Epidemiology

- Epidemiology is the method used to find the causes of health outcomes and diseases in populations.

- Is the study of distribution, pattern and causes of health-related disorders.

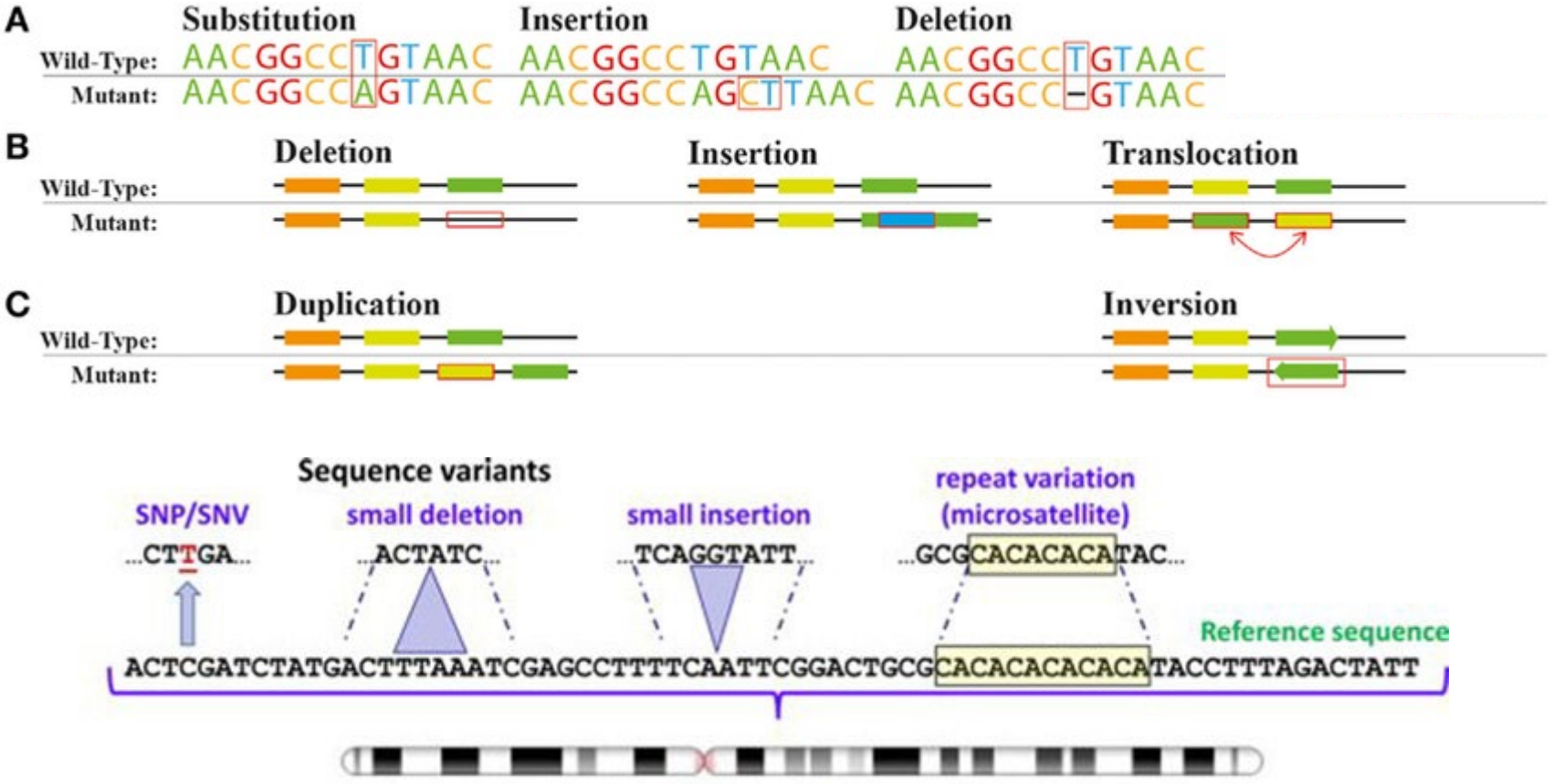
Complex Traits

- complex diseases are caused by a combination of genetic, environmental, and lifestyle factors, most of which have not yet been identified

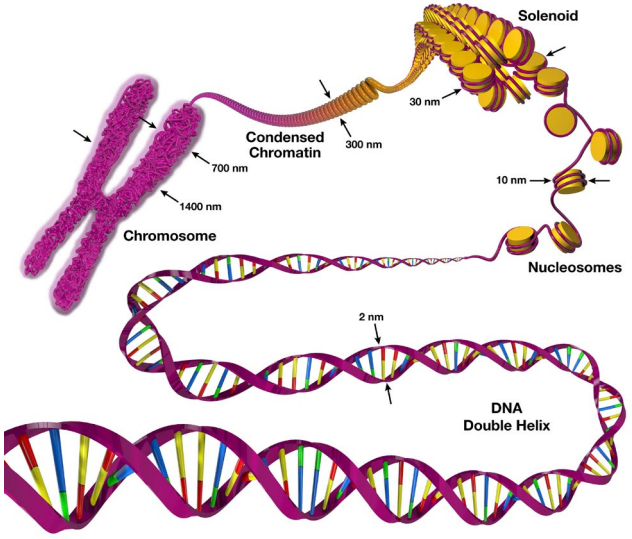
- Genetic factors represent only part of the risk associated with complex disease phenotypes

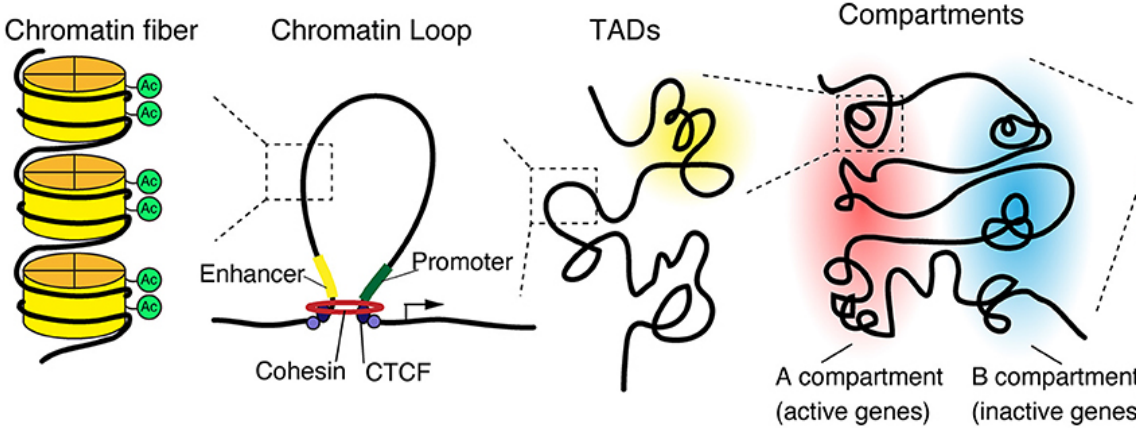
- Many polymorphysms, with small contribution, and of difficult interpretation.

Genetic Variation



Genetic variation





Single nucleotide polymorphism (SNPs)

- SNPs are point mutation in the genome
- They are identified by a rsid (e.g. rs1234)
- More than 600 million SNPs have been identified across the human genome in the world's population. A typical genome differs from the reference human genome at 4 to 5 million sites, most of which (more than 99.9%) consist of SNPs and short indels

Genes and alleles

- **Genes**: are sections of DNA that determine certain attributes or characteristics of a particular human. Genes encode for proteins that are the functional units that influence human traits.

- **Alleles**: when a gene mutation occurs, we can end up with two “versions” of the same gene. In this process, each form differs slightly in the sequence of their base DNA. These gene variants still express the same gene in a different way i.e. brown hair vs blonde hair. This difference in the versions of the same gene is known as alleles.

MAF

- Minor allele frequency (MAF) is the frequency at which the second most common allele occurs in a given population

- It tells us if a variant is common or rare in the population; variations having a frequency of greater than 1% in the population are considered common variants.

Hardy Weinberg equilibrium

- It's a model that explains genetic of populations

- It answer the question "why if an allele is dominant (e.g. brown hair) not everyone in the population has that allele?"

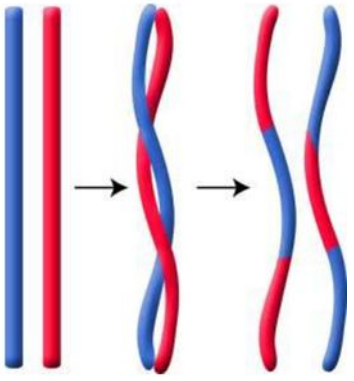
- It shows that allele frequencies remains stable through generations unless an event that destroy this equilibrium happen. This equilibrium equals to:

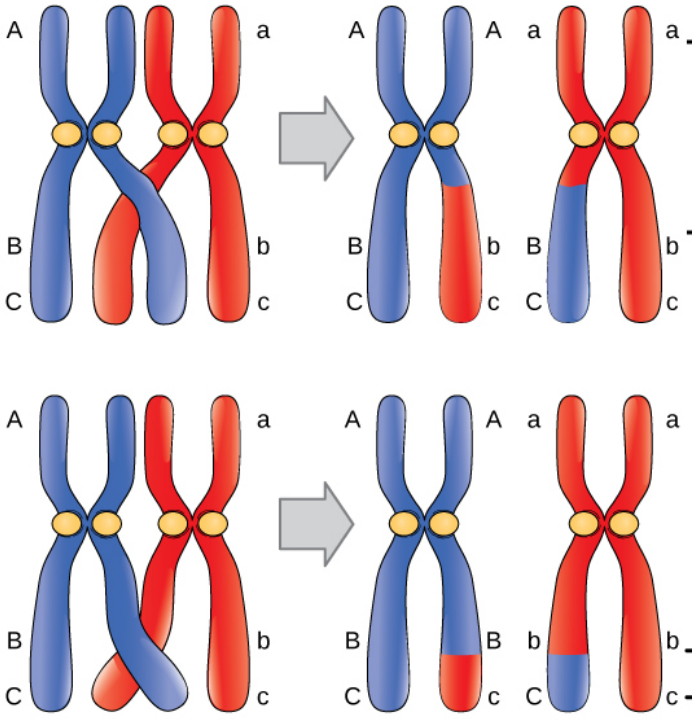
$$p^2 + 2pq + q^2$$

where p^2 is the frequency of the dominant homozygous, q^2 the frequency of the recessive homozygous and $2pq$ the frequency of heterozygous

Genetic recombination

- Remember **Crossing Over** is the exchange of genes between homologous chromosomes during **prophase I** of meiosis.
- The new combination of genes produced by crossing over and independent assortment is called **genetic recombination**





Linkage disequilibrium

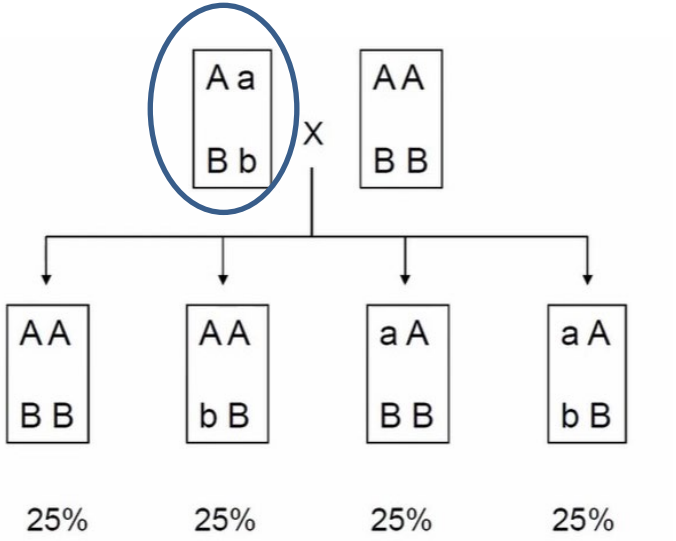
- Nucleotides in our genome are not inherited independent but in short strands of dna: this is in conflict with Mendel independent assortment law.

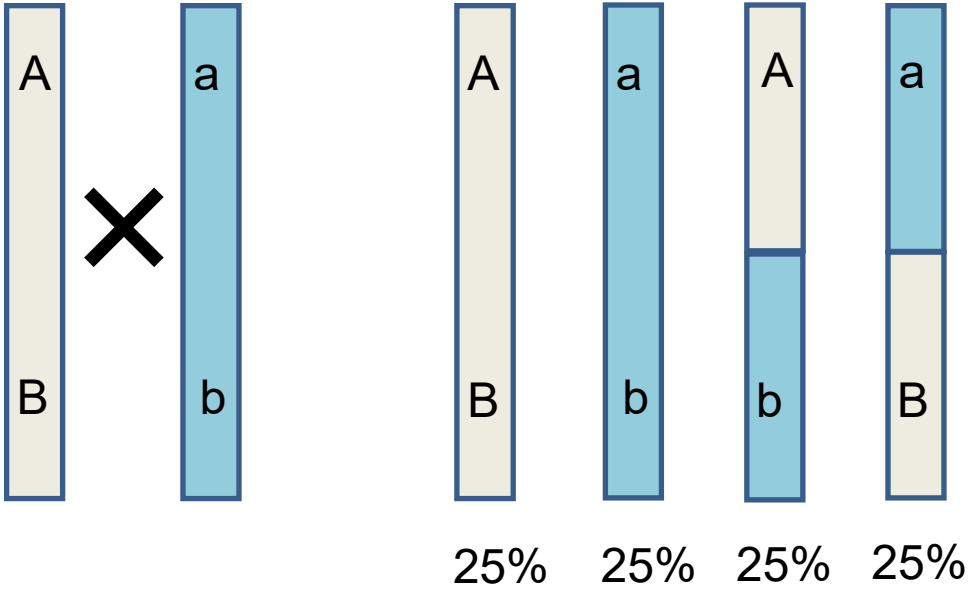
- **Linkage**: 2 Alleles are located in the same chromosome

- **Disequilibrium**: Is not following the expected frequencies as if the 2 alleles were independent (equilibrium)

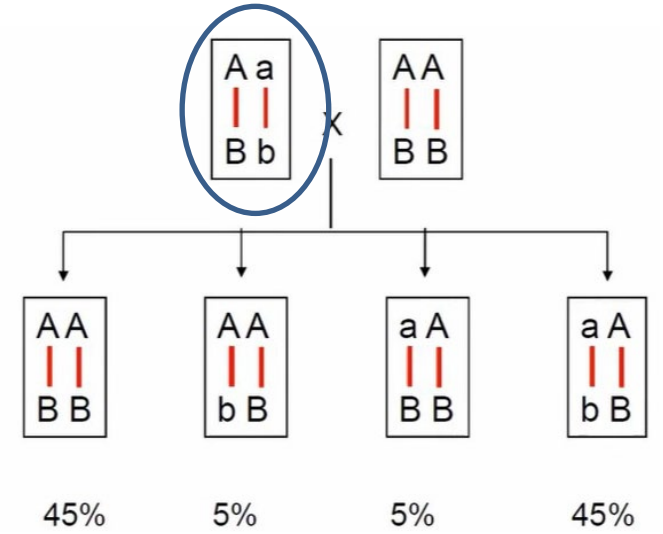
- Can be measured with Pearson correlation

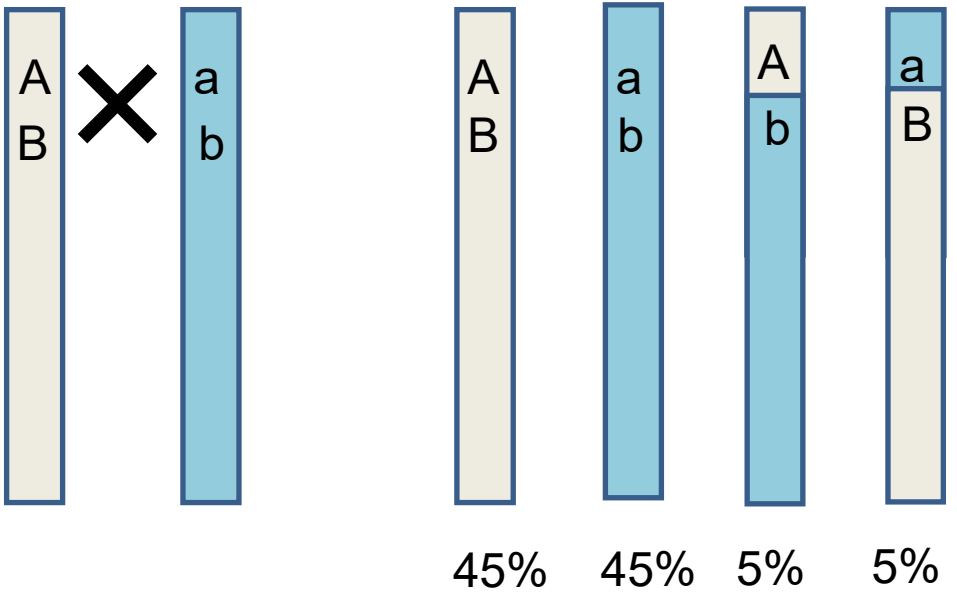
Linkage Equilibrium





Linkage Disequilibrium

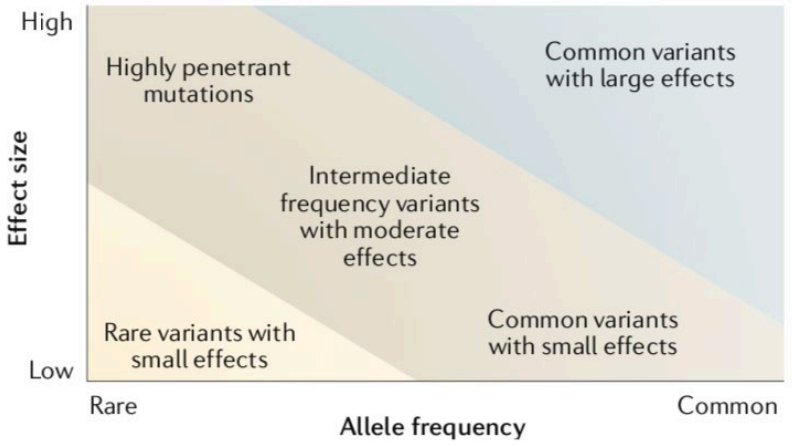




Haplotype

- Sequencing the whole genome is often costly so genomes have been sequenced with microarrays that just tags SNPs
- Trough haplotype we can obtain the whole sequence of the genome trough a statistical process called imputation

Common Variants Low effect-size

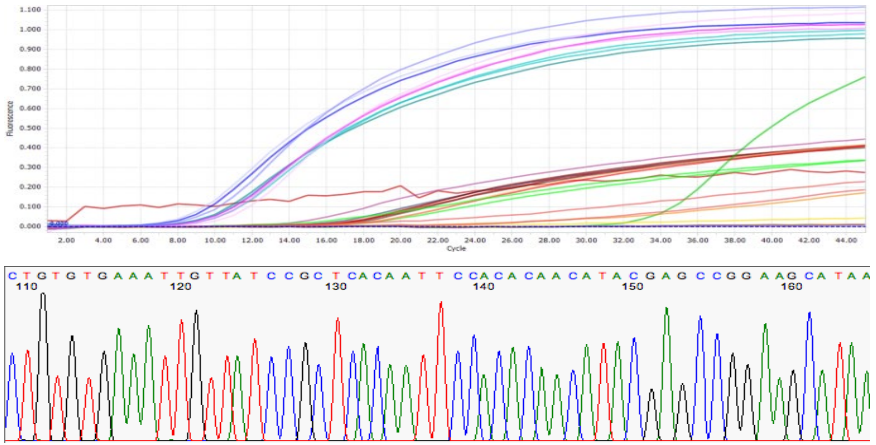


Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. In Frontiers in Genetics (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fgene.2020.00424

- Variants related to complex traits tipically have small effect size so the understanding of their role in disease etiology it's difficult

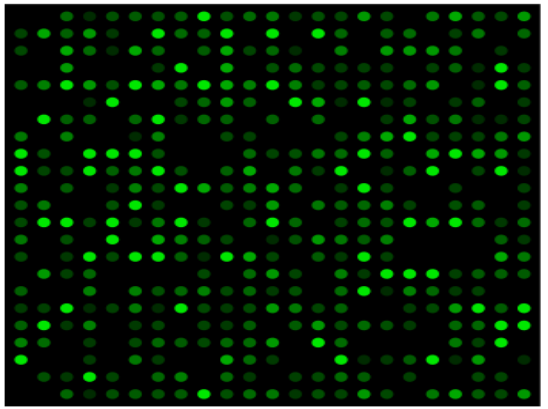
METHODS OF ASSOCIATION

Candidate gene



- Preselected variants (biased, not agnostic)
- No population structure control.
- LD not assessed.

GWAS



- “Whole” genome representation
- Needs well-defined phenotypes..
- Low resolution for small effect sized variants as in complex and polygenic diseases.
- Need large sample size

Genome Wide Association Study

- Test association of every genotyped locus variant against a phenotype.
- The association is done by regression using all subjects variants and phenotype value.
- Briefly, depending on the phenotype, the regression may be linear regression (continous or multiclass) or logistic regression (binary).
- Regression allows to include other variables (covariates) on its equation, allowing to take into account (in the association analysis) the variability of the subjects due to this covariate. This is call adjusting for the covariate)

Regression types

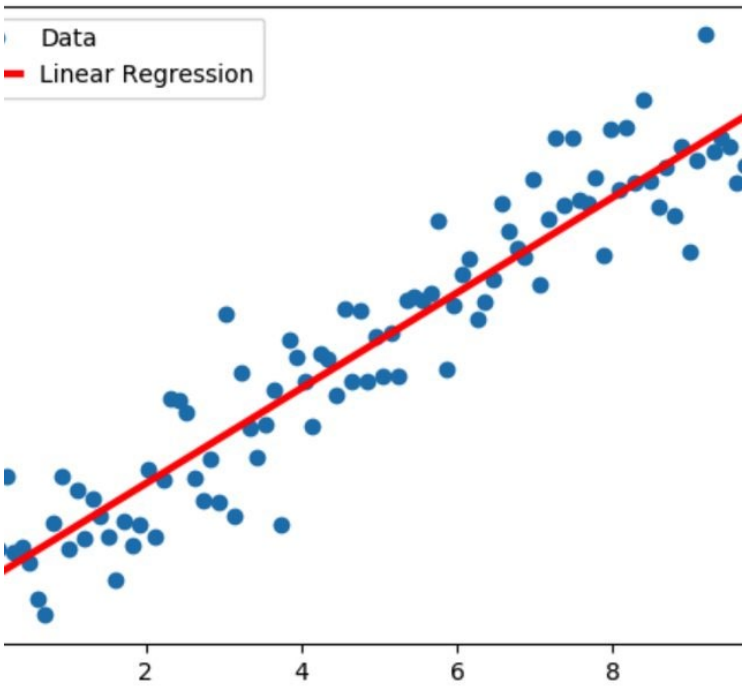
- Linear

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

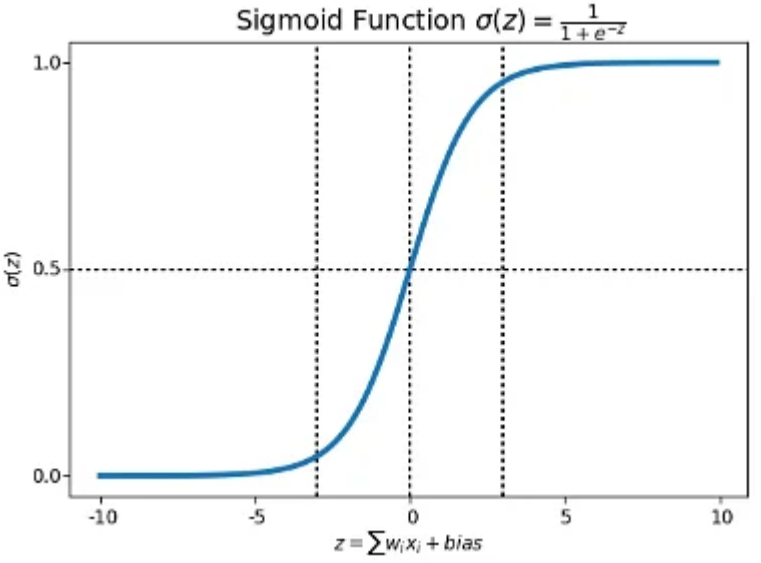


Regression types

- Logistic

$$f(x) = \frac{1}{1+e^{-(x)}}$$

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$



Statistics

- $\boldsymbol{H_0}$**:** The hypothesis of no differences between the 2 groups, the one that I want to prove wrong
- $\boldsymbol{H_1}$: The hypothesis of difference between the 2 groups, the one that I want to prove
- **p-value**: is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.
- **β (linear regression)**: the average amount by which the dependent variable increases when the independent variable increases one standard deviation and other independent variables are held constant. In other words, is the effect of the variable of interest on the outcome.
- **OR (logistic regression)**: is a comparison of an outcome given 2 different groups (control/case). If I have an OR > 1 means that in the case group the odd of the outcome is higher for that specific variable.

Types of phenotypes and covariates

- Binary
- Multiclass (categorical)
- Continous

0
1
1
1
0
1
0
1
0
1
0

0=Cases
1=Controls

2
4
3
4
2
4
2
1
4
3

1= No obesity
2= Obesity
3= Obesity with diabetis
4= Obesity with diabetis and renal failure

Different levels of a coding where each number represents a category

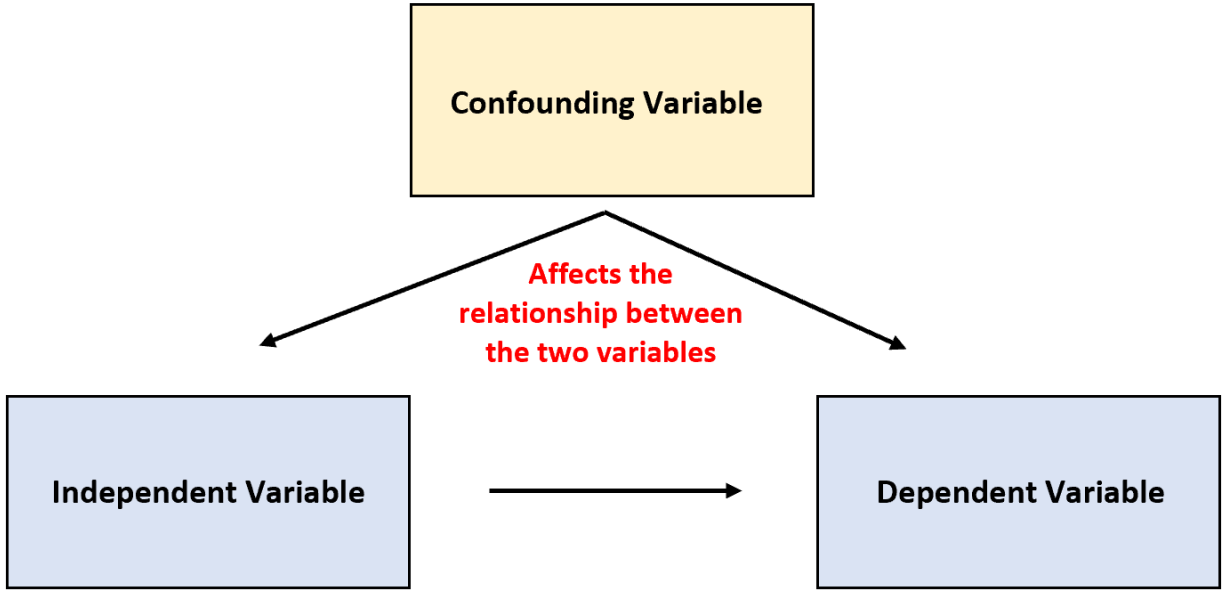
0.59
1.83
0.96
1.05
1.56
1.78
0.62
1.33
1.17

A measure that can have any value and fraction of it (blood pressure)

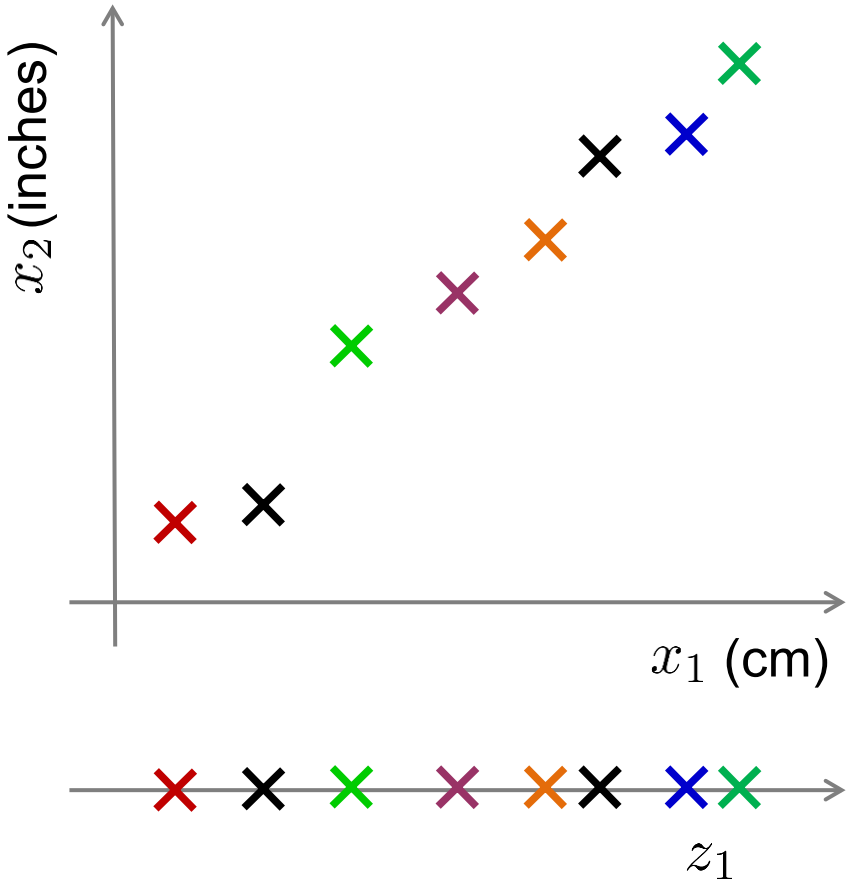
Adjusting a model: Covariates

- Adjusting analysis means to introduce in the regression model confounding factor (such as age and BMI)

- Ex if we are test genetics in association to a disease, we could detect an effect for some



Dimensionality reduction



Reduce data from 2D to 1D

$$x^{(1)} \rightarrow z^{(1)}$$

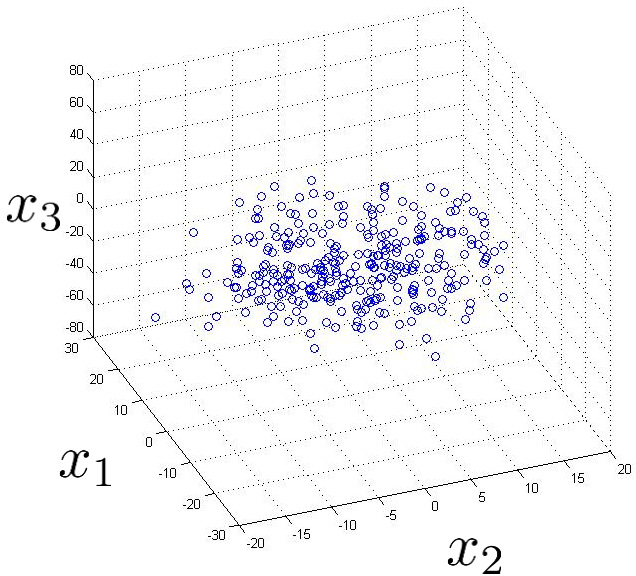
$$x^{(2)} \rightarrow z^{(2)}$$

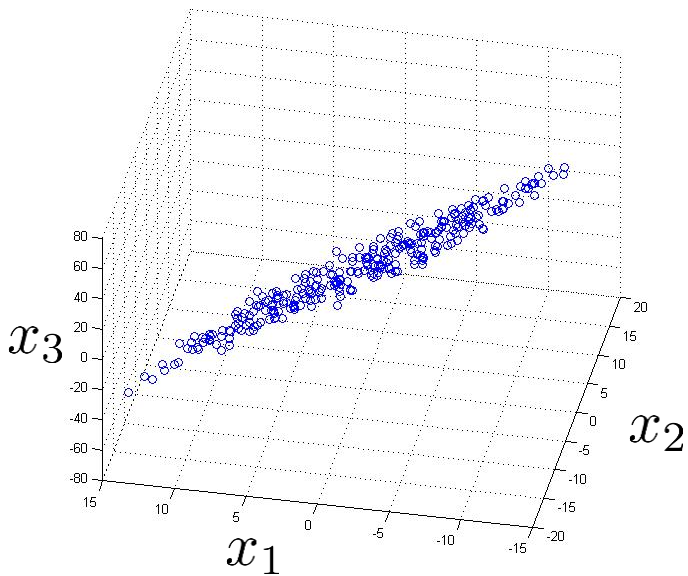
$$\vdots$$

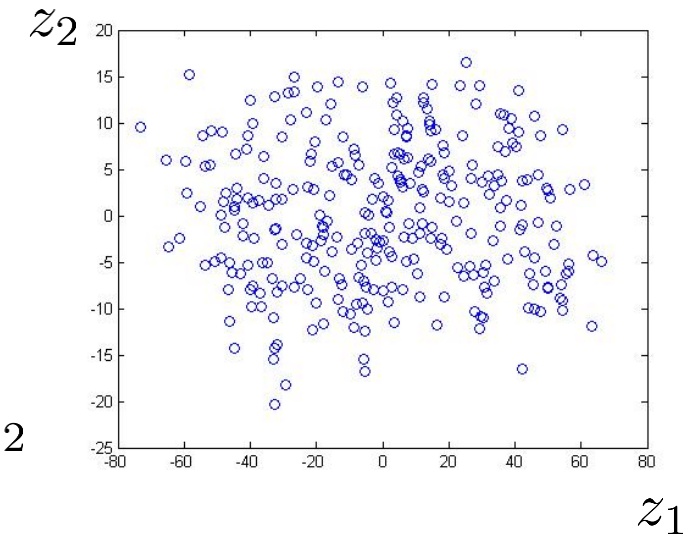
$$x^{(m)} \rightarrow z^{(m)}$$

PCA

Reduce data from multi dimensionality to high dimensionality in a way that the maximum variance lis maintained ex here 3D to 2D



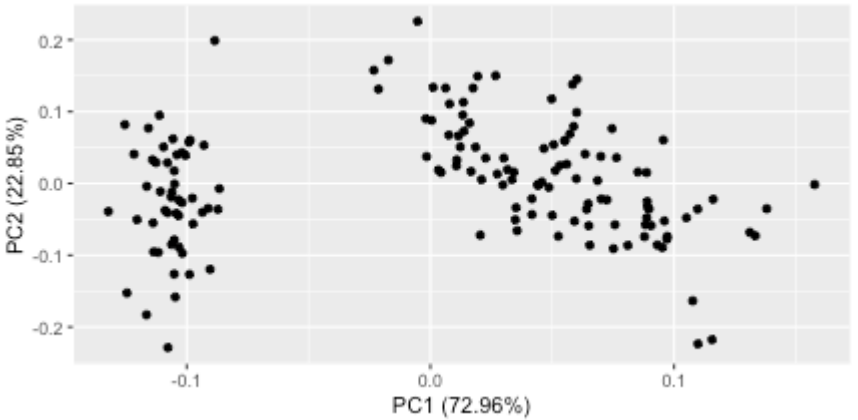




Andrew Ng

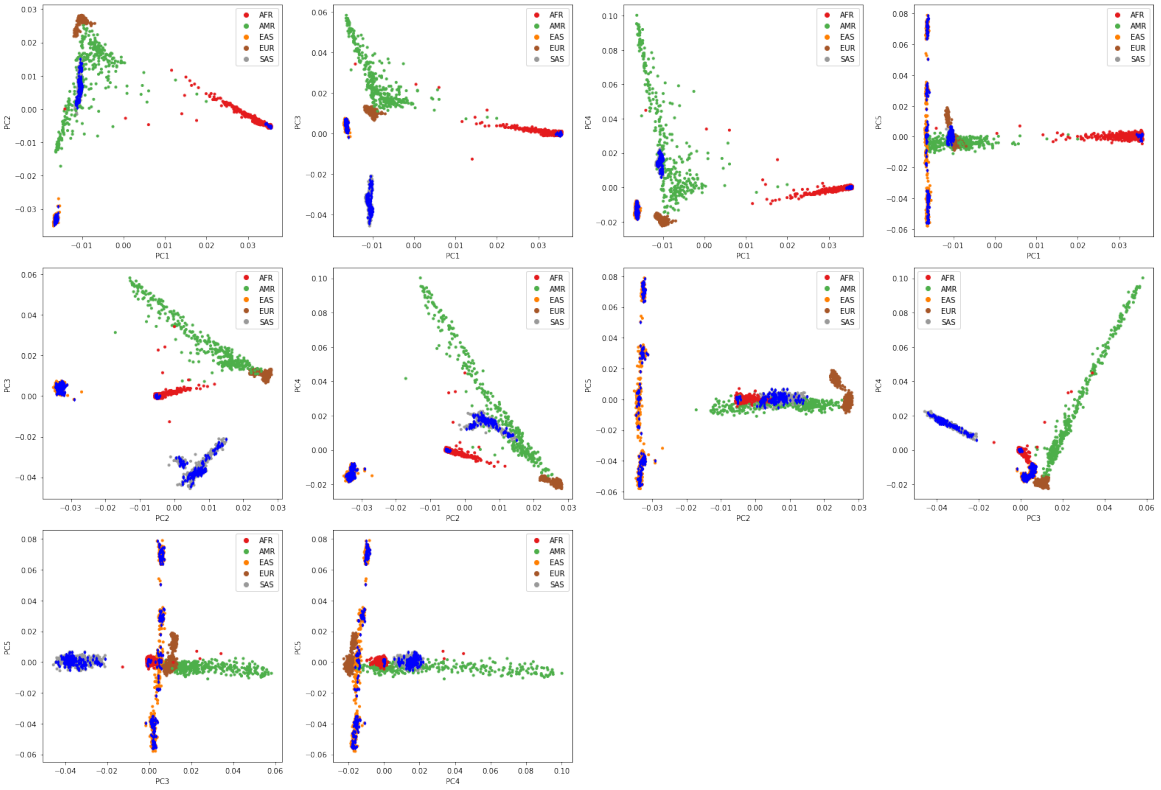
Adjusting a model: Relatedness

It is common to include in the analysis the PCA (principal components analysis) of the genetic data of the samples included in the study. This will correct for samples relatedness that could lead to a biased analysis multicollinearity



Adjusting a model: Etnicity

- Plotting our data against a genetic panel tell us if we have different ethnicities in our sample (carrying out multi etnic gwas can bias the analysis an open topic in research)
- Genetic marker for skin color might also be associated with malaria resistance because the trait is correlated with the population structure



Multiple testing

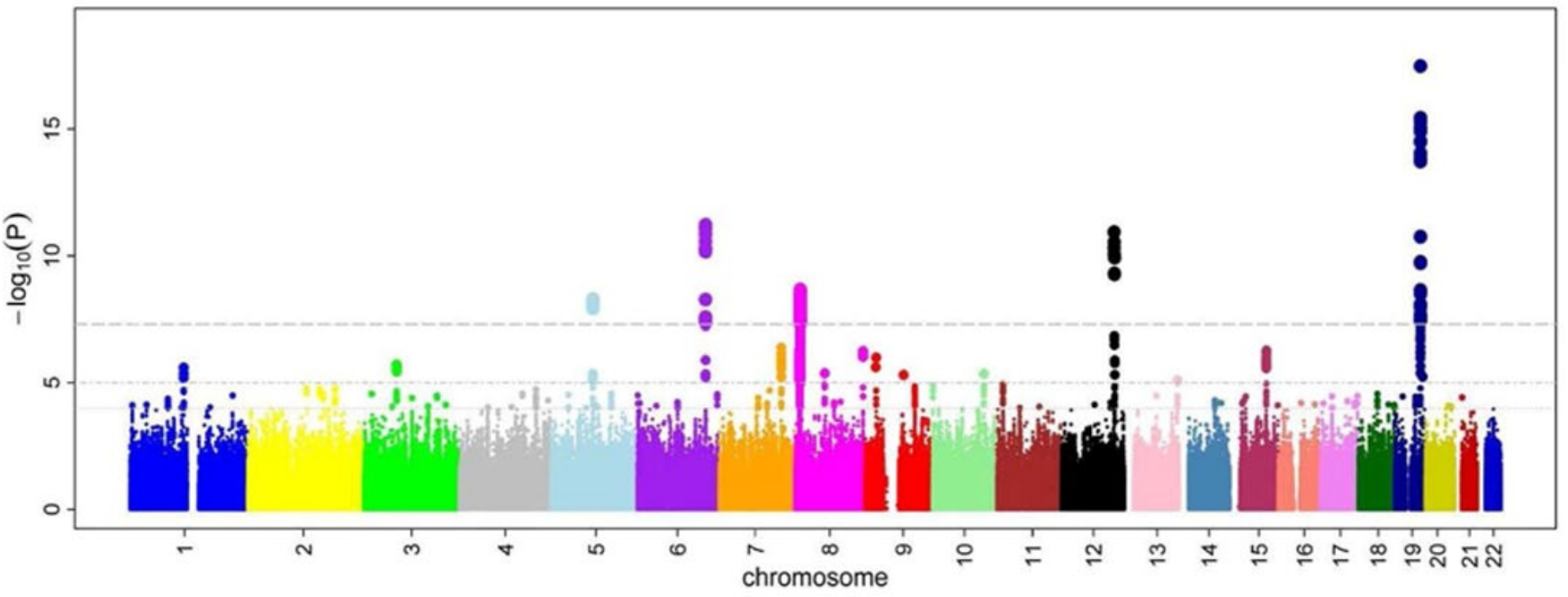
- When we are testing multiple hypothesis the probability of finding false positives by chance is increasing

- We need then to correct for multiple testing. One of the most common way is Bonferroni correction in which we lower the significance threshold.

$$B\alpha = \frac{\alpha}{n}$$

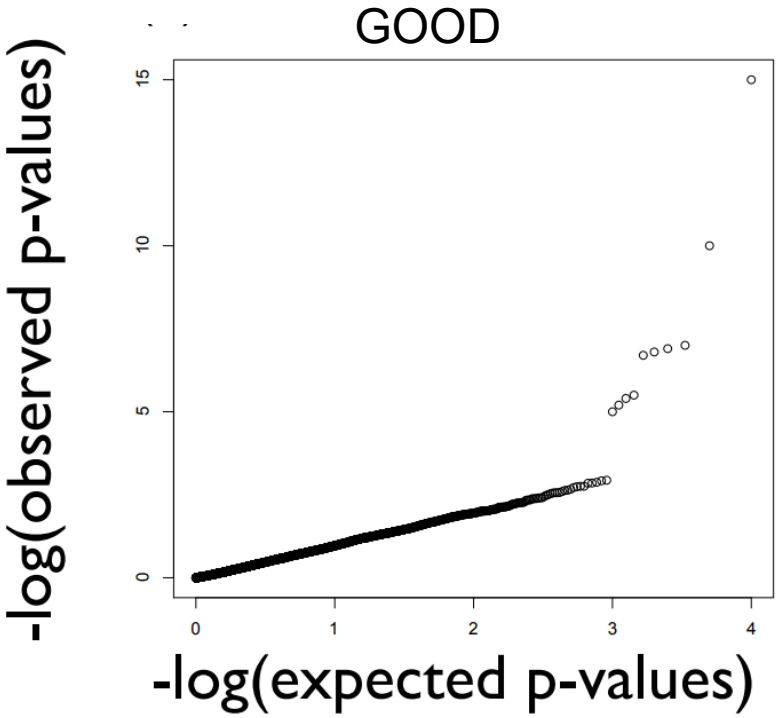
Where α is the starting significance (usually 0.05) and n is the number of tests, for GWAS, where we test million of SNPs the significance threshold becomes 5×10^{-8}

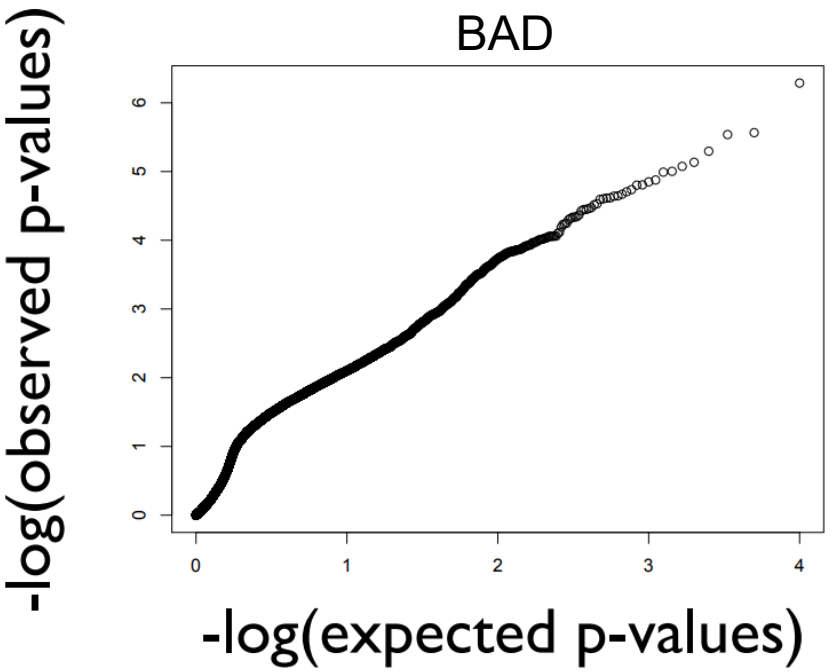
GWAS Results – Manhattan Plot



- Visualize significantly associated SNPs across the whole genome

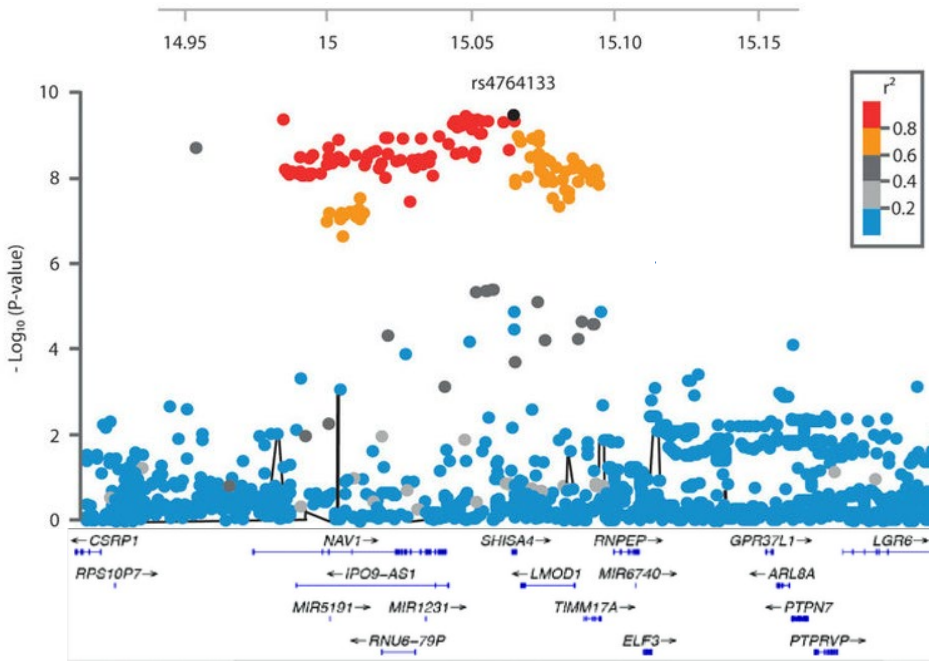
GWAS Results – QQ Plot





It tells us the quality of the GWAS, it shows how well our results are following the expected distribution of many non associated snps and few significant hits.

GWAS locus



Den Hollander, Wouter & Boer, Cindy & Hart, Deborah & Yau, Michelle & Ramos, Yolande & Metrustry, Sarah & Broer, Linda & Deelen, Joris & Cupples, L & Rivadeneira, Fernando & Kloppenburg, Margreet & Peters, Marjolein & Spector, Tim & Hofman, Albert & Slagboom, P & Nelissen, Rob & Uitterlinden, André & Felson, David & Valdes, Ana & van Meurs, Joyce. (2017). Genome-wide association and functional studies identify a role for matrix Gla protein in osteoarthritis of the hand. Annals of the Rheumatic Diseases. 76. annrheumdis-2017. 10.1136/annrheumdis-2017-211214.

- Many SNPs and genes are present among a genomic locus

- Linkage Disequilibrium (LD) make the determination of a causal variant more difficult

Over 90% of GWAS variants fall in non-coding regions of the genome and thus do not directly affect the coding sequence of a gene.

The common practice for variants in non coding regions is to assign the closest gene as the functionally relevant element

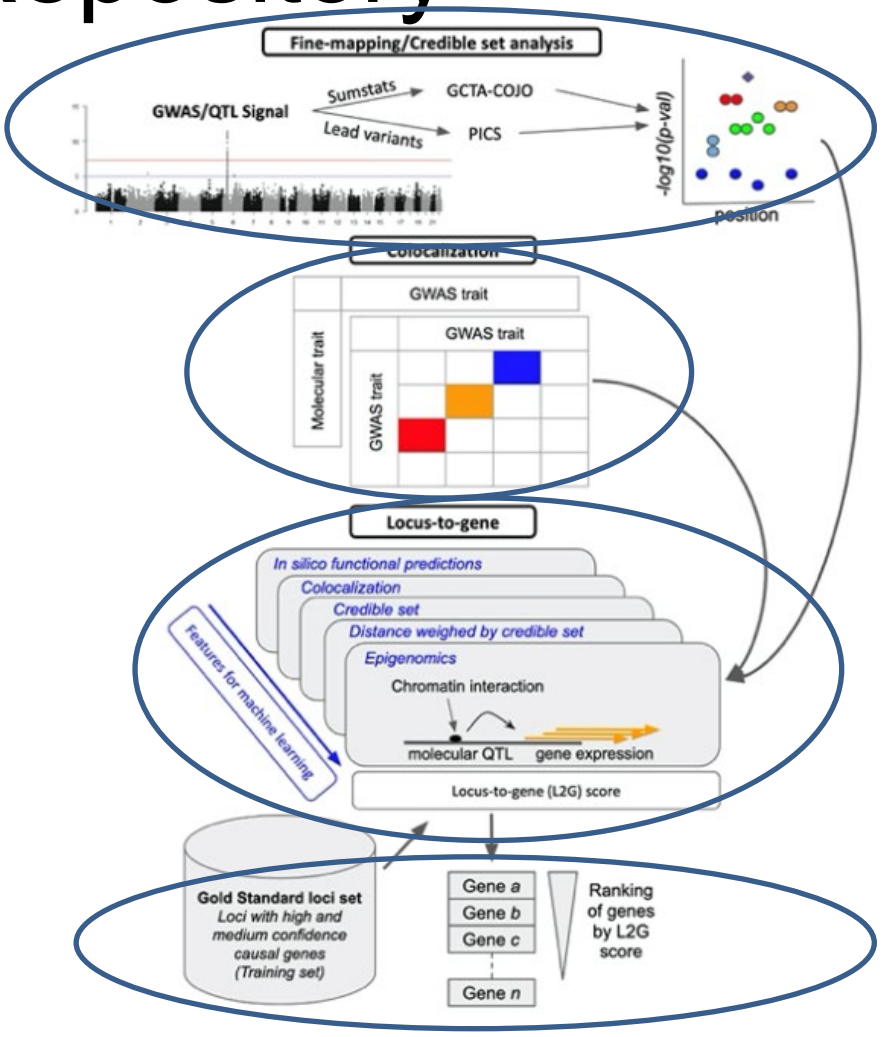
Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics, 11.*

GWAS challenges

- Definition of phenotype
- Small sample sizes
- Still expensive data
- Imputation (European based panels and ultra rare variants missing)
- Linkage Disequilibrium
- Missing heritability (small effect of variants)
- Role of epigenetics is not well investigated
- Lacking of a post-GWAS pipeline
- Multienicity
- Mapping to functional elements
- Still few data on regulatory element and chromatine structure
- Epistasis (gene-gene interaction)

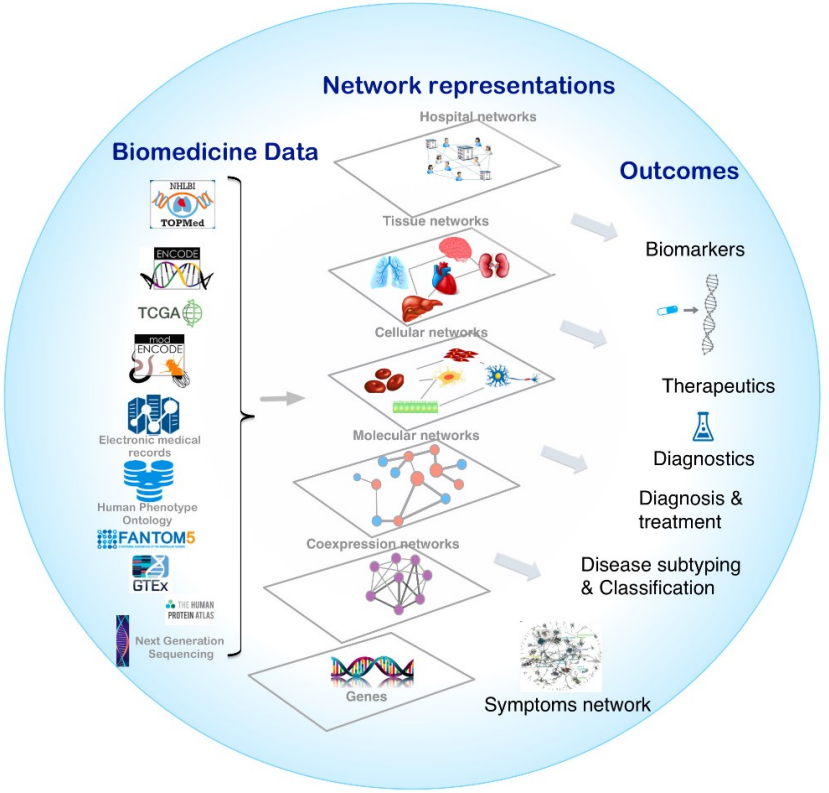
Variant-Gene Repository

- Starting from conditional analysis to independently associated loci
- Fine mapping the variants at every locus in order to assign at each variant a probability to be causal
- Colocalize the trait associated variant to a molecular trait such as gene expression
- Implement data coming from different sources
- Assign a score to each gene in the genomic locus as the result of a machine learning model trained on a gold standard dataset



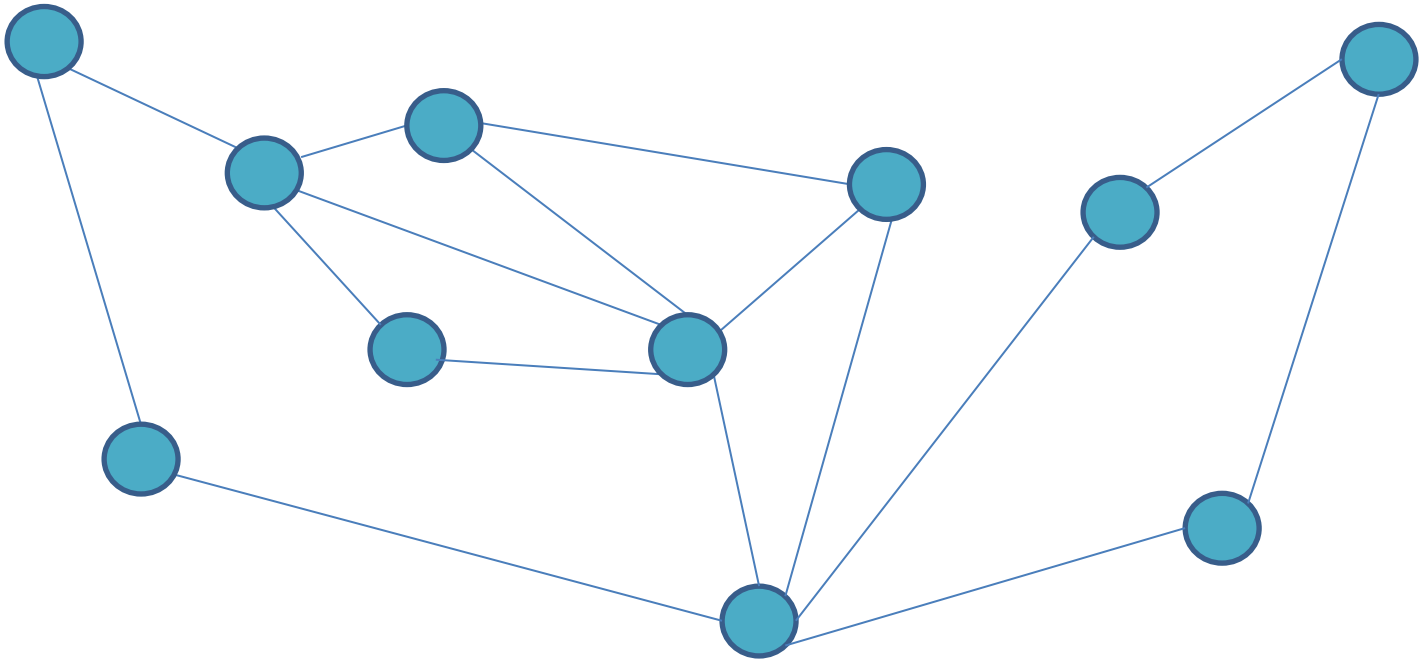
From variants to function

- Understand how variants affect the phenotype requires an understanding of disregulated pathways from a multi-omics perspective
- In this view multiple biological data ca be integrated in order to have a mechanistical view on the development of the disease
- Network based methods have proven to be effective to model different components of multiple omics data

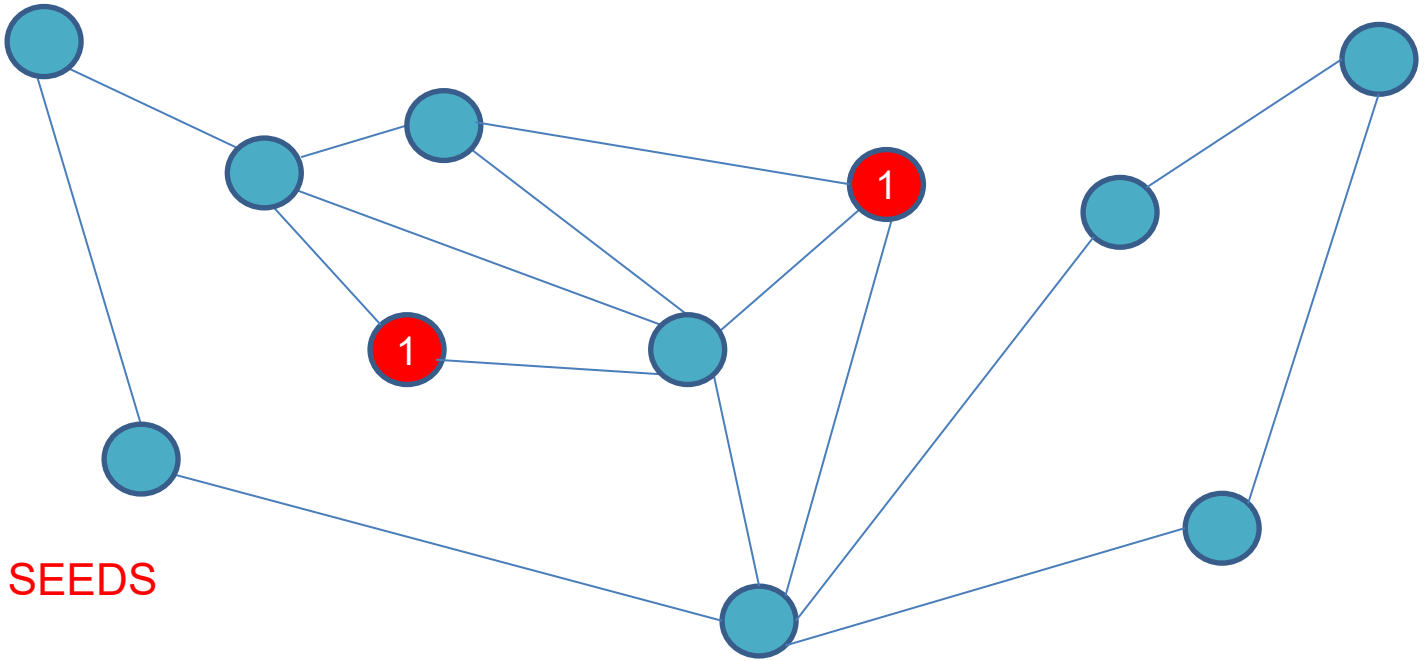


Sonawane, A. R., Weiss, S. T., Glass, K., & Sharma, A. (2019). Network medicine in the age of biomedical big data. In Frontiers in Genetics (Vol. 10, Issue APR). Frontiers Media S.A. https://doi.org/10.3389/fgene.2019.00294

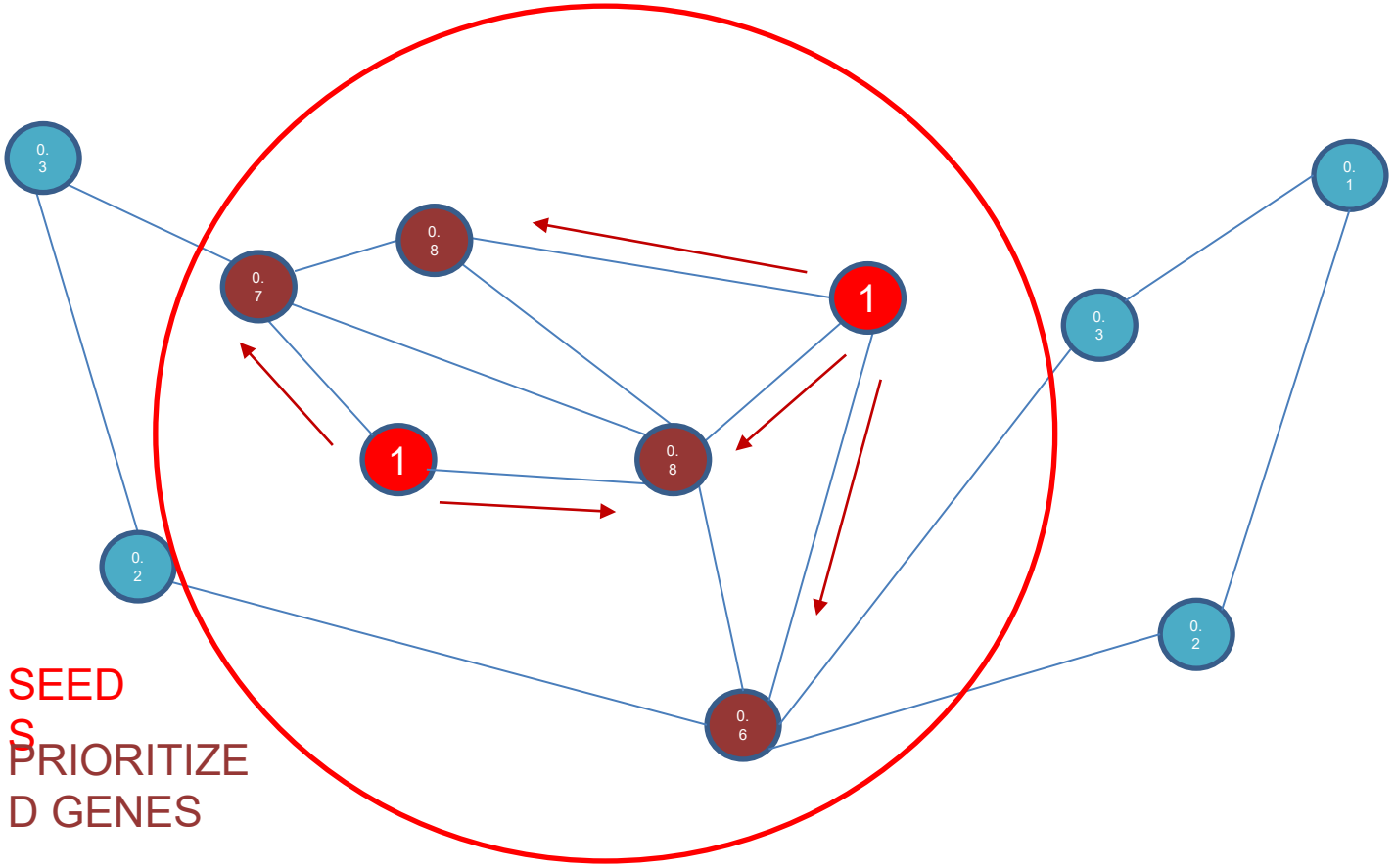
Protein Protein Interaction Network



Seeds: genes related to the disease



Seeds: genes related to the disease



Data sources

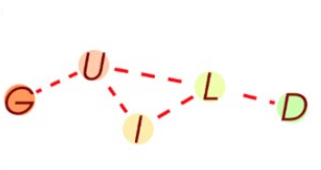
Seeds

Open Targets

Human Interactome

HIPPIE

Algorithm

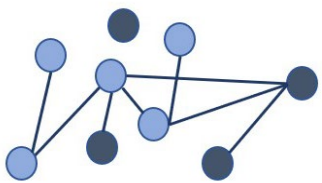


Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization

Emre Guney, Baldo Oliva

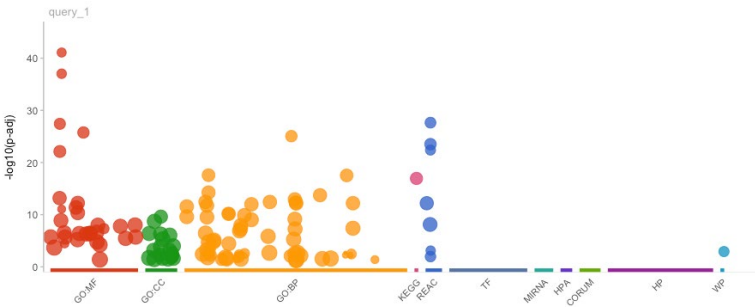
Published: September 21, 2012 • https://doi.org/10.1371/journal.pone.0043557

Funcional Enrichment



g:Profiler

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



GENEONTOLOGY
Unifying Biology

reactome

KEGG
Kyoto Encyclopedia of Genes and Genomes