# Predicting phenotypes from genotypes

**PHENOTYPE** → the set of observable characteristics of an individual resulting from the interaction of its **genotype** with the **environment**.

## MENDELIAN TRAIT

Mendelian genetics put forward the concept of dominant and recessive traits, where the **phenotypes are controlled by single genes**. These traits are known as monogenic or Mendelian traits (es. **cystic fibrosis**, etc.).

**LINKAGE ANALYSIS IN PEDIGREES**

## COMPLEX TRAIT

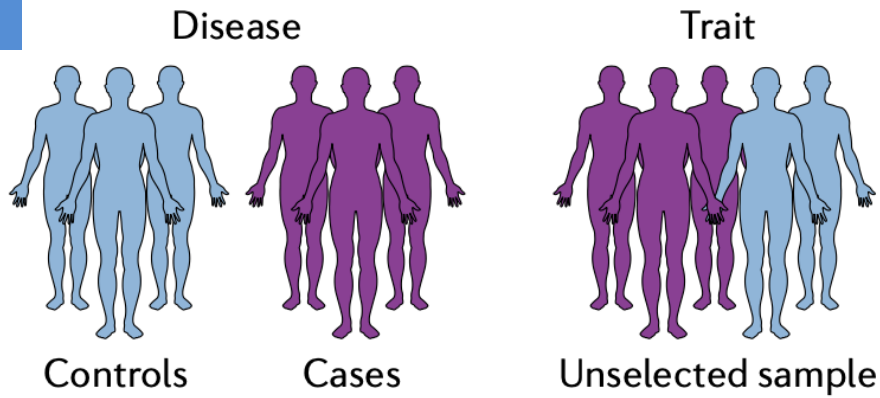
There are features or traits in human genetics which **are controlled by multiple genes** and whose inheritance does not follow the rules of Mendelian genetics. Such traits are known as complex traits (es. autism, cardiac disease, cancer, diabetes, Alzheimer's disease, and asthma). Complex traits are believed to result from **gene-gene** and **gene-environment interactions**, genetic heterogeneity, and potentially other yet unknown reasons.

**GWAS**

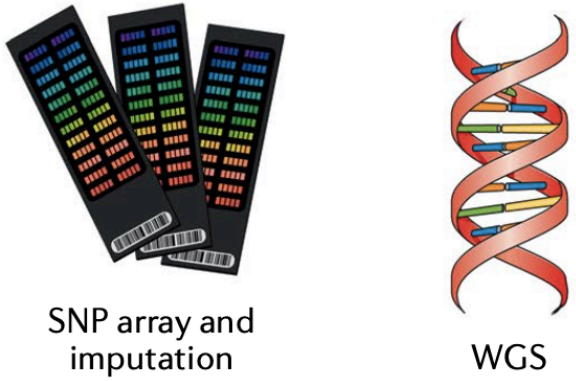
## Genome-wide association studies

Analysis of genetic markers (SNPs) to find places in the genome associated with differences in the trait of interest.

**1**



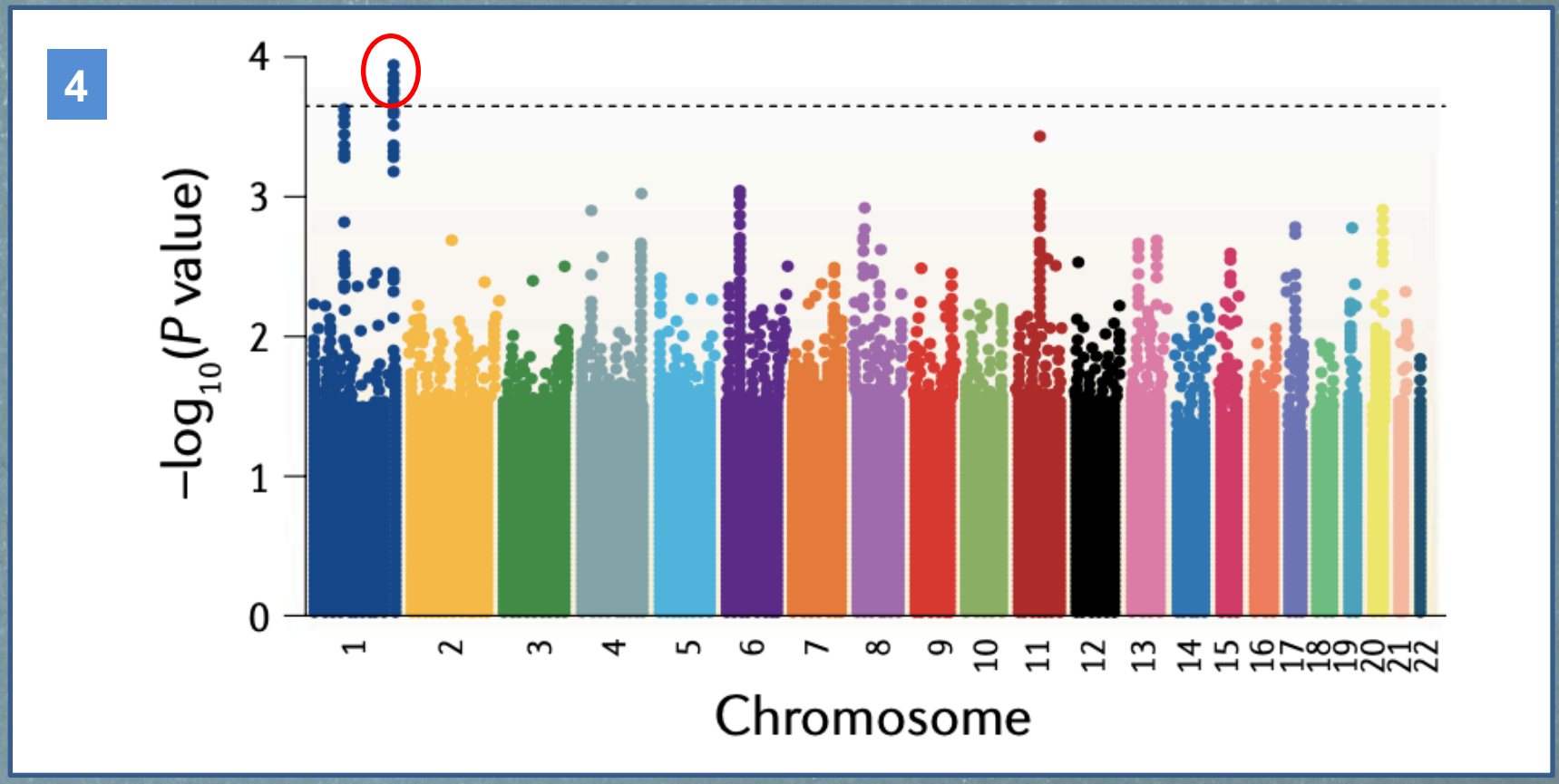

**2**




**3**

## Genome-wide association studies

Analysis of genetic markers (SNPs) to find places in the genome associated with differences in the trait of interest.




Manhattan plot: each point represents a SNP, plotted with its p-value (on a $-\log_{10}$ scale) as a function of genomic position (x-axis).

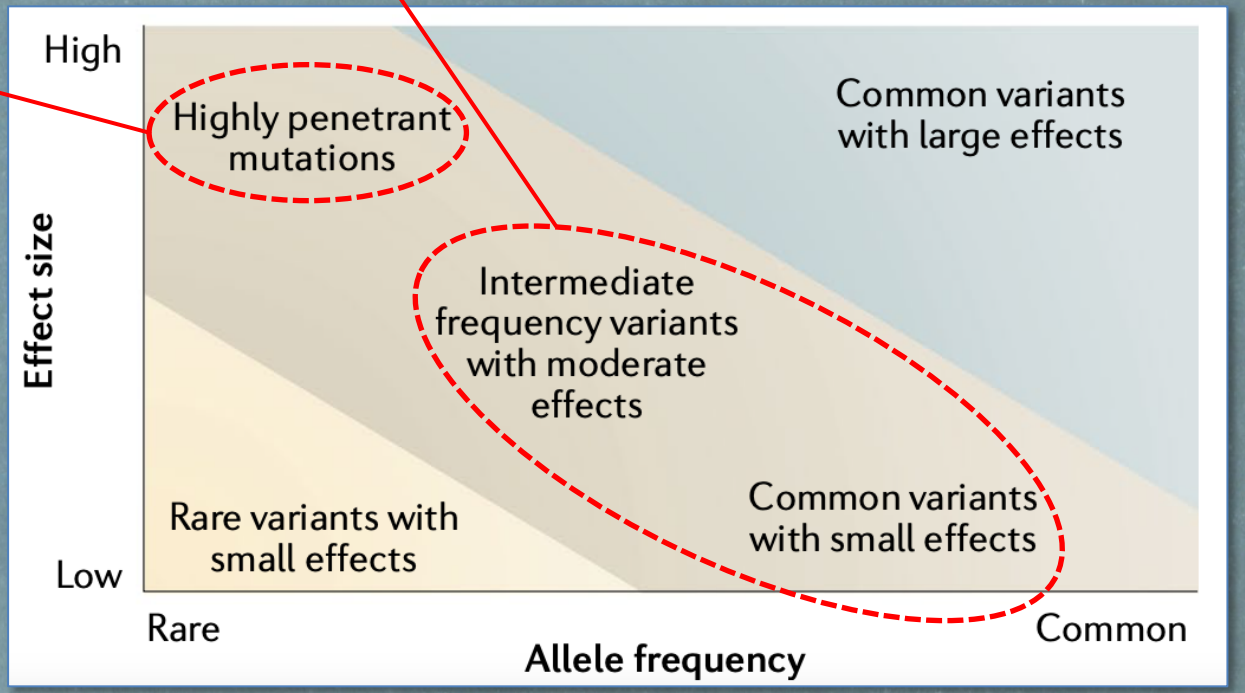
Initial excitement was somewhat tempered by the realization that GWAS loci typically have **small effect sizes** and explain only a modest proportion of trait heritability.

**Effect sizes**

The magnitudes of the effect of alleles on phenotypic values.

age-related macular degeneration (AMD): 99% of the individuals with the highest-risk genotypes (including at CFH) had AMD (*Chen, W et al., 2010*).

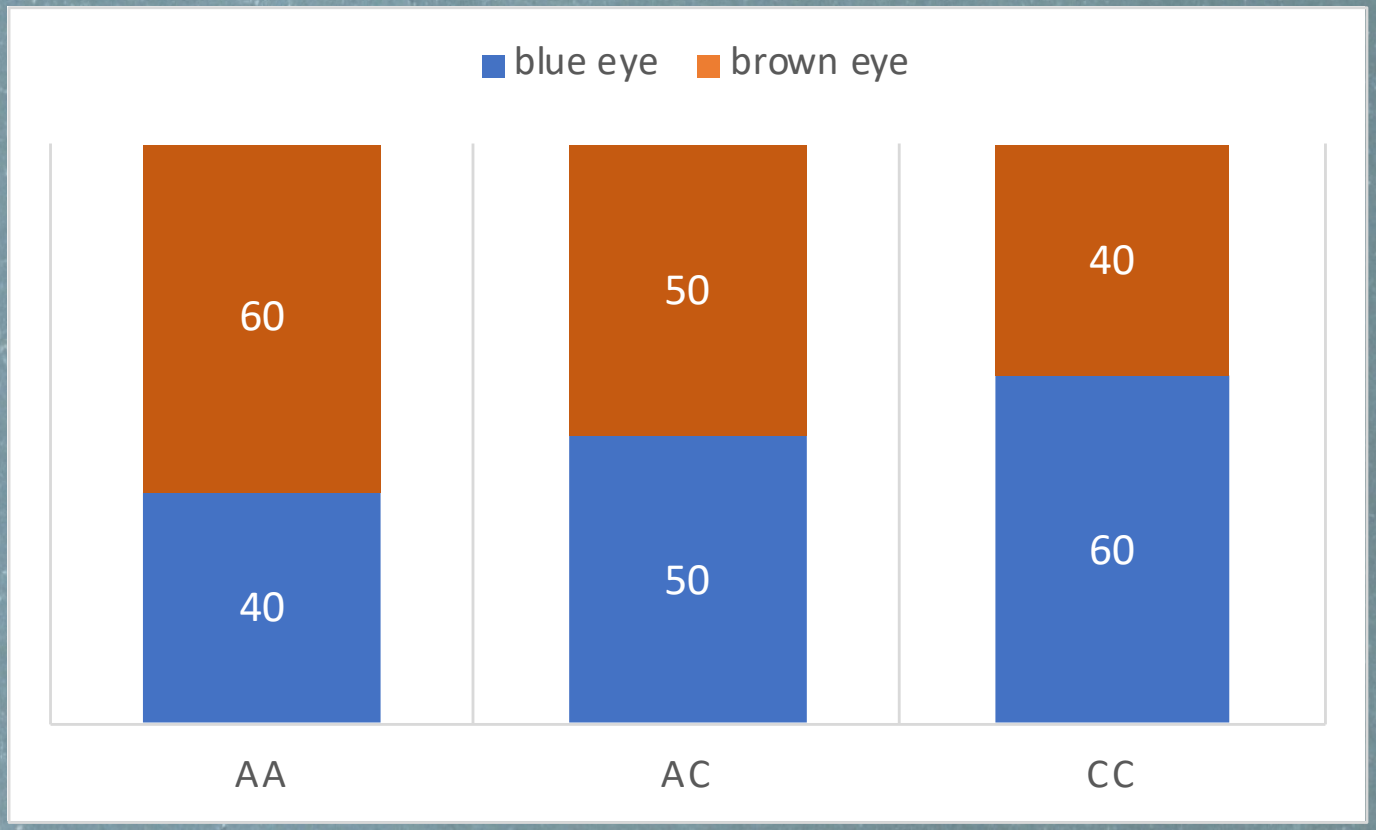
The **modest proportion of heritability** explained and **the small effect sizes** of GWAS-identified Single Nucleotide Variations (SNVs) **limit their clinical predictive value**.
**BUT THIS DEPENDS ON THE TRAIT WE ARE EXAMINING!**




*Genetic variants exist along a spectrum of allele frequencies and effect sizes. Most risk variants identified by GWAS lie within the two diagonal lines. Rare variants with small effect sizes are difficult to identify using GWAS, and common variants with large effects are unusual for common complex diseases (Tam et al., 2019).*

Can we use the genetic markers found in GWAS for genetic risk prediction?

**ASSOCIATION DOES NOT NECESSARILY MEAN PREDICTIVENESS**




This is because association studies tests each variant separately and try to answer the question: how is a single feature related to the outcome? As in the case of ancestral informative markers, we need to combine different variants in order to try to increase the prediction accuracy.

How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

| | | Condition | |
|---|---|---|---|
| | | Present | Absent |
| Test | Positive | True positive (a) | False positives (b) |
| | Negative | False negatives (c) | True negatives (d) |

The **sensitivity** of a test is the proportion of people who test positive (a) among all those who actually have the disease (a+c).

$$SENSITIVITY = \frac{True\ positive}{True\ positive + False\ negative}$$

The **specificity** of a test is the proportion of people who test negative(d) among all those who actually do not have that disease (d+b).

$$SPECIFICITY = \frac{True\ negative}{True\ negative + False\ positive}$$

How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

| | | Condition | |
|---|---|---|---|
| | | Present | Absent |
| Test | Positive | True positive (a) | False positives (b) |
| | Negative | False negatives (c) | True negatives (d) |

**Positive predictive value (PPV)** is the probability that following a positive test result, that individual will truly have that specific disease.

$$PPV = \frac{True\ positive}{True\ positive + False\ positive}$$

**Negative predictive value (NPV)** is the probability that following a negative test result, that individual will truly not have that specific disease.

$$NPV = \frac{True\ negative}{True\ negative + False\ negative}$$

How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

| | | Condition | |
|---|---|---|---|
| | | Present | Absent |
| Test | Positive | True positive (a) | False positives (b) |
| | Negative | False negatives (c) | True negatives (d) |

**Likelihood ratio positive (LR+)** is the probability of a person who has the disease testing positive divided by the probability of a person who does not have the disease testing positive.

$$\text{LR+} = \frac{p(T+|D+)}{p(T+|D-)} = \frac{Sensitivity}{1-Specificity}$$

**Likelihood ratio negative (LR-)** the probability of a person who has the disease testing negative divided by the probability of a person who does not have the disease testing negative.

$$\text{LR-} = \frac{p(T-|D+)}{p(T-|D-)} = \frac{1-Sensitivity}{Specificity}$$

## Bayes theorem and the LR

**Post-Test Odds = Pre-test Odds * LR.**

For example, let's say a patient returning from a vacation to Rio presents with a fever and joint pain. Past data tells you that 70% of patients in your practice who return from Rio with a fever and joint pain have Zika. The blood test result is positive, with a likelihood ratio of 6. To calculate the probability the patient has Zika:

Step 1: Convert the pre-test probability to odds:
0.7 / (1 – 0.7) = 2.33.

Step 2: Use the formula to convert pre-test to post-test odds:
Post-Test Odds = Pre-test Odds * LR = 2.33 * 6 = 13.98.

Step 3: Convert the odds in Step 2 back to probability:
(13.98) / (1 + 13.98) = 0.93.

**There is a 93% chance the patient has Zika.**

*From https://www.statisticshowto.com/likelihood-ratio/*

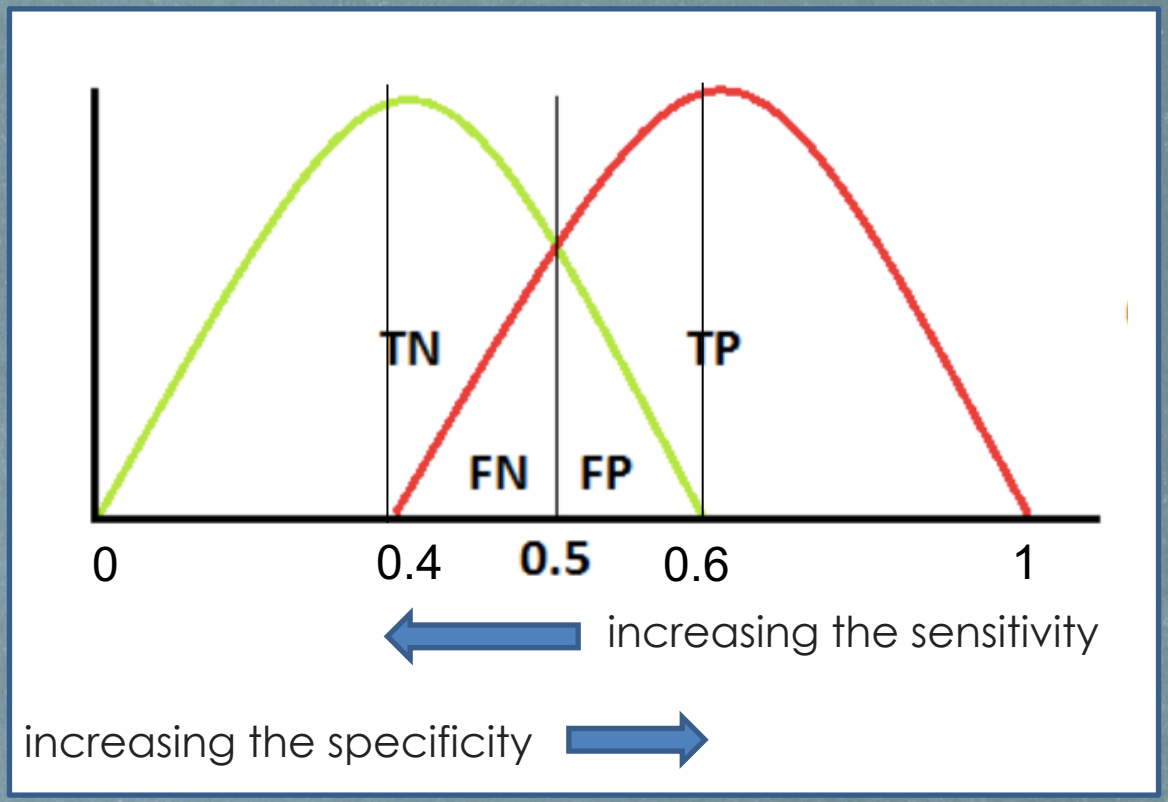
How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

| | | Condition | |
|---|---|---|---|
| | | Present | Absent |
| Test | Positive | True positive (a) | False positives (b) |
| | Negative | False negatives (c) | True negatives (d) |

Sensitivity = a/ (a+c)
Specificity = d/ (b+c)
Positive predictive value = a/ (a+b)
Negative predictive value = d/ c+d)
Positive likelihood ratio: a/ (a+c) / b (b+d) or (Sensitivity/ 1- Specificity)
Negative likelihood ratio: c/ (a+c) / d (b+d) or (1 - Sensitivity/Specificity)

How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

## ROC Curve and AUC




How can we evaluate the performances of the test at different thresholds?

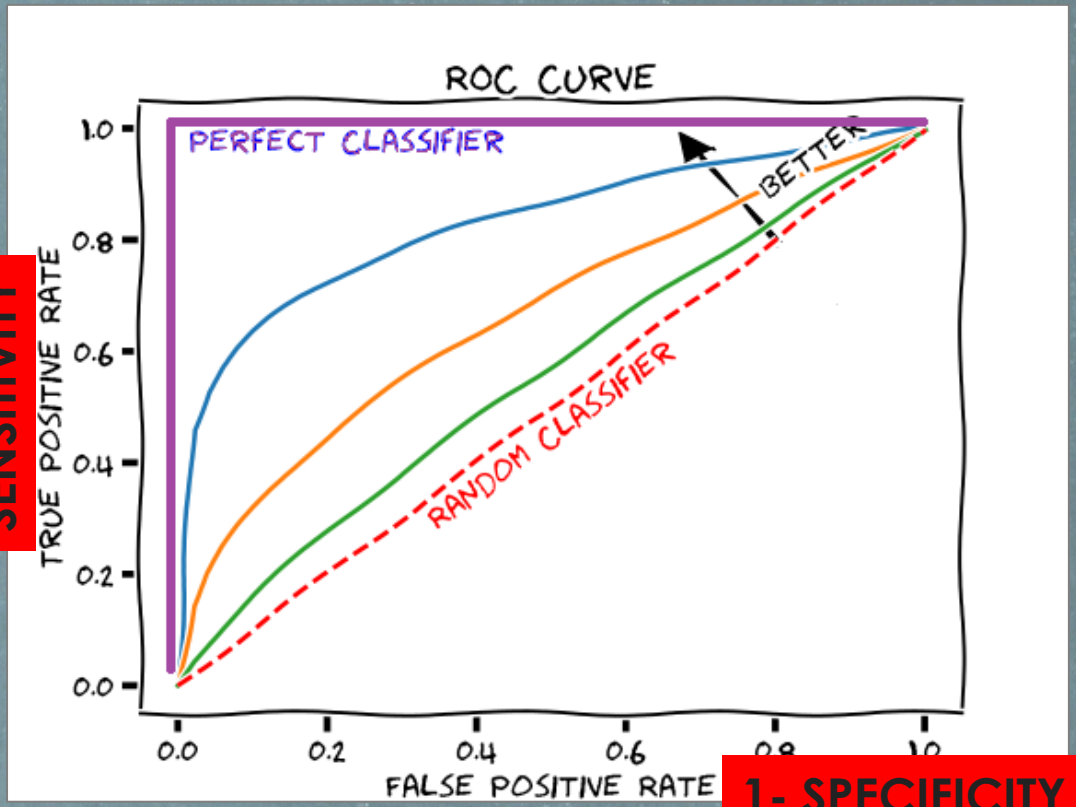
An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

## ROC Curve and AUC

On a perfect, or "gold-standard," test, all sub- jects who will develop the disease have a positive test result (sensitivity = 1), and all subjects who will not develop the disease have a negative result (specificity = 1). For composite tests, positive and negative results are defined by a cutoff value of the disease probability. The sensitivity and specificity of a composite test may differ, depending on the cutoff probability that is chosen. Therefore, the sensitivity and specificity are calculated for each possible cutoff value of the probability and plotted in a so-called **receiver-operating–characteristic (ROC) curve**. The **area under the ROC curve (AUC)** indicates the discriminative ability of a composite test. The discriminative ability is perfect if the AUC is 1, whereas an AUC of 0.50 indicates a total lack of discrimination.

*Janssens et al., Am. J. Hum. Genet., 2004*




| | | Condition | |
|---|---|---|---|
| | | Present | Absent |
| Test | Positive | True positive (a) | False positives (b) |
| | Negative | False negatives (c) | True negatives (d) |

How can we measure the **goodness of a prediction**, or in general, the **goodness of a test** (es., diagnostic test)?

## ROC Curve and AUC