

Maryam Labaf

Computational Science PhD Program

University of Massachusetts Boston

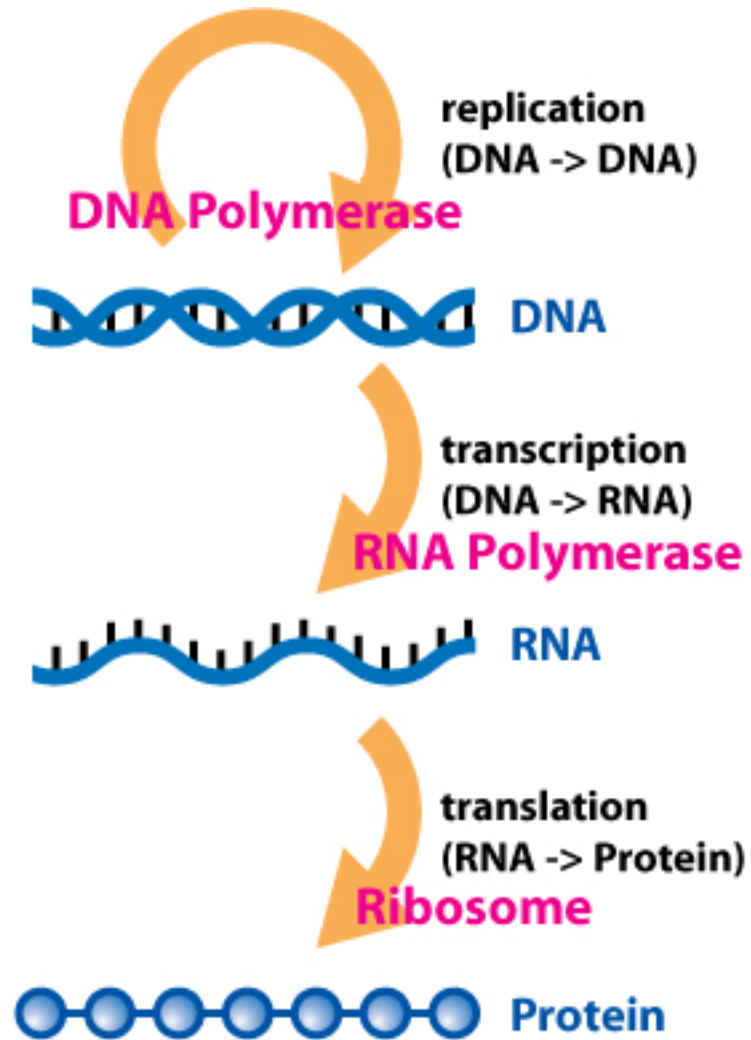
Genome Variation Analysis



Outline

- Genetic Mutation
- Biological Definitions
- Genomic Variation
- GWA Studies for Cancer
- GWAS Successes and Pitfalls
- Effect of SNPs in cis-regulatory motifs on penetrance of pathogenic mutations

Central Dogma of Molecular Biology



3' – ACAGGA – 5'



5' – UGUCCU – 3'



Cys – Pro

Genetic Mutation

Mutations: Proteins **NOT** turning out right

Where are mutations found?

- Mistake during translation



Genetic Mutation

Where are mutations found?

- Mistake during translation



- Mistake during transcription




Where are mutations found?

- DNA → RNA → PROTEIN
- 3' - CTC - 5' 5' - GAG - 3' ~~Glu~~ → Val

- DNA → RNA → PROTEIN
- 3' – CTC – 5' 5' – ~~GAG~~ – 3'
5' – GUG – 3' Val

- DNA → RNA → PROTEIN
- 3' – ~~CTC~~ – 5'
- 3' – CAC – 5' 5' – GUG – 3' Val



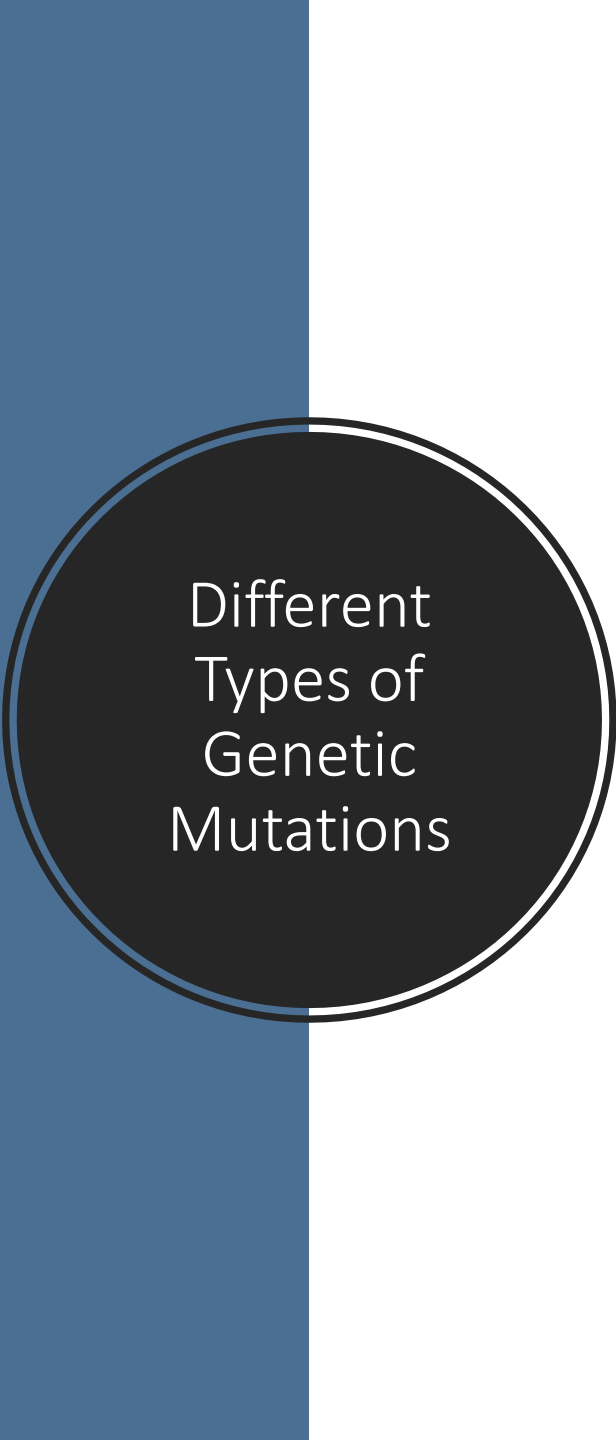
Genetic Mutation

- Mutations originates at the **DNA** level not RNA not protein
- Mutations come from:
 - Inherited
 - Spontaneous
 - DNA replication errors
 - Environmental factors
 - Completely random
- The effects of a mutation are usually found at the protein level

Different Types of Genetic Mutations

	DNA Level	RNA Level	Protein Level
No mutation	CTC CTC CTC	GAG GAG GAG	Glu – Glu - GLu
Point Mutation	CTC CTC CAC	GAG GAG GUG	Glu – Glu - Val
Frame –Shift Mutation	CTC C CTC CTC	GAG GGA GGA G	Glu – Gly - Gly

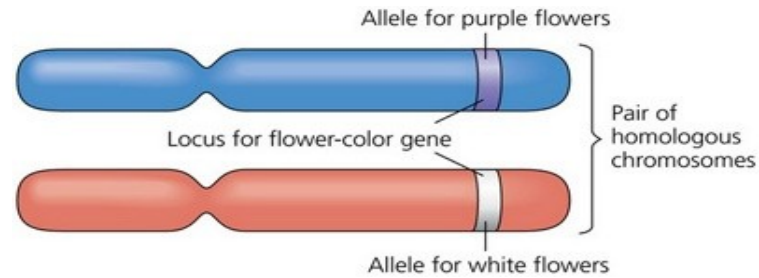
No mutation	ACA	UGU	Cys
Non-Sense Mutation	ACT	UGA	STOP!
Missense Mutation	ACC	UGG	Trp
• Silent mutations	CCA , CCG, CCT, CCC → Gly		
• Conservative mutation	Glu → Asp		
• Non-conservative mutation	Ser → Phe		



Different Types of Genetic Mutations

- **Synonymous mutations:**
 - Point mutations i.e. only one miscopied of DNA nucleotide
 - The mutated codon has the same meaning as the original codon
 - Amino acid does not change and not affects protein
 - No real role in the evolution of species
- **Nonsynonymous Mutations:**
 - Insertion or deletion of a single nucleotide in the sequence during transcription
 - Causes a frameshift mutation which throws off the entire reading frame of the amino acid sequence
 - Changes the resulting protein that is expressed
 - Be a lethal mutation if it happens near the beginning and entire protein is changed
 - Non-sense mutation - causes stop codon

Biological Definitions

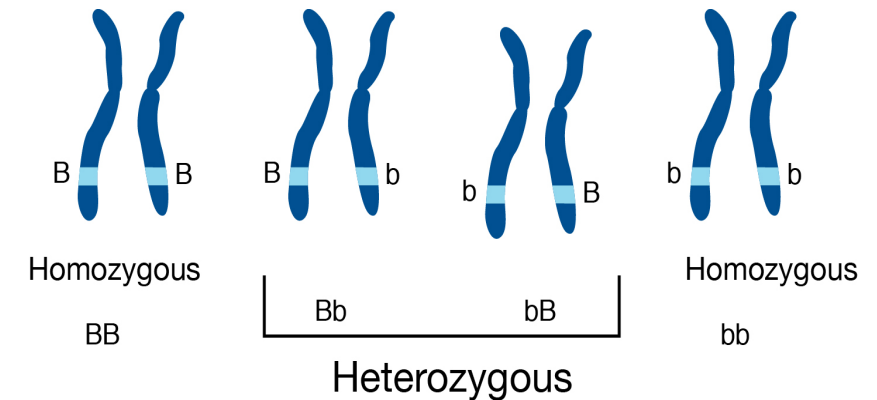


• Zygosity

- A **homozygote** possesses two identical copies of the same allele at a locus
- A **heterozygote** possesses two different alleles at a locus

• Alleles

- Variant form of a given gene
- Different alleles can lead to different phenotype



wild-type sequence

ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (orange)

ATCTTCAGCCATATGTGAAAGATGAAGTT

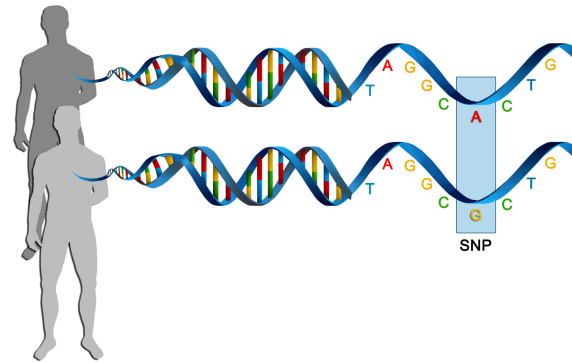
- **Indel** is a genetic difference created by insertion or deletion of a base pair or a longer DNA segment.



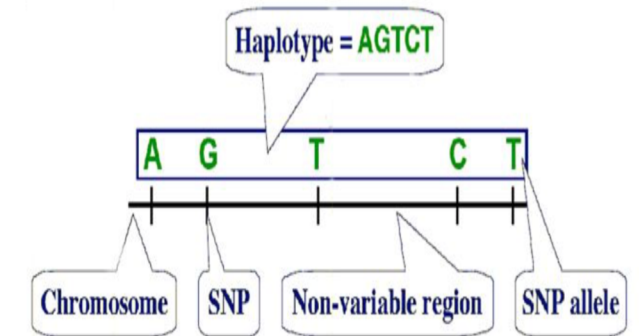
Biological Definitions

- **Rare Variant** is a genetic difference present in $<1\%$ of the alleles in the population.
- **Polymorphism** is a genetic difference present in $>1\%$ of the alleles in the population.
- **Penetrance** is the proportion of individuals carrying a particular variant (allele) of a gene (genotype) that also express an associated trait (the phenotype).
- **Expression quantitative trait loci (eQTLs)** are genomic loci that explain all or a fraction of variation in expression levels of mRNAs.

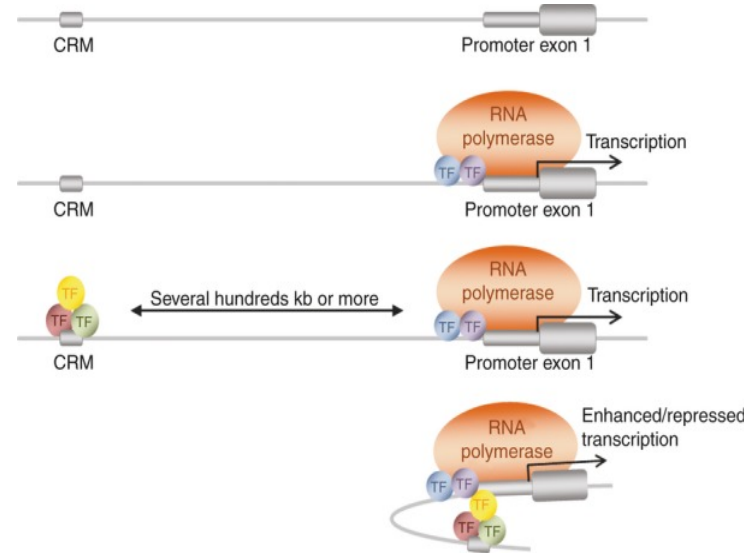
Biological Definitions



- **Single-Nucleotide Polymorphism (SNP)** is a variation in a single nucleotide that occurs at a specific position in the genome, where each variation is present to some appreciable degree within a population.

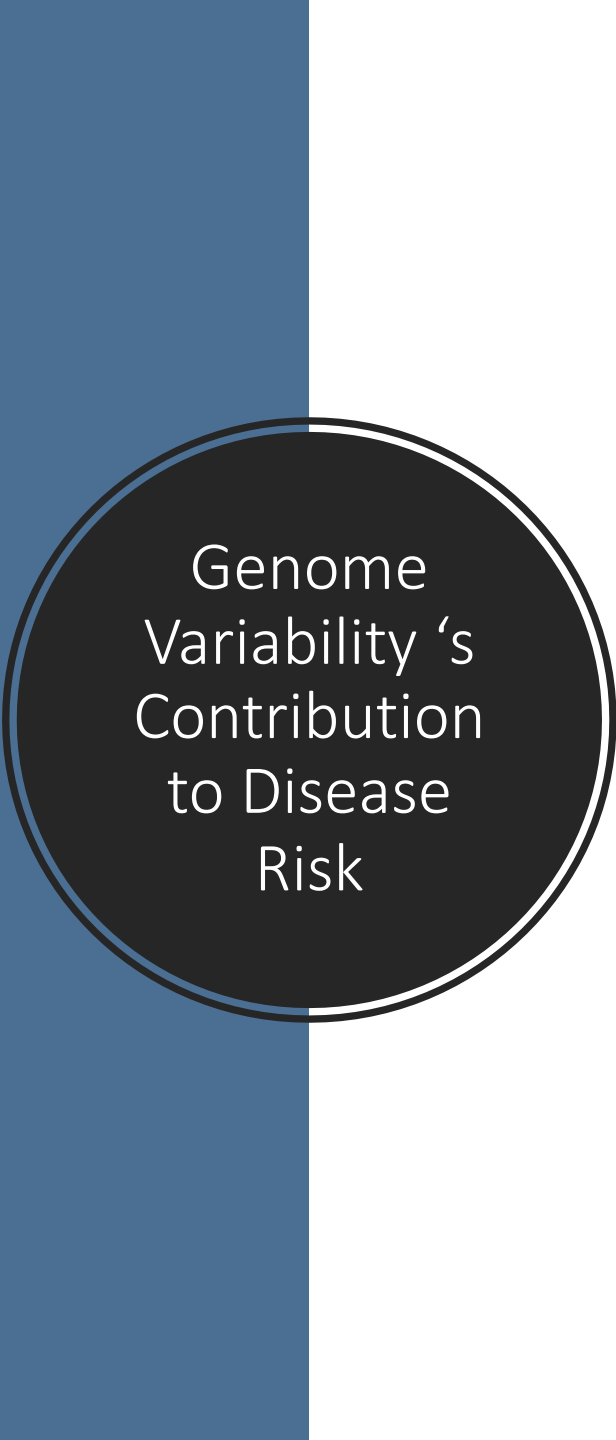


- **Haplotype** is a set of DNA variations, or polymorphisms, that tend to be inherited together.




Schematic representation of how a cis-regulatory module can enhance transcription.

- **Cis-regulatory Module** is a stretch of non-coding DNA, 100-1000 DNA bs in length, where a number of transcription factors can bind and regulate expression of nearby genes and regulate their transcription.



Genome Variability's Contribution to Disease Risk

- Discover of **genetic variants** explains variation in gene expression level which helps to differentiate disease risk among individuals.
- **Variable penetrance** (proportion of carriers with phenotype) and **variable expressivity** (severity of phenotype) are common phenomena causes individuals carrying the same variant to display highly variable symptoms.
- **How combinations of genetic variants may have joint effects on disease risk?**
- Researches were largely interested in variants that overlapped protein-coding regions of the genome.
- However, results from **Genome-Wide Association Studies** have found more than 88% of variants to be in non-coding region.

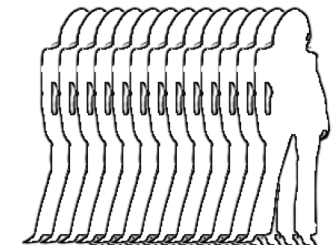
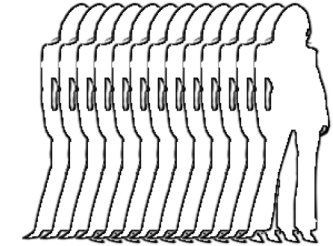


Genome-wide Association Study (GWAS)

- **GWA study** or **GWAS** is an observational study of a genome-wide set of genetic variants in different individuals to see any variant is associated with a trait.
- GWAS uses **analysis of SNPs** to find places in genome associated with differences in the trait of interest.
- Uses the **case-control setup** to compares two large groups of individuals, one healthy control group and one case group affected by disease.

Methods for GWA Studies

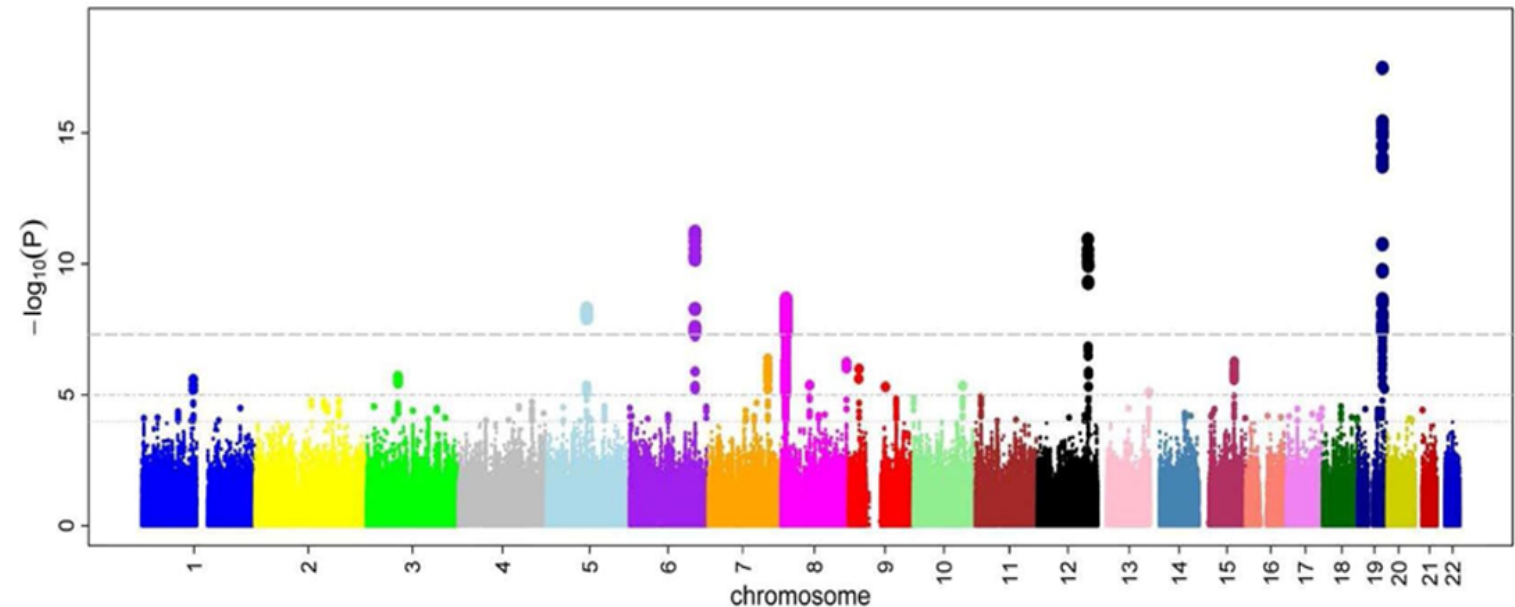
- Use a **DAN chip** to identify (genotype) their alleles at each of $\sim 10^6$ SNP genome positions
- Look for the SNPs where the two populations have **different allele frequencies**.
 - **Odds ratio**¹ is the fundamental unit for reporting effect size.
 - Allele frequency in the case group \gg the control \rightarrow odd ratio > 1
 - Use **chi-square test** to calculate the p-values for the significance of the odds ratio
- Odds ratios that are significantly different from 1 is the objective of the GWA study because this shows that a SNP is associated with disease.

	SNP1	SNP2	SNP...
	Cases	Cases	<i>Repeat for all SNPs</i>
	Count of G: 2104 of 4000	Count of G: 1648 of 4000	
	Frequency of G: 52.6%	Frequency of G: 41.2%	
	Controls	Controls	
	Count of G: 2676 of 6000	Count of G: 2532 of 6000	
	Frequency of G: 44.6%	Frequency of G: 42.2%	
	P-value: $5.0 \cdot 10^{-15}$	P-value: 0.33	

1. The odds of disease for individuals having a specific allele and the odds of disease for individuals who do not have that same allele

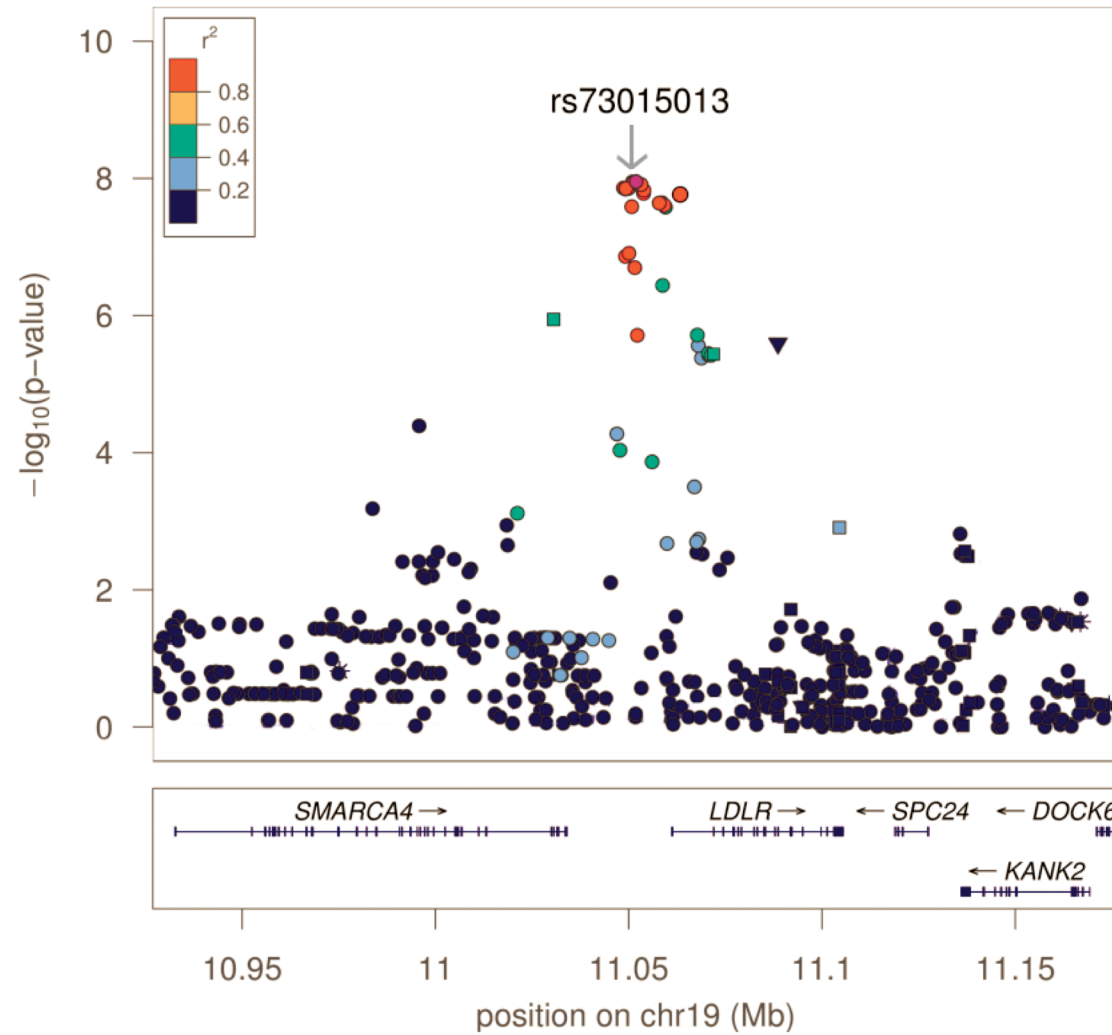
GWAS and Manhattan Plot

- Manhattan plot shows the negative logarithm of the P-values as a function of genomic location.



- An illustration of a Manhattan plot depicting several strongly associated risk loci.
- Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level.
- This example is taken from a GWA study investigating microcirculation.
- The tops indicates genetic variants that more often are found in individuals with constrictions in small blood vessels.

GWAS and Manhattan Plot



Regional association plot, showing individuals SNPs in the LDL receptor region and their association to LDL-cholesterol levels



Successes of GWAS

- Primarily goal was better understanding the biology of disease.
- Allow hundreds of thousands of polymorphism, usually SNPs, to be assessed simultaneously, i.e. investigate the entire genome.
- Better defining the relative role of genes and the environment in disease risk, assisting in risk prediction (enabling preventative and personalized medicine), and investigating [natural selection](#) and population differences.
- The GWAS approach is hypothesis-free, in that it looks at very many SNPs simultaneously rather than focusing on loci whose biology suggests that a causal relationship to the disease is likely.
- Find significant association between common genetic variants at genomic loci and complex traits in large population samples



Pitfalls of GWAS

- Dependency **to large sample size** to attain a certain level of statistical power.
- GWA studies are associated with **only a small increases risk** of the disease and have only a small predictive value.
- The median of odd ratio is 1.33 per risk-SNP, with only a few showing odds ratios above 3.0. Therefore, they do not explain much of the heritable variation.
- **Rare variants** raise the risk of disease much more than common variants – In GWAS 90~95% of variations in genetic risk factors are unexplained.
- **High false-positive rate** due to the the large number of hypothesis being tested – 500k -2M variants are being tested for associations.
- Difficulty of **identifying the few significant variants** that have clear functional implications in the mechanisms of the trait.