# Genome-Wide Association Studies

Mikhail Dozmorov

Spring 2018

# 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

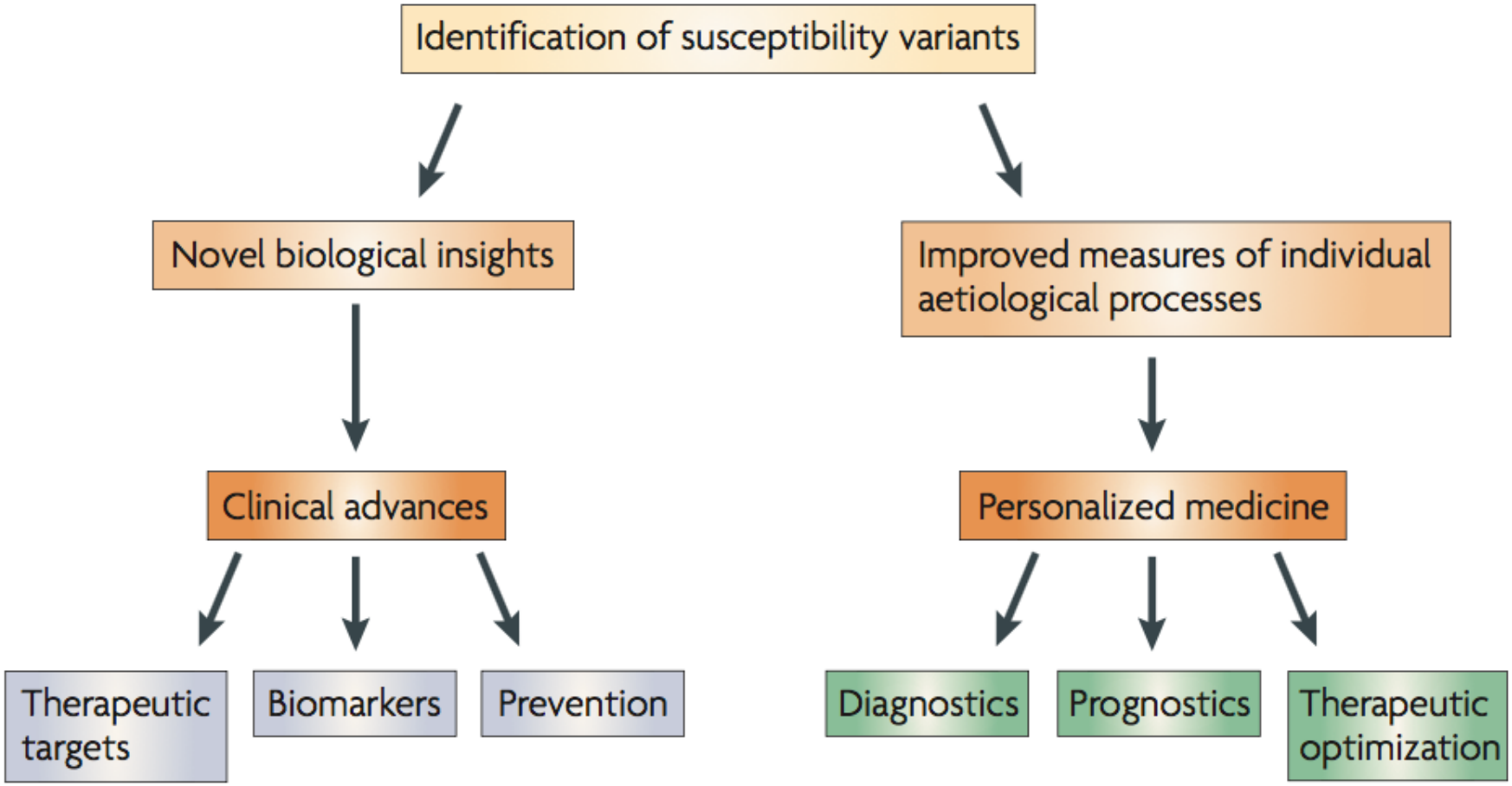
| Analysis | Purpose | Discoveries |
|---|---|---|
| GWAS | detecting trait-SNP associations | ~10,000 robust associations with diseases and disorders, quantitative traits, and genomic traits |
| Genome-wide CNV analysis | detecting trait-CNV associations | hundreds of associations with diseases and disorders |
| Genome-wide assessment of LD | quantifying genome architecture | large variation in LD in the genome |
| Estimation of SNP heritability[a] | genetic architecture | large proportion of genetic variation captured by common SNPs |
| Estimation of genetic correlation[a] | detecting and quantifying pleiotropy | pleiotropy is ubiquitous |
| Polygenic risk scores[a] | detecting pleiotropy; validating GWAS discoveries | out-of-sample prediction works as expected; detection of novel trait associations |
| Mendelian randomization[a] | 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[a] | 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 |

[a]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.

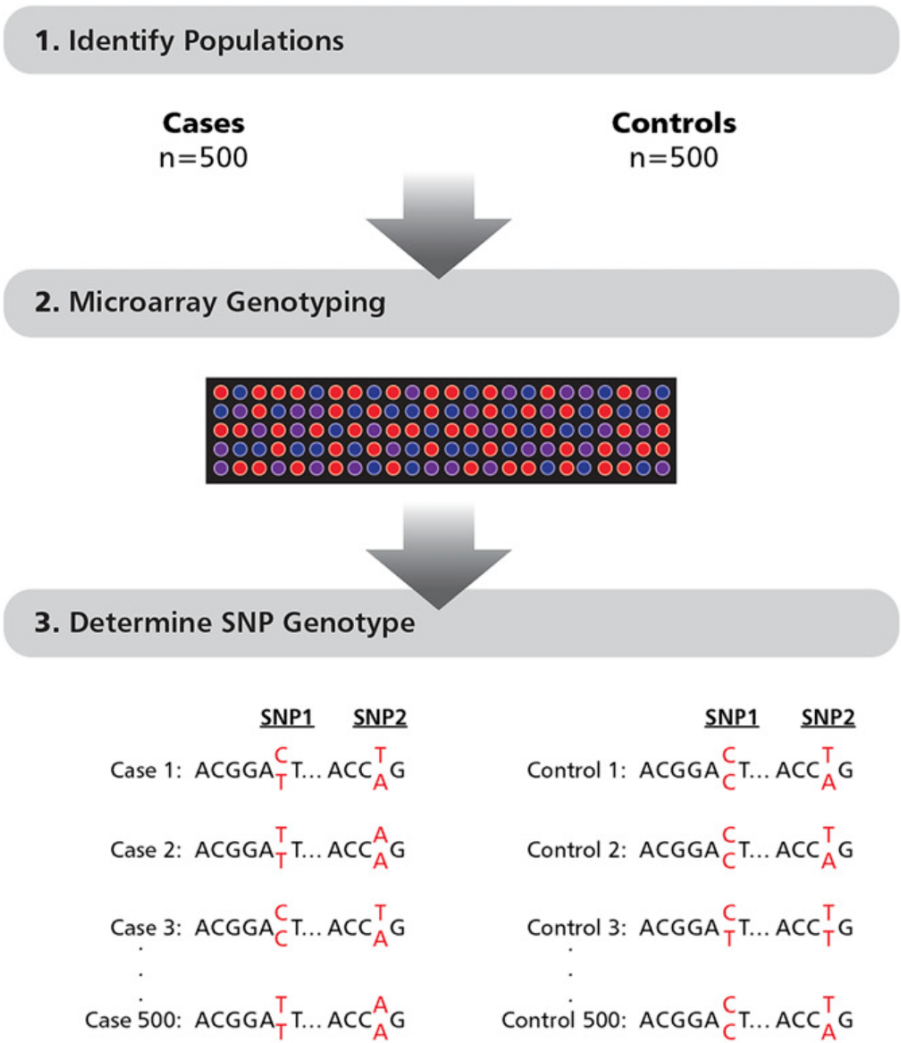
# Genetic definitions

**TABLE I Definitions of Important Genetic Terminology**

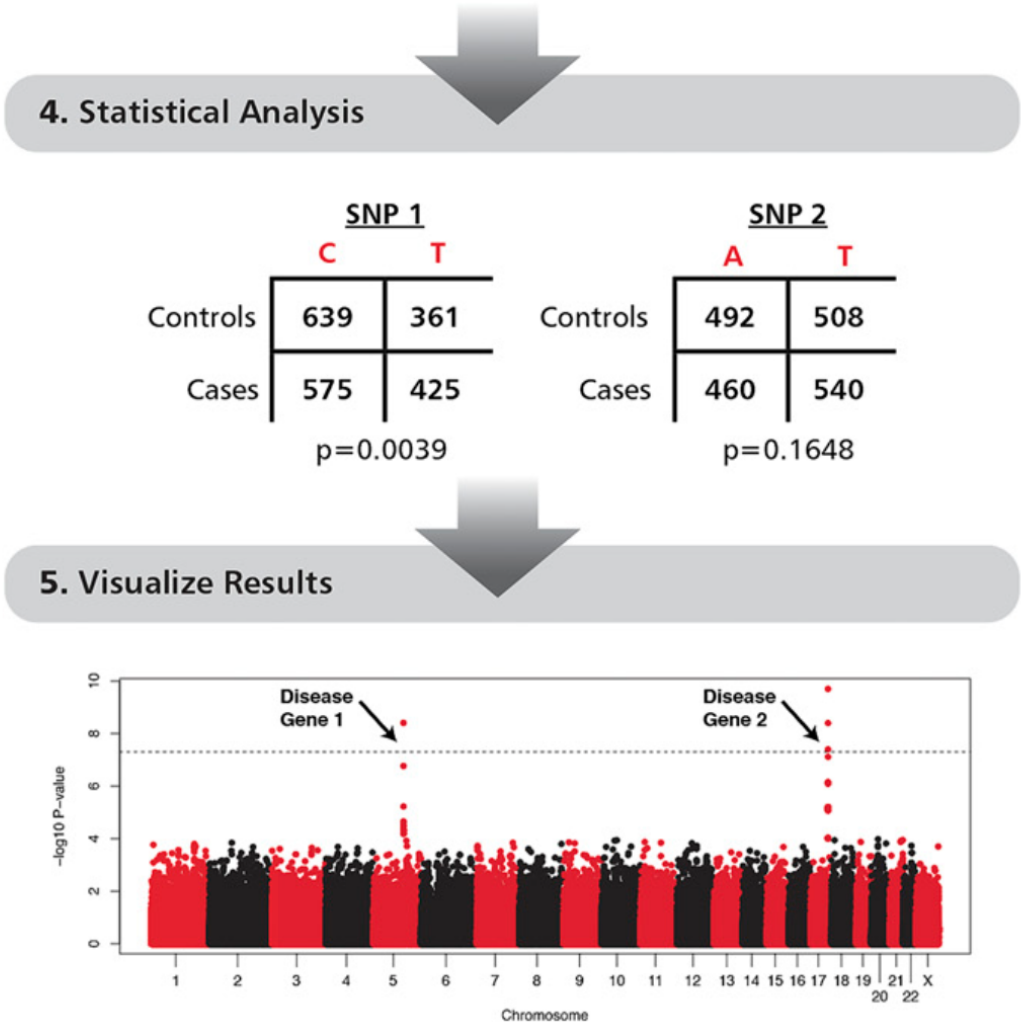
| Term | Definition |
|---|---|
| Linkage mapping | A family-based method to identify genomic regions inherited from parent to affected offspring through multiple generations; generally used to map the location of a disease-causing gene in a family |
| Positional cloning | A research strategy that combines 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

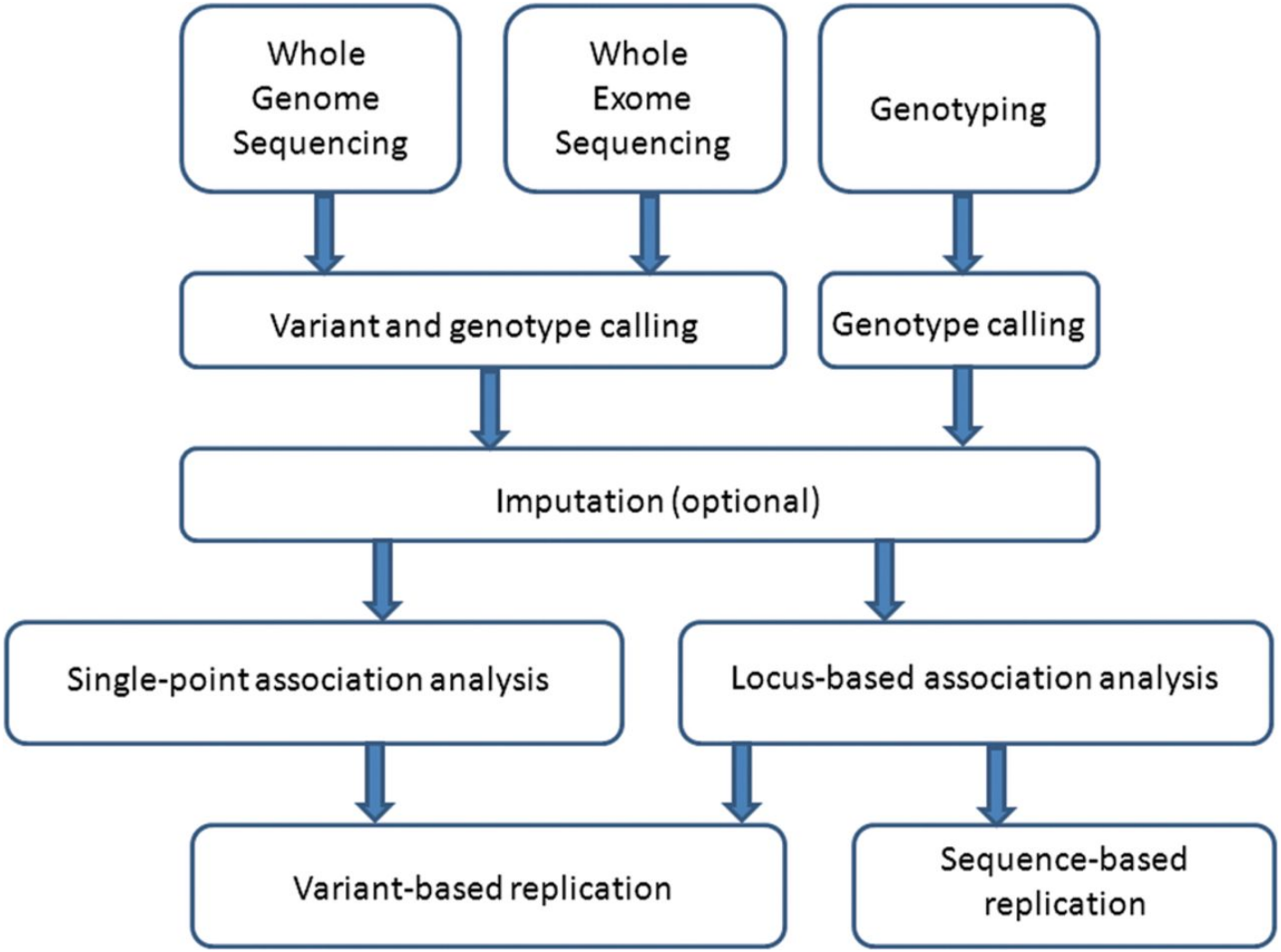
# GWAS workflow



# GWAS workflow

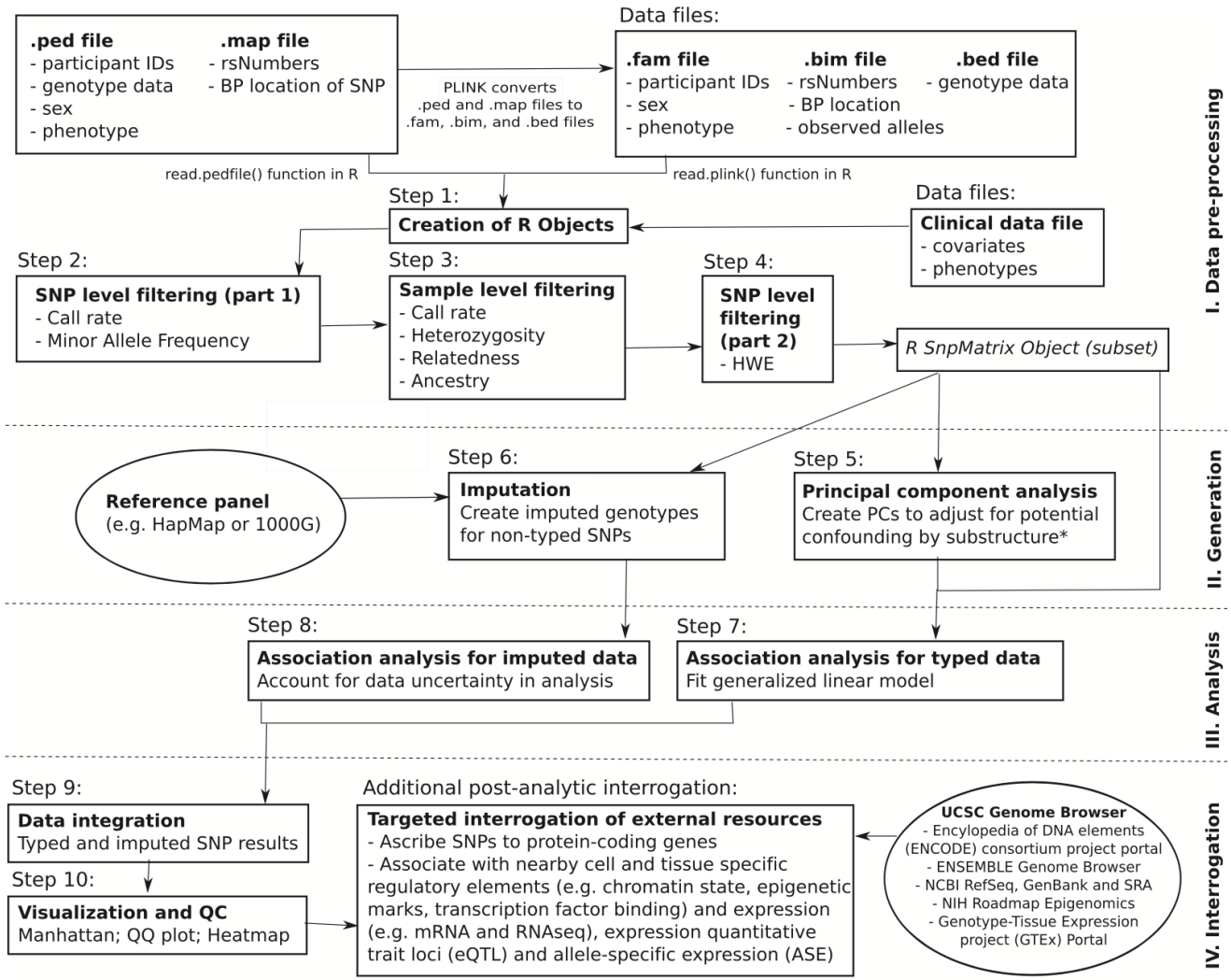


# GWAS workflow



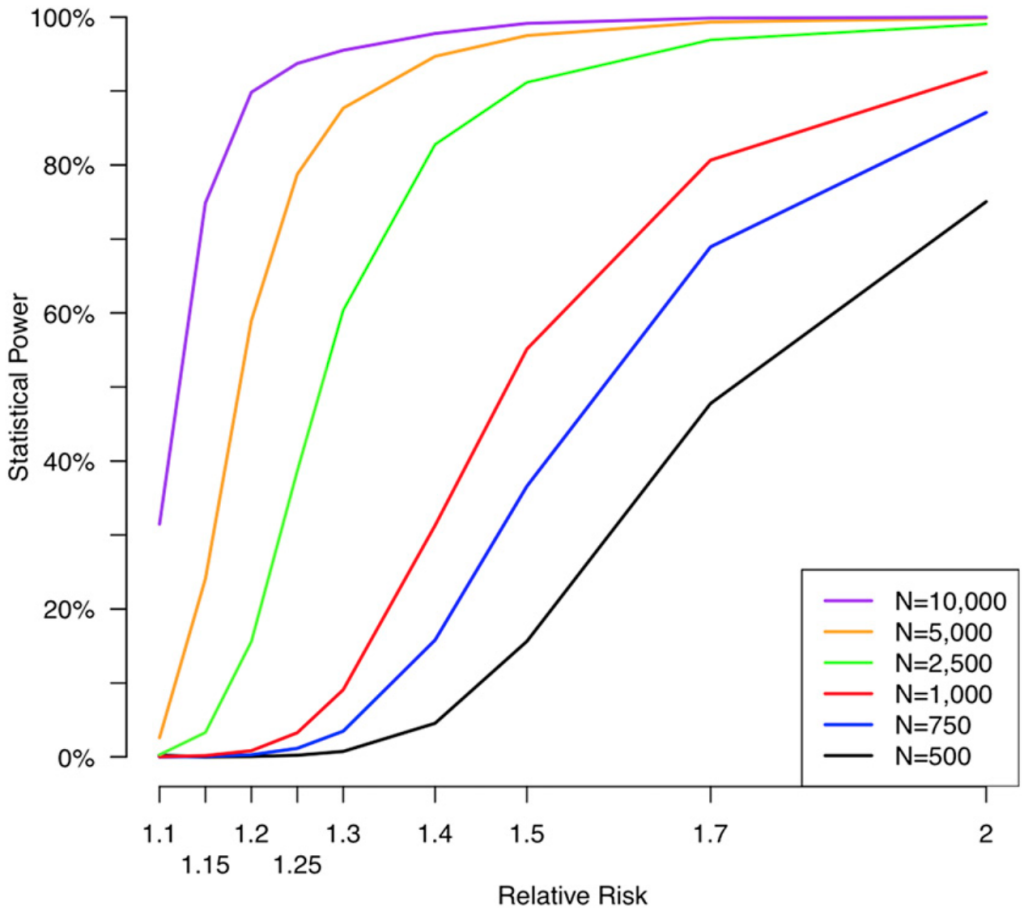

Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. Hum Mol Genet. 2013 Oct 15;22(R1):R16-21. doi: 10.1093/hmg/ddt376

# 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

# 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 million x 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 $P = 5x10^{-8}$ 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


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

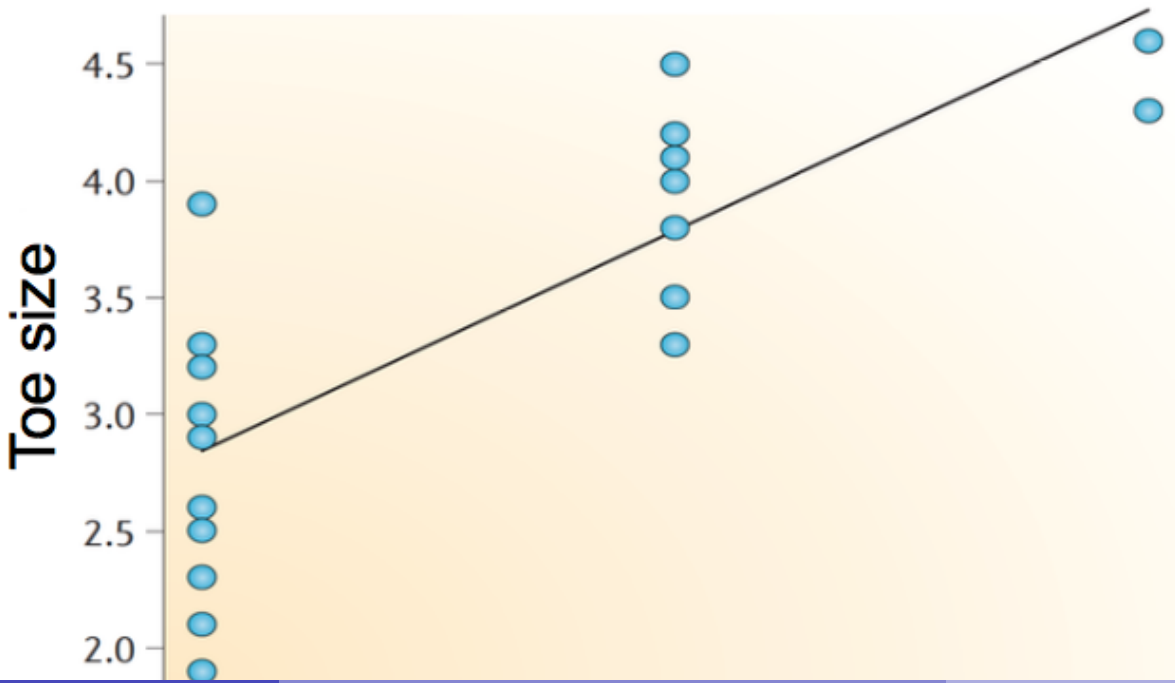
# Statistical analysis: linear regression

Two main parameters: p-value and effect size

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

$$Trait = \beta_0 + \beta_1 SNP_1$$

$$Toesize = \beta_0 + \beta_1 rs9876543$$



# Statistical analysis: linear regression

Two main parameters: p-value and effect size

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

$$Trait = \beta_0 + \beta_1 SNP_1$$

$$Toesize = \beta_0 + \beta_1 rs9876543$$

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

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

$$\frac{0.05\ P-value}{1\ million\ common\ SNPs} = 5x10^{-8}$$

- Need large effect or large sample size