

The Human Genome Project and genome-wide association

Bioinformatics · Universitat Politècnica de Catalunya

Alexandre Perera-Lluna

1 · Build a reference

2 · Describe variation

3 · Model populations

4 · Explain trait variation

5 · Test association

6 · Interpret responsibly

By the end, you should be able to calculate allele frequencies, distinguish heritability measures, choose a genotype model and explain why a GWAS hit is not automatically causal.

Milestone	Contribution
1990	International public project launched
2001	Draft human genome analyses published
2003	Project completed: high-quality reference, with gaps remaining

A shared coordinate system made it possible to connect sequence, genes, variation and phenotype. The reference was assembled from multiple donors; it was never a complete catalogue of human diversity.

NHGRI: Human Genome Project.

Project · landmark	Question	Contribution
HapMap · launched 2002	Which variants travel together?	Common SNPs and haplotypes; a foundation for GWAS
1000 Genomes / IGSR · Phase 3, 2015	Which variants occur across populations?	2,504 people, 26 populations ; reference panels for variation and imputation
ENCODE · launched 2003	Where are candidate functional elements?	Maps of transcription, chromatin and gene regulation across cell types

Reference sequence → variation catalogue → biological interpretation.

Project names link to official resources. [1000 Genomes resource paper](#) · [ENCODE overview](#).

A clinical programme

- Announced in **2012**; Genomics England established in **2013**.
- **100,000 genomes** milestone reached in **December 2018**.
- NHS participants affected by **rare disease or cancer**, with family members included.

What sequencing helps resolve

- **Rare disease:** compare relatives to identify inherited or new variants.
- **Cancer:** compare tumour and normal DNA to identify acquired variants.
- Helped establish the **NHS Genomic Medicine Service**.

100,000 genomes $\neq$ 100,000 people: one person can contribute tumour and normal samples.

Project	Scale · dated milestone	Research purpose
UK Biobank	Whole genomes from ~500,000 participants, released in 2023	Link genetic variation to health records, lifestyle, imaging and disease outcomes
All of Us · United States	>414,000 short-read genomes in the Research Hub FAQ, checked September 2026	Study health and disease in a diverse cohort with linked health information

More participants improve statistical power. Diverse recruitment and good phenotype data determine which questions we can answer.

UK Biobank: 500k release · All of Us: available genomic data. Individual-level data require authorised research access.

Milestone	What it adds	What it does not mean
T2T-CHM13, 2022	Gapless autosomes and X in this assembly	Every human genome, or a Y chromosome
Complete Y, 2023	A resolved Y chromosome sequence	All Y-chromosome diversity
Draft pangenome, 2023	47 phased diploid assemblies	Exhaustive representation of humanity

Later work includes [HPRC Release II \(2025\)](#). These are dated landmarks, not a latest-release inventory. Completeness and population representation remain distinct goals.

[Nurk et al. \(2022\)](#) · [Rhie et al. \(2023\)](#) · [Liao et al. \(2023\)](#).

- A variant is described **relative to a reference sequence**.
- Reference allele $\neq$ most frequent allele $\neq$ ancestral allele $\neq$ harmless allele.
- Record the **assembly, chromosome, position and alleles**; for transcripts, include accession and version.
- Different reference choices can change coordinates and variant representation.

Do not combine datasets until their builds and allele orientations have been checked.

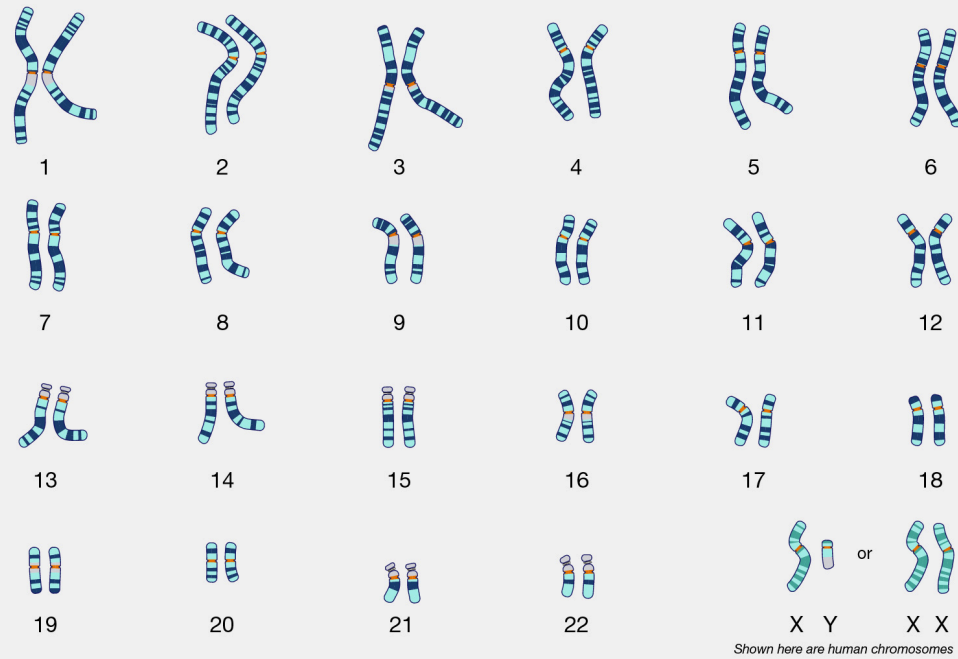
HGVS: reference sequences.

“Percent identical” depends on **what is aligned** and **which differences count**.

Comparison	Questions to ask
Two human genomes	SNPs only, or also indels and structural variants?
Human and another species	Orthologous aligned sequence, genes, or whole genomes?
Gene counts	Which annotation release and gene definition?

The source slide’s single percentages and fixed gene counts hid these choices. Genome similarity is not a ranking of biological complexity.

An ideogram



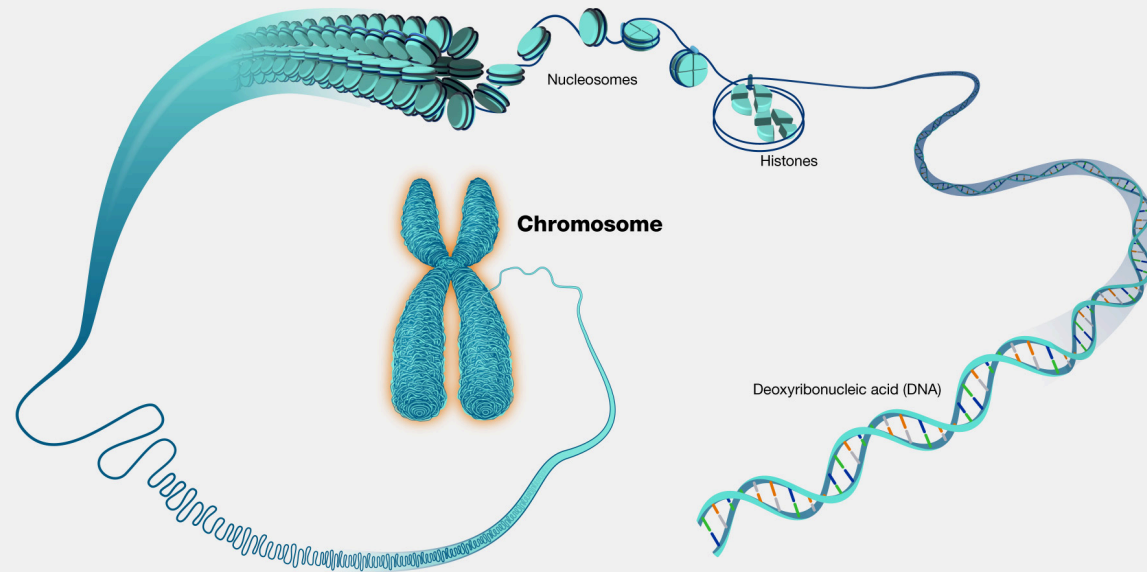
A metaphase spread



Pairs ordered by size and banding; the 23rd pair is XX or XY.

Chromosomes are photographed and classified.

Figures: [National Human Genome Research Institute](#) — public domain (credit “National Human Genome Research Institute”); G-banded karyogram via [Wikimedia Commons](#).



Keep three terms separate

- **Homologues:** maternal and paternal versions of a chromosome.
- **Sister chromatids:** replicated copies of one chromosome.
- **Centromere:** region that supports chromosome segregation.

An X-shaped chromosome is replicated; not every chromosome looks like an X.

Figure: NIH/NHGRI.

Term	Meaning
Locus	A genomic position or region
Allele	One sequence version at a locus
Genotype	Allele combination carried by an individual
Haplotype	Alleles on the same physical chromosome copy
Allele frequency	Fraction of sampled chromosome copies carrying an allele
Prevalence	Fraction of people with a specified condition at a specified time

Diploid autosomal examples have two copies per person. Ploidy can differ elsewhere.

In **100 diploid people**, observe 49 AA, 42 Aa and 9 aa.

$$p = f(A) = \frac{2(49) + 42}{2(100)} = 0.70, \qquad q = f(a) = 0.30.$$

Quantity	Value
Frequency of allele a	30% of chromosome copies
Frequency of genotype Aa	42% of people
People carrying at least one a	51% of people

These are three different quantities. Counting alleles does **not** require Hardy–Weinberg equilibrium.

For two alleles, $p + q = 1$.

Random union of gametes gives:

$$f(AA) = p^2, \quad f(Aa) = 2pq, \quad f(aa) = q^2.$$

At $p = 0.7$: **49%, 42%, 9%**.

This is an expected distribution, not a claim that every population satisfies it.

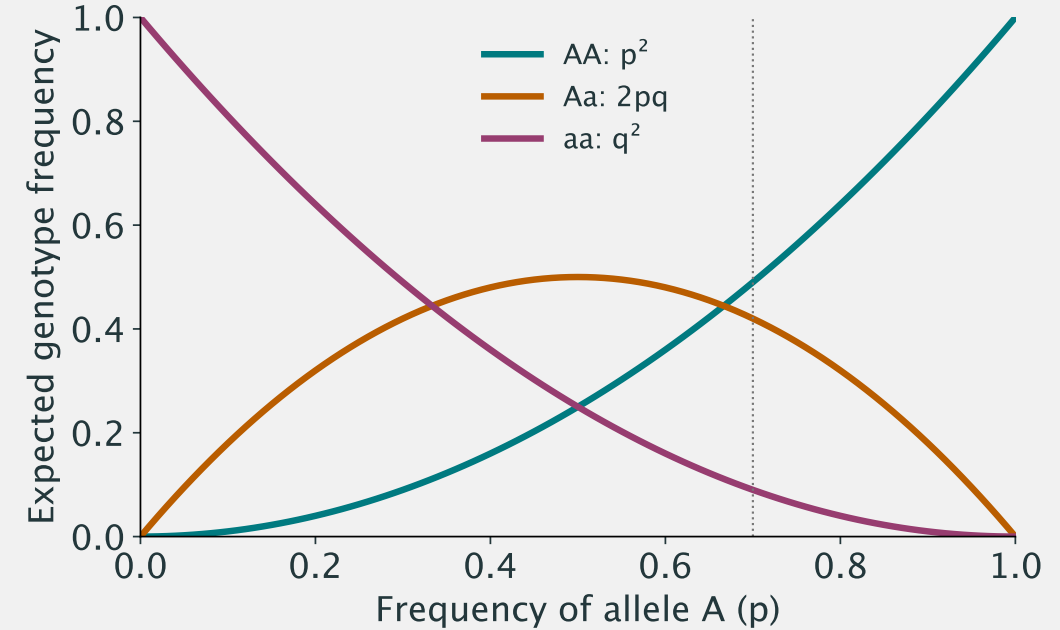


Figure: original plot of Hardy–Weinberg equations.

Possible reason	Interpretation
Genotyping error	A technical problem to investigate
Population mixture or inbreeding	Genotype proportions may depart from random mating
Selection or disease ascertainment	Particular genotypes may be overrepresented
Sampling variation	A test can reject by chance

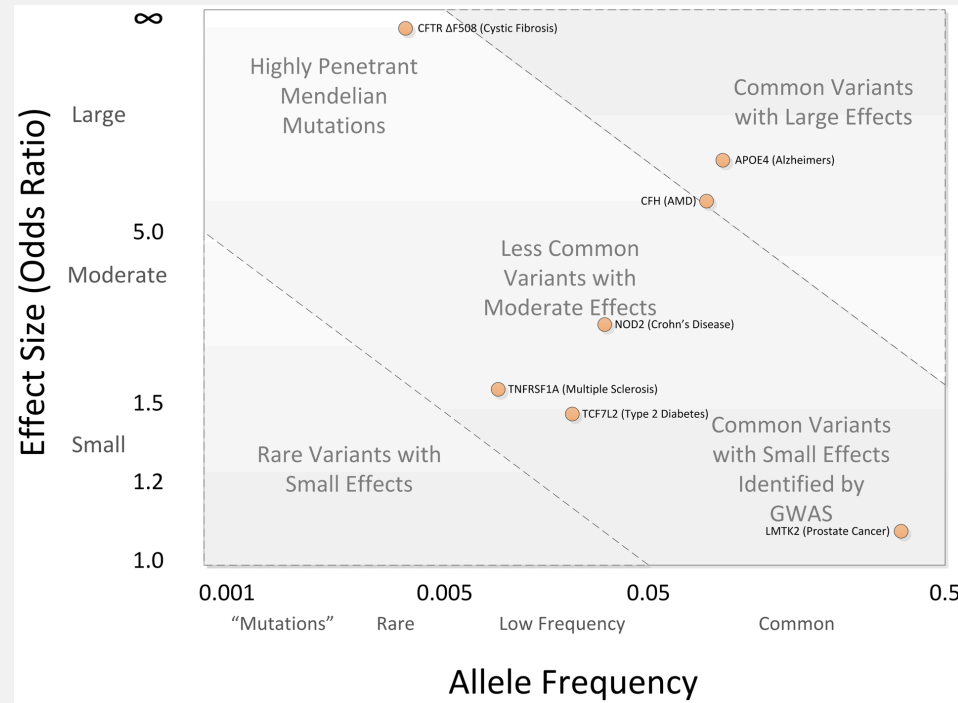
For case–control GWAS, HWE filtering is often assessed in **controls**. Choose thresholds for the study; do not treat every departure as an error.

PLINK: Hardy–Weinberg tests and filters.

The **common disease–common variant hypothesis** motivated early GWAS.

- Many common variants have small marginal effects.
- Rare variants can have large **or** small effects.
- A common disease can contain monogenic subtypes and a polygenic majority.
- Frequency and effect size do not determine each other by a universal rule.

Frequency, effect size and penetrance describe different properties.



Read this as a conceptual map

- Rare, large-effect variants: family studies and sequencing can be informative.
- Common, small-effect variants: large association studies are often needed.
- Low frequency and small effect: especially difficult to detect.

Figure: Bush & Moore (2012), PLOS Computational Biology.

- Locate and annotate genes and regulatory elements.
- Identify variants in Mendelian disease and complex-trait studies.
- Compare species and investigate evolutionary history.
- Design molecular assays and gene-editing experiments.
- Study gene expression, drug response and tumour evolution.

Sequence is a starting point: functional interpretation requires biological and clinical evidence.

How many loci?

Monogenic: one locus has a major role in a disorder.

Polygenic: many loci contribute to trait variation.

How do effects relate?

Pleiotropy: one genetic factor affects multiple traits.

Epistasis: the effect at one locus depends on genotype at another.

Example	Key gene(s)	Important qualification
Cystic fibrosis	CFTR, chromosome 7	Usually recessive; alleles and modifiers affect severity
Familial hypercholesterolaemia	Often LDLR, chromosome 19	APOB, PCSK9 and other mechanisms also occur

Penetrance, expressivity, environment and other genes can modify a monogenic condition.

[NCBI: LDLR](#) · [NCBI: CFTR](#).

Many variants

+

Environment

→

Variation in a trait

- Each locus can follow Mendelian transmission while the **overall trait** is complex.
- Height is quantitative; disease can be recorded as binary despite an underlying continuous liability.
- A bell-shaped trait distribution is a useful model in some settings, not a necessary consequence of “many genes”.

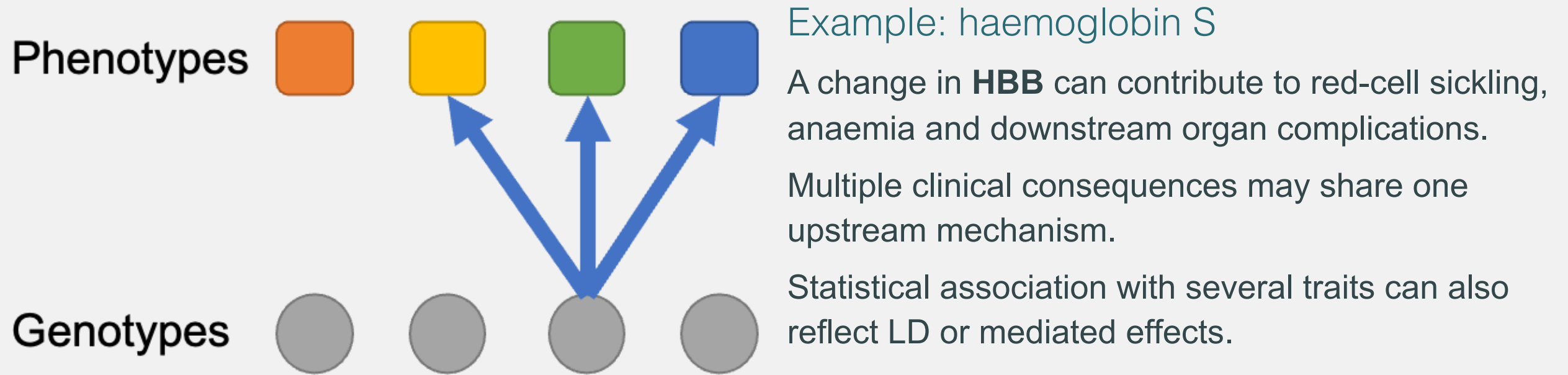


Figure: supplied lecture. [Sickle cell disease, GeneReviews](#).

An interaction model can include:

$$Y = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \gamma G_1 G_2 + \varepsilon.$$

- $\gamma \neq 0$ means departure from additivity **on this model's scale**.
- Biological pathway interaction and statistical interaction are not identical.
- Interaction estimates need adequate sample sizes and careful multiple-testing control.

A pathway can involve interacting proteins even when an additive statistical model predicts well.

$$H^2 = \frac{V_G}{V_P} \quad \text{and} \quad h^2 = \frac{V_A}{V_P}.$$

Measure	Numerator
Broad-sense H^2	Total genetic variance V_G
Narrow-sense h^2	Additive genetic variance V_A

Both use **phenotypic variance** in the denominator. Neither is the proportion of one person's phenotype "caused by genes".

Under a simplified model with uncorrelated components:

$$P = \mu + A + D + I + E, \quad V_P = V_A + V_D + V_I + V_E.$$

- A : additive genetic effects; D : dominance; I : epistasis; E : residual environment and measurement.
- Correlations among components add covariance terms.
- Gene–environment interaction needs an explicit term; it cannot be added to the phenotype model and then omitted from its variance.

High heritability does not mean that an intervention cannot change the trait.

Yang et al. (2017): interpreting SNP heritability.

Pair, assuming no inbreeding	Kinship ϕ	Additive relationship $r = 2\phi$
Individual with self	1/2	1
Parent–child or full siblings	1/4	1/2
Grandparent–grandchild; half siblings	1/8	1/4
Aunt/uncle–niece/nephew	1/8	1/4
First cousins	1/16	1/8

Kinship is the probability that two sampled alleles are identical by descent. Shared environment and nonadditive effects prevent a general rule $h^2 = \text{correlation}/r$.

A GWAS tests genetic variants across the genome for association with a phenotype.

A signal can arise from a causal variant, a correlated marker in LD, population structure, batch effects or chance.

A small p-value is evidence against a model's null hypothesis—not proof of a causal gene or a large effect.

WTCCC (2007): a landmark multi-disease GWAS.

Phenotype	Example	Common starting model
Binary outcome	Cases versus controls	Logistic regression
Quantitative trait	Height, blood lipid concentration	Linear regression
Molecular trait	Gene expression or protein abundance	QTL model appropriate to the measurement

Case–control sampling does not directly estimate population disease prevalence. An eQTL is associated with expression; it is not automatically a disease mechanism.

PLINK 2 association models.

1 · Define phenotype and sampling

2 · Genotype / sequence + quality control

3 · Align alleles, phase and impute

4 · Fit association models + covariates

5 · Check calibration + replicate

6 · Fine-map and investigate function

Document genome build, allele coding, ancestry composition, sample overlap and each exclusion step. These choices are part of the analysis, not housekeeping.

Samples	Variants
Missingness and contamination	Call rate and batch differences
Duplicates and relatedness	Allele alignment and strand ambiguity
Phenotype and covariate consistency	Allele counts/frequency
Population structure	HWE in an appropriate comparison group
Consent and permitted use	Imputation quality

Thresholds depend on study design and data quality. Rare-variant tests require special attention to low counts.

[PLINK 2 filtering documentation.](#)

Observed markers

+

Reference haplotypes

→

Genotype probabilities

- Phasing estimates which alleles share a chromosome copy.
- Imputation uses shared haplotype patterns to infer untyped variants.
- Dosage is an **expected allele count** between 0 and 2, not always an integer.
- Reference-panel match, allele frequency and local LD affect accuracy.

Use imputation quality and uncertainty in downstream analysis.

$$Y_i = \beta_0 + \beta_G G_i + \boldsymbol{\gamma}^T \mathbf{C}_i + \varepsilon_i.$$

- G_i : 0, 1 or 2 copies of the chosen effect allele, or imputed dosage.
- $\mathbf{C}_i$: covariates such as age, study batch and ancestry components.
- β_G : mean trait change per additional effect allele under the additive model.
- Test $H_0 : \beta_G = 0$; report effect estimate, uncertainty and effect allele.

Changing which allele is counted reverses the sign of the estimated effect.

PLINK 2: linear and logistic regression.

$$\log \frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)} = \beta_0 + \beta_G G_i + \boldsymbol{\gamma}^T \mathbf{C}_i.$$

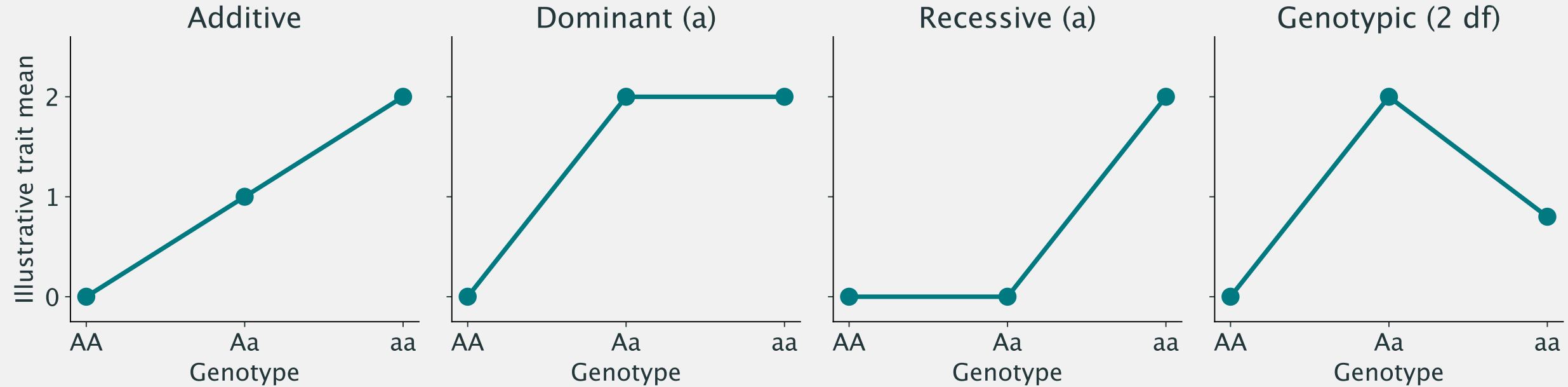
$$\text{Odds ratio per allele} = e^{\beta_G}.$$

An odds ratio is not an absolute risk and is not generally a risk ratio. Baseline risk and ascertainment matter. Very low allele counts or separation may require Firth or other suitable methods.

[PLINK 2 association documentation.](#)

Model for allele a	AA	Aa	aa	SNP degrees of freedom
Additive dosage	0	1	2	1
Dominant	0	1	1	1
Recessive	0	0	1	1
Genotypic	baseline	indicator 1	indicator 2	2

“Dominant” and “recessive” here describe a fitted contrast for a chosen allele; they are not established modes of disease inheritance.



Compare the mean at **Aa** with the means at AA and aa. Curves are illustrative, not patient results.

Figure: original schematic plots; axes and allele coding explicitly labelled.

$$Y_i = \beta_0 + \beta_1 1(G_i = 1) + \beta_2 1(G_i = 2) + \boldsymbol{\gamma}^T \mathbf{C}_i + \varepsilon_i.$$

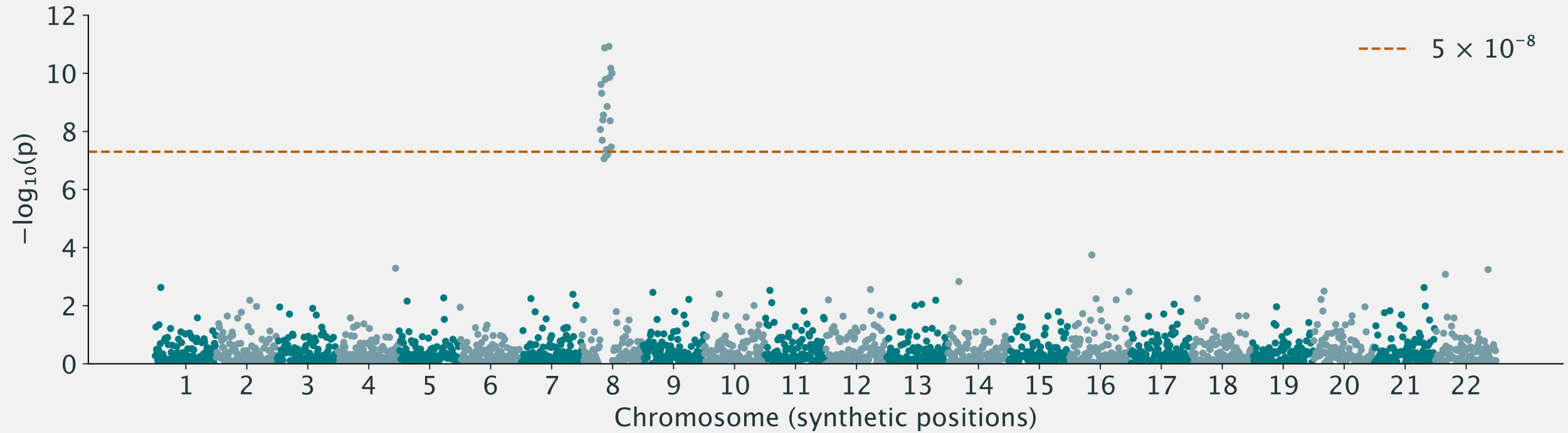
Joint null: $H_0 : \beta_1 = \beta_2 = 0$.

- No requirement that the heterozygote mean lies halfway between homozygotes.
- More flexibility, but usually less power than a correctly specified one-df model.
- For a binary trait, apply this linear predictor to the **logit**, not directly to a 0/1 response.

One million tests at a nominal $p < 0.05$ would yield about **50,000 false positives** under the global null.

$$\alpha_{\text{Bonferroni}} = \frac{0.05}{10^6} = 5 \times 10^{-8}.$$

This is a conventional common-variant GWAS benchmark. LD makes tests correlated; sequencing, ancestry, variant sets and multiple phenotypes can require different corrections. A discovery still needs replication and a credible analysis.



Each dot is a variant. A cluster can reflect many correlated markers tagging **one** signal.

Figure: reproducible synthetic data, generated for teaching; not a real disease GWAS.

Suppose ancestry groups differ in allele frequency **and** disease prevalence for environmental or sampling reasons.

A pooled genotype–disease association can then appear without a causal allele effect.

- Inspect ancestry structure, phenotype distribution and study batches.
- Include appropriate covariates and consider mixed models.
- Genetic ancestry is a statistical description, not a substitute for social or environmental variables.

Price et al. (2006).

Relatives share alleles across the genome, so their residuals may be correlated.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathbf{u} \sim N(0, \sigma_g^2 \mathbf{K}).$$

$\mathbf{K}$ represents genetic similarity. Mixed models can account for relatedness and sample structure; binary outcomes need appropriate extensions.

Family-based designs can also be deliberately used for robust within-family inference. They are not inherently flawed GWAS designs.

Kang et al. (2010).

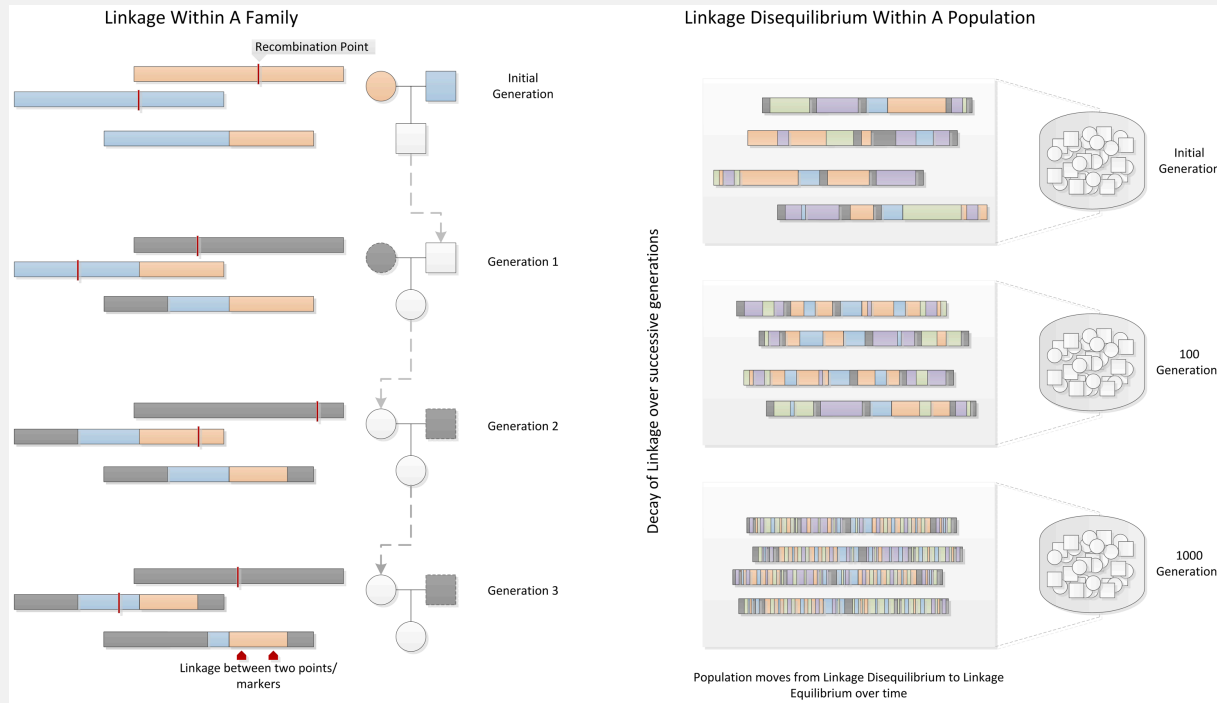


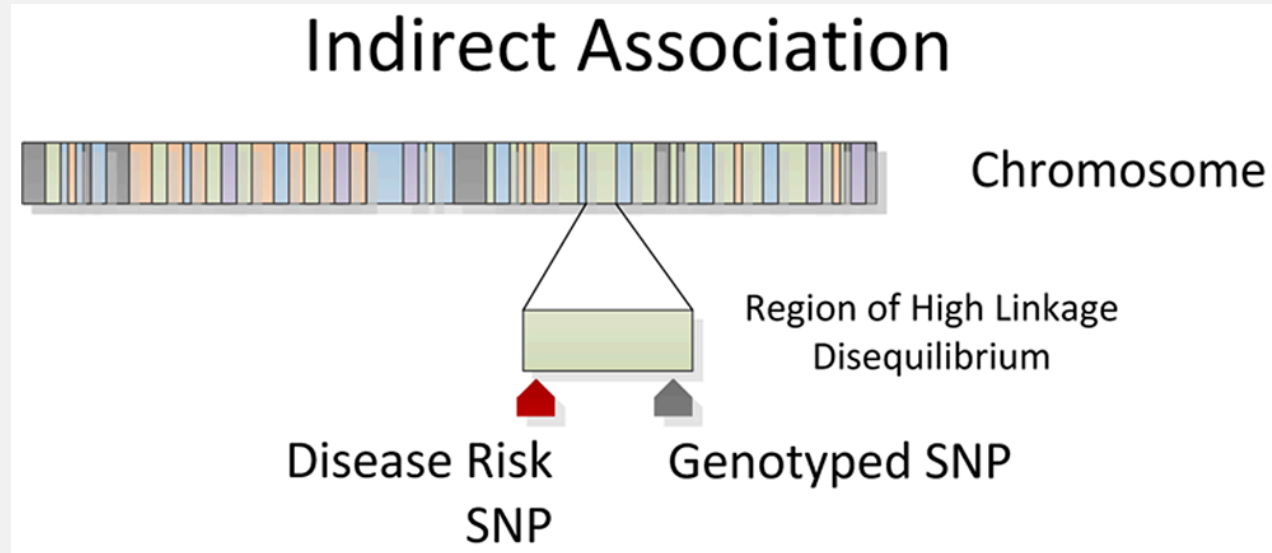
Figure: Bush & Moore (2012), PLOS Computational Biology.

Distinguish the concepts

Linkage: nearby loci tend to cosegregate in families.

LD: alleles at different loci occur together non-randomly in a population.

Recombination, demography and selection shape LD.



A hit identifies a region to investigate

- A lead SNP can be a proxy.
- LD patterns differ between populations.
- The closest gene is not automatically causal.
- Fine-mapping and functional evidence help distinguish alternatives.

Figure: Bush & Moore (2012), PLOS Computational Biology.

For alleles A and B at two loci:

$$D = f(AB) - p_A p_B, \quad r^2 = \frac{D^2}{p_A(1 - p_A)p_B(1 - p_B)}.$$

If $p_A = p_B = 0.5$ and $f(AB) = 0.4$, then $D = 0.15$ and $r^2 = 0.36$.

If haplotypes AB and ab each have frequency 0.5, $r^2 = 1$: perfect tagging in that population. If the loci are independent, $D = 0$.

Single-variant GWAS usually tests one marker at a time, adjusting for covariates.

A joint predictive model instead combines many markers:

$$\widehat{Y} = \widehat{\beta}_0 + \sum_{j=1}^m \widehat{\beta}_j G_j.$$

There are 2^m possible subsets of m markers. Exhaustive subset search quickly becomes infeasible; LD also makes coefficients difficult to estimate without suitable constraints.

For distinct unordered pairs among m markers:

$$\text{Number of pairs} = \binom{m}{2} = \frac{m(m-1)}{2}.$$

For $m = 1,000$, this gives **499,500** interaction terms, in addition to main effects.

Use $\sum_{j < k} \gamma_{jk} G_j G_k$ to count each pair once. The source expression $m!/(m-2)!$ counted each pair twice.

Approach	Main idea
Ridge / Bayesian shrinkage	Stabilise many small, correlated effects
LASSO / elastic net	Shrink coefficients, possibly selecting a subset
Trees, boosting, neural networks	Model flexible patterns; need careful validation
Biological priors / pathway models	Use external knowledge without assuming it proves causality

Keep relatives and overlapping samples from leaking across evaluation splits. Tune in training data; report performance in independent data.

$$\text{PGS}_i = \sum_j \hat{\beta}_j G_{ij}.$$

- Score performance depends on discovery sample size, LD, ancestry, phenotype and environment.
- Evaluate **discrimination and calibration** in the intended population.
- A score is not an absolute risk without an appropriate risk model.
- Clinical benefit and fairness require evidence beyond association significance.

Martin et al. (2019).

“Missing heritability”: which gap are we discussing?

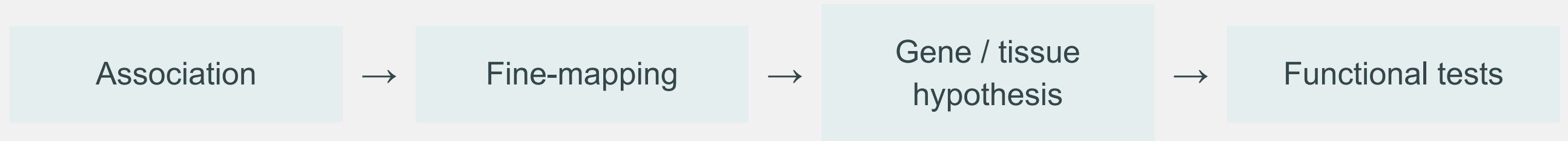
Quantity	What contributes?
Significant-hit variance explained	Variants that pass a chosen discovery threshold
SNP-based heritability	Additive effects tagged by the analysed variant set
Pedigree-based estimate	Familial resemblance under a specified model

These need not agree. Many effects are too small to pass single-variant significance, even when their combined contribution can be estimated.

Yang et al. (2017).

- Rare or structural variants may be poorly tagged by the analysed markers.
- Many weak associations lie below the discovery threshold.
- Phenotype noise and environmental differences change what is estimated.
- Family models can be affected by shared environment and their assumptions.
- Nonadditive effects and gene–environment interactions can matter, depending on the trait.

Do not label all unexplained variance “missing genetic effects”.



Replication checks whether the association survives another sample. Colocalisation can test whether two signals may share a causal variant. Experiments can test a proposed mechanism. Each step adds evidence; none should be silently replaced by “nearest gene”.

- Reference genomes provide coordinates; human variation is broader than one reference.
- Allele frequency, genotype frequency and prevalence are different quantities.
- Heritability is a population-specific variance ratio.
- GWAS tests association; LD and confounding shape its signals.
- Good inference needs quality control, appropriate models, replication and honest uncertainty.

- [NHGRI: Human Genome Project](#)
- [Nurk et al. \(2022\): complete human genome assembly](#)
- [Liao et al. \(2023\): draft human pangenome](#)
- [Bush & Moore \(2012\): GWAS teaching chapter](#)
- [Yang et al. \(2017\): SNP-based heritability](#)
- [PLINK 2: association analysis](#)

