Polymorphisms and Association Tests

Phenotype = Genetics + Environment

Tiago Bresolin, Ph.D.
Assistant Professor
Department of Animal Sciences
University of Illinois Urbana-Champaign

Agenda

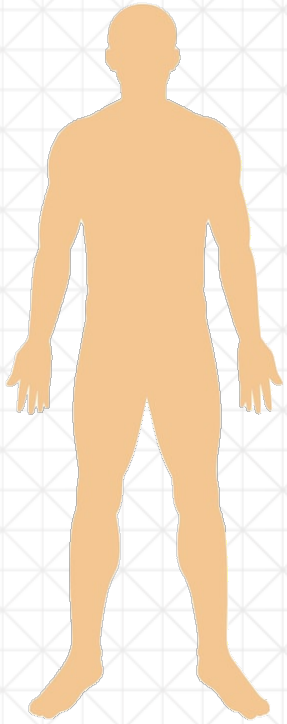
- Basics of polymorphism
- Foundation of genome association
  - ✓ Single marker association
  - ✓ Genome-wide association
- Statistical methods for genome association
- From association to prediction
- Genomics in the era of AI

Basics in genetics – step back

Paternal Chromosome

Maternal Chromosome

23 pairs



Paternal chromosome

… A C G C G C T A T A G C G A …

… T G C G C G A T A T C G C T …

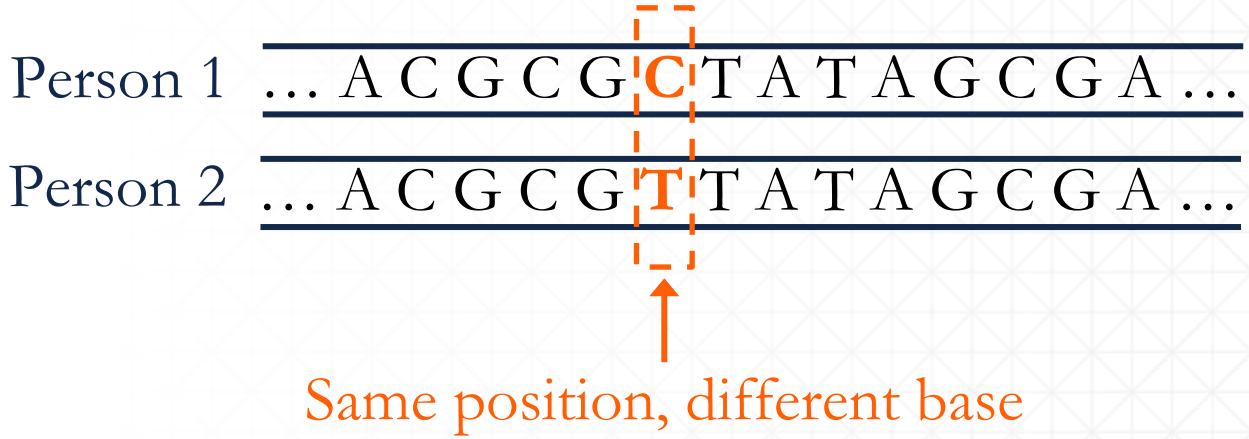
Maternal chromosome

… A C G C G T T A T A G C G A …

… T G C G C A A T A T C G C T …

What is a polymorphism?

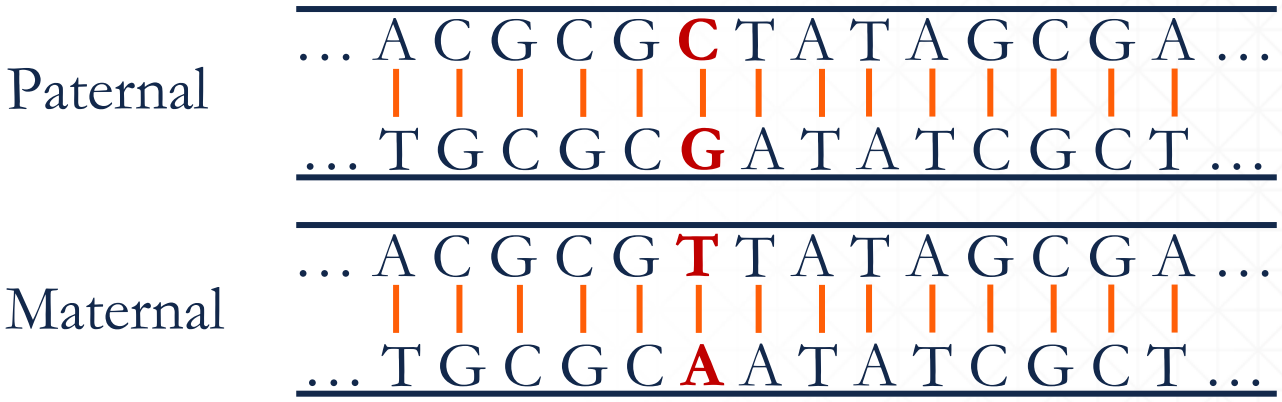
- A polymorphism is a position in the genome where the DNA sequence or nucleotide differs among individuals in a population



- To be a polymorphism, the minor allele frequency (MAF) should be at least 1%
  - ✓ Above that 1% line, a variant is "common", and we call it a polymorphism
  - ✓ Below it, we say that it is a "rare variant"

What is a polymorphism?

- Each person carries two copies, so three genotypes are possible



A/A homozygous

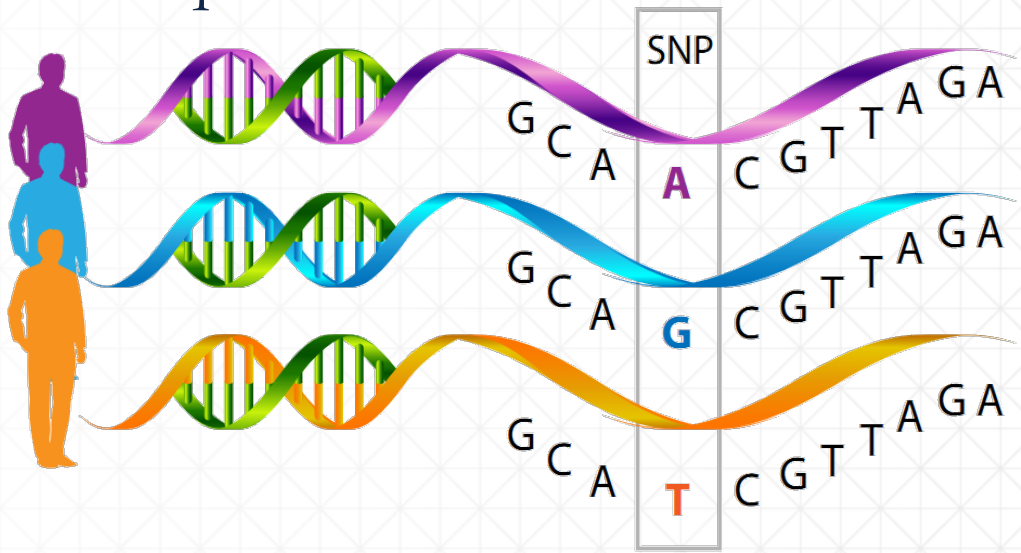
A/G heterozygous

G/G homozygous

- The **alleles** are the alternative bases at a given position (here, C and T), and the **genotype** is the pair a person carries.

Types of polymorphism!

- Indels: small insertions or deletions of a few bases
- STRs/microsatellites: short sequences repeated a variable number of times (the basis of DNA fingerprinting).
- CNVs and larger structural variants: duplicated, deleted, or inverted segments spanning thousands to millions of bases.
- Single nucleotide polymorphism (SNP)
  - ✓ A single base change in your DNA sequence



How do we identify these SNPs/genotypes?

- Whole genome sequencing (WGS) - recap

Extra DNA and library preparation

Sequence high-quality DNA

Assemble with short or long reads

Perform variant calling

VCF data report

Whole genome sequencing (WGS)

Reference genome … A C G C G **C** T A T A G C G A …

Sample NA12878
… A C G C G **C** T A T A G C G A … Paternal
… A C G C G **T** T A T A G C G A … Maternal

VCF report:

```
[HEADER LINES]
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10001014 . C T 843.77 PASS . GT:AD:DP:GQ:PL 0/1:14,13:27:99:872,0,810
20 10001019 . T G 364.77 PASS . GT:AD:DP:GQ:PL 0/0:33,0:33:99:0,99,1188
20 10001298 . T A 884.77 PASS . GT:AD:DP:GQ:PL 1/1:0,30:30:89:913,89,0
```

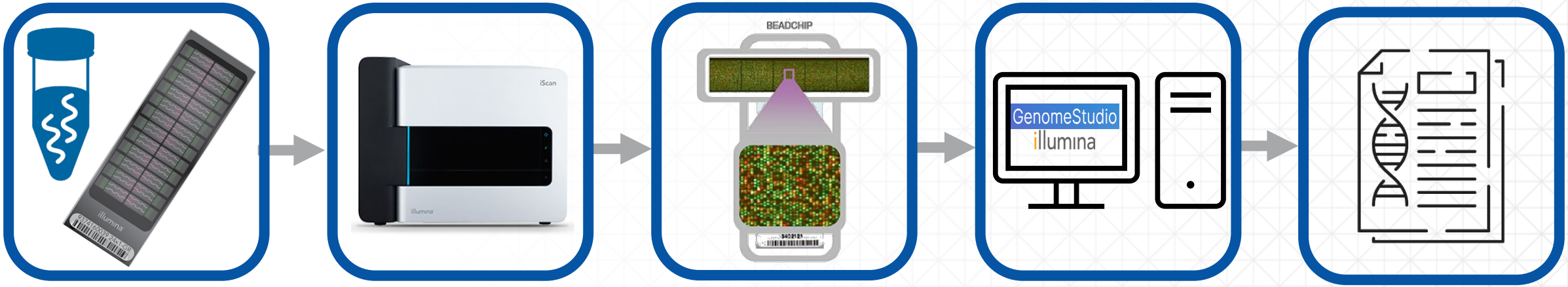
Genotype codes

01	1	AB
00	0	AA
11	2	BB

- 0/0 : the sample is homozygous reference
- 0/1 : the sample is heterozygous, carrying one copy each of the REF and ALT alleles
- 1/1 : the sample is homozygous alternate

How do we identify these SNPs/genotypes?

- SNP-chip or genotype arrays



Extra DNA and SNP-chip preparation

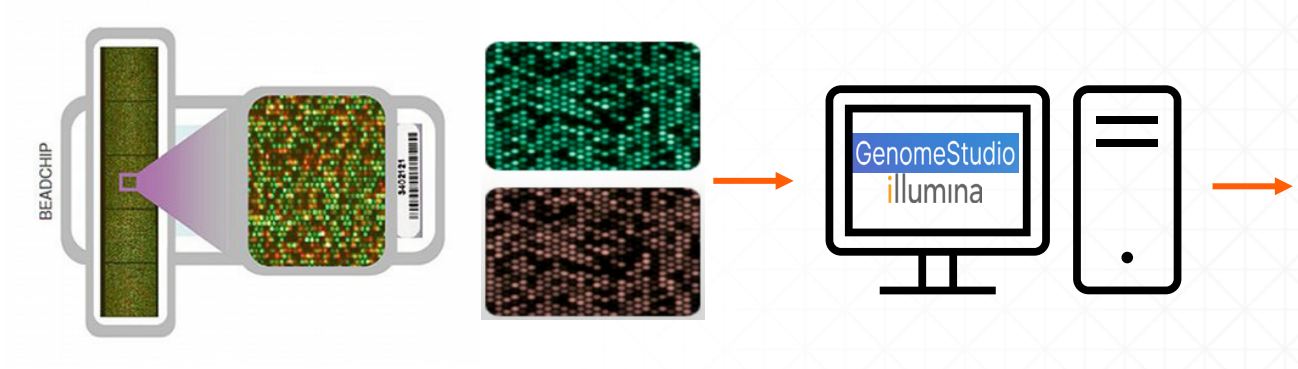
Genotyping: read SNP-chip arrays

Color map as the iScan output

Convert color map data to genotypes

Final reports

SNP-chip genotyping – BeadChip Illumina technology




```
[Header]
GSGT Version  1.9.4
Processing Date    2/12/2016 10:38 AM
Content       90K-Ind_Full_A_PublicMasked.bpm
Num SNPs  74677
Total SNPs    74677
Num Samples   121
Total Samples 121
File      1 of 121
[Data]
SNP Name  Sample ID Allele1 - AB  Allele2 - AB  GC Score
ARS-BFGL-BAC-10975 185  B    B    0.7180
ARS-BFGL-BAC-11025 185  A    B    0.8382
ARS-BFGL-BAC-11044 185  -    -    0.0000
ARS-BFGL-BAC-11193 185  A    A    0.9108
ARS-BFGL-BAC-11215 185  B    B    0.4560
ARS-BFGL-BAC-11276 185  B    B    0.9270
ARS-BFGL-BAC-11748 185  A    A    0.6585
ARS-BFGL-BAC-11750 185  A    A    0.9318
ARS-BFGL-BAC-11783 185  B    B    0.9169
ARS-BFGL-BAC-1180  185  B    B    0.8295
ARS-BFGL-BAC-11805 185  A    B    0.7605
ARS-BFGL-BAC-11867 185  B    B    0.7094
ARS-BFGL-BAC-11913 185  B    B    0.8399
ARS-BFGL-BAC-12159 185  B    B    0.9366
```

Generating the genotype data

Final Report

```
[Header]
GSGT Version  1.9.4
Processing Date    2/12/2016 10:38 AM
Content       90K-Ind_Full_A_PublicMasked.bpm
Num SNPs  74677
Total SNPs    74677
Num Samples   121
Total Samples 121
File      1 of 121
[Data]
SNP Name  Sample ID Allele1 - AB  Allele2 - AB  GC Score
ARS-BFGL-BAC-10975 185  B    B    0.7180
ARS-BFGL-BAC-11025 185  A    B    0.8382
ARS-BFGL-BAC-11044 185  -    -    0.0000
ARS-BFGL-BAC-11193 185  A    A    0.9108
ARS-BFGL-BAC-11215 185  B    B    0.4560
ARS-BFGL-BAC-11276 185  B    B    0.9270
ARS-BFGL-BAC-11748 185  A    A    0.6585
ARS-BFGL-BAC-11750 185  A    A    0.9318
ARS-BFGL-BAC-11783 185  B    B    0.9169
ARS-BFGL-BAC-1180  185  B    B    0.8295
ARS-BFGL-BAC-11805 185  A    B    0.7605
ARS-BFGL-BAC-11867 185  B    B    0.7094
ARS-BFGL-BAC-11913 185  B    B    0.8399
ARS-BFGL-BAC-12159 185  B    B    0.9366
```

	ARS-BFGL-BAC-10975	ARS-BFGL-BAC-11025	ARS-BFGL-BAC-11044	ARS-BFGL-BAC-11193
185	BB	AB	--	AA
194	BB	AA	AB	AA
195	BB	AB	AA	AA
196	BB	AA	BB	AA
197	BB	AA	AA	AA
198	AA	--	AA	AA
199	BB	AB	AA	AB
200	BB	AA	AB	AA
201	BB	AA	BB	AA
202	BB	AA	BB	AA
203	BB	AA	BB	AB
186	BB	AA	AA	AA

	ARS-BFGL-BAC-10975	ARS-BFGL-BAC-11025	ARS-BFGL-BAC-11044	ARS-BFGL-BAC-11193
185	2	1	*NA*	0
194	2	0	1	0
195	2	1	0	0
196	2	0	2	0
197	2	0	0	0
198	0	*NA*	0	0
199	2	1	0	1
200	2	0	1	0
201	2	0	2	0
202	2	0	2	0
203	2	0	2	1
186	2	0	0	0

Clean genotype data for downstream analysis

	ARS-BFGL-BAC-10975	ARS-BFGL-BAC-11025	ARS-BFGL-BAC-11044	ARS-BFGL-BAC-11193
185	2	1	NA	0
194	2	0	1	0
195	2	1	0	0
196	2	0	2	0
197	2	0	0	0
198	0	NA	0	0
199	2	1	0	1
200	2	0	1	0
201	2	0	2	0
202	2	0	2	0
203	2	0	2	1
186	2	0	0	0

Sample Metrics

Call Rate

- % of loci that a genotype was called for the sample

LogR Dev

- Deviation in the normalized intensity value, which may indicate contamination

SNP Metrics

GC Score

- Metric that reflects the distance of a called SNP from the center of the cluster

Call Rate

- % of samples that a genotype was called for the SNP

Minor Allele frequency

- Frequency of the second allele in the population

What comes next?

- Math calculations only work on numeric values!
- How we deal with missing values?

	ARS-BFGL-BAC-10975	ARS-BFGL-BAC-11025	ARS-BFGL-BAC-11044	ARS-BFGL-BAC-11193	ARS-BFGL-BAC-11215	ARS-BFGL-BAC-11276	ARS-BFGL-BAC-11748	ARS-BFGL-BAC-11750	ARS-BFGL-BAC-11783
185	2	1	NA	0	2	2	0	0	2
194	2	0	1	0	2	2	0	1	2
195	2	1	0	0	NA	1	0	1	2
196	2	0	2	0	2	2	0	1	2
197	2	0	0	0	2	2	0	0	2
198	0	NA	0	0	NA	1	NA	1	1
199	2	1	0	1	2	1	0	0	2
200	2	0	1	0	2	1	0	0	2
201	2	0	2	0	2	2	0	0	1
202	2	0	2	0	NA	2	NA	0	2
203	2	0	2	1	2	1	0	0	1
186	2	0	0	0	2	1	0	0	1

Imputation

Estimating missing data or missing information using study data

Data can be missing for three reasons:

1) Genotyping assay failure is likely for at least some loci in some samples
   - Less than 1%

2) Economically desirable
   - Over 80% if a selective genotyping strategy is used
   - For example:
     - Whole-genome sequence = ~$150 - $500 (5M+)
     - High-density SNP chip: ~$100 (700K)
     - Low-density SNP chip: ~ $50 (100K)

3) Combine data from different chips for association analysis

illumina®

affymetrix
Biology for a better world

Genotype imputation fundamentals

Reference population → high-density SNP data

Sample 1 . . . 1 0 0 2 0 0 1 0 0 1 1 0 1 2 0 0 1 0 0 0 1 0 0 1 1 1 1 0 1 0 1 1 1 0 . . .
Sample 2 . . . 1 1 0 0 0 1 2 1 0 1 1 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 1 0 0 1 1 0 0 0 . . .
Sample 3 . . . 0 1 0 1 1 0 0 0 0 0 1 0 1 0 0 0 1 0 1 1 1 0 1 1 0 2 2 1 1 0 0 1 1 1 . . .
Sample 4 . . . 0 0 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 2 0 1 1 0 0 1 1 0 0 . . .
Sample 5 . . . 1 1 1 0 0 1 2 2 0 1 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 0 0 0 1 2 1 0 1 1 . . .
Sample 6 . . . 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 1 0 0 1 1 0 0 0 1 0 0 0 0 0 1 1 0 . . .

Target population → low-density SNP data

Sample 1 . . . 1 **? ? ? ? ?** 1 0 **? ? ?** 0 1 2 0 **? ? ?** 0 **?** 1 **? ?** 1 2 0 **? ? ? ? ? ?** 1 0 . . .
Sample 2 . . . 0 **? ? ? ? ?** 1 1 **? ? ?** 0 0 0 0 **? ? ?** 0 **?** 1 **? ?** 1 1 1 **? ? ? ? ? ?** 1 1 . . .

Imputed Population → high-density SNP data

Sample 1 . . . 1 **1 1 1 0 0** 1 0 **0 0 1** 0 1 2 0 **1 1 0** 0 **1** 1 **1 2** 1 2 0 **0 1 1 1 1 0** 1 0 . . .
Sample 2 . . . 0 **0 0 1 1 1** 1 1 **0 1 0** 0 0 0 0 **0 0 1** 0 **1** 1 **0 2** 1 1 1 **0 0 0 1 1 1** 1 1 . . .

Haplotype-based imputation

Haplotypes are the two sequences of alleles that have been inherited together from the individual's parents

With the data clean, now we can start mining this data to ask biological questions, such as:

1) Which SNP is associated with the target phenotype?

2) How much does a SNP explain of the target trait variation?

Foundation of genome association

- What is genome association analysis?
  - ✓ Genome association analysis is a technique used to find genetic variants (e.g., SNPs) associated with a phenotype.

phenotype	SNPs
y_1	10101020200001010202000010101010200101010011111…
y_2	10100001010101000101000200220010101000000020100 10…
y_3	10101020200001010202000010101010200101010011111…
.	
.	
.	
y_n	10101020200001010202000010101010200101010011111…

Find genetic variants associated with the trait of interest

Foundation of genome association

- We want to find variants associated with the phenotype of interest to …

Better understand the biological mechanisms underlying the trait under study

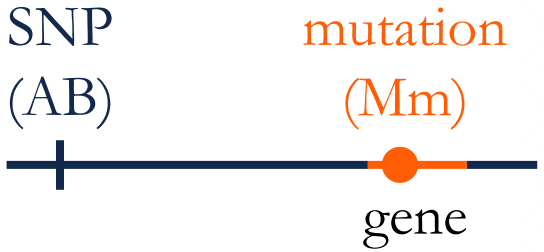
Identity potential genes associated with diseases for prevention

Generate information that can lead to the development of new drugs/vaccines

Develop genomic strategies for improving health outcomes via risk prediction and precision medicine

Foundation of genome association

❖ What is the relational behind association analysis using SNPs?



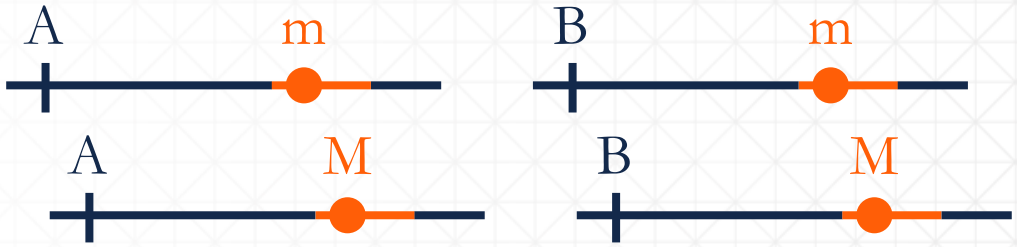
The causal mutation **affects** the target trait, but it is not in the SNP data

SNP **does not affect** the target trait, but it is in the SNP data

Linkage equilibrium

✓ SNP and mutation are independent

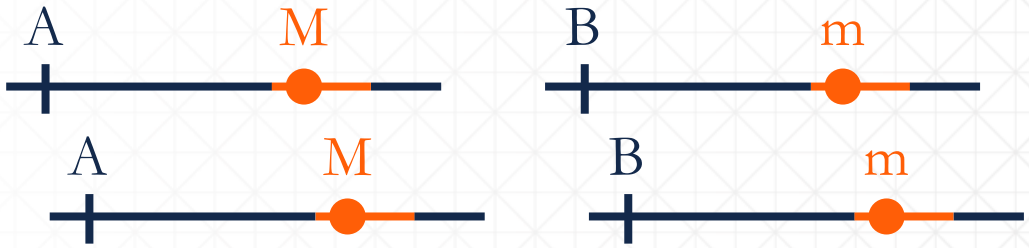
✓ SNP does not provide info about the status of the mutation



Linkage disequilibrium

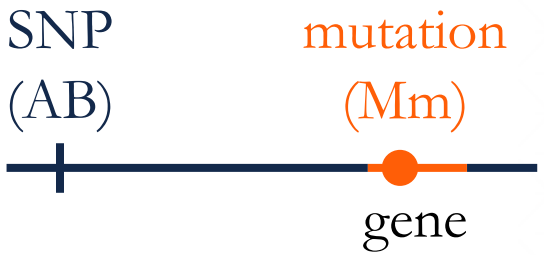
✓ non-random association of alleles at different loci

✓ SNP provides info about the status of the mutation



Foundation of genome association

❖ What is the relational behind association analysis using SNPs?



The causal mutation **affects** the target trait, but it is not in the SNP data

SNP **does not affect** the target trait, but it is in the SNP data

If the SNP is in linkage disequilibrium with the mutation, then the status of the SNP (A or B) provides information about the status of the mutation (M = wild type or m = mutant allele)

If the mutation affects the trait under study, then the SNP will be significantly associated with that trait

Foundation of genome association

- Single marker regression
  - ✓ Evaluate the association of one genetic variant (SNP) at a time

phenotype = non-genetic effects + genetic variant + error

BMI

=

Activities

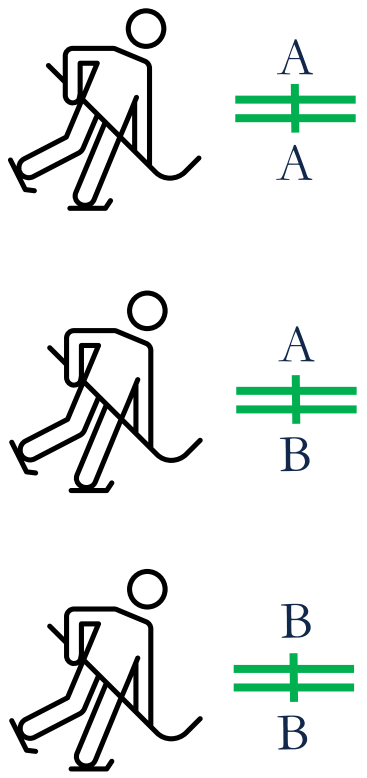
+

SNP data

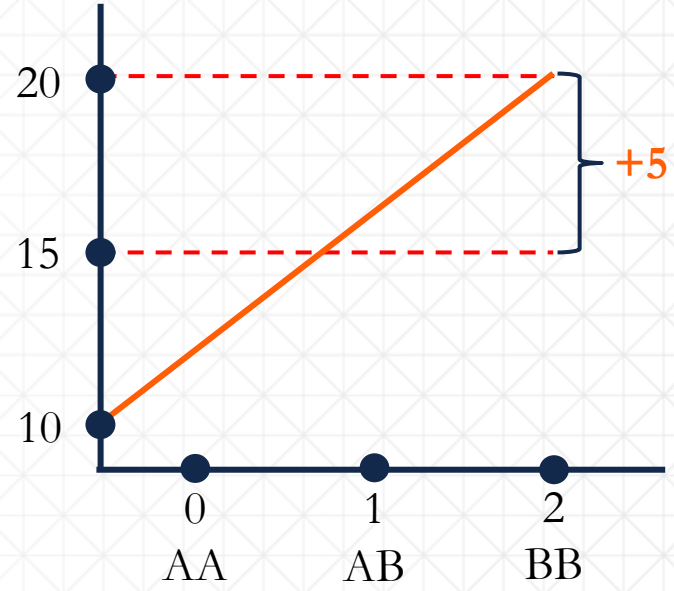
Foundation of genome association

- Single marker regression

phenotype = non-genetic effects + genetic variant (SNP_1) + error



SNP genotype	Phenotype
AA (0)	+10
AB (1)	+15
BB (2)	+20



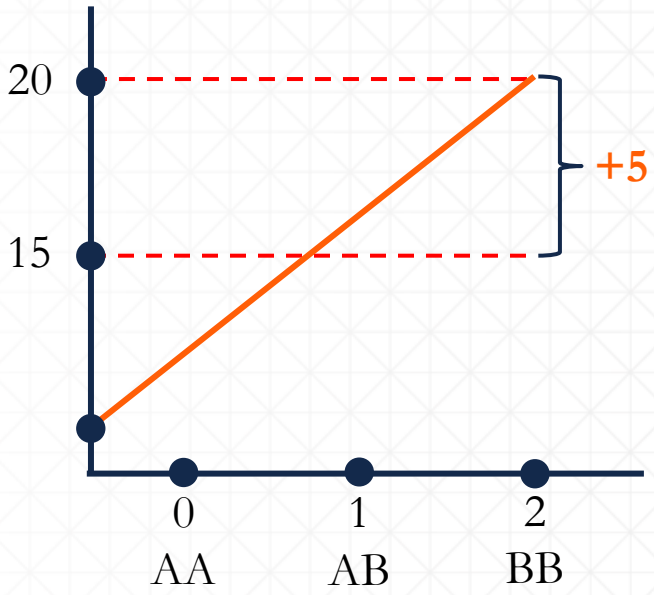
What is the substitution effect of allele B? **+5**

Foundation of genome association

- Single marker regression

  phenotype = non-genetic effects + genetic variant (SNP_1) + error

SNP genotype	Phenotype (average)
AA (0)	+10
AB (1)	+15
BB (2)	+20



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

$x = 0, 1, \text{or } 2$

Regression interpretation: changing one copy of allele A by B increases +5 units in the trait

Foundation of genome association

- In practice → genome-wide association analysis (GWAS)
  - ✓ Consider ***m*** genetic variants (SNPs), then we fit the model as:

phenotype = non-genetic effects + SNP_1 + SNP_2 + … + SNP_m + error

$$y_n = \beta_0 + \beta_1 SNP_1 + \beta_2 SNP_2 + \cdots + \beta_m SNP_m + e_i$$

phenotype

intercept or mean

$\beta_1, \ldots, \beta_m$
allele substitution effects estimated simultaneously

error or residual

AA (0), AB (1), BB (2)

SNP genotype	Phenotype (average)
AA (0)	+10
AB (1)	+15
BB (2)	+20

Foundation of genome association

- GWAS considering 400,000 SNPs and 10,000 samples

$$y_n = \beta_0 + \beta_1 SNP_1 + \beta_2 SNP_2 + \cdots + \beta_m SNP_m + e_i$$

Rewrite as:

$$Y = W_{SNP}\beta_{SNP} + e$$

$$12_1 = 0 + \beta_1 AA(0) + \beta_2 AB(1) + \cdots + \beta_{400000} AA(0) + e_i$$

$$18_2 = 0 + \beta_1 AB(1) + \beta_2 BB(2) + \cdots + \beta_{400000} BB(2) + e_i$$

$$16_{10000} = 0 + \beta_1 AB(1) + \beta_2 AB(1) + \cdots + \beta_{400000} AA(0) + e_i$$

$$\begin{bmatrix} 12 \\ 18 \\ \cdot \\ \cdot \\ \cdot \\ 16 \end{bmatrix} = \begin{bmatrix} 0\ 1\ \ldots\ 0 \\ 1\ 2\ \ldots\ 2 \\ \cdot \\ \cdot \\ \cdot \\ 1\ 1\ \ldots\ 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{400000} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_{400000} \end{bmatrix}$$

Solving this linear equation

$$\hat{\beta} = (W'W)^{-1}W'Y$$

Effect for each SNP

Foundation of genome association

- GWAS considering thousands of SNPs we fit the model as:

phenotype = non-genetic effects + SNP_1 + SNP_2 + … + SNP_m+ error

$$\boldsymbol{y = X\beta + W_{SNP}\beta_{SNP} + e}$$

$\boldsymbol{\beta}$ represents the non-genetic (environmental) fixed effects

$\boldsymbol{X}$ relates the fixed effect to the phenotype $\boldsymbol{y}$

$\boldsymbol{\beta_{SNP}}$ represents the SNP effects

$\boldsymbol{W_{SNP}}$ represent the SNP information coded as 0 (AA), 1 (AB), and 2 (BB)

This model does not account for population structure!

Foundation of genome association

- GWAS – output

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	SNP_8	SNP_9	...	SNP_M
y_1	1	0	1	1	1	0	0	1	2	...	0
y_2	1	2	0	0	1	1	1	0	1	...	0
y_3	0	1	1	0	0	1	1	0	0	...	0
.											
.											
.											
y_n	0	1	1	0	0	1	1	1	2	...	0

SNP ID	Effect
001	-0.003
002	-0.002
003	+0.010
004	+0.001
005	-0.098
006	+0.003
007	-0.002
008	+0.002
009	+0.676
.	.
.	.
.	.
M	+0.080

Foundation of genome association

❖ Association significance test

SNP ID	Effect
001	**-0.003**
002	-0.002
003	+0.010
004	+0.001
005	-0.098
006	+0.003
007	-0.002
008	+0.002
009	+0.676
.	.
.	.
.	.
M	+0.080

✓ Is this SNP significantly associated with the trait?

- Hypothesis test for the effect (slope in the regression model)

✓ p-value:

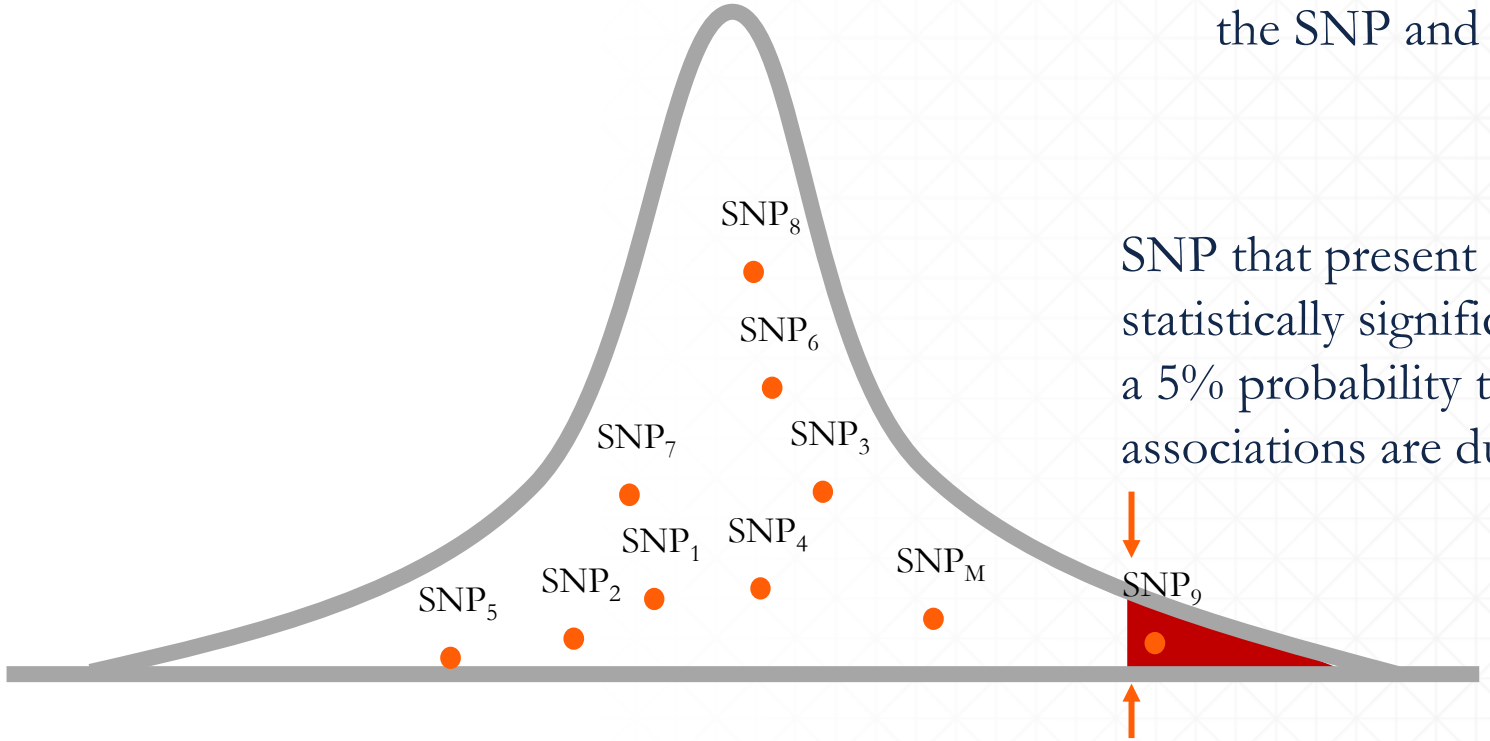
- A measure of the likelihood that the association found in the distribution of data points was due to random chance, given that there is no association between the SNP and the phenotype of interest

Foundation of genome association

❖ Association significance test: p-value

Null Hypothesis (H_0): There is no relationship or association between the SNP and phenotype of interest

Alternative Hypothesis (H_1): This is what you would believe if the null hypothesis is concluded to be untrue; it suggests that there is a relationship or effect between the SNP and the phenotype



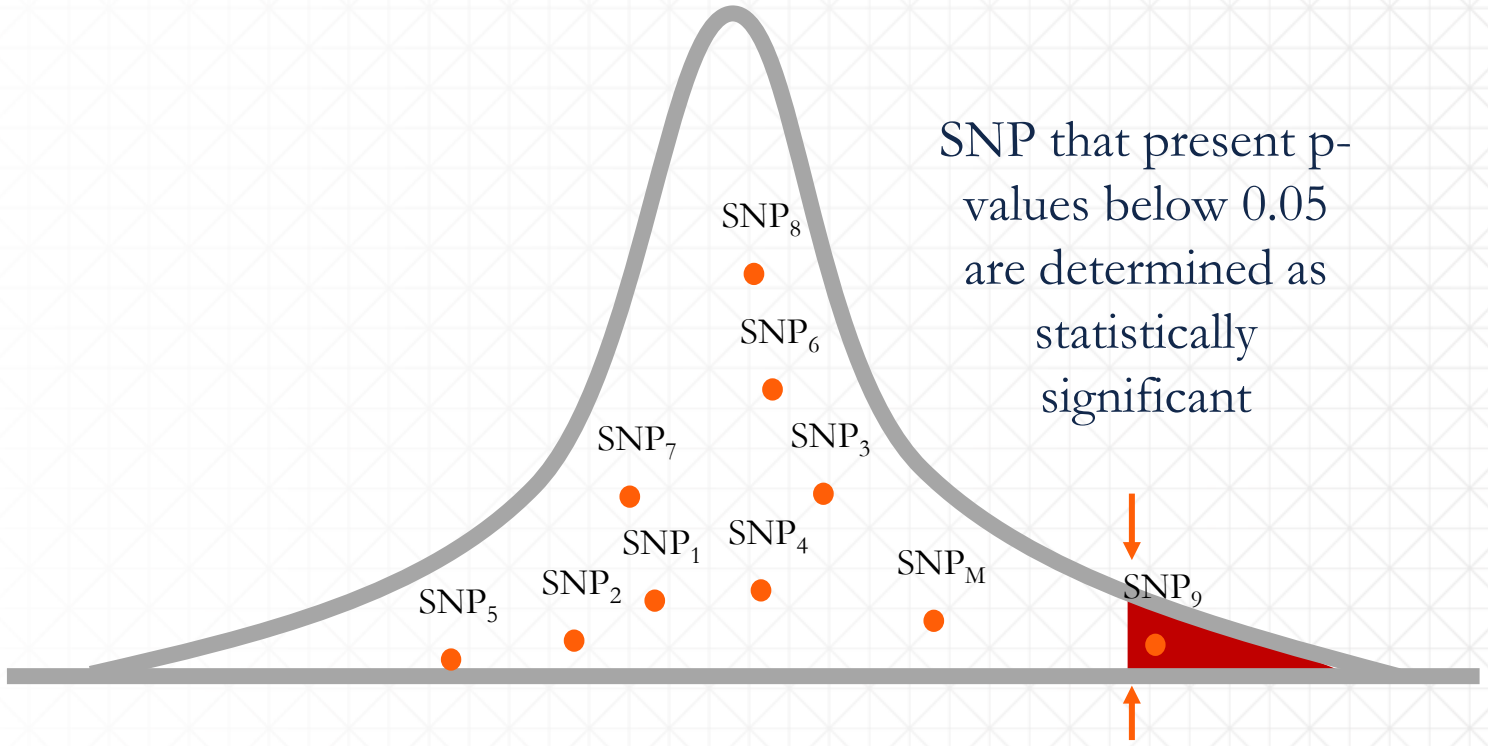
SNP that present p-values below 0.05 are determined statistically significant. This means that there is less than a 5% probability that the observed differences or associations are due to random chance alone.

Even with a threshold of 0.05, there is still a 5% chance of producing a false positive result

Foundation of genome association

- GWAS – output

SNP ID	Effect	p-value
001	-0.003	0.43
002	-0.002	0.74
003	+0.010	0.21
004	+0.001	0.98
005	-0.098	0.03
006	+0.003	0.52
007	-0.002	0.65
008	+0.002	0.76
009	+0.676	5E-11
.	.	.
.	.	.
.	.	.
M	+0.080	0.02



Foundation of genome association

❖ Association significance test

✓ When working with many SNPs, the number of false positives increases significantly, decreasing the statistical power of the study

✓ Multiple test corrections:

✓ Bonferroni correction (BF):

$$BF = \frac{\alpha}{m}$$

α = is the desired significance level

m = is the number of hypotheses or SNPs

Foundation of genome association

- GWAS – output

SNP ID	Effect	P-value
001	-0.003	0.43
002	-0.002	0.74
003	+0.010	0.21
004	+0.001	0.98
005	-0.098	0.03
006	+0.003	0.52
007	-0.002	0.65
008	+0.002	0.76
009	+0.676	5E-11
.	.	.
.	.	.
.	.	.
M	+0.080	0.02

→ p-value < BF (5E-5)

Bonferroni correction threshold

$$BF = \frac{\alpha}{m}$$

$$BF = \frac{0.05}{1000}$$

$$BF = 5E\text{-}5$$

Let's have a break!

How to estimate the SNP effects?

- Problem:
  - ✓ Unique estimates of marker effects do not exist
- Solution 1 (classical):
  - ✓ Indirect (e.g., GBLUP) or direct effect estimation (e.g., ridge regression)
- Solution 2 (Bayesian):
  - ✓ Direct effect estimation assigning a "prior distribution" on the marker effects (e.g., BayesA, BayesB, BayesC, BayesR, …)
- Solution 3 (Artificial Intelligence) – **Be careful with assumptions**
  - ✓ ML, GANs, VAE, Diffusion models, LLM

Genomic best linear unbiased prediction (GBLUP)

ID	P	SNP_1	SNP_2	SNP_3	⋯	SNP_M
1	10	0	1	1	⋯	0
2	12	1	0	2	⋯	1
3	20	0	2	0	⋯	0
4	10	1	1	0	·	1
5	11	1	1	0	·	0
6	8	1	0	0	·	1
7	12	0	1	0	·	0
8	16	1	1	1	·	1

$$\boldsymbol{y = X\beta + Zu + e}$$

$$\begin{bmatrix} X'X & X'Z \\ Z'Z & Z'Z + \lambda G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Genomic relationship matrix

GV
+2
-0.1
+4
+0.1
-0.2
+1
+5
+2

What is the limitation?

ONLY use individuals that have SNP and phenotype information to predict BV/ SNP effects

SNP	Effect	P-value
001	-0.003	0.43
002	-0.002	0.74
003	+0.010	5E-11
.	.	.
.	.	.
.	.	.
M	+0.080	0.02

SNP effects

$$\hat{a} = (M'M)^{-1}(M'\hat{u})$$

Ridge-regression genomic best linear unbiased prediction (RR-GBLUP)

Equivalent models

$$y = X\beta + Zu + e$$

$$y = X\beta + \sum_{i=1}^{m} M_i a_i + e$$

SNP effects estimated in the training population

SNP	Effect	P-value
001	-0.003	0.43
002	-0.002	0.74
003	+0.010	5E-11
.	.	.
.	.	.
.	.	.
M	+0.080	0.02

What is the limitation?

ONLY use individuals that have SNP and phenotype information to estimate the SNP effects

In reality, what we want is …

ID	P	SNP_1	SNP_2	SNP_3	…	SNP_M
1	8	?	?	?	…	?
2	10	?	?	?	…	?
3	12	?	?	?	…	?
4	20	?	?	?	.	?
5	10	?	?	?	.	?
6	11	?	?	?	.	?
7	8	1	1	0	.	1
8	12	1	1	0	.	0
9	16	1	0	0	.	1
10	9	0	1	0	.	0
11	16	1	1	1	.	1
12	14	0	0	1	.	0
13	?	1	1	0	.	1
14	?	0	2	0	.	1
15	?	2	1	0	.	0
16	?	1	0	1	…	1

Long historical data with **phenotypes** observations for millions of individuals

Short historical data with **phenotypes** observations and **genotypes** (SNPs) data

Individuals with **ONLY genotype** (SNPs) information

Combine all this information to predict the genomic values

Single-step genomic best linear unbiased prediction (ssGBLUP)

ID	P	SNP_1	SNP_2	SNP_3	…	SNP_M	GV
1	8	?	?	?	…	?	+2
2	10	?	?	?	…	?	+1
3	12	?	?	?	…	?	-0.3
4	20	?	?	?	.	?	+0.1
5	10	?	?	?	.	?	+0.3
6	11	?	?	?	.	?	+5
7	8	1	1	0	.	1	+4
8	12	1	1	0	.	0	-0.6
9	16	1	0	0	.	1	-0.2
10	9	0	1	0	.	0	+1
11	16	1	1	1	.	1	+2
12	14	0	0	1	.	0	-0.4
13	?	1	1	0	.	1	+0.6
14	?	0	2	0	.	1	-0.08
15	?	2	1	0	.	0	+8
16	?	1	0	1	…	1	-0.9

The H genetic relationship matrix

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

$$\begin{bmatrix} X'X & X'Z \\ Z'Z & Z'Z + \lambda H^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

GV for all individuals in the H matrix

SNP effects

$$\hat{a} = (M'M)^{-1}(M'\hat{u})$$

Christensen and Land (2010)

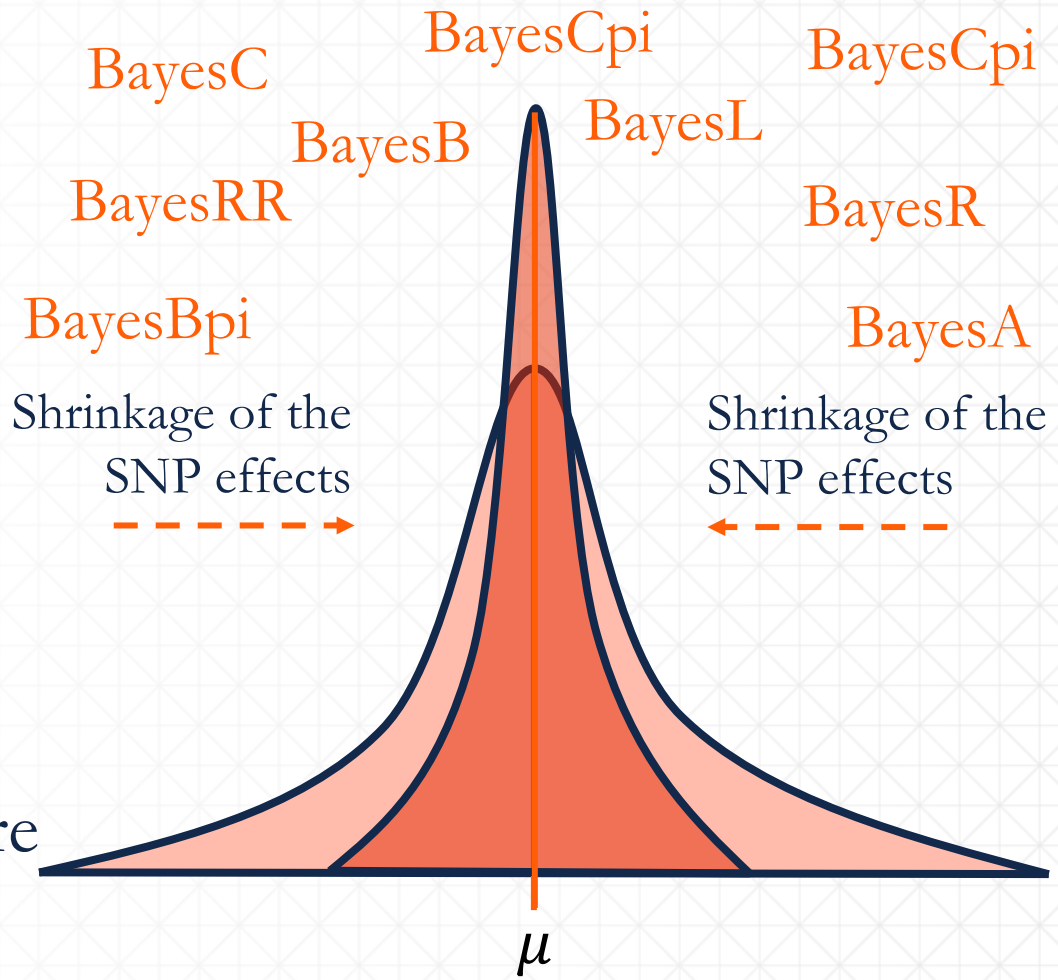
Bayesian methods

- The Bayes alphabet:
  - ✓ Same as the idea of RR-GBLUP

$$y = X\beta + \sum_{i=1}^{m} M_i a_i + e$$

$$a_i \mid \sigma_i^2 \sim N(0, \sigma_i^2)$$

Variance for each SNP



Bayesian methods allow some SNPs to have larger effects and shrink others more strongly.

Bayesian methods

❖ Advantage

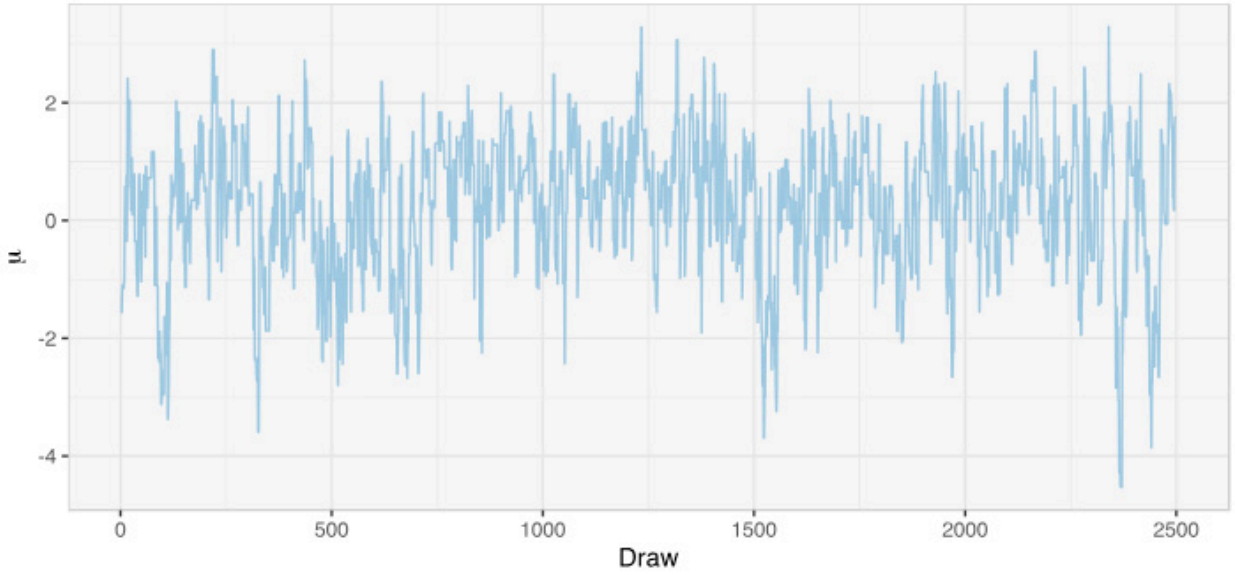
LARGER p

ID	P	SNP_1	SNP_2	SNP_3	…	SNP_M
1	10	0	1	1	…	0
2	12	1	0	2	…	1
3	20	0	2	0	…	0
4	10	1	1	0	.	1
5	11	1	1	0	.	0
6	8	1	0	0	.	1
7	12	0	1	0	.	0
8	16	1	1	1	.	1

Smaller n

❖ Limitations

- ✓ Markov Chain Monte Carlo Sampling
- ✓ Convergence challenges
- ✓ Computationally intensive
- ✓ **ONLY** individuals who have SNP and phenotype information re used to estimate the SNP effects



Genome-wide association analysis (GWAS) – in summary

- **Most popular:** single marker regression
  - ✓ Individuals must have both phenotype and genotypic records
  - ✓ Fits one SNP at a time

- **Single-step analysis:**
  - ✓ Combines all the available phenotypic, pedigree, and genotypic information
  - ✓ Fits all the SNPs simultaneously
  - ✓ Identify regions that explain a significant amount of genetic variance

- **Bayesian methods**
  - ✓ Originally developed for genomic prediction, then extended for gene mapping

- **Haplotype-based analysis**

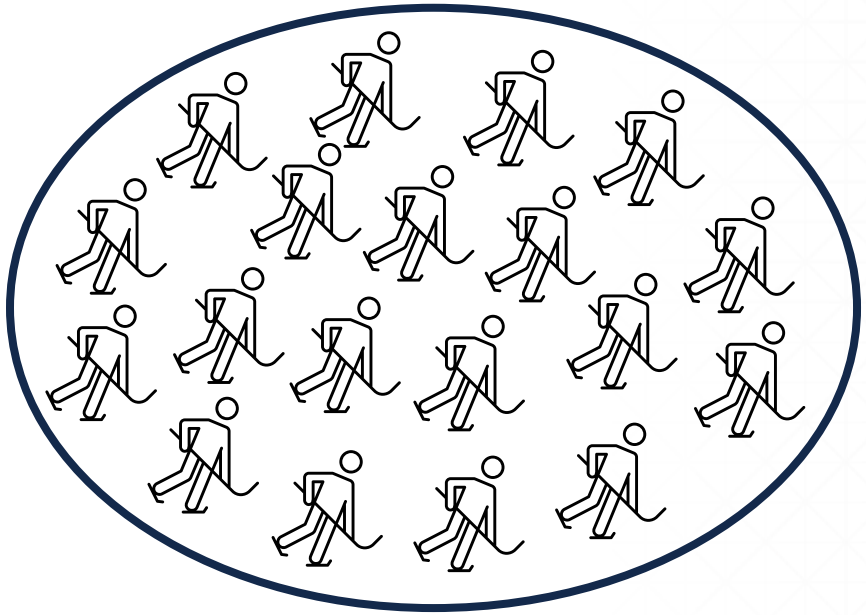
GWAS – final remarks

- **Interpretation of the findings**
  - ✓ Significant SNPs explain only a small proportion of the genetic variance
  - ✓ Significant SNPs, in most cases, are not the causal variants
  - ✓ The identification of causal variants is often a complex process
- **Replication/validation**
  - ✓ Validate the findings in an independent population
- We use GWAS to **point us** toward genes involved in the expression of a trait
  - ✓ GWAS **cannot** identify a specific gene involved in the expression of a trait
  - ✓ GWAS **can** tell you where to start looking for a gene involved in the expression of a trait

From association to prediction

- You already learned how to estimate the SNP effect
  - ✓ You can use any method we covered so far
  - ✓ Prediction just applies them to a new person's genotype

Training population



Phenotypes + Genotypes (SNPs)

RR-GBLUP

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{M_i a_i} + \boldsymbol{e}$$

SNP effects

SNP	Effect
001	-0.003
002	-0.002
003	+0.010
.	.
.	.
.	.
M	+0.080

From association to prediction – good approach

- Predicting target trait

Validation population

Only Genotypes

SNP information for the validation population coded as 0, 1, and 2

ID	SNP_1	SNP_2	SNP_3	…	SNP_M
1	0	1	0	…	1
2	1	0	2	…	1
3	0	2	1	…	2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N	2	0	1	…	0

X

SNP effects estimated in the training p

SNP	Effect
001	-0.003
002	-0.002
003	+0.010
.	.
.	.
.	.
M	+0.080

=

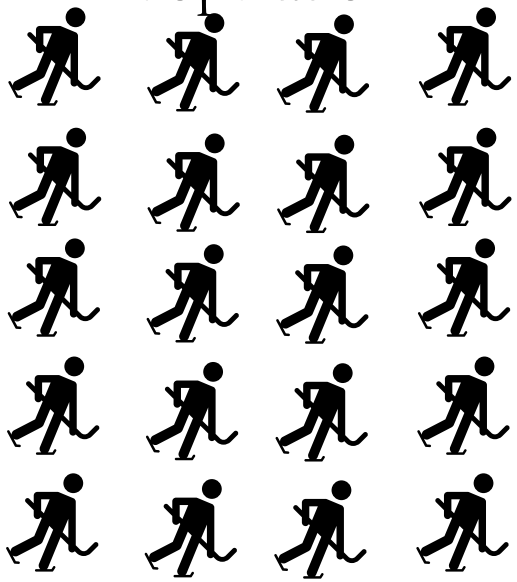
Predicted target trait

GV
+2
-0.1
+4
.
.
.
+5

From association to prediction – better approach

- GBLUP approach

Population



ID	P	SNP_1	SNP_2	SNP_3	…	SNP_M	GV
1	10	0	1	1	…	0	+2
2	12	1	0	2	…	1	-0.1
3	20	0	2	0	…	0	+4
4	10	1	1	0	.	1	+0.1
5	11	1	1	0	.	0	-0.2
6	8	1	0	0	.	1	+1
7	12	0	1	0	.	0	+5
8	16	1	1	1	.	1	+2
9	?	0	0	1	.	0	+1
10	?	1	1	0	.	1	+2
11	?	0	2	0	.	1	-0.3
12	?	2	1	0	.	0	+0.2
13	?	1	0	1	…	1	+0.1

We predict the genomic values:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

$$\begin{bmatrix} X'X & X'Z \\ Z'Z & Z'Z + \lambda G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Genetic relationship matrix

We want to estimate the genomic values for all individuals that are in the genetic relationship matrix in a single step!

From association to prediction – best approach ssGBLUP

ID	P	SNP_1	SNP_2	SNP_3	…	SNP_M	GV
1	8	?	?	?	…	?	+2
2	10	?	?	?	…	?	+1
3	12	?	?	?	…	?	-0.3
4	20	?	?	?	.	?	+0.1
5	10	?	?	?	.	?	+0.3
6	11	?	?	?	.	?	+5
7	8	1	1	0	.	1	+4
8	12	1	1	0	.	0	-0.6
9	16	1	0	0	.	1	-0.2
10	9	0	1	0	.	0	+1
11	16	1	1	1	.	1	+2
12	14	0	0	1	.	0	-0.4
13	?	1	1	0	.	1	+0.6
14	?	0	2	0	.	1	-0.08
15	?	2	1	0	.	0	+8
16	?	1	0	1	…	1	-0.9

The H genetic relationship matrix

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

$$\begin{bmatrix} X'X & X'Z \\ Z'Z & Z'Z + \lambda H^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

GV for all individuals in the H matrix

However, it requires family relationship information to build A matrix

From association to prediction - summary

- What it buys you and what limits it
  - ✓ The output is a single score per person, which enables stratification, but its accuracy has hard ceilings.

Population sorted by polygenic score

low risk — **high-risk tail**

↑ flag the top few % → earlier / intensified screening or intervention

Uses

- A continuous per-person score ranks people and flags the high-risk tail (e.g., top 5%).
- Combine with established clinical risk factors an integrated risk estimate, not genetics alone.

Limits

- Accuracy is capped by training size and trait heritability. You can't predict beyond what is heritable.
- Portability: a score trained in one ancestry predicts worse in others (effects & LD differ).

Where classical association testing stops

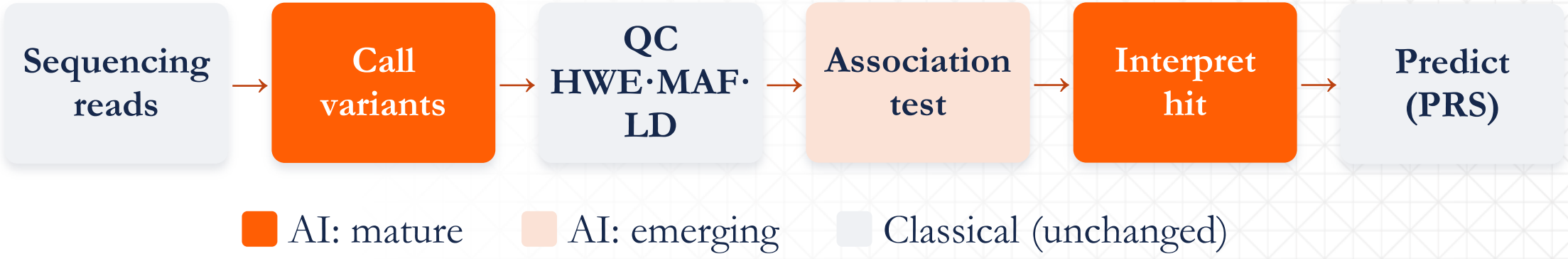
- The mixed model is powerful, but it has three hard limits:
  - ✓ It tests one SNP at a time, assuming additive, independent effects
  - ✓ It tells you where the signal is, not what the variant does, especially for the ~95% of GWAS hits that fall outside protein-coding regions
  - ✓ Prediction accuracy plateaus as sample size grows, because linear models can't capture epistasis or non-linear gene $\times$ environment effects

- The model $\boldsymbol{y = X\beta + W_{SNP}\beta_{SNP} + e}$ is the state-of-the-art

- AI enters exactly where association tests can't go further.

Where does AI enter the pipeline?

❖ The classical workflow is unchanged. AI plugs in at specific stations.

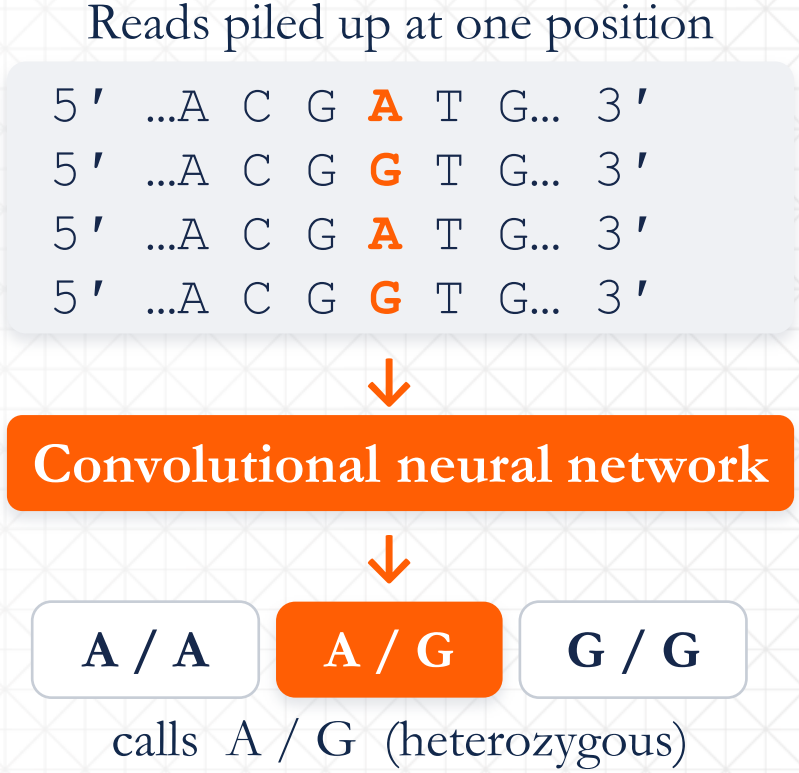


BEYOND THE PIPELINE - A FOURTH USE: GENERATING DATA

Privacy-preserving sharing · boosting underrepresented ancestries. AGs can stand in for or augment the real data feeding the pipeline above.

AI in polymorphism: calling the SNP

- Every genotype begins as billions of short, error-prone reads.
- Classical calling (GATK, bcftools): statistical likelihoods plus hand-tuned filters.
- DeepVariant reframes calling as image recognition, and beats the prior state of the art with no hand-coded genomics rules.



Why we care: a miscalled SNP is a **false polymorphism**. It corrupts your HWE check, MAF filter, and every association downstream.

Why this matters for your data

Across platforms

Illumina, PacBio HiFi, Oxford Nanopore, you retrain the model, you don't rebuild the method.

Across species

Trained on human truth sets, it calls variants in other mammals, livestock rides on human reference data.

Low coverage

The biggest accuracy gains are where low-pass sequencing + imputation lives, cost-sensitive genotyping.

Even the basecalling

In a nanopore, the raw electrical signal (A/C/G/T) step is itself a neural network.

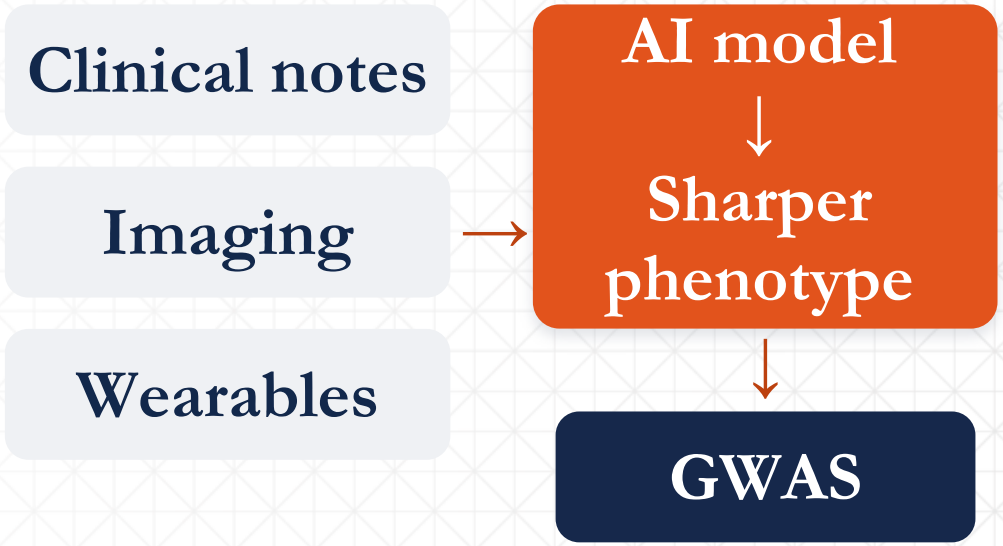
Related callers you'll see: DeepTrio (families) · DNAscope · Clair3 (nanopore)

Start with a better phenotype

- GWAS power is capped by phenotype quality
  - ✓ The cheapest win is to sharpen the Y, not just the X.

$$\rightarrow \boxed{y} = X\beta + W_{SNP}\beta_{SNP} + e$$

  - ✓ NLP on clinical notes → cleaner case/control status
  - ✓ Deep learning on images (retina, ECG, MRI) → quantitative traits no human reads off the scan
  - ✓ Wearables → continuous digital biomarkers



AI-derived disease **liabilities** uncovered novel loci that conventional case/control GWAS missed — and combining the two was more powerful than either alone.

AI inside the association test

To be honest, for discovery, the linear mixed model still wins. ML is already inside the tools and adds where additivity can't.

1 Train

A neural net classifies case vs. control from genotypes.

2 Rank

Explainability (layer-wise relevance) ranks the SNPs that drove the call.

3 Test

Classical hypothesis tests run on the shortlist only.

DeepCOMBI - the network is a filter; the statistics still adjudicate.

Already in the tools. regenie uses ridge regression to build whole-genome predictors, then tests each SNP (BOLT-LMM: Bayesian mixture prior).

Bayesian approaches. BayesB / BayesC already assumes most SNP effects are zero and let the data pick the few that matter - same spirit as ML feature selection.

From a SNP to a mechanism

- Most GWAS hits are non-coding SNPs. Thus, a Manhattan peak is a location, not a function.

Worked example: lactase persistence (rs4988235). A single non-coding **C→T** change in a regulatory enhancer (inside *MCM6*) keeps the *LCT* lactase gene switched on into adulthood. It spread in dairying populations.

AlphaGenome

Nature, 2026

Input ~1 Mb of sequence → predicts expression, chromatin & splicing at single-base resolution. Scores how one base change shifts them.

Enformer

sequence model

Predicts how a single SNP changes nearby gene expression directly from the DNA sequence.

DNA language models

Evo2 · Nucleotide Transformer

Score a single nucleotide's likely impact from learned evolutionary constraint, no labeled assay needed.

Coding SNPs are the minority — for those, AlphaMissense scores the effect. **But for SNPs, the action is non-coding.**

A fourth use: generative synthetic genomes

- Calling and association use real genotypes. Generative models manufacture new ones. Artificial genomes (AGs) that resemble real people but belong to none.

Real cohort genotypes → **Generative model**
VAE · GAN · RBM · diffusion · LM → **Artificial genomes (AGs)**

Why it's useful

- Privacy-preserving sharing
- Augment under-represented
- Imputation panels
- Benchmarking

Where it breaks

- Rare alleles under-represented - often the variants you need
- LD & fine structure may drift - training on synthetic data can bake in artifacts

We already build reference populations by simulation - generative models are an alternative.

What AI does not change!

Correlation ≠ causation

A stronger signal isn't a causal claim; fine-mapping and Mendelian randomization are still required.

Portability bias

Models trained on European-ancestry data export poorly, now affecting the functional models too, not just PRS.

Validation is mandatory

Explainability is not replication. Nothing here is a validated, standalone clinical decision.

The linear model isn't dead

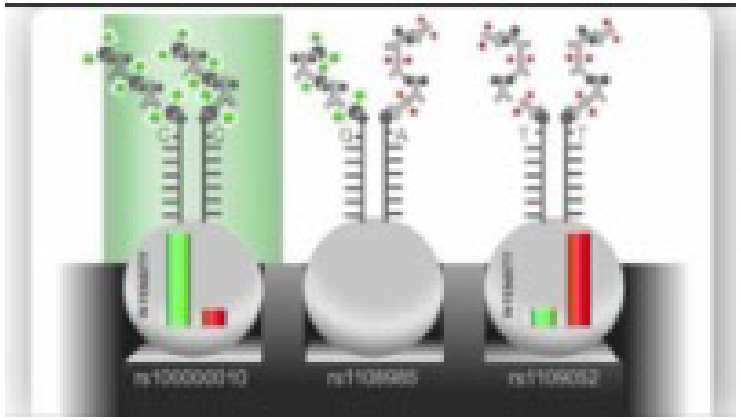
For PRS, GBLUP, and ridge still match or beat deep learning at realistic sample sizes.

AI changed what we can observe and interpret. It did not change what counts as evidence.

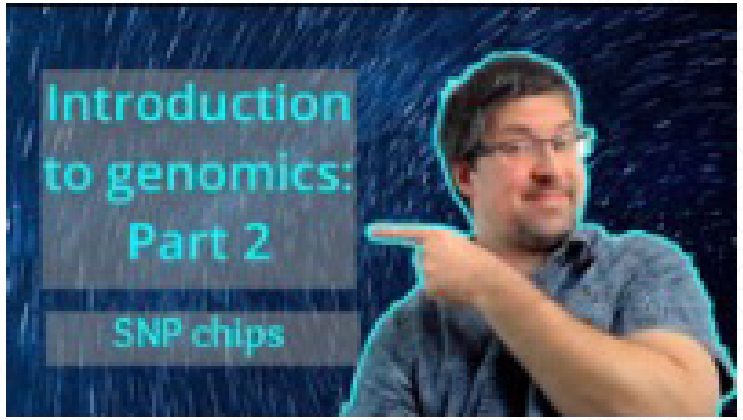
Future resources

- Additional resources to learn more about genotype data and genotype imputation

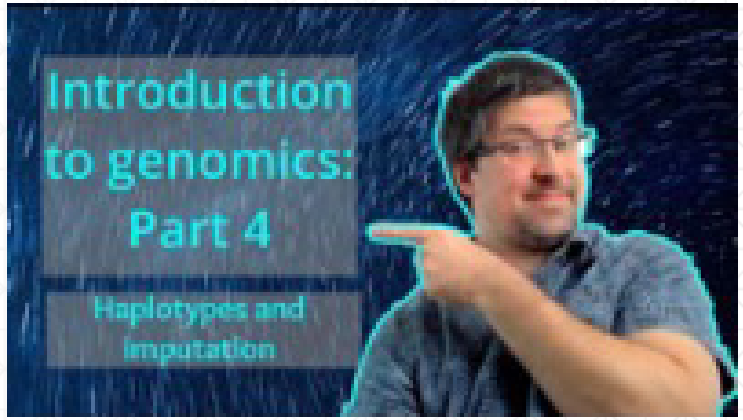
Illumina BeadChip



SNP data

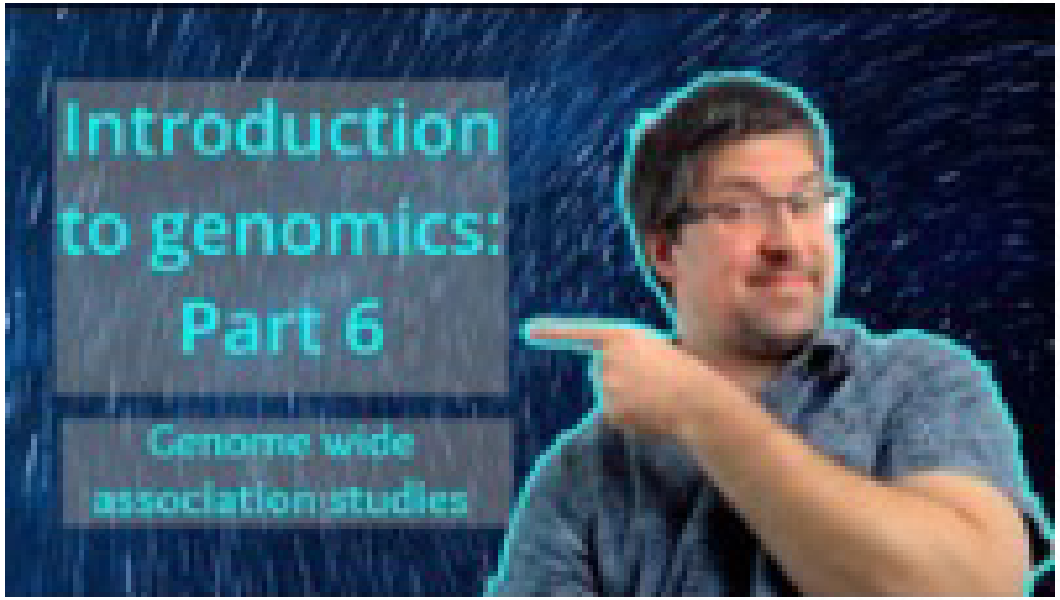


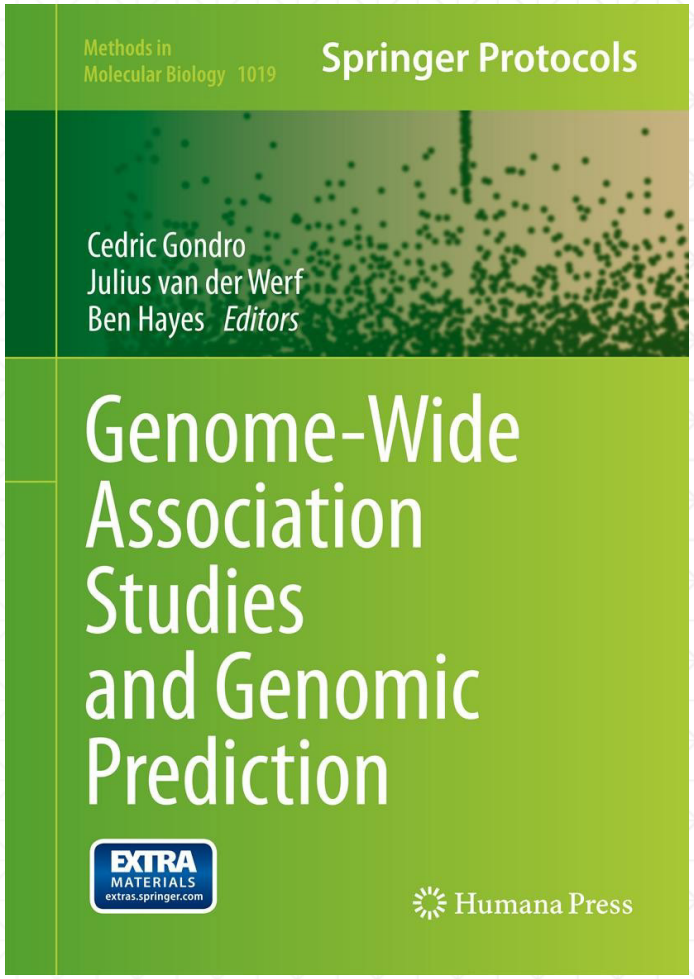
Genotype imputation



Future resources

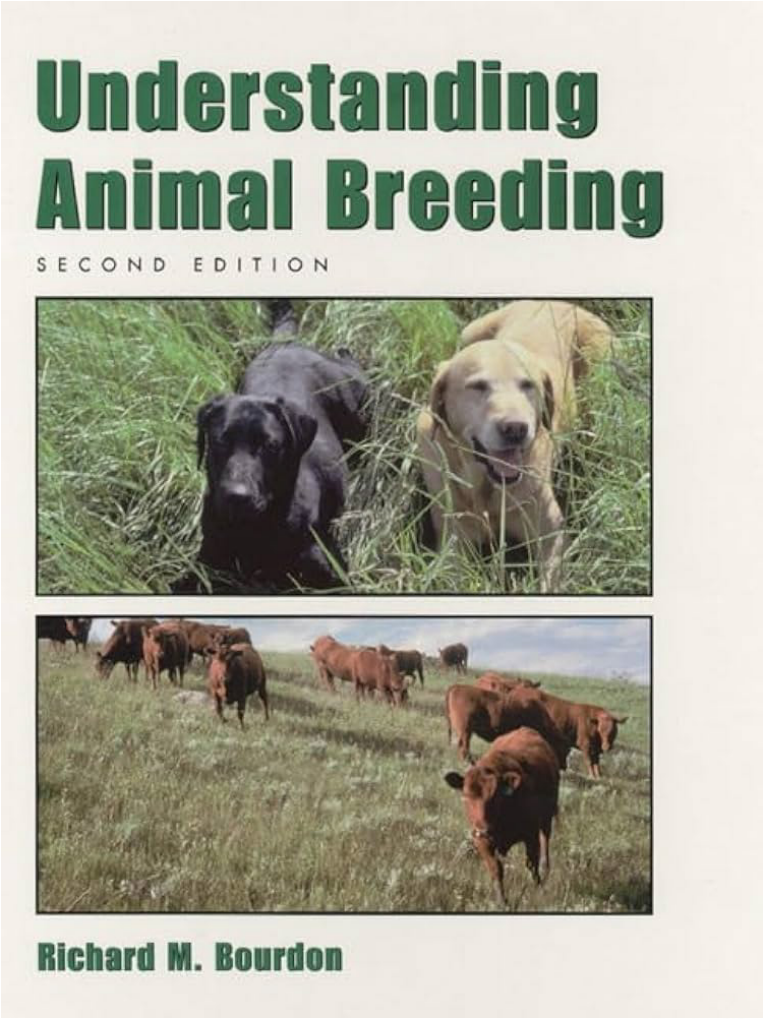
- Additional resources to learn more about GWAS

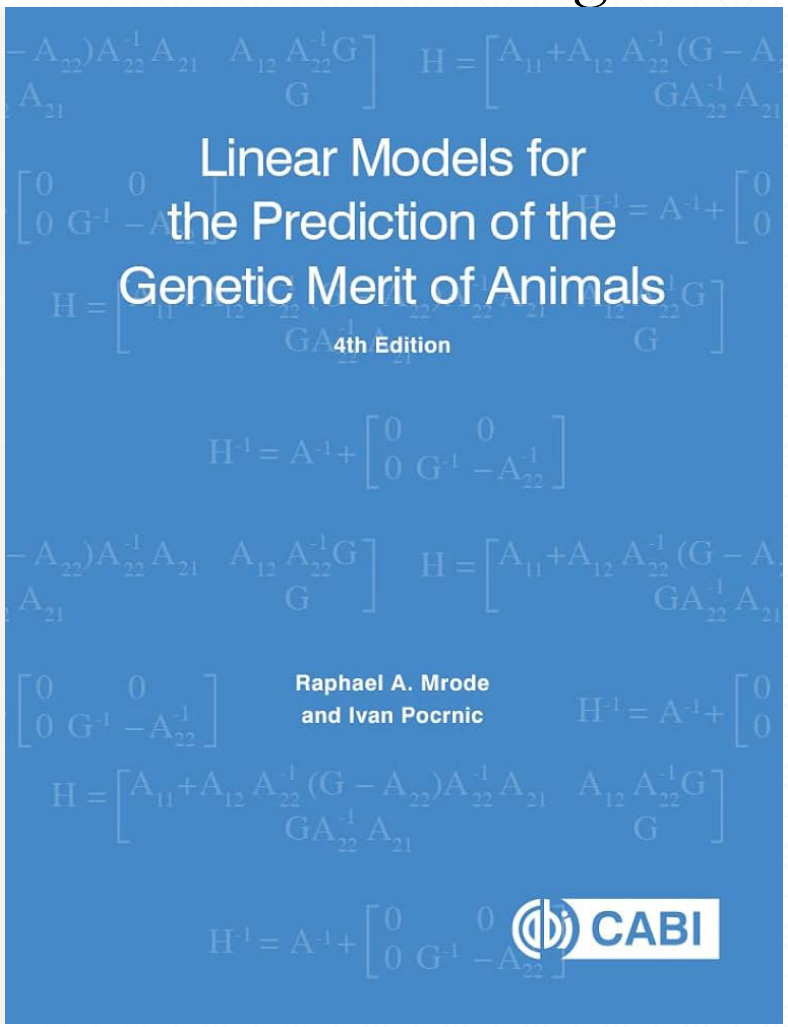


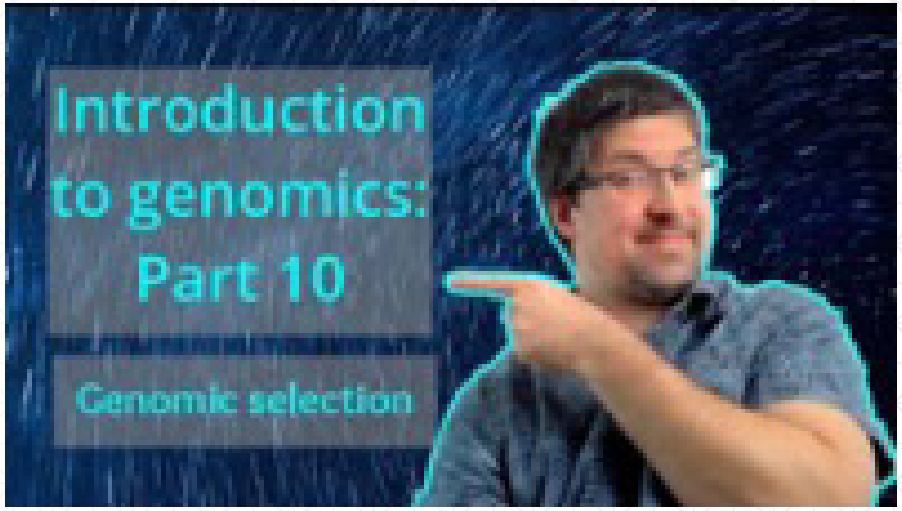


Future resources

- Additional resources to learn more about genomic selection







@iphenogenomics

Thank you!

Dr. Tiago Bresolin
bresolin@illinois.edu

ILLINOIS
PHENOMICS
AND GENOMICS
RESEARCH GROUP