

# Genome-Wide Association Studies

Mikhail Dozmorov

2021-05-03

# Definitions

- **ASSOCIATION STUDY** - A genetic variant is genotyped in a population for which phenotypic information is available (such as disease occurrence, or a range of different trait values). If a correlation is observed between genotype and phenotype, there is said to be an association between the variant and the disease or trait.
- **QUANTITATIVE TRAIT** - A biological trait that shows continuous variation (such as height) rather than falling into distinct categories (such as diabetic or healthy). The genetic basis of these traits generally involves the effects of multiple genes and gene–environment interactions. Examples of quantitative traits that contribute to disease are body mass index, blood pressure and blood lipid levels
- **CANDIDATE GENE** - A gene for which there is evidence of its possible role in the trait or disease that is under study.

# The Role of GWAS SNP Arrays in Human Genetic Discoveries

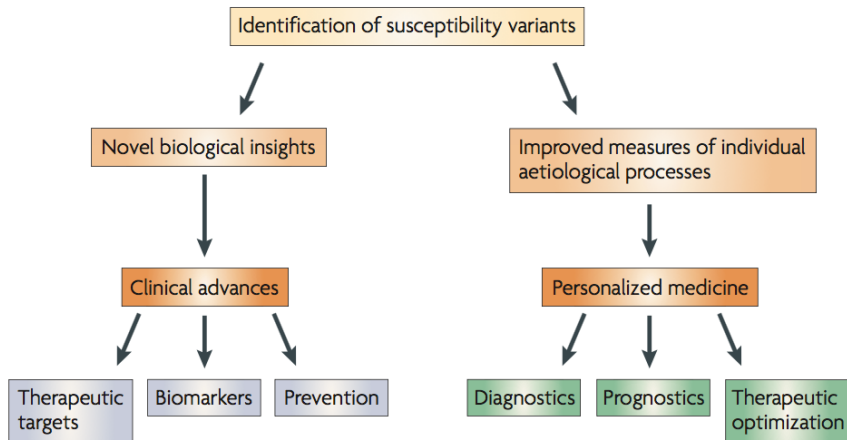
<b>Analysis</b>	<b>Purpose</b>	<b>Discoveries</b>
GWAS	detecting trait-SNP associations	~10,000 robust associations with diseases and disorders, quantitative traits, and genomic traits
Genome-wide CNV analysis	detecting trait-CNV associations	hundreds of associations with diseases and disorders
Genome-wide assessment of LD	quantifying genome architecture	large variation in LD in the genome
Estimation of SNP heritability <sup>a</sup>	genetic architecture	large proportion of genetic variation captured by common SNPs
Estimation of genetic correlation <sup>a</sup>	detecting and quantifying pleiotropy	pleiotropy is ubiquitous
Polygenic risk scores <sup>a</sup>	detecting pleiotropy; validating GWAS discoveries	out-of-sample prediction works as expected; detection of novel trait associations
Mendelian randomization <sup>a</sup>	testing causal relationships	replication of known causal relationships; empirical evidence of observational associations that are not causal
Population differences in allele frequencies	reconstructing human population history; detecting selection	genetic structure can mimic geographical structure; evidence of natural selection
Trait GWAS with -omics GWAS <sup>a</sup>	fine-mapping; detecting target genes; function	two-thirds of GWAS-associated loci implicate a gene that is not the nearest gene to the most associated SNP

<sup>a</sup>These analyses can be performed with GWAS summary statistics.

# Genetic disorders

- *Mendelian* diseases - caused by changes in a single genes
- Most heritable diseases in humans are *multigenic* or *complex* rather than Mendelian
- In a complex genetic model, many genes and possible environmental factors collectively increase the risk of disease in a population, but each gene individually contributes minor to modest effects.

# Genome-Wide Association Studies (GWAS)



McCarthy, Mark I., Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. "Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges." *Nature Reviews. Genetics* 9, no. 5 (May 2008): 356–69. <https://doi.org/10.1038/nrg2344>.

# Genetic definitions

**TABLE I** Definitions of Important Genetic Terminology

Term	Definition
Linkage mapping	A family-based method to identify genomic regions inherited from parent to affected offspring through multiple generations; generally used to map the location of a disease-causing gene in a family
Positional cloning	A research strategy that combines linkage mapping, gene identification, and sequencing to identify mutations in a likely disease-causing gene
Single nucleotide polymorphism (SNP)	Single base sequence changes that are common throughout the genome and are useful markers for disease gene mapping in large populations
Genome-wide association study (GWAS)	A research strategy that involves searching the entire genome to identify polymorphisms, primarily SNPs, that are associated with disease risk
Qualitative traits	Phenotypes described as discrete dichotomous values (e.g., either diseased or healthy); generally tested using chi-square contingency tables
Quantitative traits	Phenotypes that vary in degree and may be described by numerical measurements (e.g., Cobb angle for scoliosis); may be tested using analysis of variance
Replication	Independent validation of significant GWAS associations; provides additional evidence for significant associations to ensure that the association is not an artifact of the design, method, or populations used in the original study

Paria N, Copley LA, Herring JA, Kim HK, Richards BS, Sucato DJ, Rios JJ, Wise CA. The impact of large-scale genomic methods in orthopaedic disorders: insights from genome-wide association studies. *J Bone Joint Surg Am.* 2014 Mar 5;96(5):e38. doi: 10.2106/JBJS.M.00398

# GWAS workflow

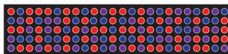
## 1. Identify Populations

**Cases**  
n=500

**Controls**  
n=500



## 2. Microarray Genotyping



## 3. Determine SNP Genotype

	SNP1	SNP2		SNP1	SNP2
Case 1:	ACGGA <sub>T</sub> T...ACC <sub>A</sub> T	G	Control 1:	ACGGA <sub>C</sub> T...ACC <sub>A</sub> T	G
Case 2:	ACGGA <sub>T</sub> T...ACC <sub>A</sub> T	G	Control 2:	ACGGA <sub>C</sub> T...ACC <sub>A</sub> T	G
Case 3:	ACGGA <sub>C</sub> T...ACC <sub>A</sub> T	G	Control 3:	ACGGA <sub>T</sub> T...ACC <sub>T</sub> T	G
⋮			⋮		
Case 500:	ACGGA <sub>T</sub> T...ACC <sub>A</sub> T	G	Control 500:	ACGGA <sub>C</sub> T...ACC <sub>A</sub> T	G

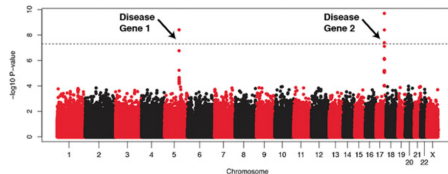
# GWAS workflow

## 4. Statistical Analysis

		<u>SNP 1</u>				<u>SNP 2</u>	
		C	T			A	T
Controls		639	361	Controls		492	508
Cases		575	425	Cases		460	540

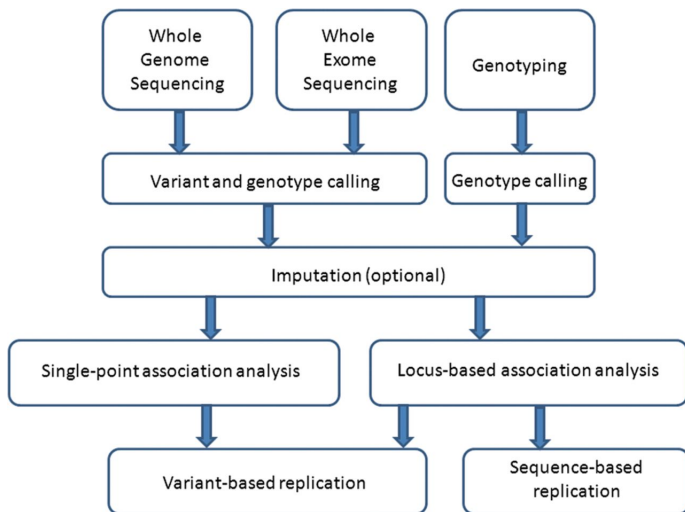
$p=0.0039$                        $p=0.1648$

## 5. Visualize Results



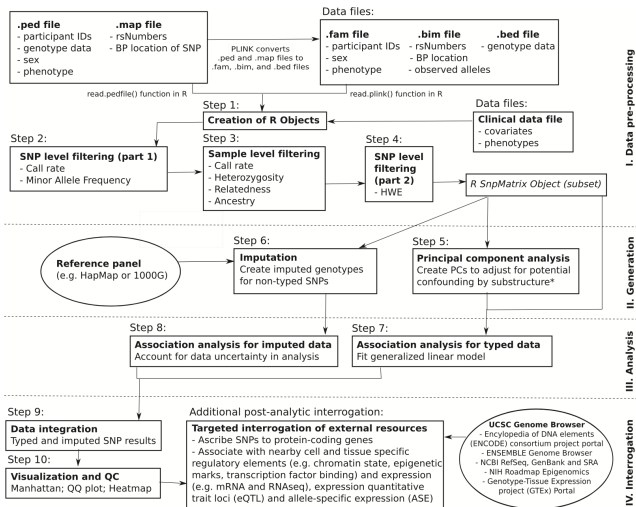


# GWAS workflow



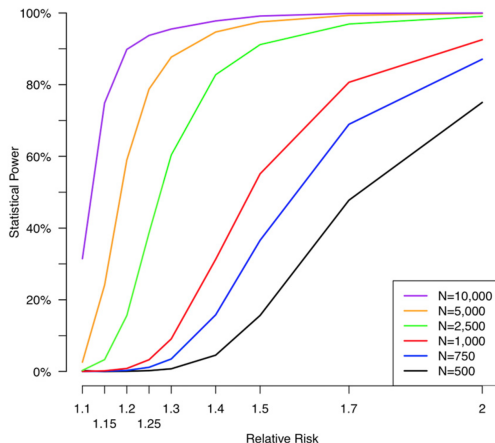
Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. *Hum Mol Genet.* 2013 Oct 15;22(R1):R16-21. doi: 10.1093/hmg/ddt376

# GWAS workflow



Reed, Eric, Sara Nunez, David Kulp, Jing Qian, Muredach P. Reilly, and Andrea S. Foulkes. "A Guide to Genome-Wide Association Analysis and Post-Analytic Interrogation." *Statistics in Medicine* 34, no. 28 (December 10, 2015): 3769–92.

# GWAS power



Online tools, such as PAWE (Power Analysis With Errors, <http://www.jurgott.org/linkage/pawe3d.zip>) and the Genetic Power Calculator (<http://zzz.bwh.harvard.edu/gpc/>), are available.

[http://journals.lww.com/jbjsjournal/Citation/2014/03050/The\\_Impact\\_of\\_Large\\_Scale\\_Genomic\\_Methods\\_in.16.aspx](http://journals.lww.com/jbjsjournal/Citation/2014/03050/The_Impact_of_Large_Scale_Genomic_Methods_in.16.aspx)

## GWAS significance level

- The first is the need to adjust for multiple testing and the probability of chance associations. For example, for a significance level ( $\alpha$ ) of 0.05, a typical GWAS involving 1 million SNPs will generate  $1 \text{ million} \times 0.05 = 50,000$  SNPs with  $p < 0.05$  as a result of chance.

Genome-wide significance levels for the GWAS era were estimated to be at  $P = 5 \times 10^{-8}$  based on the number of independent common-frequency variants across the genome calculated based on the European population data from the HapMap Project

# GWAScatalog - the Catalog of Published Genome-Wide Association Studies

GWAS / Diagram

This diagram shows all SNP-trait associations with  $p$ -value  $\leq 5.0 \times 10^{-6}$ , published in the GWAS Catalog.

Filter the diagram ^

Filter by trait

Clear

Apply

Show SNPs for

- Digestive system disease 714
- Cardiovascular disease 412
- Metabolic disease 379
- Immune system disease 359
- Nervous system disease 308
- Liver enzyme measurement 87
- Lipid or lipoprotein measurement 410
- Inflammatory marker measurement 310
- Hematological measurement 2108



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

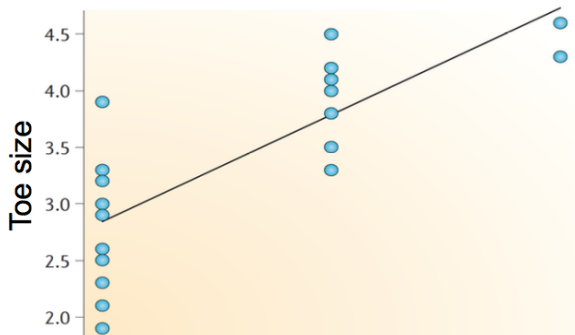
# Statistical analysis: linear regression

Two main parameters: p-value and effect size

$$y = \beta_0 + \beta_1 x$$

$$\textit{Trait} = \beta_0 + \beta_1 \textit{SNP}_1$$

$$\textit{Toesize} = \beta_0 + \beta_1 \textit{rs9876543}$$



# Statistical analysis: linear regression

Two main parameters: p-value and effect size

$$y = \beta_0 + \beta_1 x$$

$$\textit{Trait} = \beta_0 + \beta_1 \textit{SNP}_1$$

$$\textit{Toesize} = \beta_0 + \beta_1 \textit{rs9876543}$$

$$\textit{Toesize} = \beta_0 + \beta_1 \textit{rs9876543} + \beta_2 \textit{sex} + \beta_3 \textit{age} + \beta_4 \textit{age}^2 + \beta_5 \textit{BMI}$$

Assumptions

Trait is normally distributed for each genotype, with a common variance

- Subjects independent (e.g. unrelated)

# Odds ratio

- Surrogate measure of effect of allele on risk of developing disease

Allele	A	C	Total
Case	860	1140	2000
Control	1000	1000	2000
Total	1860	2140	4000

Odds of C allele given case *status* =  $CaseC/CaseA$

Odds of C allele given control *status* =  $ControlC/ControlA$

$$OddsRatio = \frac{CaseC/CaseA}{ControlC/ControlA} = \frac{1140/860}{1000/1000} = 1.33$$



# Multiple testing

- Genotype and test > 300K – 5M SNPs
- Correct for the multiple tests

$$\frac{0.05 \text{ } P \text{ - value}}{1 \text{ million common SNPs}} = 5 \times 10^{-8}$$

- Need large effect or large sample size