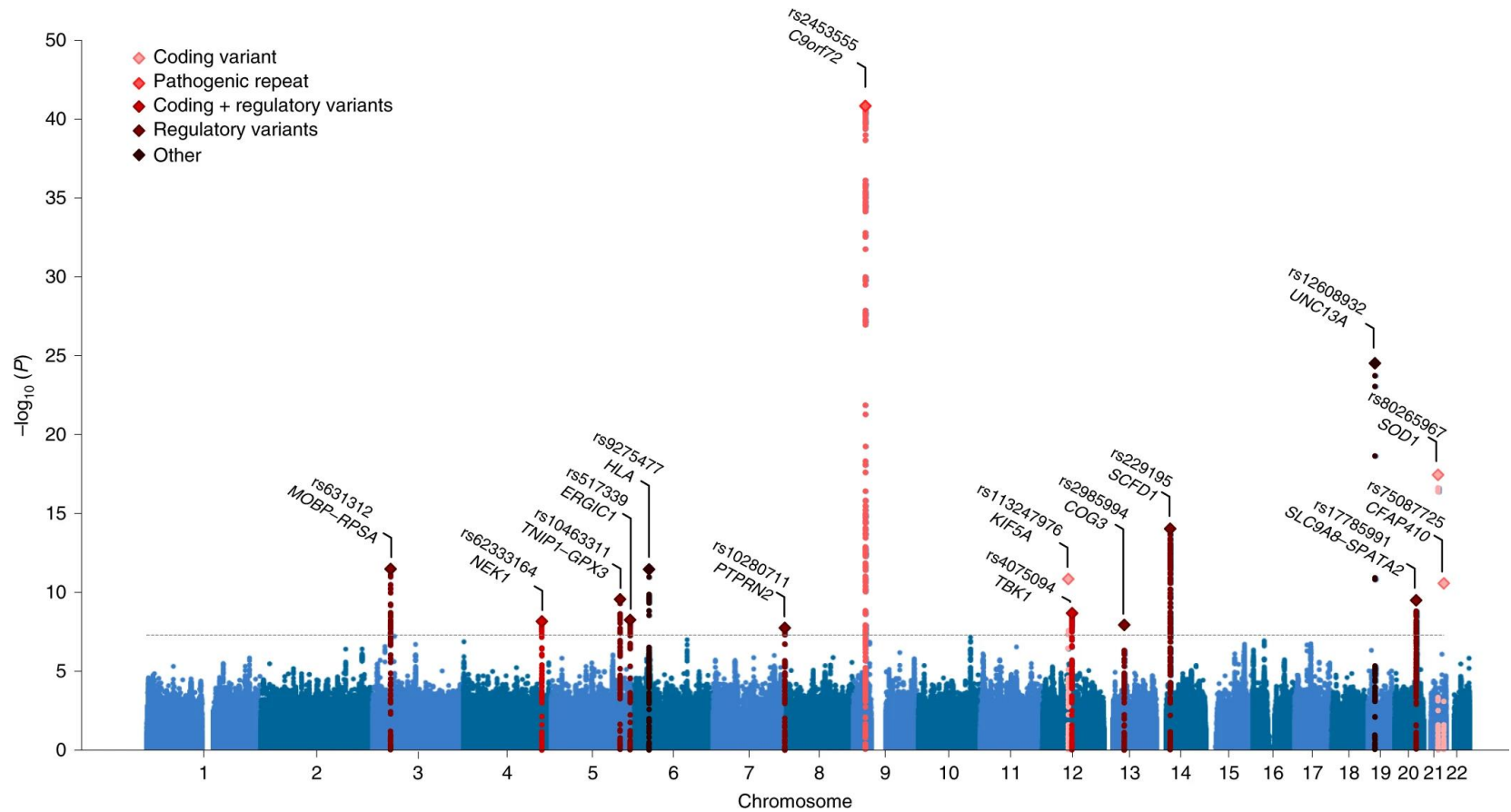
AJHG 2023

## Article

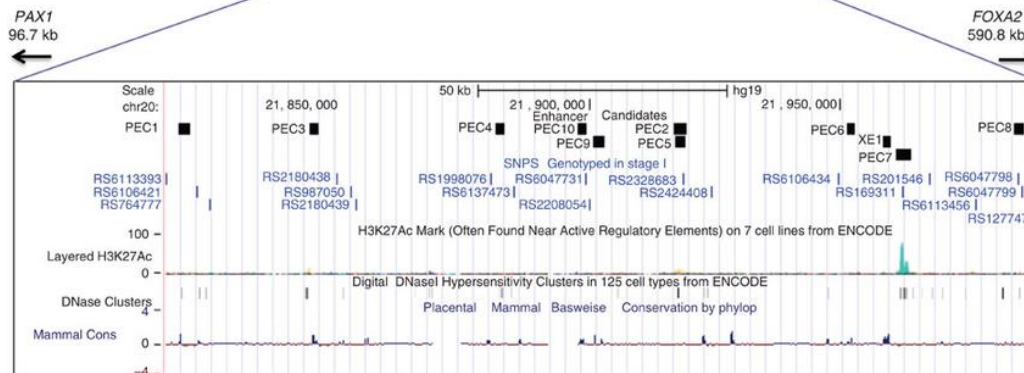
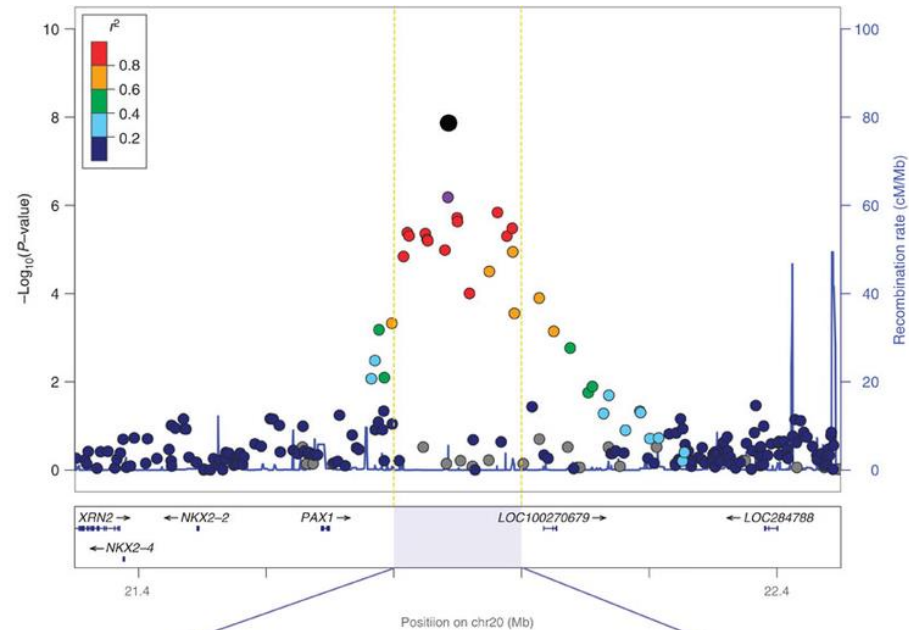
### A saturated map of common genetic variants associated with human height

Yengo et al., Nature 2022

# Manhattan plot: GWAS of amyotrophic lateral sclerosis



# Regional association plot: GWAS of idiopathic scoliosis in females



# GWAS sample sizes vs. loci

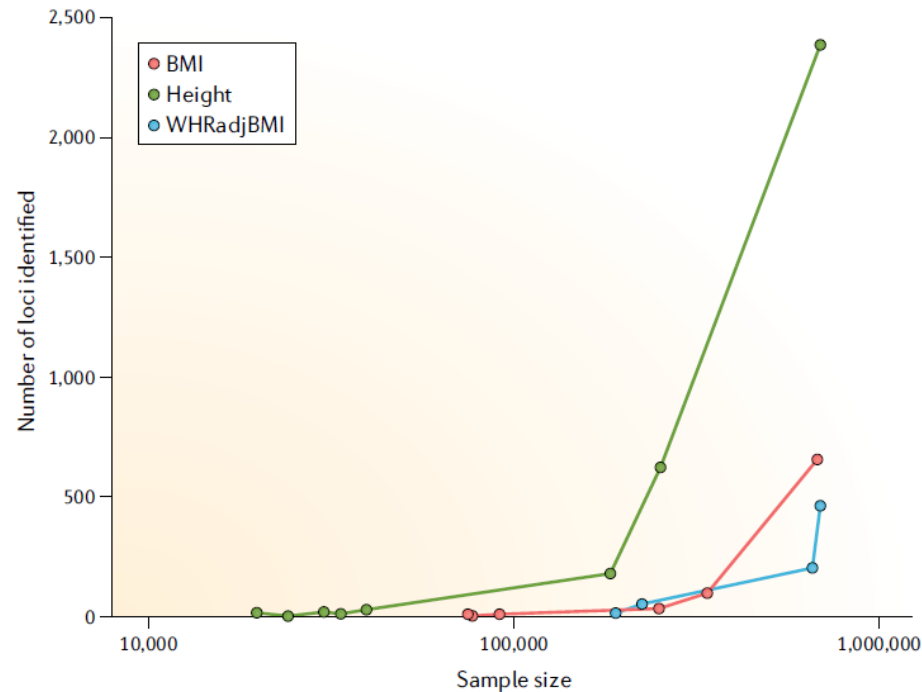


Fig. 2 | **Number of loci identified as a function of GWAS sample size.** A plot of the number of independent or near-independent genome-wide significant loci ( $P < 5 \times 10^{-8}$ ) reported from genome-wide association studies (GWAS) in European or predominantly European populations for three anthropometric traits: body mass index (BMI)<sup>22,103,104,182,333,334</sup>, height<sup>22,335-341</sup> and waist-to-hip ratio adjusted for BMI (WHRadjBMI)<sup>167,342-344</sup>. For each trait, there is a threshold sample size above which the rate of locus discovery accelerates. The identification of risk loci has yet to plateau for these traits.

# Challenges and limitations

## **Confounding (correlation is not causation)**

- Population stratification can introduce false positives
- causal variants in linkage disequilibrium (LD)

## **Technical:**

- Sample sizes growing, software solutions imperfect and bugs occur

## **From GWAS to biology:**

- Identifying causal variants challenging
- Identifying corresponding genes/mechanisms even harder

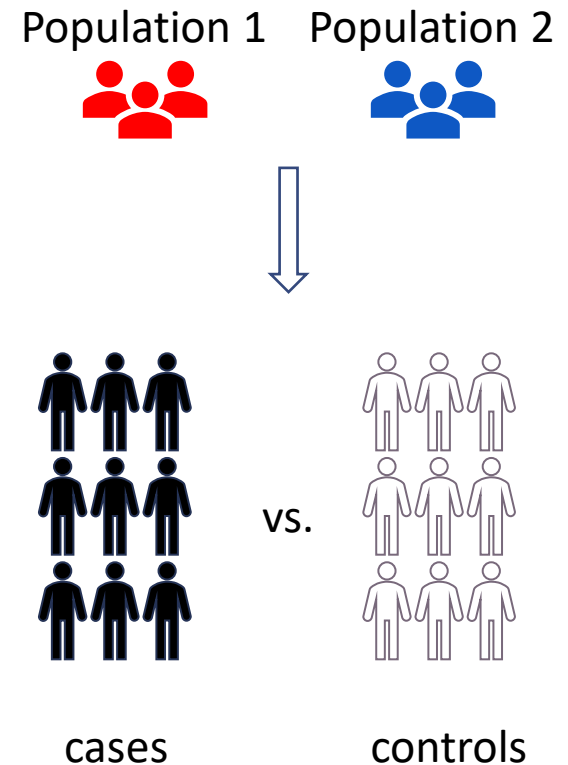
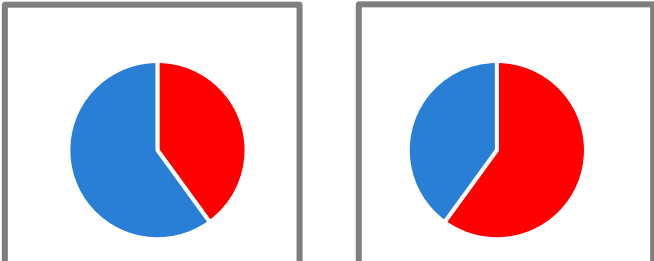
## Benefits and limitations of genome-wide association studies

*Vivian Tam<sup>1</sup>, Nikunj Patel<sup>1</sup>, Michelle Turcotte<sup>1</sup>, Yohan Bossé<sup>2,3</sup>, Guillaume Paré<sup>1,4</sup> and David Meyre<sup>1,4,5\*</sup>*

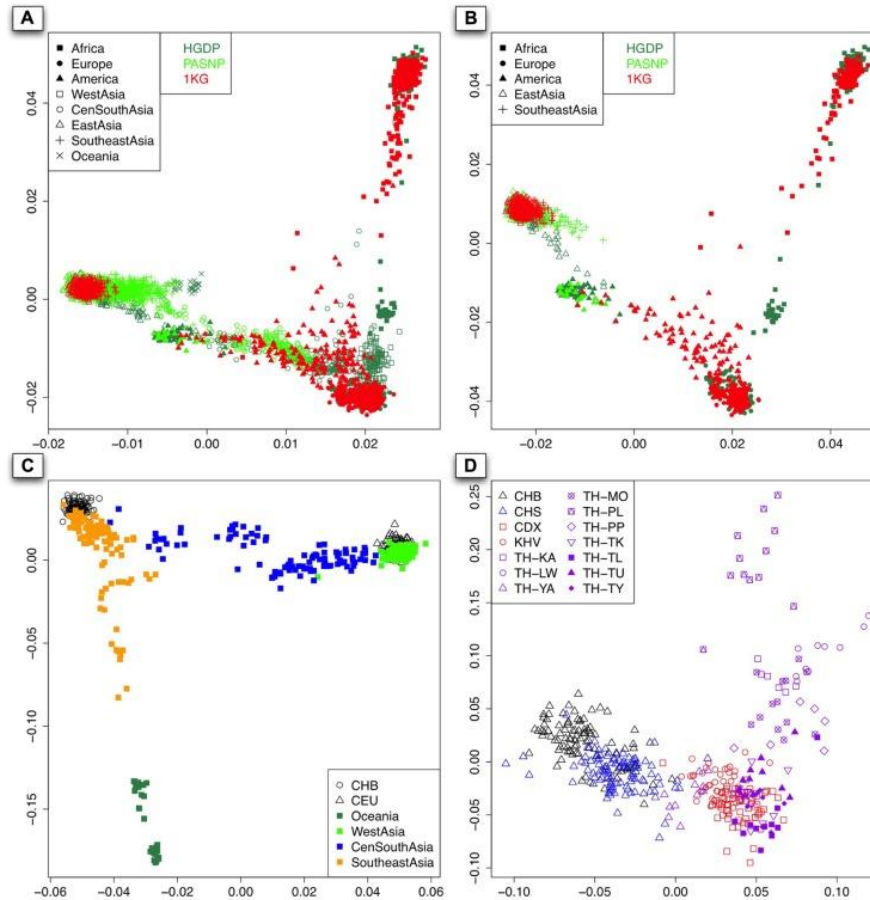
# Confounding: population stratification

- Allele frequency differences between populations due to genetic drift and gene flow
- Sampling from different populations can lead to false-positive association findings

genotype	# case	# control
AA	421	319
AB	469	505
BB	110	176



# Confounding: population stratification



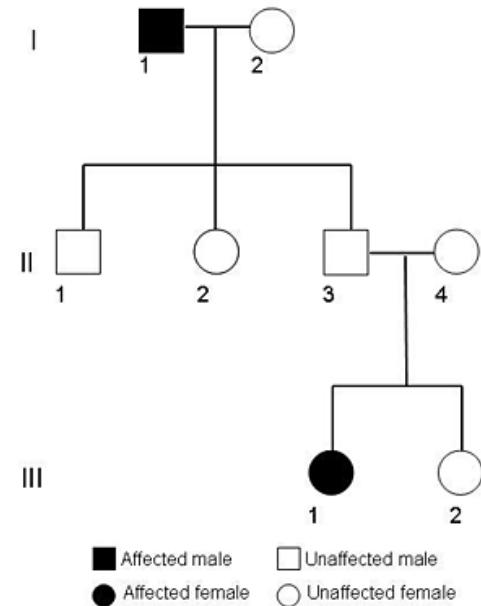
Common approach: principal component analysis (PCA)-based covariate adjustment

Lu and Xu, *Frontiers in Genetics* 2013



# Confounding: population stratification

- Other approach: **family-based association studies**
- Classical transmission disequilibrium test (TDT) in 1993
- TDT tests for both linkage and association
- Robust against
  - population stratification
  - misspecified phenotype distributions
  - ascertainment based on phenotypes
- Extended to Family-Based Association Test (FBAT) framework **by Nan Laird et al.**



Volume 50, Issue 4  
April 2000



RESEARCH ARTICLES | APRIL 28 2000

## A Unified Approach to Adjusting Association Tests for Population Admixture with Arbitrary Pedigree Structure and Arbitrary Missing Marker Information

Subject Area: # Genetics

Daniel Rabinowitz; Nan Laird

*Hum Hered* (2000) 50 (4): 211–223.

<https://doi.org/10.1159/000022918> [Article history](#)

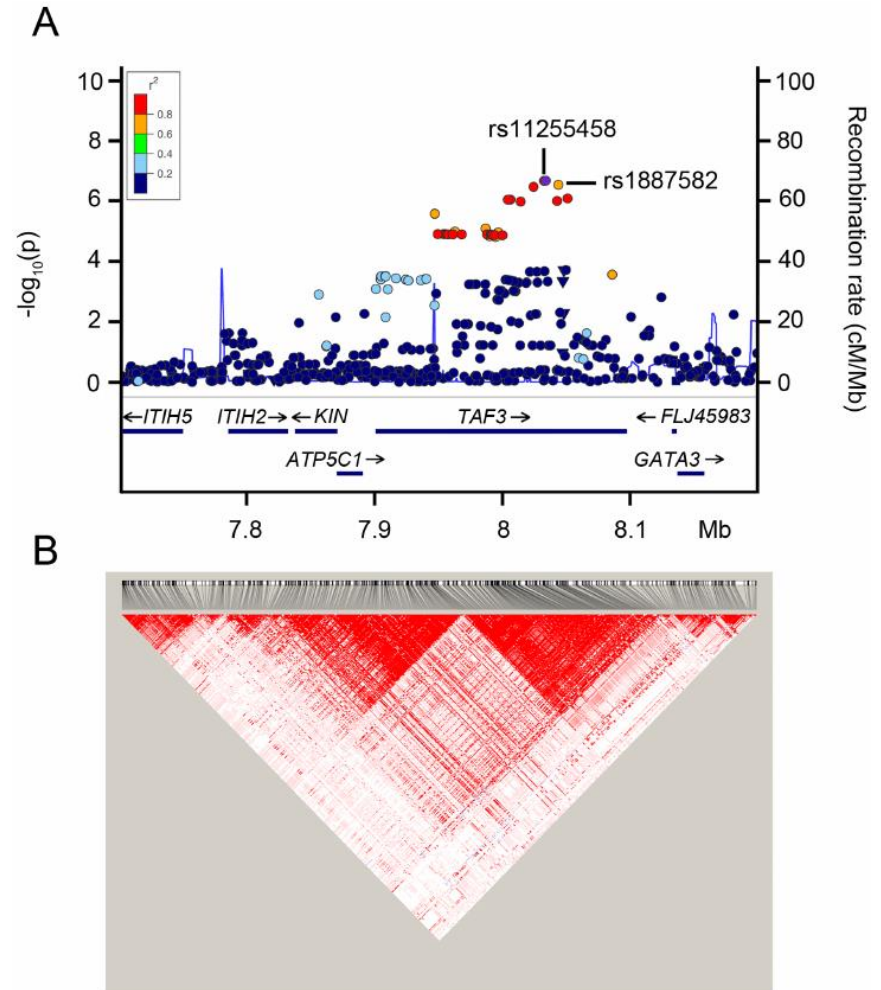
[Share](#) [Tools](#)

Rabinowitz and Laird, Human Heredity 2000

# Confounding due to LD

- Signal cluster of associations due to LD
- Causal variant cannot be identified directly

*LD: non-random association between alleles of genetic loci in proximity due to linkage.*



# Confounding due to LD

- Finemapping approaches: identifying most likely causal variant (statistically)
- Resolution level lower bounded by LD structure



From genome-wide associations to candidate causal variants by statistical fine-mapping

Daniel J. Schaid<sup>1</sup>\*, Wenan Chen<sup>2</sup> and Nicholas B. Larson<sup>1</sup>

Table 1 | Commonly used Bayesian fine-mapping software

Software	Trait type <sup>a</sup>	Input covariates <sup>b</sup>	Uses summary statistics?	Maximum number of causal variants <sup>c</sup>	Input annotation?	Causal search	Main output	Refs
BIMBAM v1.0	qt and binary	No	No	Fixed	No	Exhaustive	Bayes factor	113,114
mvBIMBAM v1.0.0	mqt	No	Yes	1	No	Exhaustive	Bayes factor	115,116
SNPTEST v2.5.4-beta3	qt, binary, mqt and multinomial	No	No	1	No	Exhaustive	Bayes factor	117
piMASS v0.9	qt and binary	No	No	Computed	No	MCMC	Bayes factor and PIP	45
BVS v4.12.1	Binary	Yes	No	Computed	Yes	MCMC	Bayes factor and PIP	97,118,119
FM-QTL	qt	No	No	Computed	Yes	MCMC	Bayes factor and PIP	96
DAP v1.0.0	qt	Yes	Yes	1, fixed and computed	Yes	Exhaustive	Bayes factor and PIP	52
Fine-mapping	Multinomial	Yes	No	Computed	No	Greedy	PIP	30
Trinculo	Multinomial	Yes	No	Computed	No	Greedy	Bayes factor and PIP	30,120
BayesFM	Binary	Yes	No	20	No	MCMC	PIP	30
ABF	qt and binary <sup>d</sup>	Yes	Yes	1	No	Exhaustive	Bayes factor	121
fgwas v0.3.6	qt and binary <sup>d</sup>	No	Yes	1	Yes	Exhaustive	Bayes factor and PIP	95
CAVIAR/eCAVIAR	qt and binary <sup>d</sup>	No	Yes	Fixed	No	Exhaustive	$\rho$ probability confidence set and PIP	46,101
PAINTOR v3.0	qt, binary <sup>d</sup> and mqt	No	Yes	Fixed and computed	Yes	Exhaustive and MCMC	Bayes factor and PIP	54,57,71
CAVIARBF v0.2.1	qt and binary <sup>d</sup>	No	Yes	Fixed	Yes	Exhaustive	Bayes factor and PIP	47,94
FINEMAP v1.1	qt and binary <sup>d</sup>	No	Yes	Fixed	No	Shotgun stochastic search	Bayes factor and PIP	53
JAM in R2BGLiMS v0.1	qt and binary <sup>d</sup>	No	Yes	Fixed and computed	No	Exhaustive and MCMC	Bayes factor and PIP	55

# Technical errors: software implementations

- Versatile gene-based association study VEGAS/VEGAS2: Gene-based association test using summary statistics and LD information from reference panel (Liu et al. 2010, Mishra and MacGregor 2015)
- Finding: top-x% implementation incorrect, can be fixed by minor modification (single bracket misplaced)
- can introduce false positives

$n$  SNPs mapped to gene

z-scores  $z_1, \dots, z_n$  from single variant analysis

Two options: top-SNP or top-x%

$$T = \sum_{\text{top } x\%} z_o^2(i)$$

where

$$z_o^2(1) \geq z_o^2(2) \geq \dots \geq z_o^2(n)$$

## Reporting Correct $p$ Values in VEGAS Analyses

Julian Hecker,<sup>1</sup> Anna Maaser,<sup>2,3</sup> Dmitry Prokopenko,<sup>1</sup> Heide Loehlein Fier,<sup>1,4</sup> and Christoph Lange<sup>4,5</sup>

<sup>1</sup>Institute of Genomic Mathematics, University of Bonn, Bonn, Germany

<sup>2</sup>Institute of Human Genetics, University of Bonn, Bonn, Germany

<sup>3</sup>Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany

<sup>4</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

<sup>5</sup>Department of Medicine, Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

# From GWAS to biology

- the majority of GWAS associations remain mechanistically uncharacterized
- Most GWAS hits in noncoding regions of the genome, uncertainty about which gene is responsible for their biological effects

## REVIEW

### From variant to function in human disease genetics

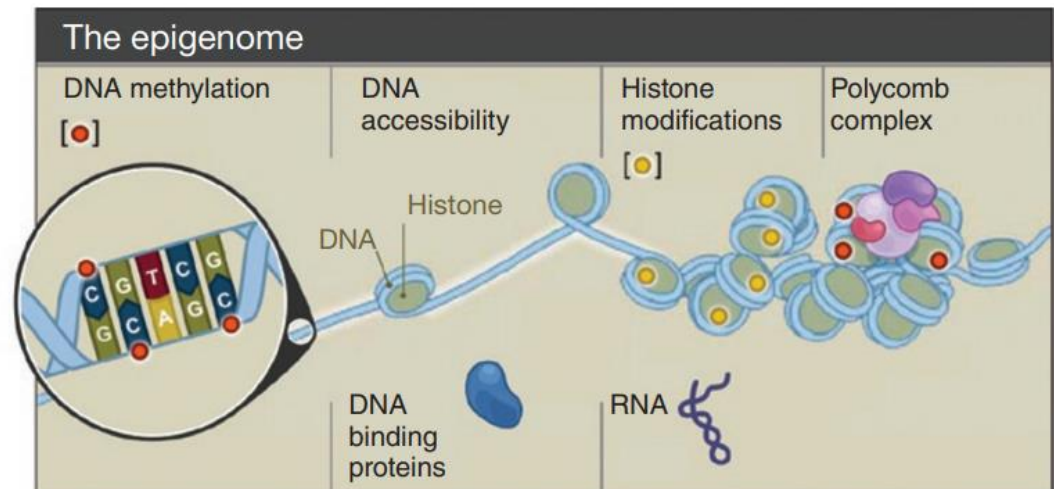
Tuuli Lappalainen<sup>1,2\*</sup> and Daniel G. MacArthur<sup>3,4,5\*</sup>

Over the next decade, the primary challenge in human genetics will be to understand the biological mechanisms by which genetic variants influence phenotypes, including disease risk. Although the scale of this challenge is daunting, better methods for functional variant interpretation will have transformative consequences for disease diagnosis, risk prediction, and the development of new therapies. An array of new methods for characterizing variant impact at scale, using patient tissue samples as well as in vitro models, are already being applied to dissect variant mechanisms across a range of human cell types and environments. These approaches are also increasingly being deployed in clinical settings. We discuss the rationale, approaches, applications, and future outlook for characterizing the molecular and cellular effects of genetic variants.

# From GWAS to biology: functional annotations

- Projects such as ENCODE (Nature 2020), Gencode (Nucleic Acids Res 2021), and Roadmap Epigenomics enabled predictions of most severe class gene-disrupting variants
- For most variant classes, accurate predictions difficult

## Roadmap Epigenomics





Bernstein et al., Nature Biotechnology 2013

# From GWAS to biology: QTL

- Another approach to reveal unknown putative gene regulatory effects: molecular quantitative trait locus (molQTL) mapping
- molQTL: genetic variation associated with molecular traits (gene expression, splicing and chromatin accessibility).
- most common: cis-eQTLs (cis-associated QTLs associated to gene expression levels)
- Limited by available tissues and cell types
- New directions through single cell sequencing and spatial transcriptomics

Primer | [Published: 25 January 2023](#)

## Molecular quantitative trait loci

[François Aguet](#) , [Kaur Alasoo](#), [Yang I. Li](#), [Alexis Battle](#), [Hae Kyung Im](#), [Stephen B. Montgomery](#) & [Tuuli Lappalainen](#) 

[Nature Reviews Methods Primers](#) **3**, Article number: 4 (2023) | [Cite this article](#)

**3448** Accesses | **1** Citations | **56** Altmetric | [Metrics](#)



# From GWAS to biology: general SNP to gene strategies

Computational gene prioritization tools

Closest gene not necessarily causally involved



## Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity

Steven Gazal<sup>1,2,3,4</sup>, Omer Weissbrod<sup>3,4</sup>, Farhad Hormozdiari<sup>3,4</sup>, Kushal K. Dey<sup>3,4</sup>, Joseph Nasser<sup>4</sup>, Karthik A. Jagadeesh<sup>3,4</sup>, Daniel J. Weiner<sup>4</sup>, Huwenbo Shi<sup>3,4</sup>, Charles P. Fulco<sup>4,5,10</sup>, Luke J. O'Connor<sup>4</sup>, Bogdan Pasanici<sup>6</sup>, Jesse M. Engreitz<sup>4,7,8</sup> and Alkes L. Price<sup>3,4,9</sup>

Article

<https://doi.org/10.1038/s41588-023-01443-6>

## Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases

Received: 14 August 2020

Accepted: 9 June 2023

Published online: 13 July 2023

Check for updates

Elle M. Weeks<sup>1,37</sup>, Jacob C. Ulirsch<sup>1,2,32,37</sup>, Nathan Y. Cheng<sup>1</sup>, Brian L. Trippe<sup>3,4</sup>, Rebecca S. Fine<sup>1,5,6,33</sup>, Jenkai Miao<sup>1,5</sup>, Tejal A. Patwardhan<sup>1,7</sup>, Masahiro Kanai<sup>1,8,9,10</sup>, Joseph Nasser<sup>1</sup>, Charles P. Fulco<sup>1,3,4</sup>, Katherine C. Tashman<sup>1</sup>, Francois Aguet<sup>1</sup>, Taibo Li<sup>1,11</sup>, Jose Ordovas-Montanes<sup>1,12,13,14</sup>, Christopher S. Smillie<sup>1,3</sup>, Moshe Biton<sup>1,15,35</sup>, Alex K. Shalek<sup>1,16,17,18,19</sup>, Ashwin N. Ananthakrishnan<sup>20</sup>, Ramnik J. Xavier<sup>1,15,20,21</sup>, Aviv Regev<sup>1,18,22,36</sup>, Rajat M. Gupta<sup>1,23,24</sup>, Kasper Lage<sup>1,25</sup>, Kristin G. Ardlie<sup>1,26</sup>, Joel N. Hirschhorn<sup>1,5,6,27</sup>, Eric S. Lander<sup>1,28,29</sup>, Jesse M. Engreitz<sup>1,30,31</sup> & Hilary K. Finucane<sup>1,8</sup>

Gazal et al., Nature Genetics 2022

Weeks et al., Nature Genetics 2023



# Next steps and outlook

rejhe@channing.harvard.edu

- Biobanks (large sample sizes)
- Incorporation of non-European genetic ancestries (diversity)
- Integrative multi-omics analyses  
example: NHLBI Trans-Omics for Precision Medicine (TOPMed) initiative
- Single cell, spatial transcriptomics
- Cell models, animal models, gene perturbation approaches (difficult for complex genetics)

## ARTICLE

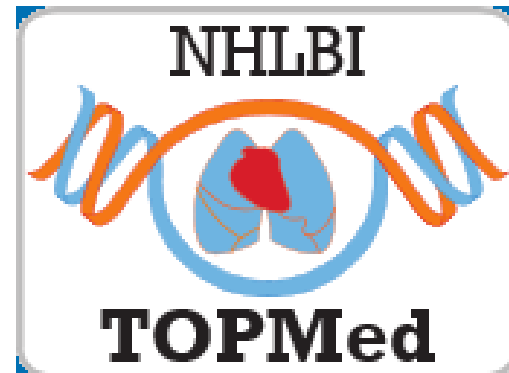
OPEN

<https://doi.org/10.1038/s41586-018-0579-z>

### The UK Biobank resource with deep phenotyping and genomic data

Clare Bycroft<sup>1,13</sup>, Colin Freeman<sup>1,13</sup>, Desislava Petkova<sup>1,12,13</sup>, Gavin Band<sup>1</sup>, Lloyd T. Elliott<sup>2</sup>, Kevin Sharp<sup>2</sup>, Allan Motyer<sup>3</sup>, Damjan Vukcevic<sup>3,4</sup>, Olivier Delaneau<sup>5,6,7</sup>, Jared O'Connell<sup>8</sup>, Adrian Cortes<sup>1,9</sup>, Samantha Welsh<sup>10</sup>, Alan Young<sup>11</sup>, Mark Effingham<sup>10</sup>, Gil McVean<sup>1,11</sup>, Stephen Leslie<sup>3,4</sup>, Naomi Allen<sup>11</sup>, Peter Donnelly<sup>1,2,14</sup> & Jonathan Marchini<sup>1,2,14\*</sup>

Nature 2018



*"Drug targets with human genetic evidence of disease association are twice as likely to lead to approved drugs and are even more likely to be approved when the exact casual gene(s) is known."*

Gerks, Thorp, Gerring, Nature Genetics 2022

citing Nelson et al., Nature Genetics 2015