Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

Haoyang Zeng
Lecture 16
April 9, 2019

Identifying genetic variants causal for traits and diseases

http://mit6874.github.io

Today's lecture

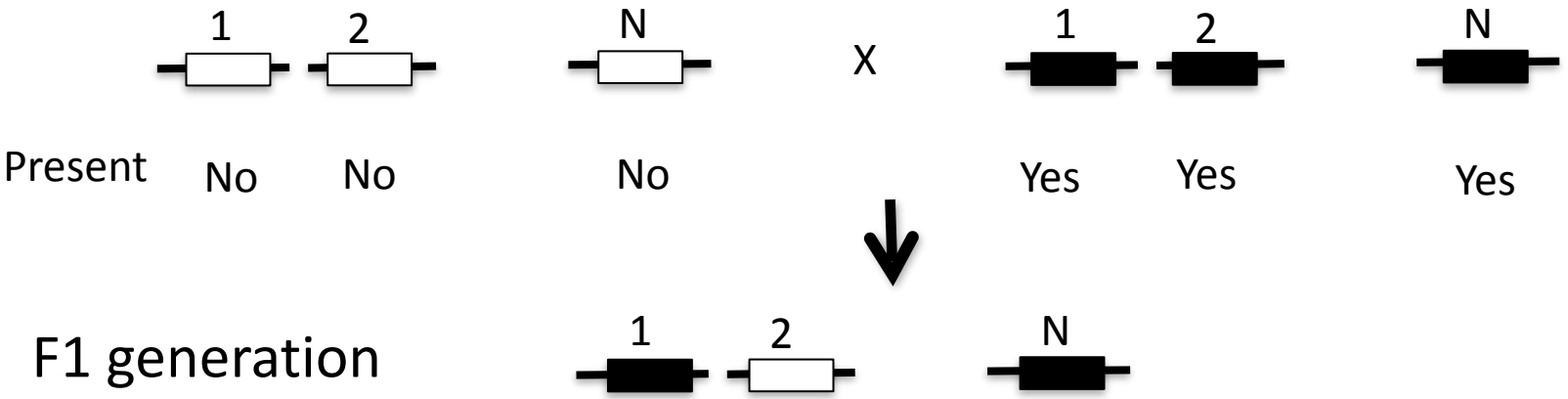
- Fundamentals of heritability
- Genome-wide association study (GWAS)
- Predicting functional variants using machine learning

Part 1 - Fundamentals of heritability

Genotype to Phenotype

- Genotype
  - Complete genome sequence (or an approximation)
  - Can be defined by markers at specific genomic sites that describe differences with a defined reference genome
- A phenotype is defined by one or more traits
  - Non-quantitative trait (dead/alive, etc.)
  - Quantitative Trait
    - Fitness (growth rate, lifespan, etc.)
    - Morphology (height, etc.)
    - Gene expression
- Quantitative Trait Loci – Genetic marker that is associated with a quantitative trait
  - eQTL – marker associated with gene expression

Binary haploid genetic model



Example Phenotypes

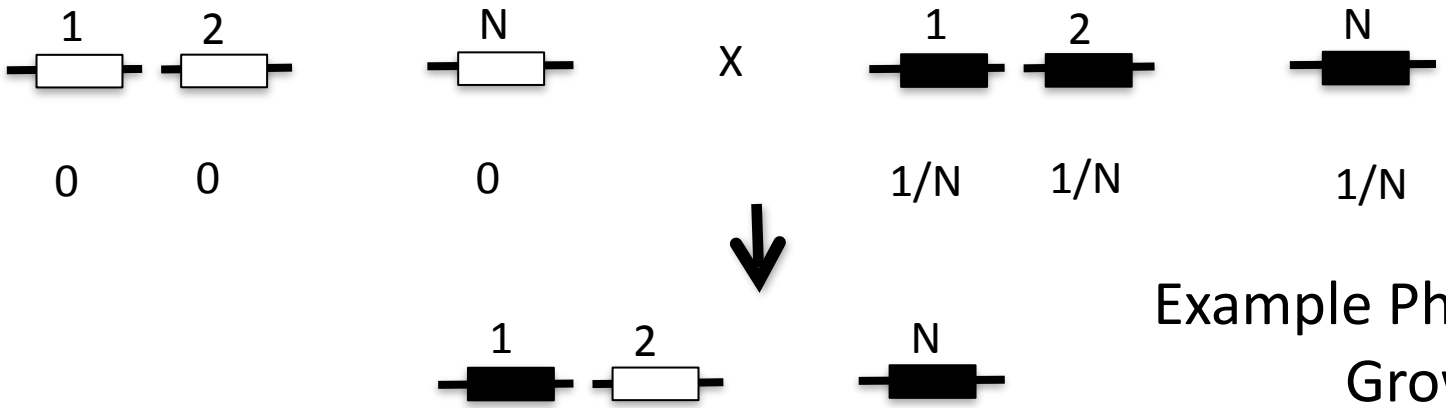
Alive/Dead in a specific environment

Resistant to a specific virus

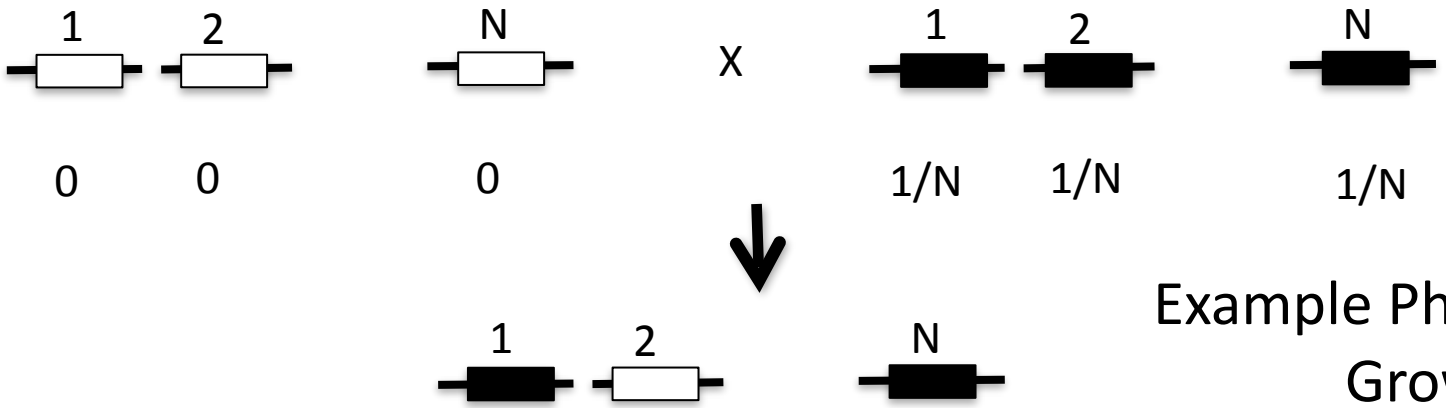
Suppose we tested 128 F1s, 16 resistant.
What is your estimate of N?

N is estimated by $\log_2$ (# F1s tested / # F1s with phenotype)

Quantitative haploid genetic model



Quantitative haploid genetic model



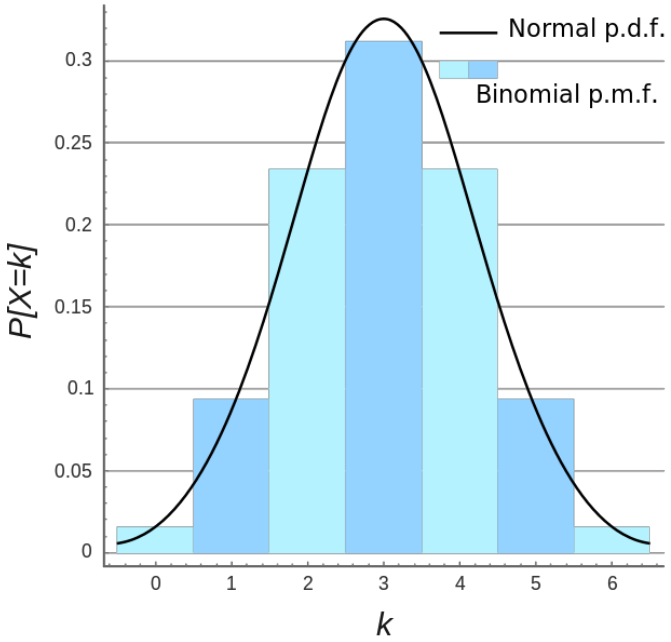
Example Phenotype Growth Rate

$$y = x / N$$

$$p(x, N) = \binom{N}{x} (1 - .5)^{N-x} .5^{x}$$

$$E[x] = N/2 \quad E[y] = 1/2$$

$$\sigma_x^2 = N/4 \quad \sigma_y^2 = 1/(4N)$$



Phenotype is a function of genotype plus an environmental component

- i – individual in [1 .. N]
- g_i – genotype of individual i
- p_i – quantitative phenotype of individual i (single trait)
- e_i – environmental contribution to p_i

Phenotype is a function of genotype plus an environmental component

- i – individual in [1 .. N]
- g_i – genotype of individual i
- p_i – quantitative phenotype of individual i (single trait)
- e_i – environmental contribution to p_i

$$p_i = f(g_i) + e_i \qquad \sigma_p^2 = \frac{1}{N}\sum_{i=1}^{N}(p_i - \mu_p)^2$$

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + 2\sigma_{ge}^2 \qquad E[e_i] = 0 \qquad E[e^2] = \sigma_e^2$$

g and e assumed or made independent yields $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$

Why two heritabilities?

- Broad-sense
  - Describes the upper bound for phenotypic prediction by an optimal arbitrary model
  - Reveals complexity of molecular mechanism
- Narrow-sense
  - Describes the upper bound for phenotypic prediction by a linear model
  - Describes relative resemblance and utility of family disease history
  - Efficient genetic mapping studies

Key caveats

- Heritability is a property of population (segregating allele frequencies) and environment (noise component)
- “Heritability” in practice may refer to either broad- or narrow-sense (or an implicit assumption that they are the same)
- Estimation is difficult (matching environments and avoiding confounding)

H^2 - Broad Sense heritability

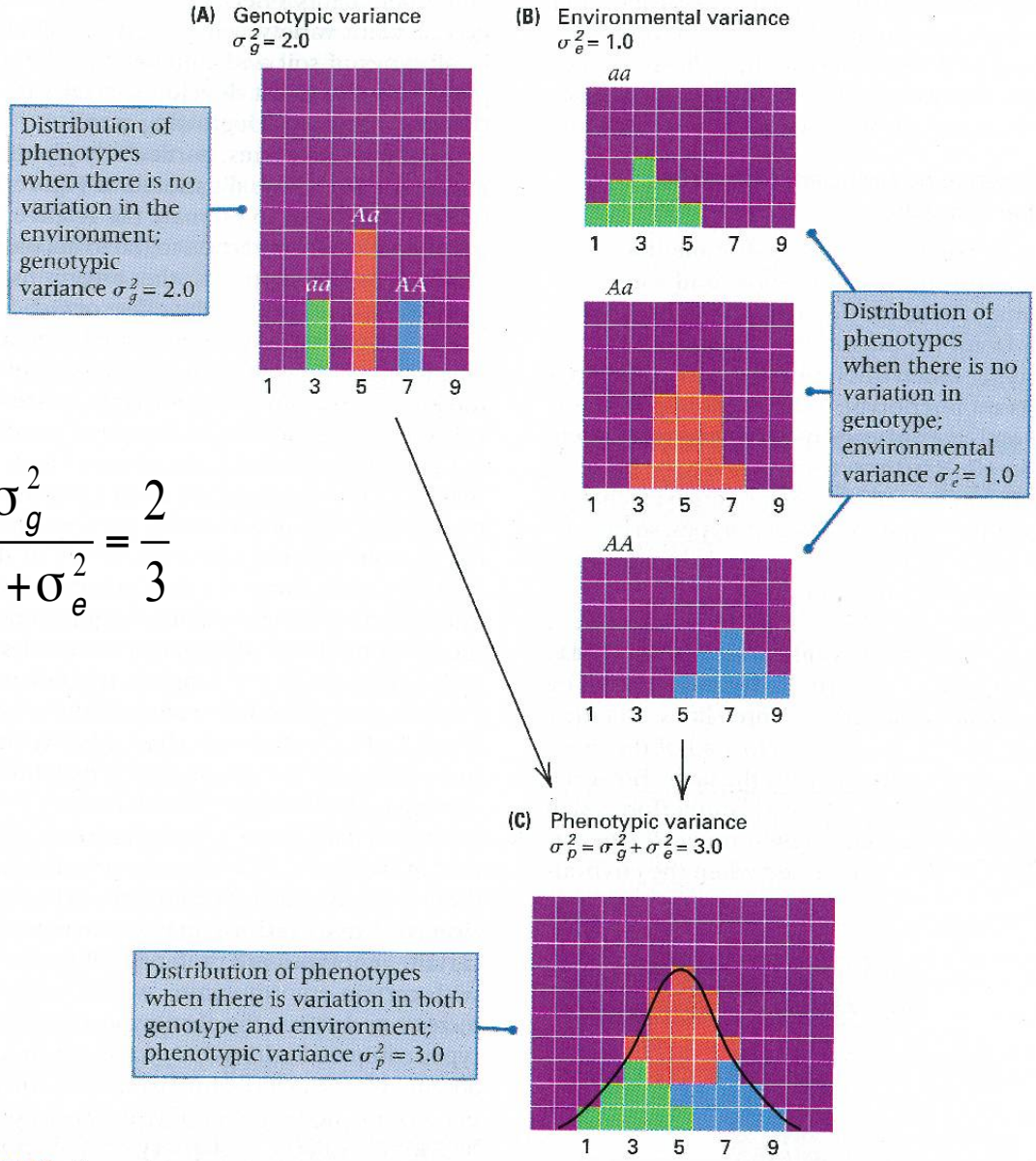
- Fraction of phenotypic variance explained by genetic component

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_p^2 - \sigma_e^2}{\sigma_p^2}$$

- Can estimate σ_e^2 from identical twins or clones.

Broad heritability of a trait is fraction of phenotypic variance explained by genetic causes

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{2}{3}$$



Additive model of phenotype

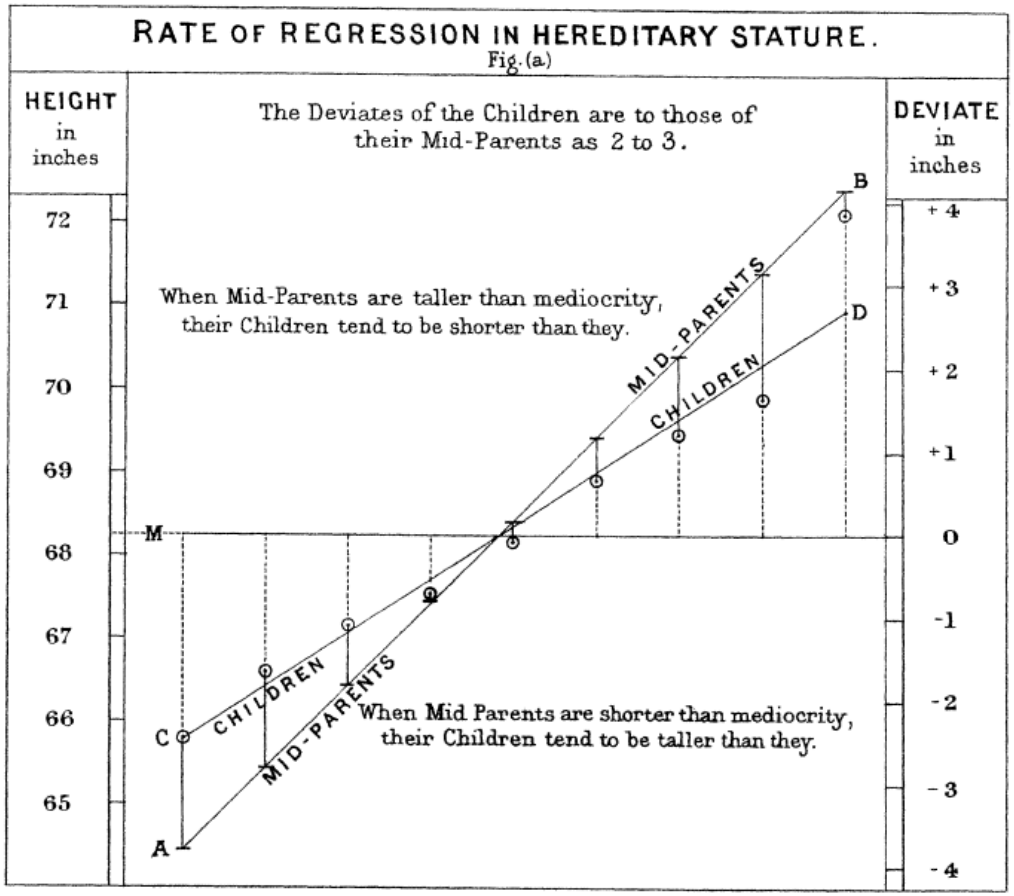
g_{ij} is marker j for individual i with values {0,1}
Quantitative trait loci (QTLs) are discovered for each trait

$$f_a(g_i) = \sum_{j \in QTL} \beta_j g_{ij} + \beta_0$$

$$E\left[f_a(g_i) \right] = \frac{f_a(p_1)}{2} + \frac{f_a(p_2)}{2}$$

Children tend to midpoint of parents for additive traits as they are expected to get an equal number of loci from each parent

Historical heritability example



Galton, “Regression towards mediocrity in hereditary stature” (1886)

h^2 - Narrow Sense heritability

- Fraction of phenotypic variance explained by an additive model of markers
- $f_a(g_i)$ is additive model of genotypic components in g_i
- Difference between heritability explained by additive model and general model is one source of “missing heritability” in current studies

h^2 - Narrow Sense heritability

- Fraction of phenotypic variance explained by an additive model of markers
- $f_a(g_i)$ is additive model of genotypic components in g_i
- Difference between heritability explained by additive model and general model is one source of “missing heritability” in current studies

$$p_i = f_a(g_i) + e_i \qquad \sigma_a^2 = \sigma_p^2 - \frac{1}{N}\sum_{i=1}^{N}\left(p_i - f_a(g_i)\right)^2$$

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

Example trait heritabilities

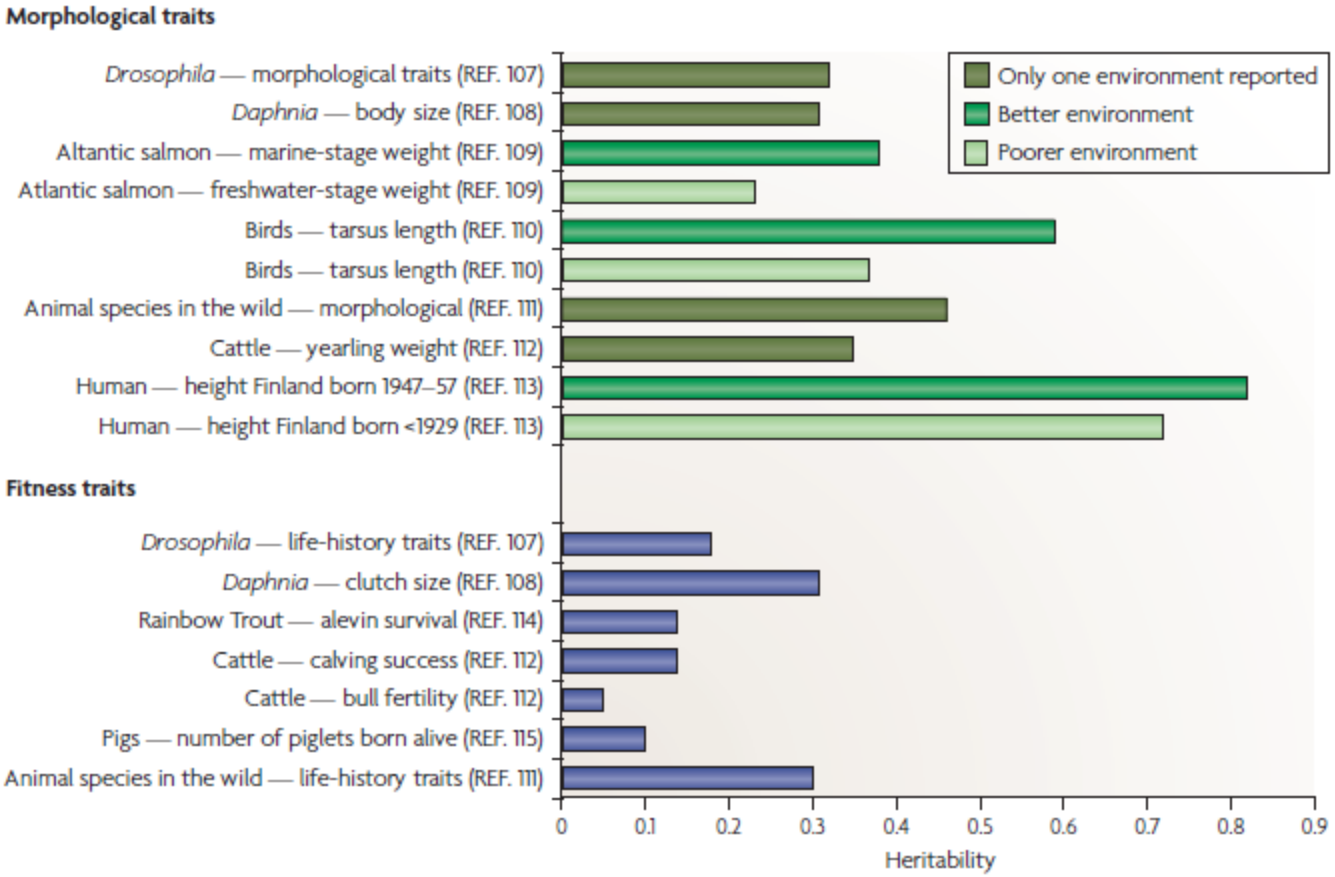
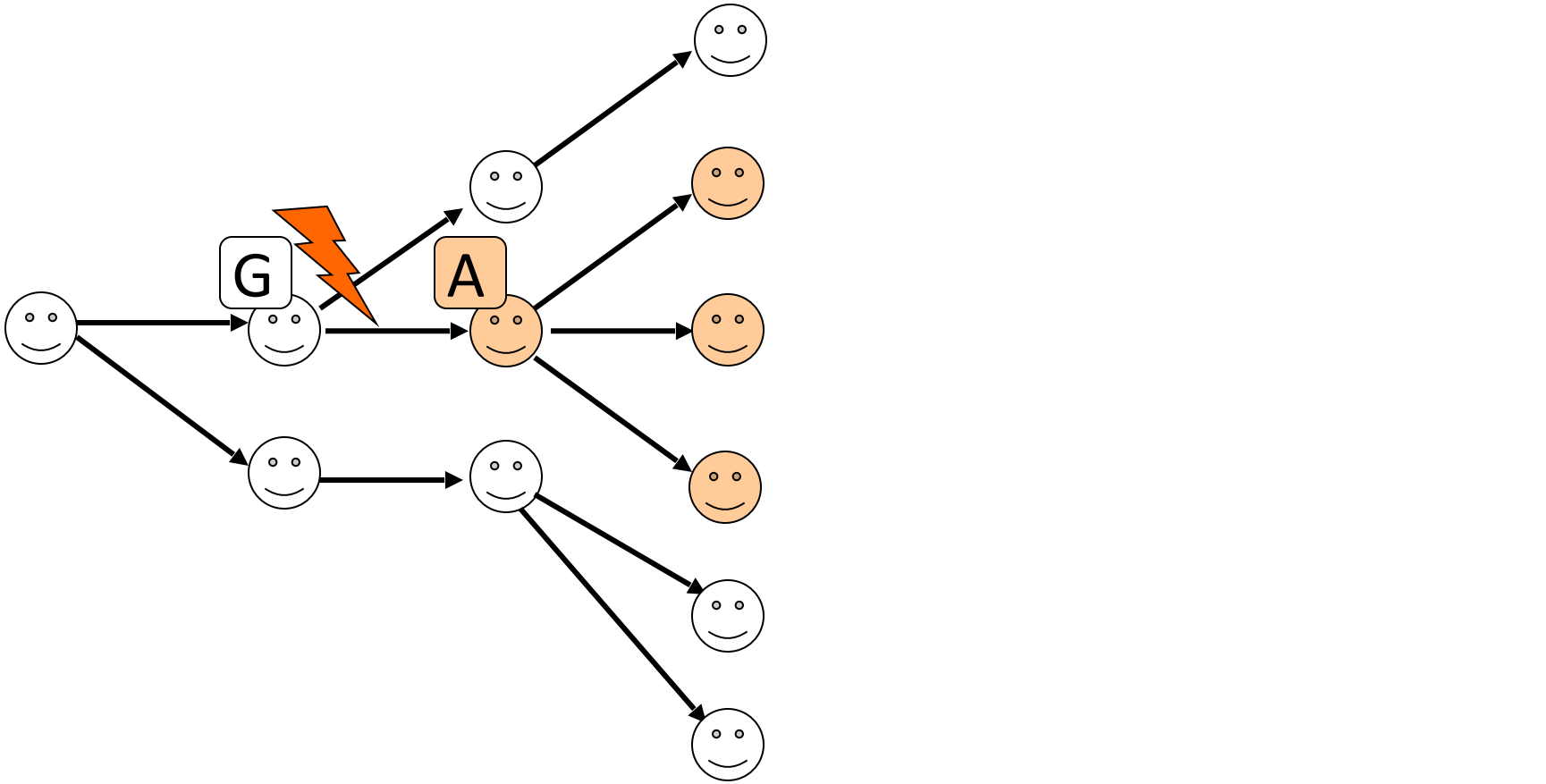


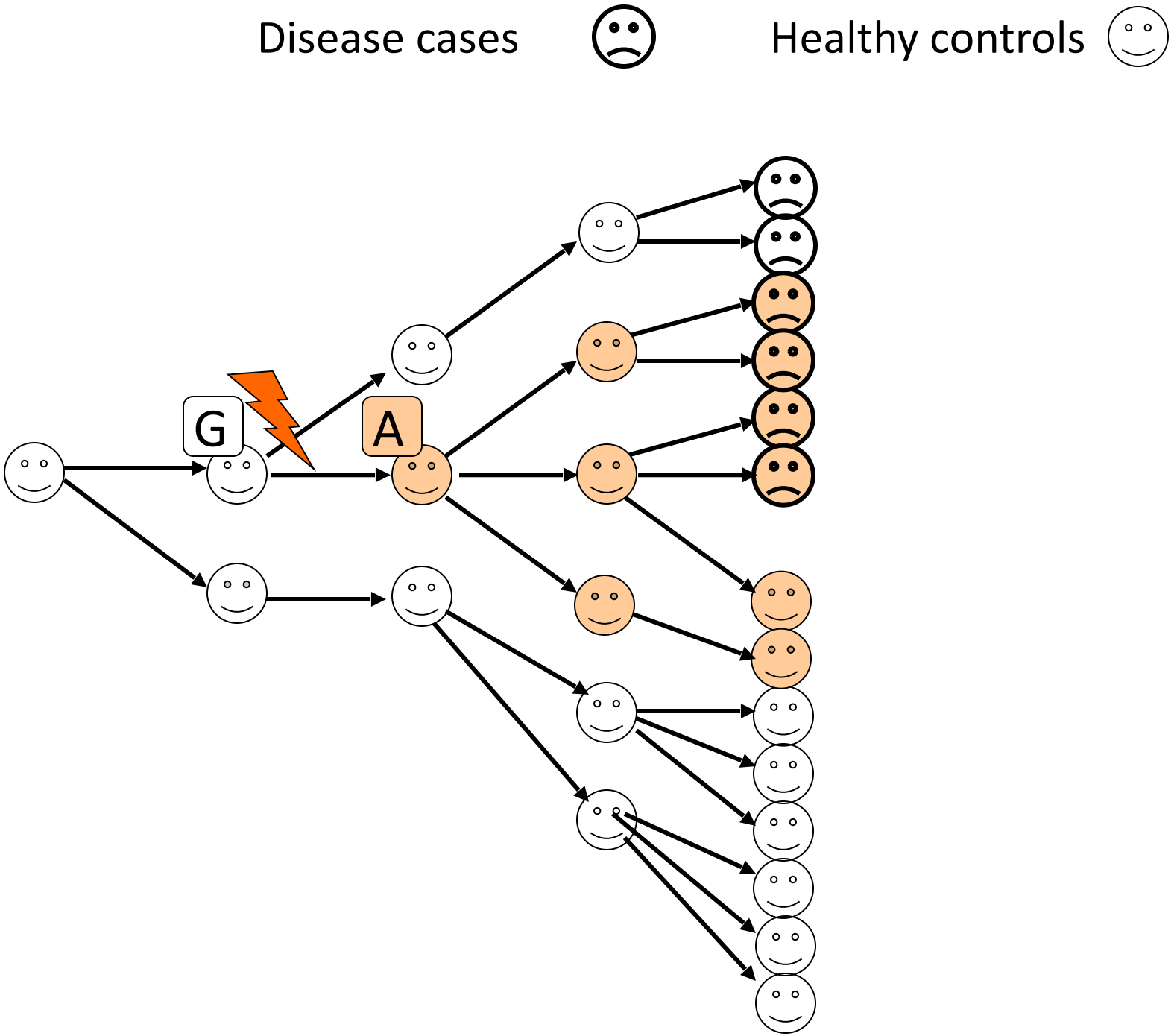
Figure 1 | **Examples of estimates of heritabilities of morphological and fitness traits.** Where possible, the estimates of heritability were taken from Reviews, and are the mean across a number of studies. The examples show that, on average, heritability estimates are larger for morphological traits than for fitness-related traits, and that heritability tends to be larger in better environments when compared with poorer environments.

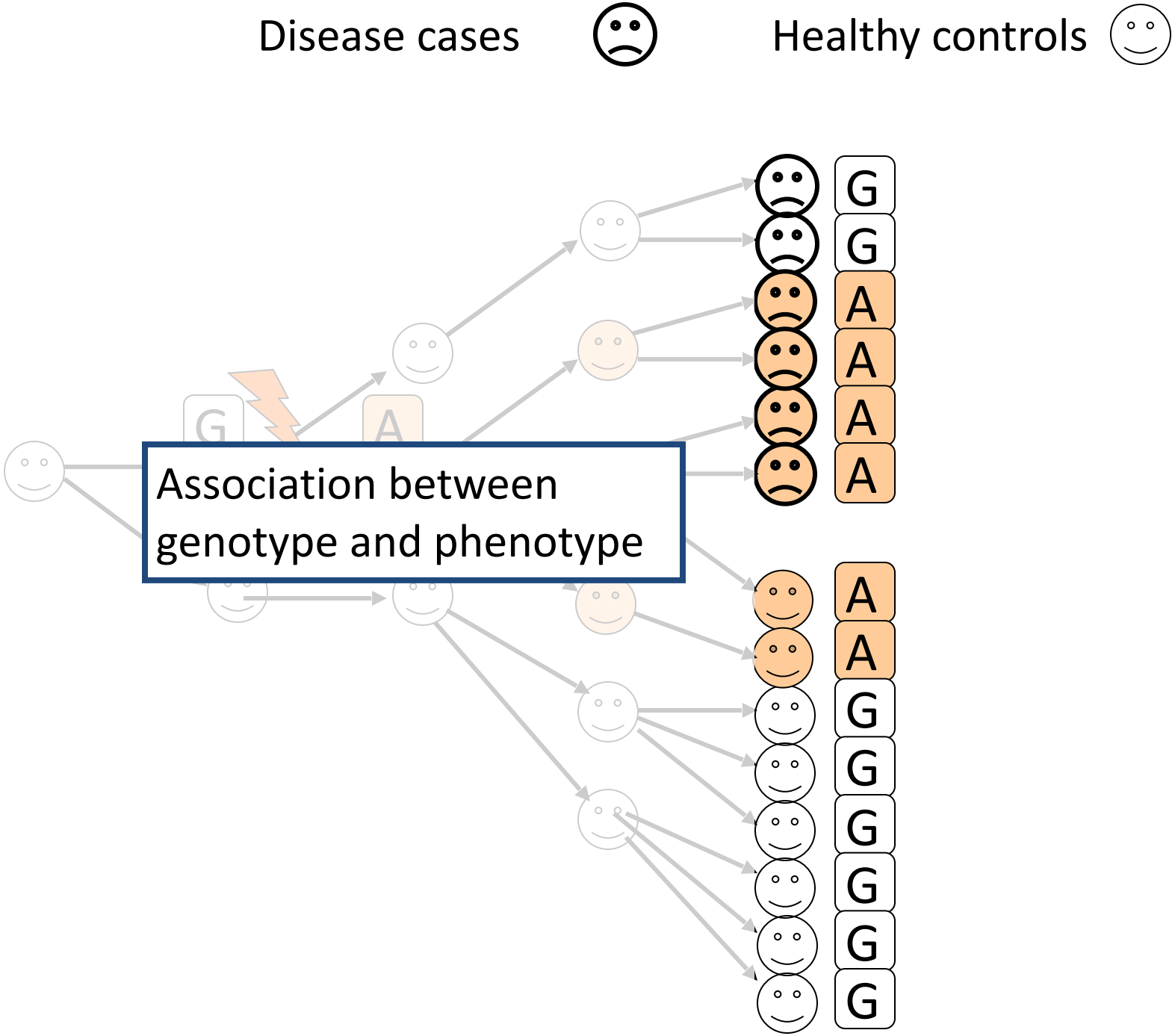
h^2 from Visscher et al. 2008

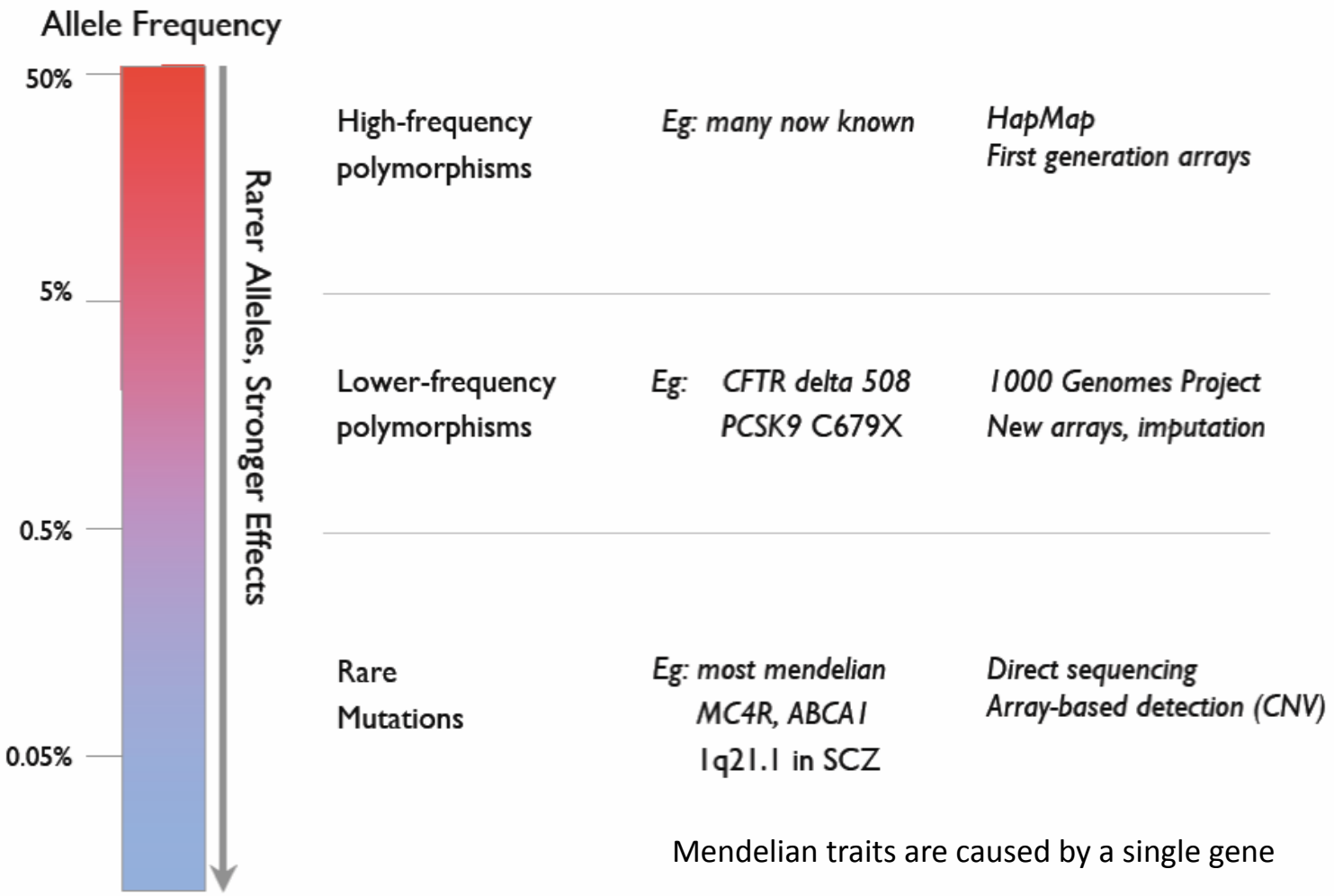
Part 2 - Genome-wide association study (GWAS)

nimation: Itsik Pe’er, Columbia




Disease cases
Healthy controls
G
A


Disease cases
Healthy controls
G
A
Association between genotype and phenotype
G
G
A
A
A
A
A
G
G
G
G
G
G



Mendelian traits are caused by a single gene

Slide courtesy of David Altshuler, HMS/Broad

Contingency Tables – χ^2 test

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	a	b	a+b
T	c	d	c+d
Total Alleles	a+c	b+d	a+b+c+d

$$E_1 = \frac{(a+b)(a+c)}{(a+b+c+d)} \qquad X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Df = (2 rows-1)x(2 columns-1) = 1

AMD=Age-related Macular Degeneration

SNP rs1061170

1238 individuals with AMD and 934 controls

2172 individuals / 4333 alleles

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	1522 (a)	670 (b)	2192
T	954 (c)	1198 (d)	2152
Total Alleles	2476	1868	4344

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

$X^2 = 279$ Df = (2 rows-1)x(2 columns-1) = 1

P-value = 1.2×10^{-62}

Contingency Tables – Fisher's Exact Test

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	a	b	a+b
T	c	d	c+d
Total Alleles	a+c	b+d	a+b+c+d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Sum all probabilities for observed and all more extreme values with same marginal totals to compute probability of null hypothesis

SNP rs1061170

1238 individuals with AMD and 934 controls

2172 individuals / 4333 alleles

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	1522 (a)	670 (b)	2192
T	954 (c)	1198 (d)	2152
Total Alleles	2476	1868	4344

$$p(a,b,c,d) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

$$p-value = \sum_{i=0}^{670} p(1522+i, 670-i, 954-i, 1198+i)$$

Does the affected or control group exhibit Population Stratification?

- Population stratification is when subpopulations exhibit allelic variation because of ancestry
- Can cause false positives in an association study if there are SNP differences in the case and control population structures
- Control for this artifact by testing control SNPs for general elevation in χ^2 distribution between cases and controls

Linkage Disequilibrium (LD) between two loci L1 and L2 in gametes

At locus L1

p_A probability L1 is A

q_a probability L1 is a

At locus L2

p_B probability L2 is B

q_b probability L2 is b

	L2 B	L2 b
L1 A	$P_{AB} = p_A p_B + D$	$P_{Ab} = p_A q_b - D$
L1 a	$P_{aB} = q_a p_B - D$	$P_{ab} = q_a q_b + D$

D = Measure of linkage disequilibrium

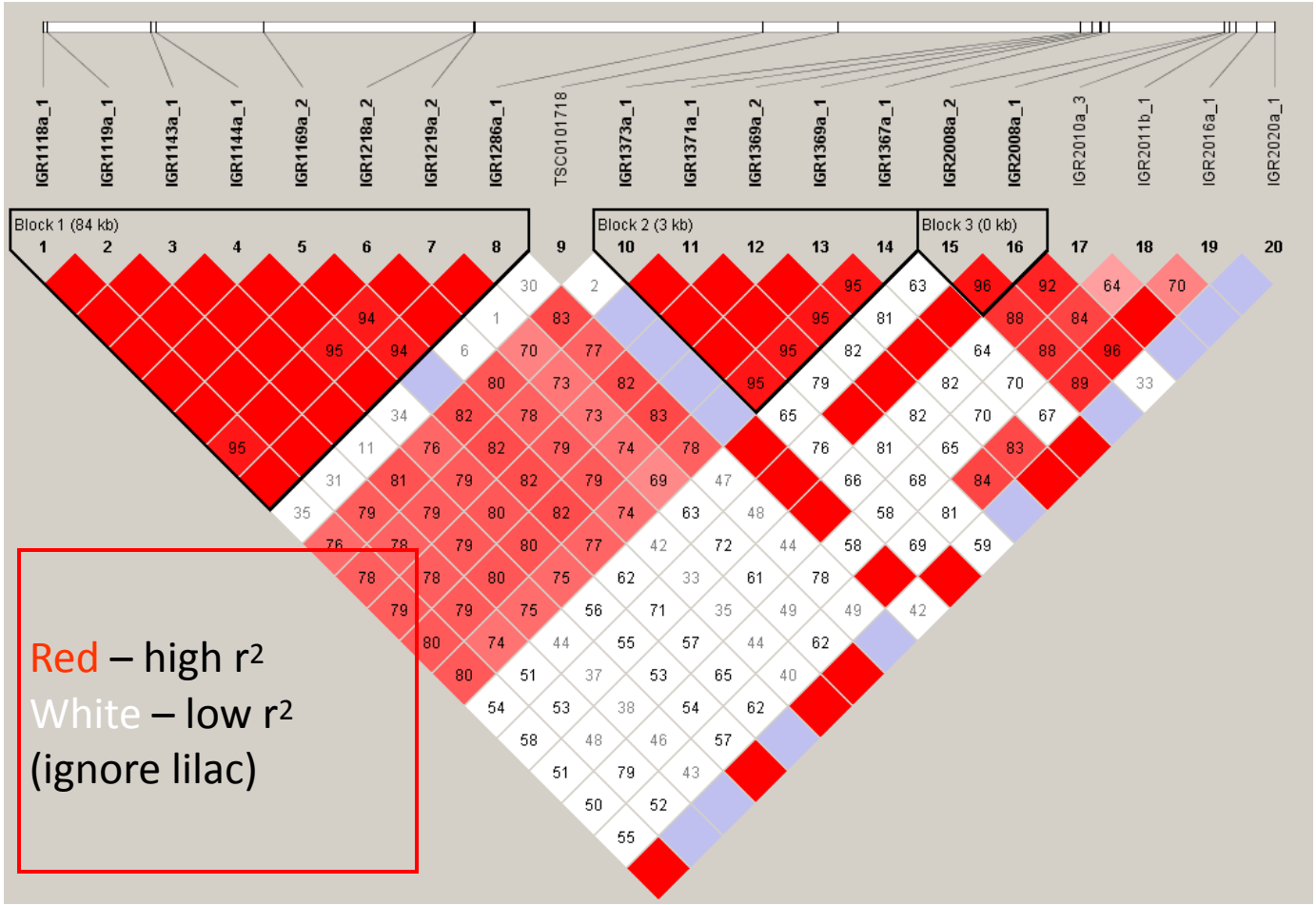
= 0 when L1 and L2 are in equilibrium

$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$

$r^2 = D^2 / (p_A q_a p_B q_b)$ Example $r^2 = .69$ when P_{AB} and $P_{ab} = .3$, P_{Ab} and $P_{aB} = .2$

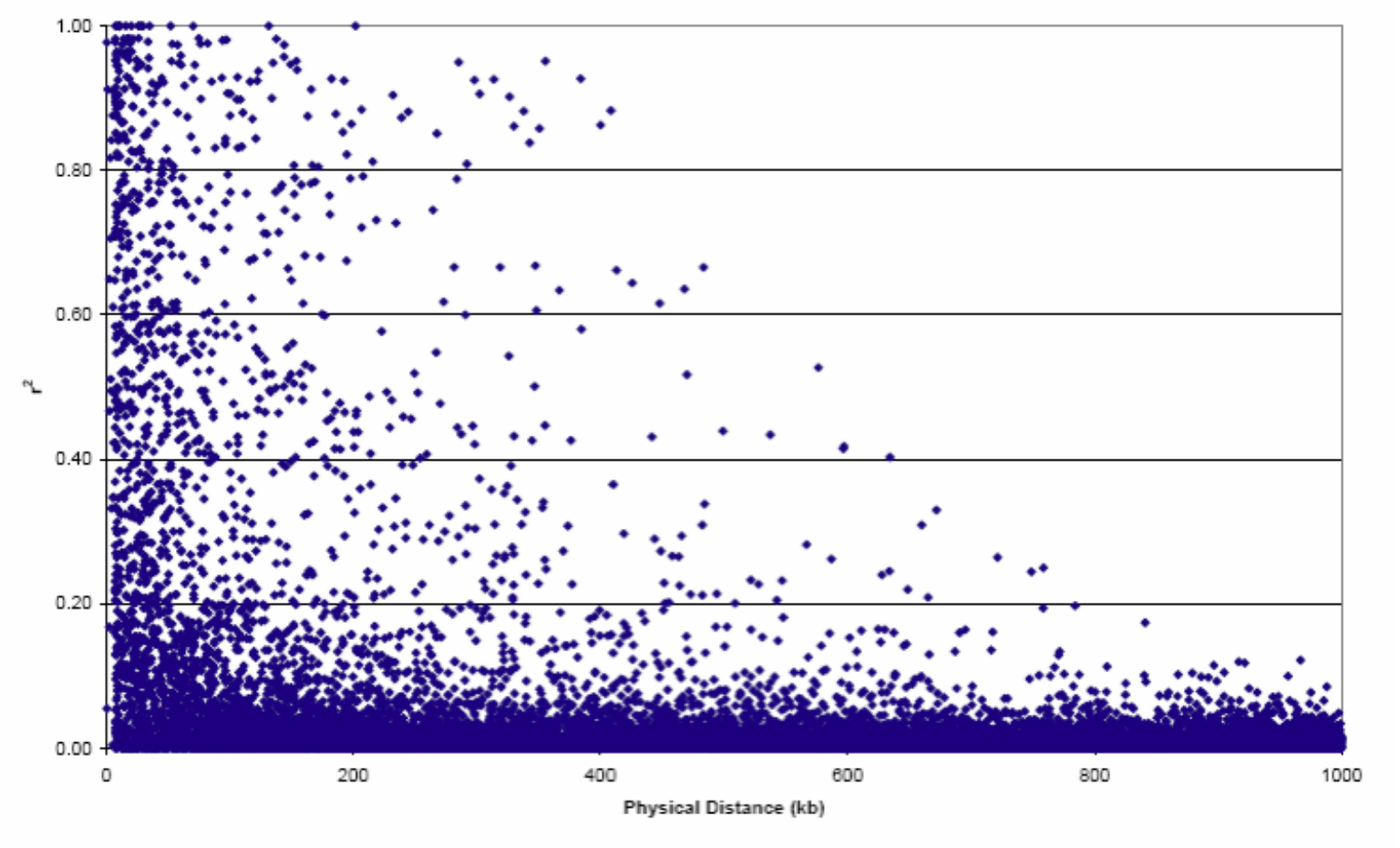
r is [0,1] and is the correlation coefficient between allelic states in L1 and L2

LD organizes the genome into haplotype blocks

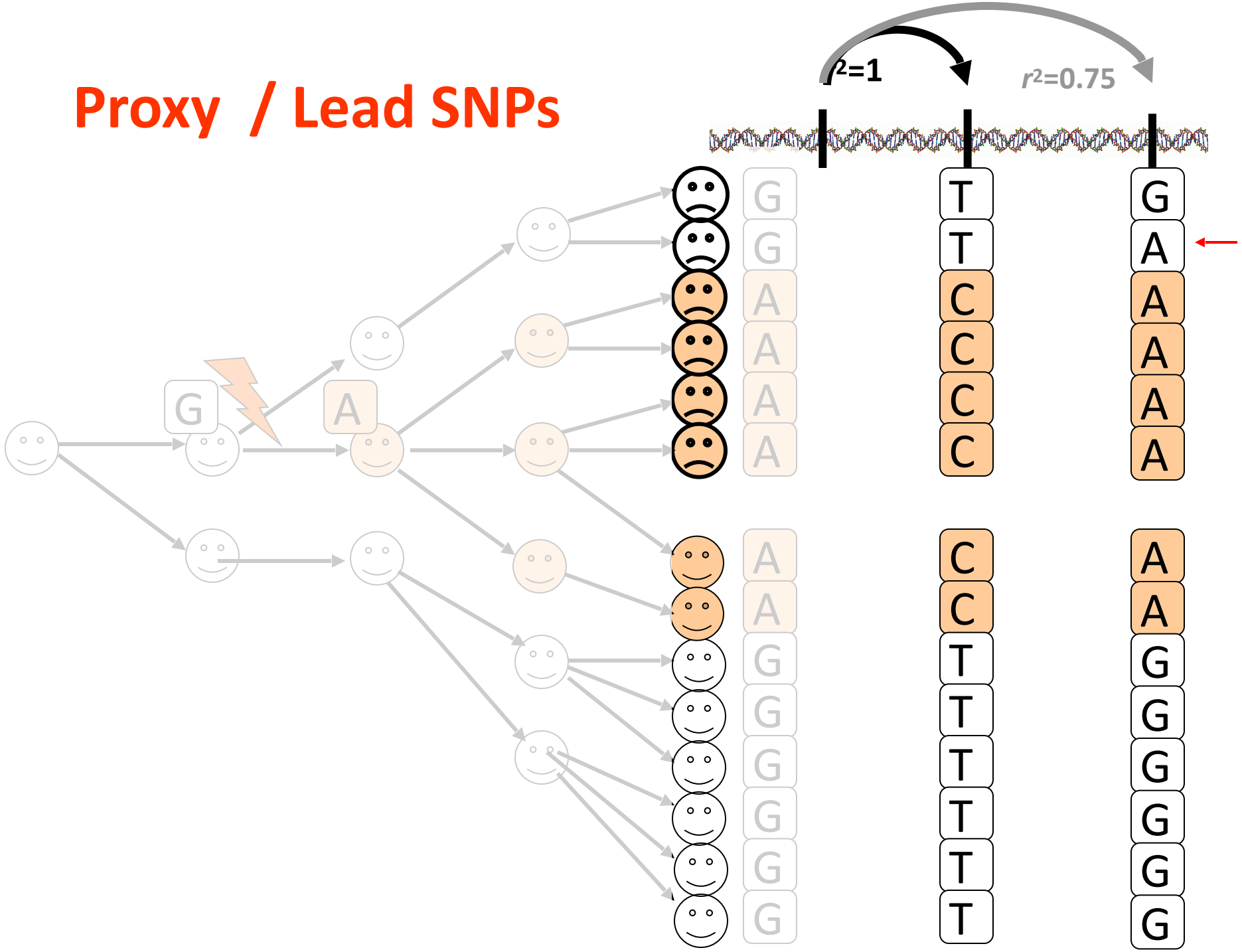


Human genome 5q31 region (associated with Inflammatory Bowel Disease)

r^2 from human chromosome 22



Proxy / Lead SNPs



Part 3 - Predicting functional variants using machine learning

Why do we want to interpret the functional consequence of a variant?

- Narrow the pool of candidate variants in GWAS to lower the statistical burden from multi-hypothesis testing
- Identify the causal variant from all that are in strong linkage disequilibrium with the lead variant
- Understand the pathological mechanism of causal variants

Two general approaches

- Annotation-based
- Ab initio from DNA sequence

Annotation-based prediction

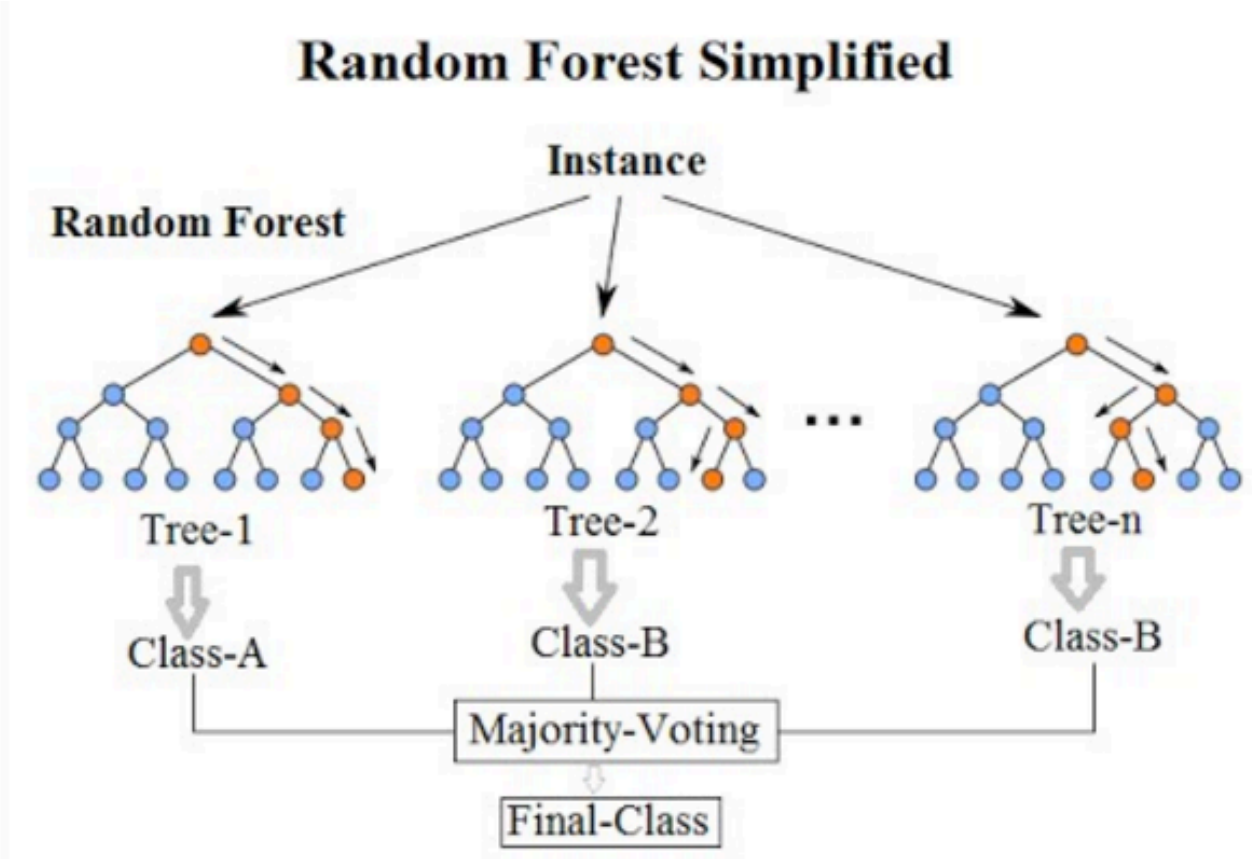
- Use functional annotations in the proximal regions as features
- The functional annotations are from functional assays (eg. ChIP-seq), gene annotation, evolutionary annotation, etc.
- Representative methods
  - CADD (Kircher et al. Nature Genetics 2014)
  - GWAVA (Ritchie et al. Nature Methods 2014)

GWAVA

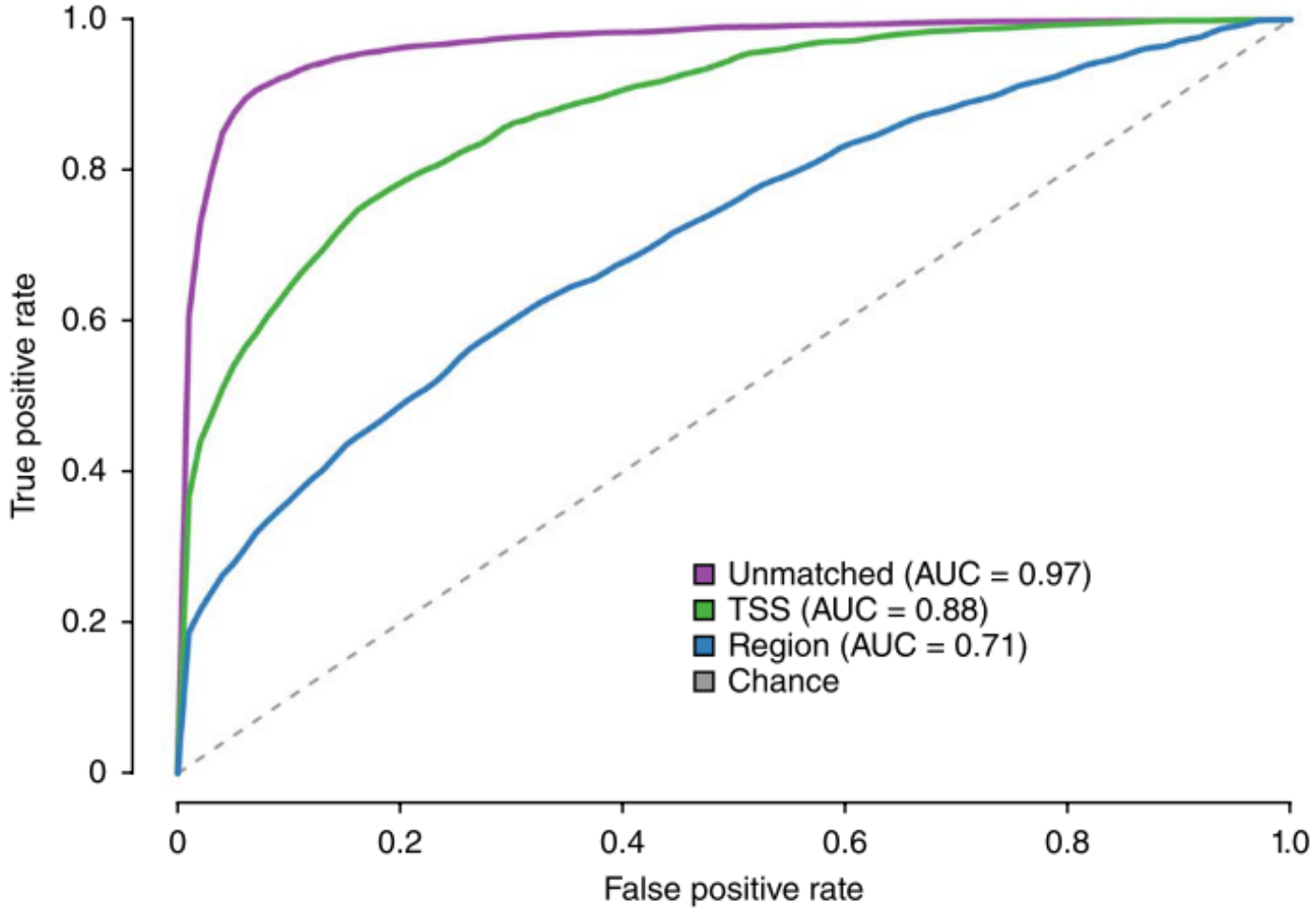
- Annotations included in the model
  - Open chromatin (DNase I Hypersensitivity)
  - Transcription factor binding (ChIP-seq peak calls for 124 TFs)
  - Histone modification (ChIP-seq peak calls for 12 HM)
  - RNA polymerase binding (ChIP-seq peak calls)
  - CpG island
  - Genome segmentation (predicted by Seaway and ChromHMM)
  - Conservation (genomic evolutionary rate profiling for mammals)
  - Human variation (Mean heterozygosity and mean derived allele frequency)
  - Genic context (distance to the nearest TSS/splice site/gene region)
  - Sequence context (G+C content, is CpG, in repeat sequence)

GWAVA

- Random forest as the computational method



GWAVA can classify regulatory mutations from controls



Limitations of annotation-based methods

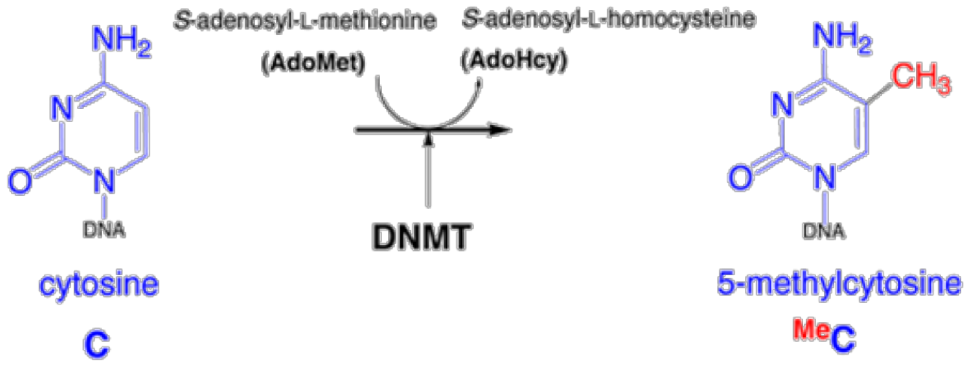
- The majority of the annotations are specific to the proximal region, not the variant itself
  - False positives arise if a variant resides in an important region, but has no functional consequence
  - For a new patient, all the functional assays have to be redone to be able to make accurate predictions

Ab initio prediction from DNA sequence

- Train a computational model that predicts functional signal from DNA sequence
- For a variant, produce the *predicted* functional signal of the proximal region for both the reference and the alternate allele of the variant
- Train a classifier to predict functional variant from the *predicted* functional change to the proximal region

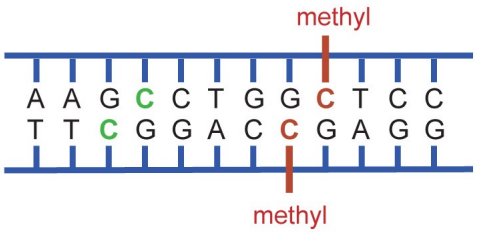
Example: CpGenie (Zeng et al. 2017) for DNA methylation prediction

- Establishment and maintenance of tissue-specific expression profiles
- X-chromosome inactivation
- Genomic imprinting
- Transposable element silencing
- Cell differentiation
- Inflammatory processes

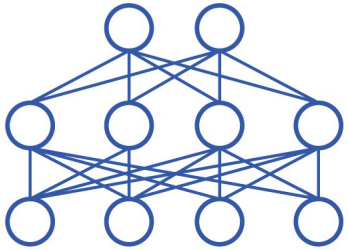


Sequence-based methylation model helps in various analysis of non-coding sequence variants

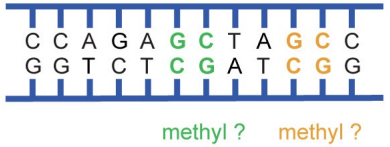
High throughput DNA methylation sequencing data



Modeling by deep convolutional neural network (CpGenie)



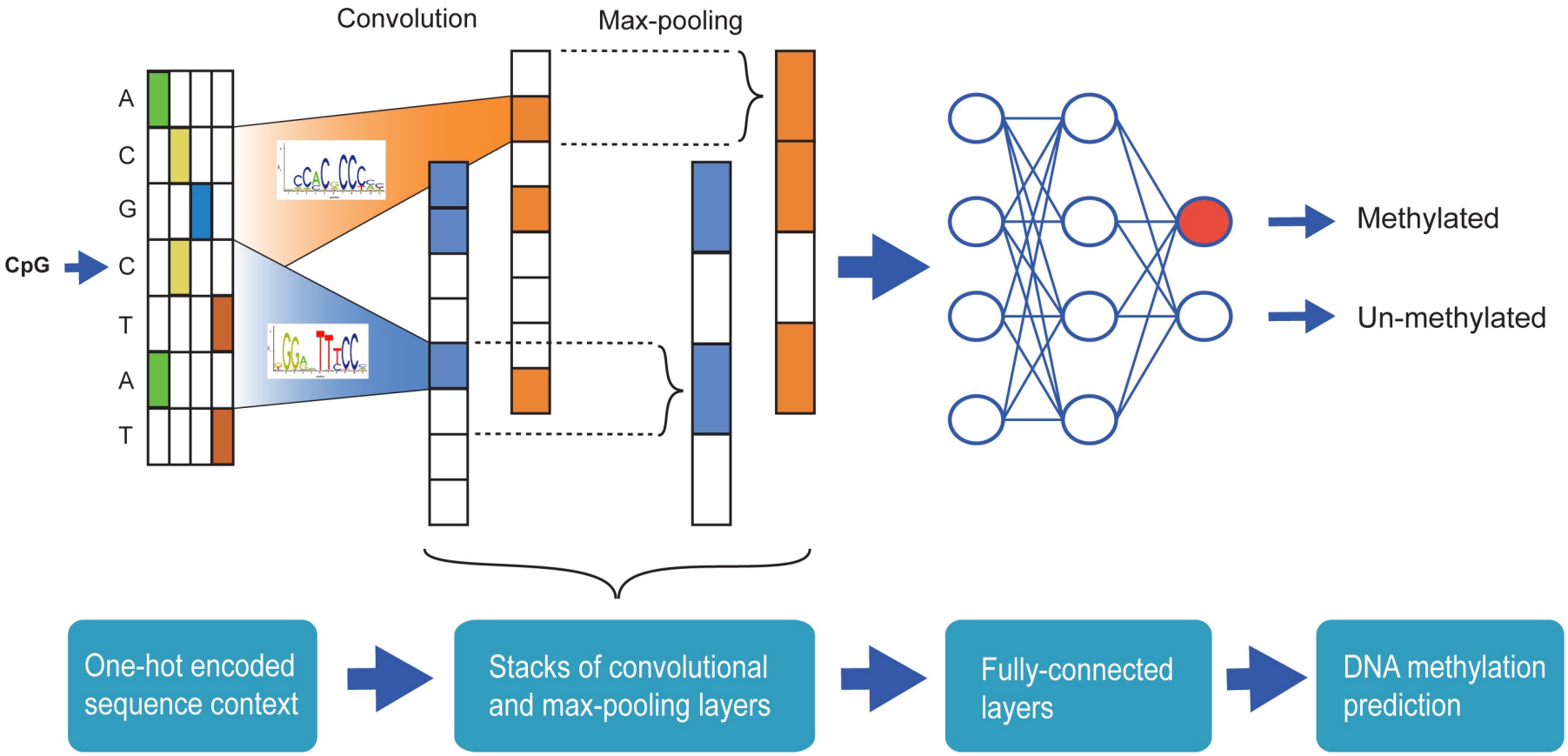
In sillicon DNA methylation prediction at CpG resolution



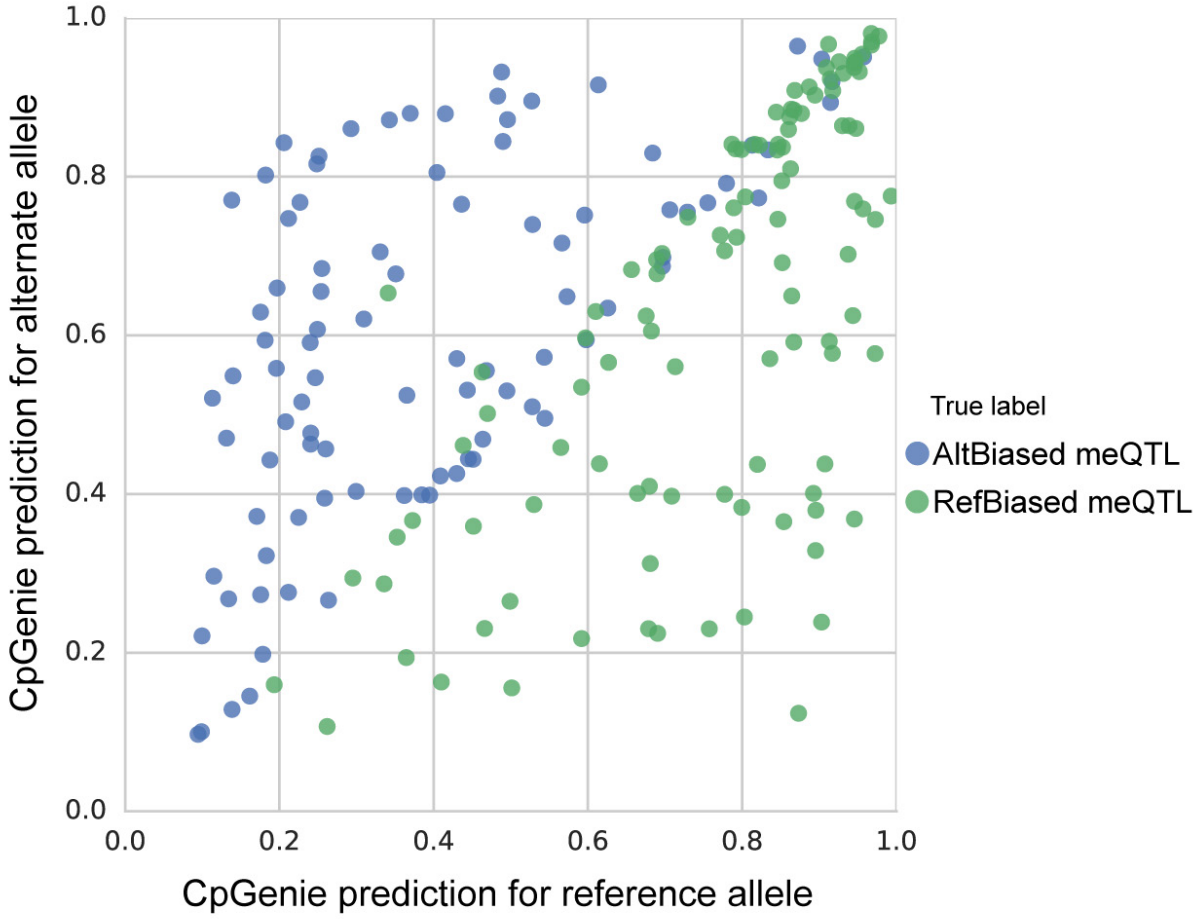
Interpreting functional non-coding mutation

Prioritizing disease-associated mutation

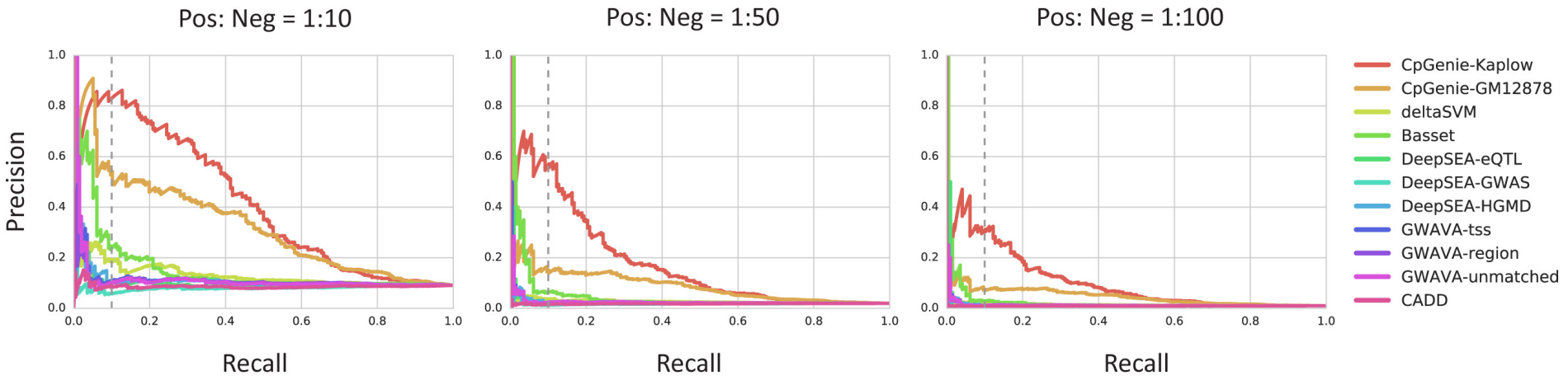
Convolutional neural network for predicting methylation level of a CpG site



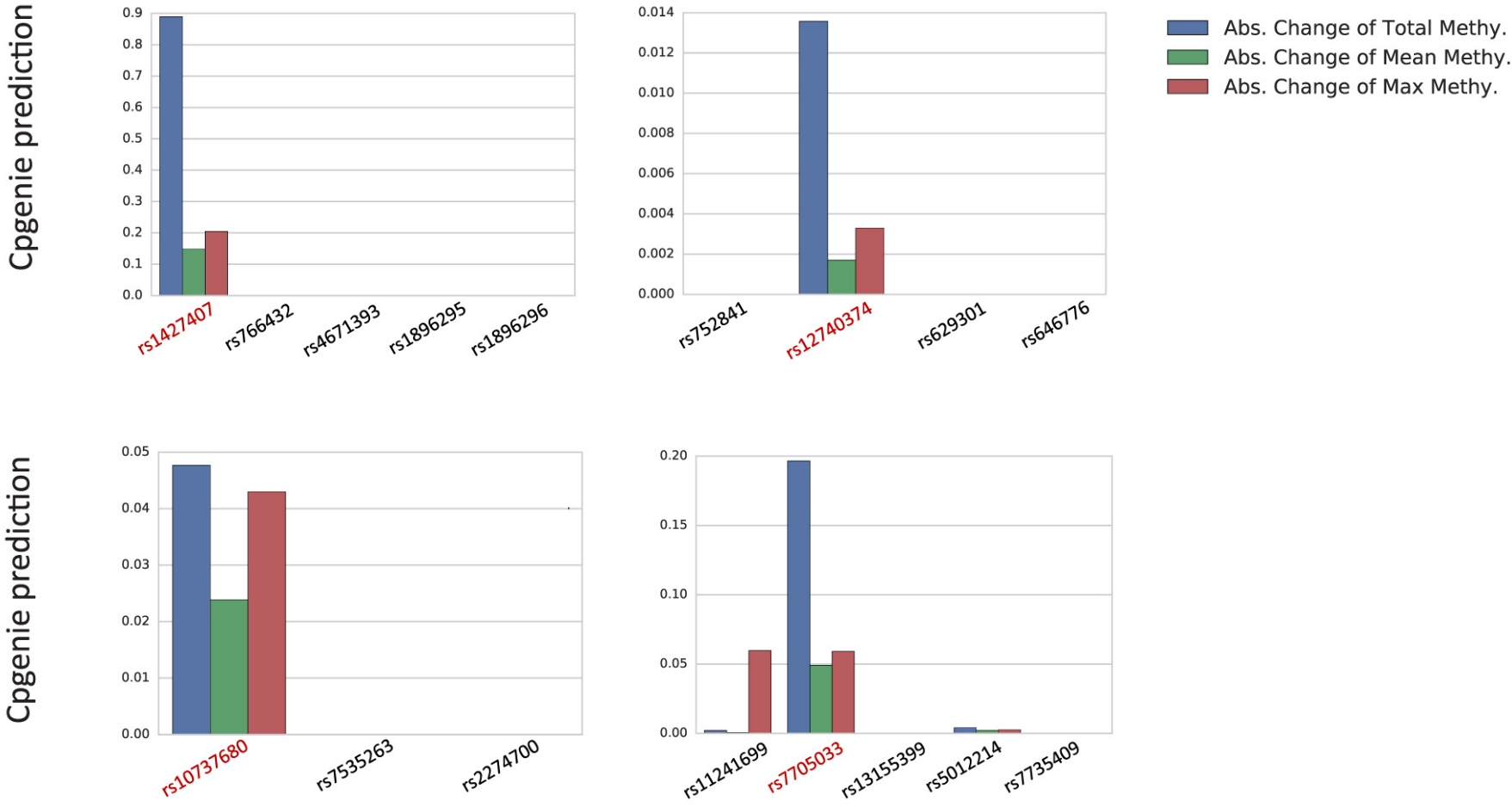
CpGenie accurately predicts the direction of allelic-change of DNA methylation



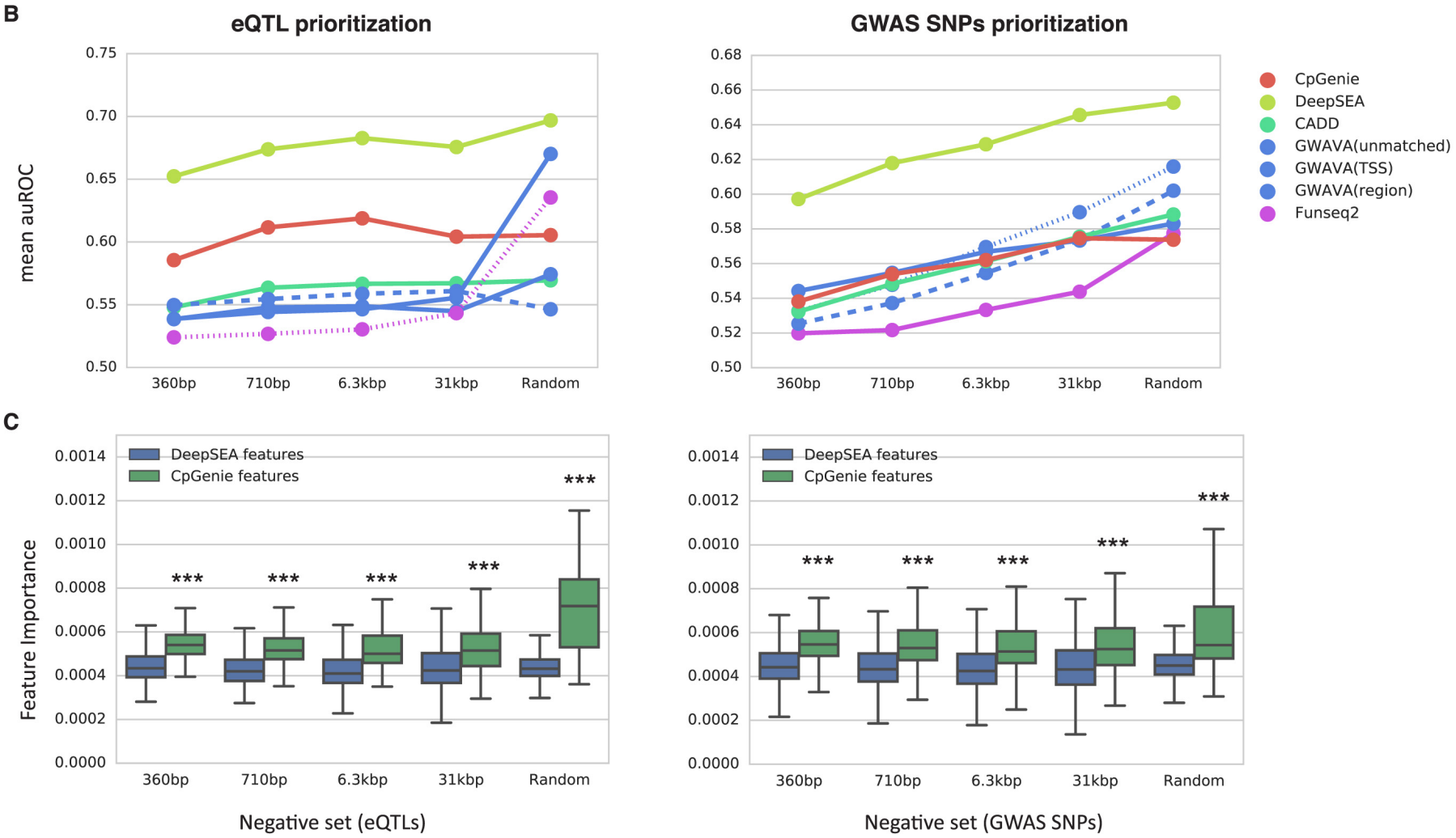
CpGenie outperforms existing methods in classifying meQTL from non-meQTLs



Predicted change in DNA methylation helps identify causal variants from those in strong LD



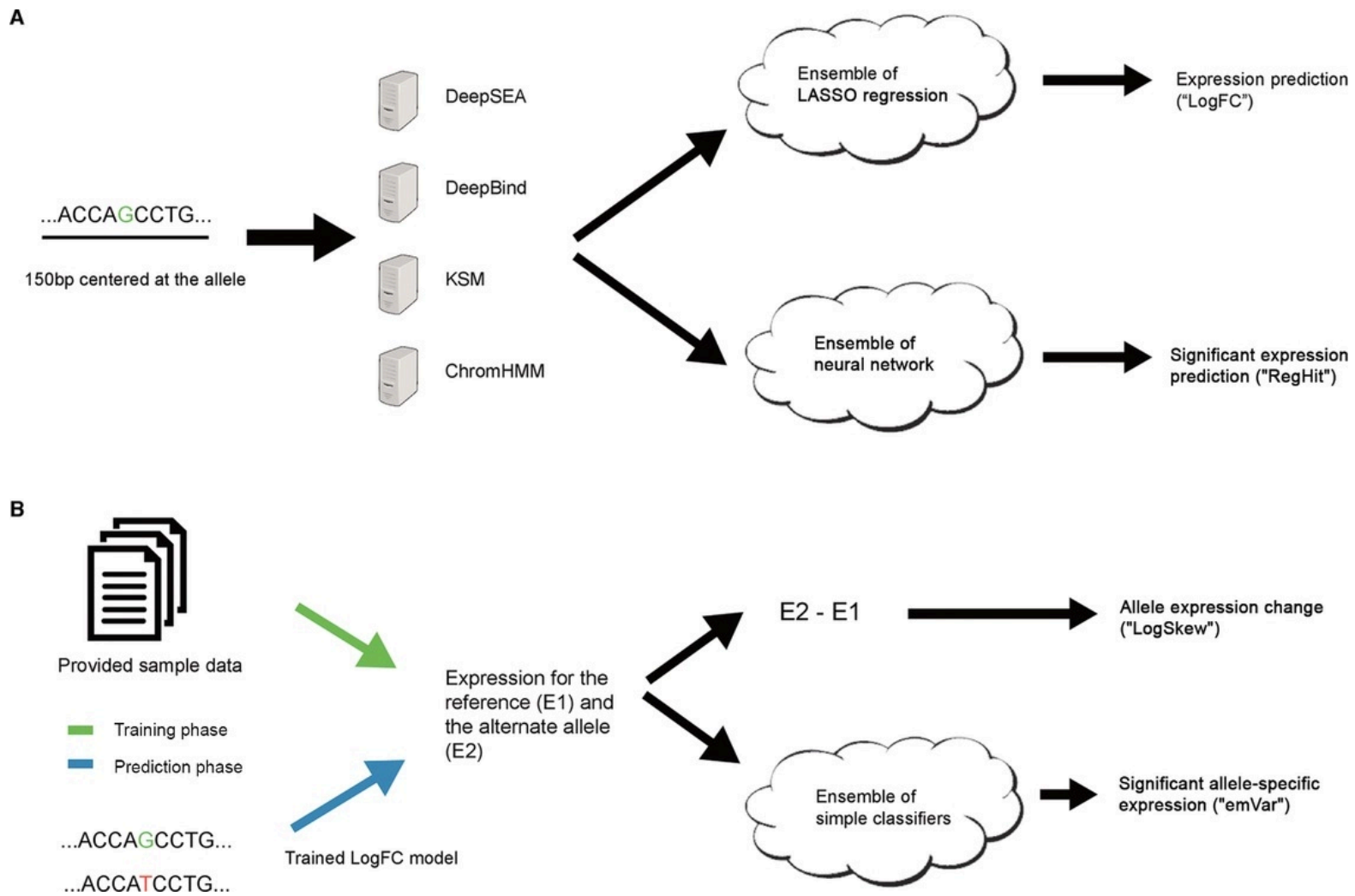
CpGenie's methylation predictions serve as important features for eQTL and GWAS SNP prioritization



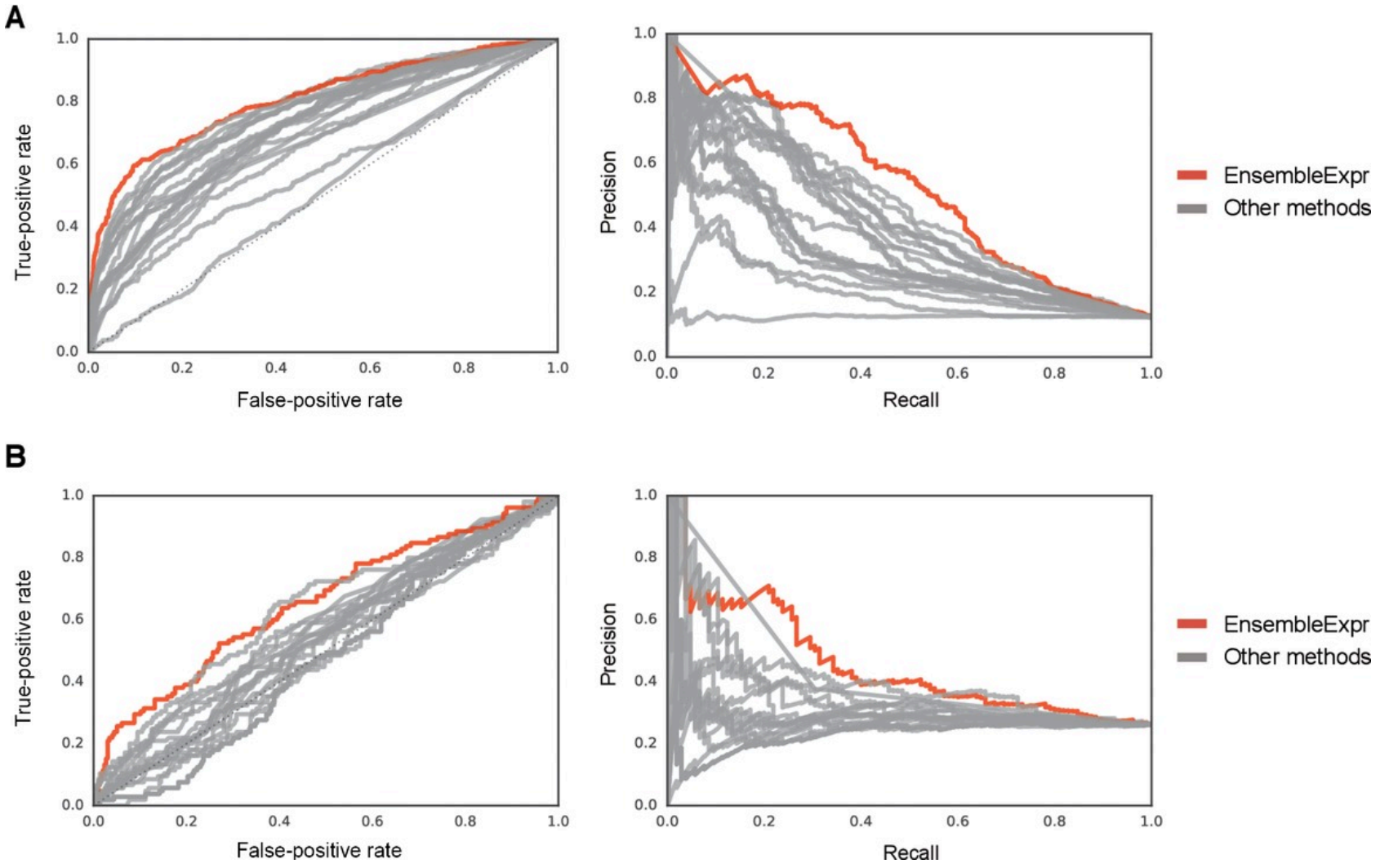
Example2: EnsembleExpr (Zeng et al. 2017) for eQTL prediction

- Dataset
  - Expression level of the reference and alternate allele of 3000+ genetic variants measured by massive parallel reporter assay (MPRA)
- Task
  - Expression prediction
    - Predict the expression level of each DNA sequence example
    - Predict which ones are significant (as defined by a quantile compared to the population)
  - eQTL prediction
    - Predict the expression difference between the ref. and alt. allele of a variant
    - Predict which variants show significant difference in expression between alleles (eQTL)

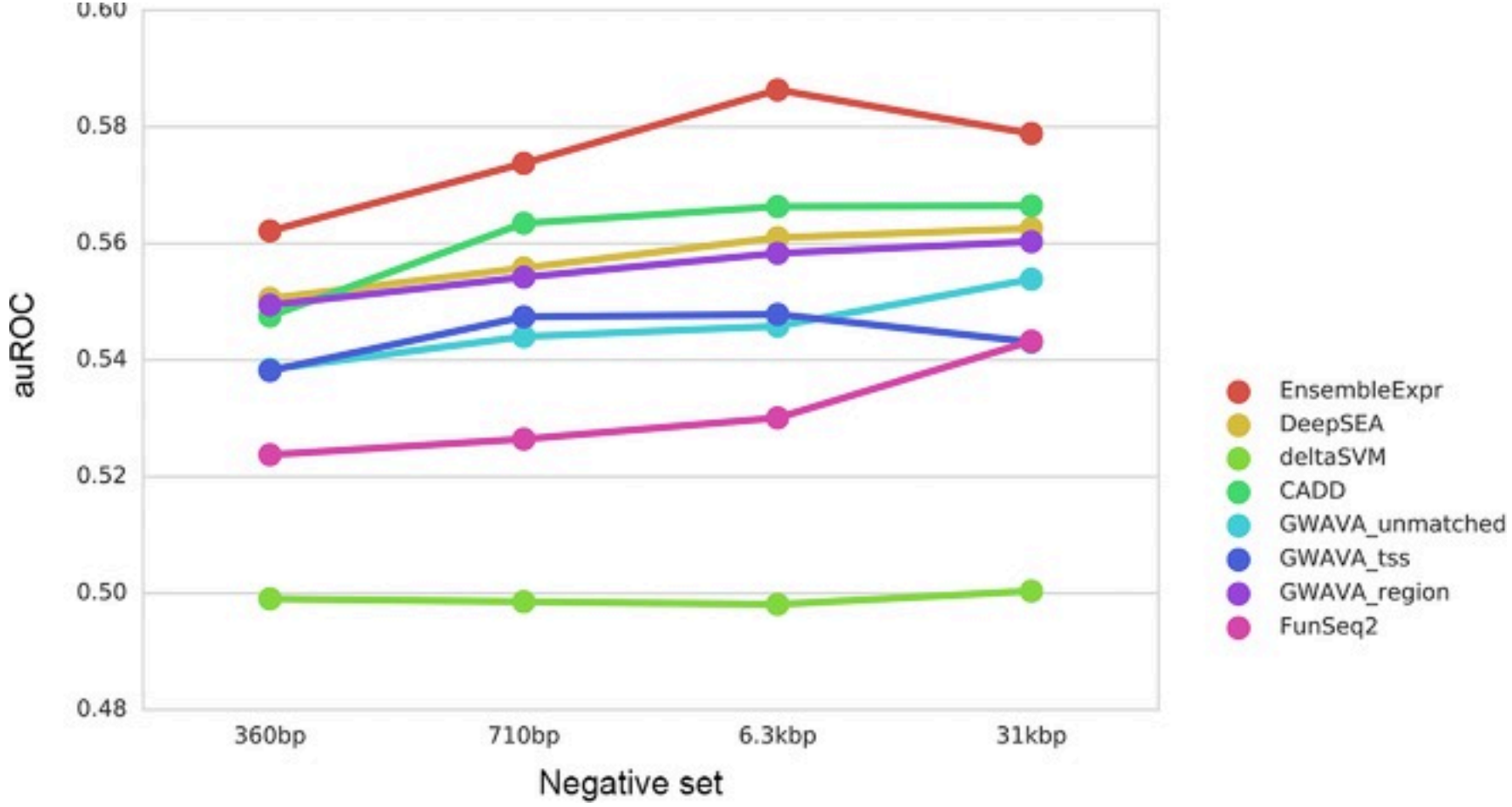
Example2: EnsembleExpr (Zeng et al. 2017) for eQTL prediction



EnsembleExpr achieved the best performance in the CAGI4 eQTL challenge



EnsembleExpr outperformed existing methods for eQTL prediction



Summary of today's lecture

- Fundamentals of heritability
- Genome-wide association study (GWAS)
- Predicting functional variants using machine learning