



Introduction to Bioinformatics

Instructor :Sakhaa Alsaedi

TA: Ebtihal Hani

Sakhaa.Alsaedi@kaust.edu.sa

Day 2: Calling variants in diploid systems
10th January 2024





Introduction to Bioinformatics

Exploring the Central Dogma of Molecular Biology

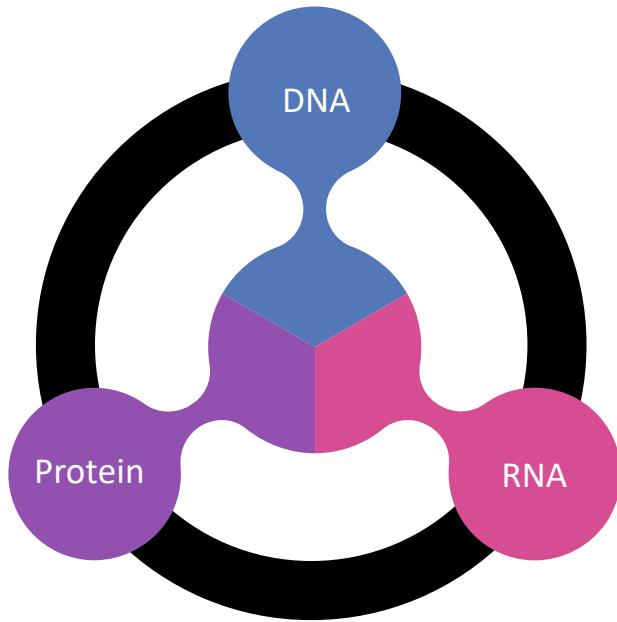
Part 1: Introduction



What is Molecular Biology?

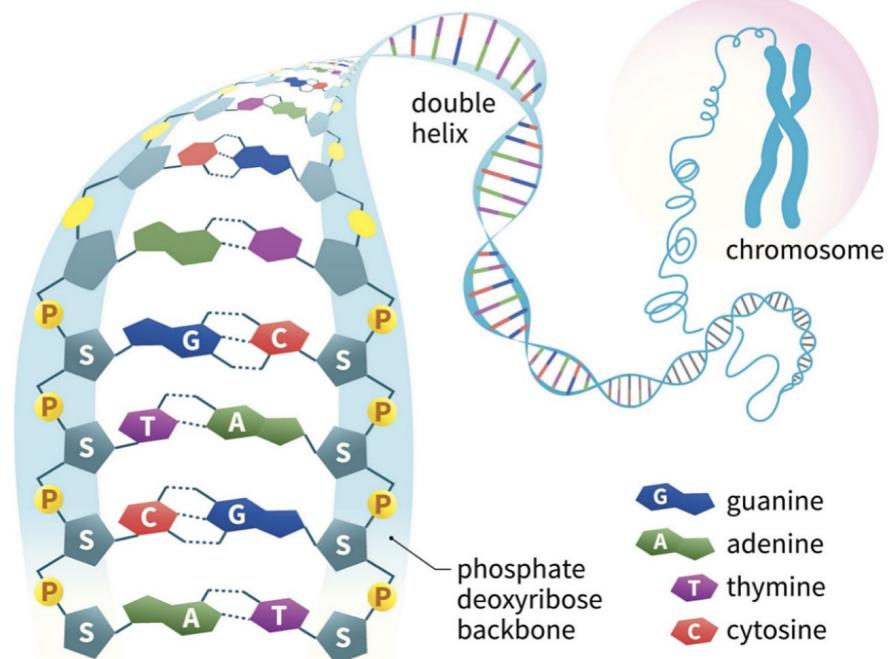
The study of structure and function of nucleic acids and proteins of biological molecules to understand the molecular mechanisms underlying various biological processes.

The Central Dogma

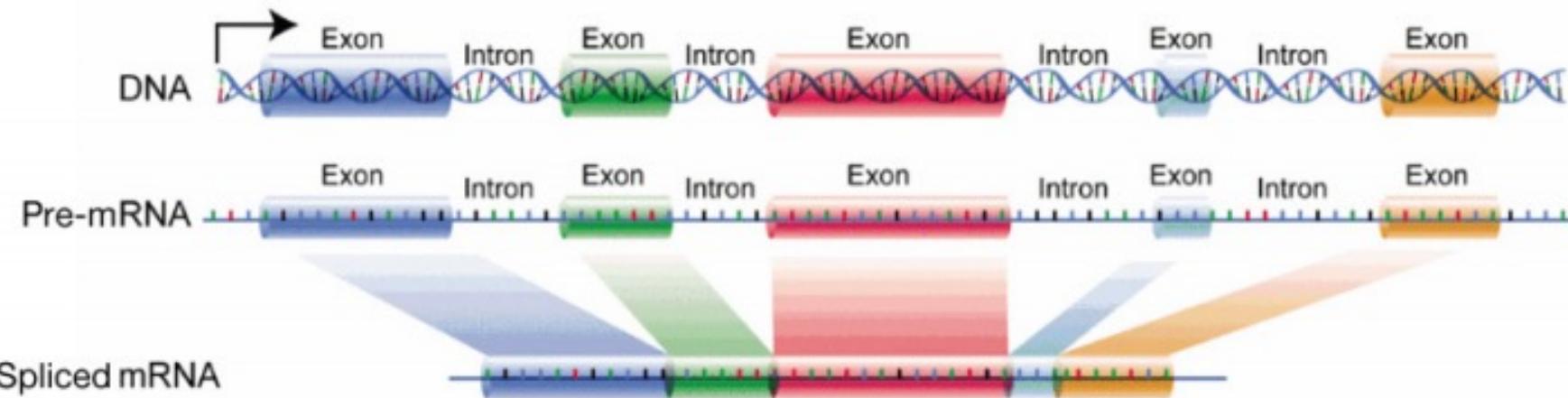


Human Genome

- Cell → Nucleus
- Nucleus → 23 Chromosomes
- Chromosomes → DNA (Genome)
- DNA → genes (Coding –noncoding)
- Genes → Nucleotide acid



Quick Summary of Nucleotide Structure



The Central Dogma Main Processes



DNA
Replication



RNA
Transcription



RNA
Translation

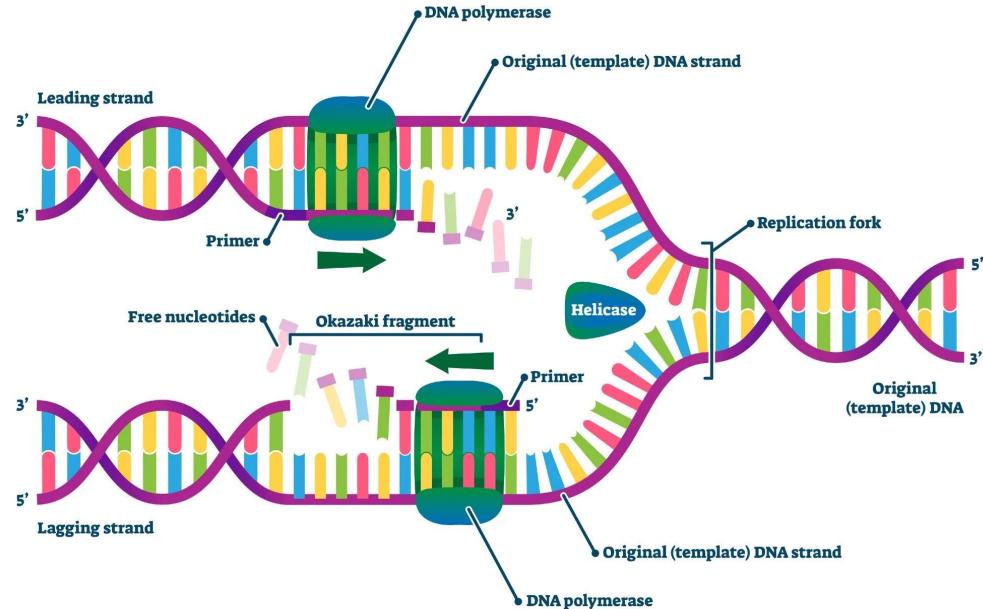




DNA Replication and Repair: Understanding how DNA is copied (replication) and maintained (repair) is crucial for maintaining genetic stability

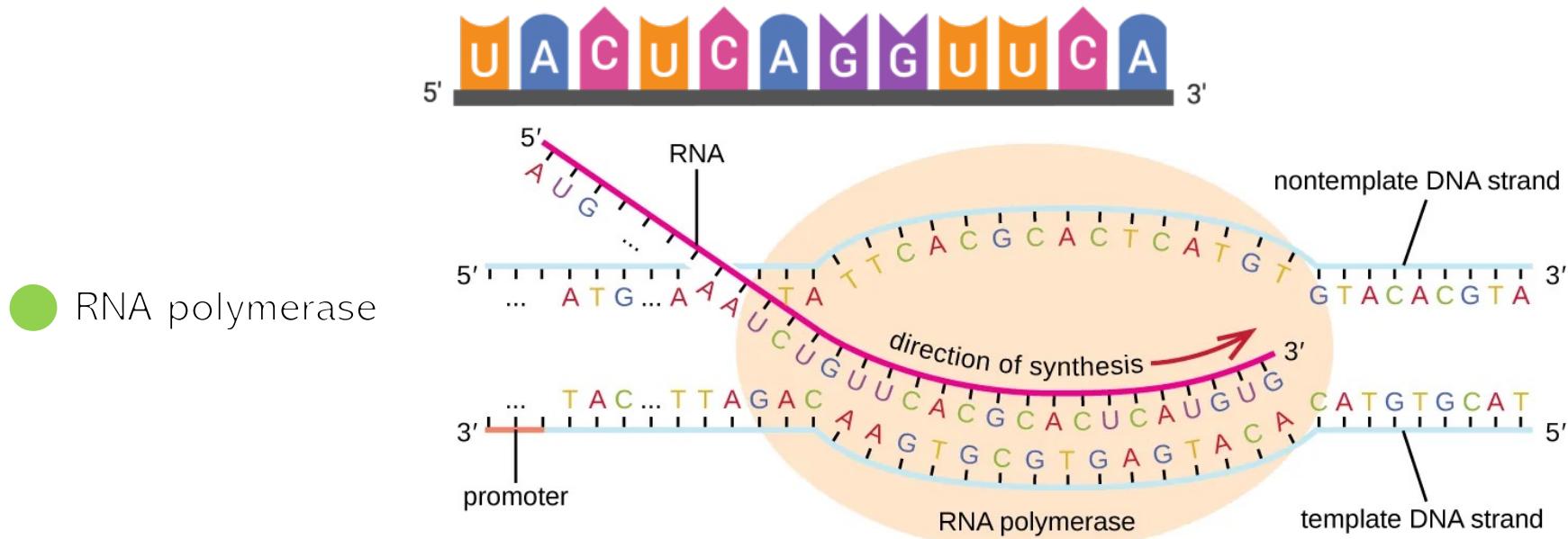


- Helicase
- DNA polymerase





Transcription: Molecular biologists study the process by which genetic information encoded in DNA is transcribed into RNA molecules, particularly messenger RNA (mRNA), which serves as a template for protein synthesis

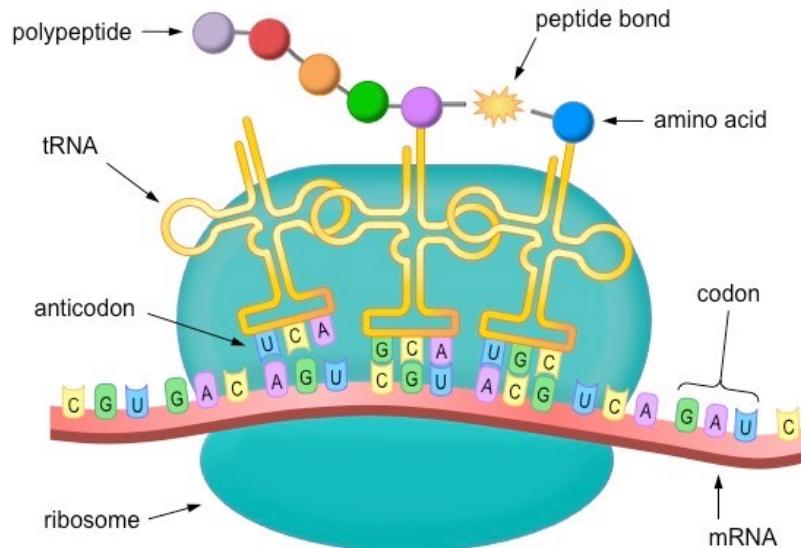




Translation: This refers to the process by which the information in mRNA is used to synthesize proteins, involving the interaction between mRNA, ribosomes, transfer RNA (tRNA), and amino acids.



- Ribosome
- tRNA

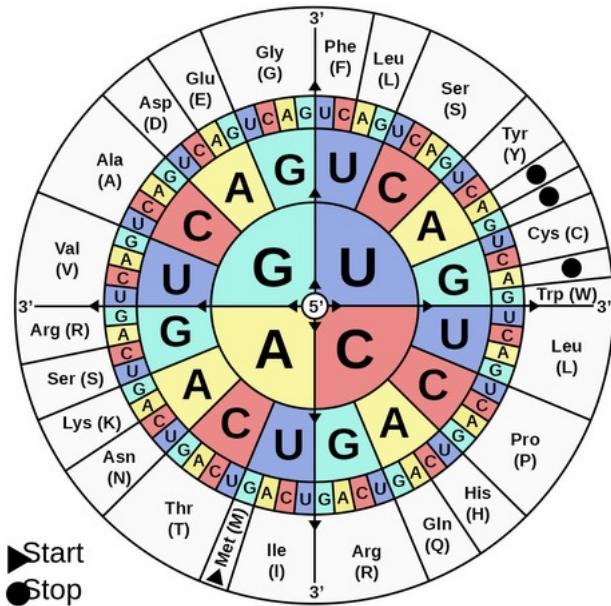


FIRST POSITION OF CODON

	T	C	A	G
T	ttt ttc tta ttg ctt ctc cta ctg att atc ata atg gtt gtc gta gtg	tct tcc tca tcg cct ccc cca ccg act acc aca acg gct gcc gca gcg	tat tac taa tag cat cac caa cag aat acc aca aag gat gac gaa gag	tgt tgc tga tgg cgt cgc cga cgg agt agc aga agg ggt ggc gga ggg
C				
A				
G				

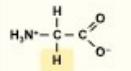
DNA CODE

SECOND POSITION OF CODON

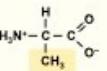


Amino Acids Types

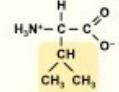
NON-POLAR



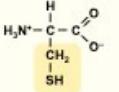
Glycine
(Gly / G)



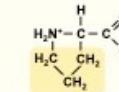
Alanine
(Ala / A)



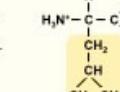
Valine
(Val / V)



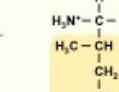
Cysteine
(Cys / C)



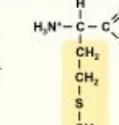
Proline
(Pro / P)



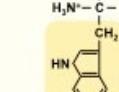
Leucine
(Leu / L)



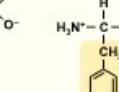
Isoleucine
(Ile / I)



Methionine
(Met / M)

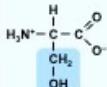


Tryptophan
(Trp / W)

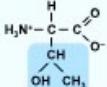


Phenylalanine
(Phe / F)

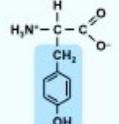
POLAR



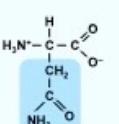
Serine
(Ser / S)



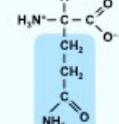
Threonine
(Thr / T)



Tyrosine
(Tyr / Y)

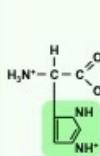


Asparagine
(Asn / N)

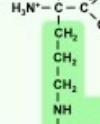


Glutamine
(Gln / Q)

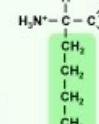
POSITIVE



Histidine
(His / H)

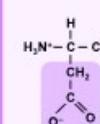


Arginine
(Arg / R)

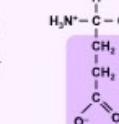


Lysine
(Lys / K)

NEGATIVE

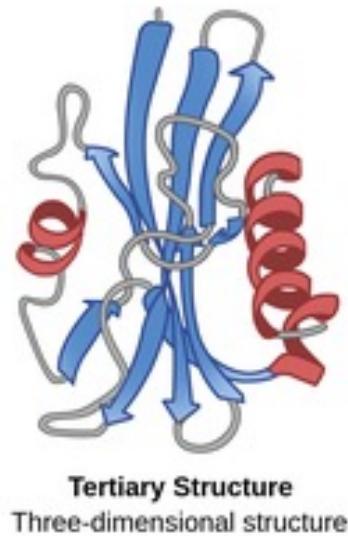
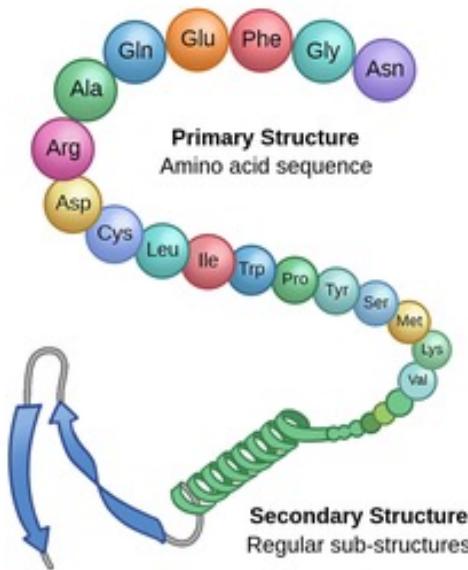


Aspartic Acid
(Asp / D)

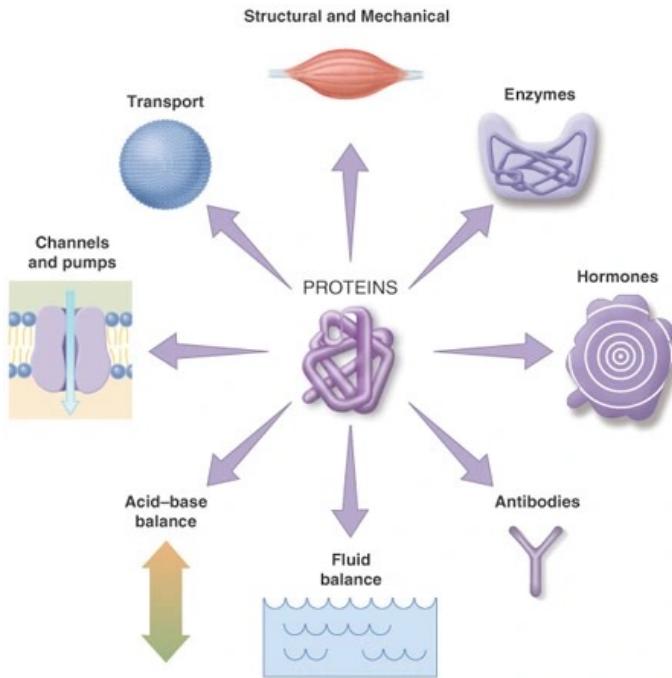


Glutamic Acid
(Glu / E)

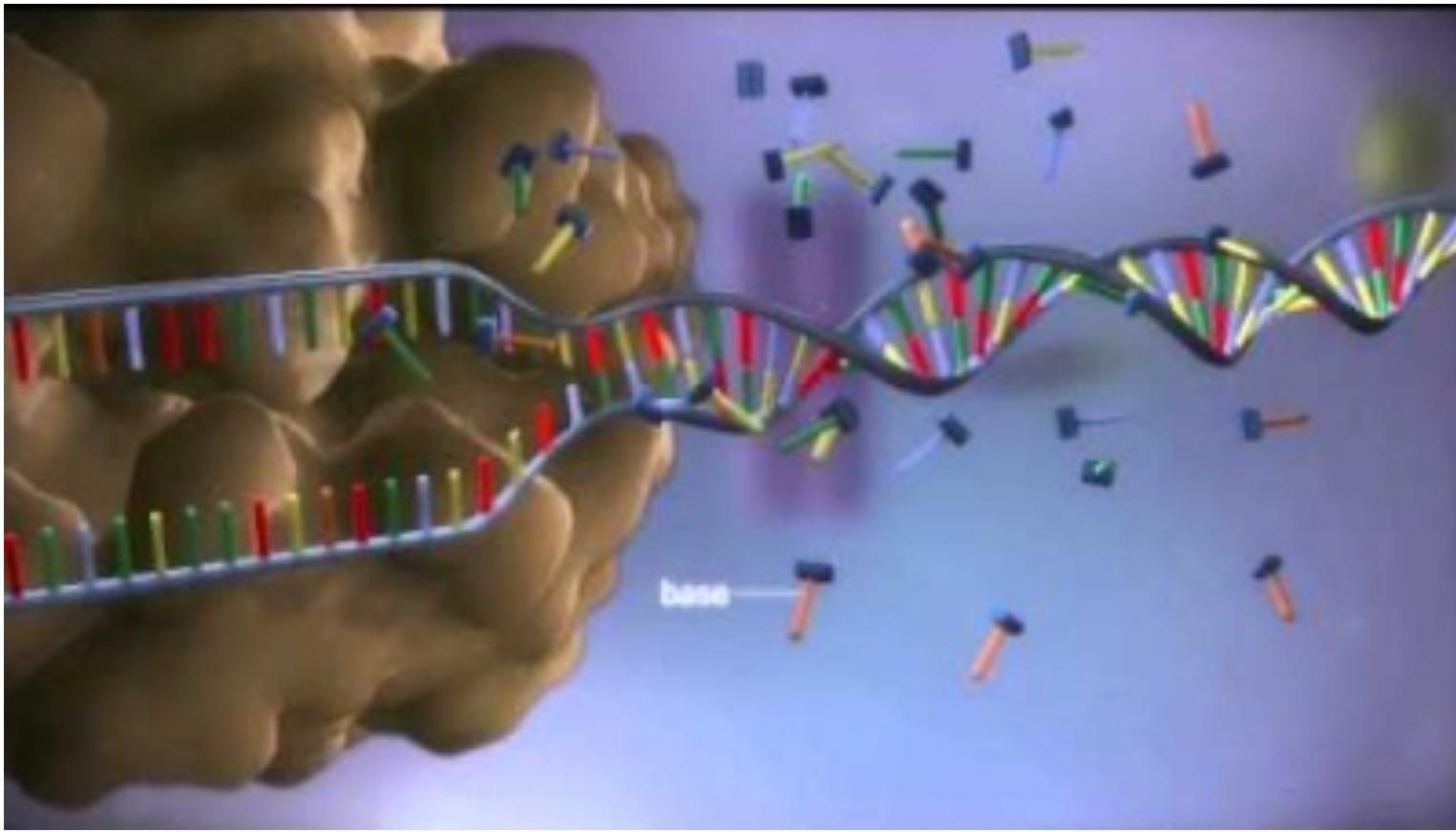
Protein Folding and Structure



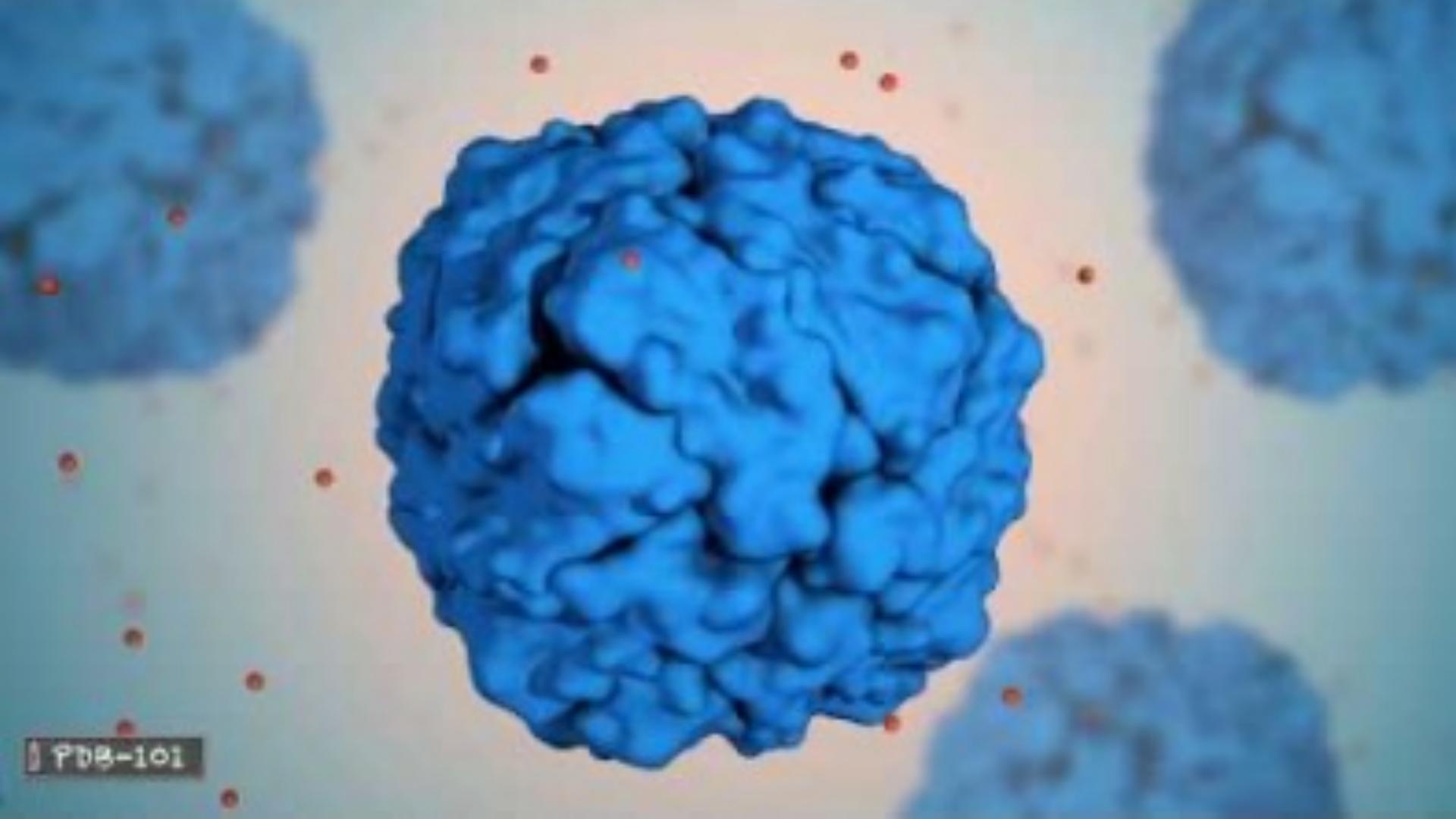
Protein Functions



Function	Description
Antibody	Antibodies bind to specific foreign particles, such as viruses and bacteria, to help protect the body.
Enzyme	Enzymes carry out almost all of the thousands of chemical reactions that take place in cells. They also assist with the formation of new molecules by reading the genetic information stored in DNA.
Messenger	Messenger proteins, such as some types of hormones, transmit signals to coordinate biological processes between different cells, tissues, and organs.
Structural component	These proteins provide structure and support for cells. On a larger scale, they also allow the body to move.
Transport/storage	These proteins bind and carry atoms and small molecules within cells and throughout the body.



<https://www.youtube.com/watch?v=gG7uCskUOrA&t=1s>



PVB-101



Introduction to Bioinformatics

Unraveling the Human Genome: Basics and Beyond

Part 2: Basic Genome



What is Basic Genome?

- What is Haplotype, Genotype, and Phenotype?
- What is SNP, Mutation, Alleles, and Variants?
- How to calculate allele frequency?
- What is homozygosity and heterozygosity?



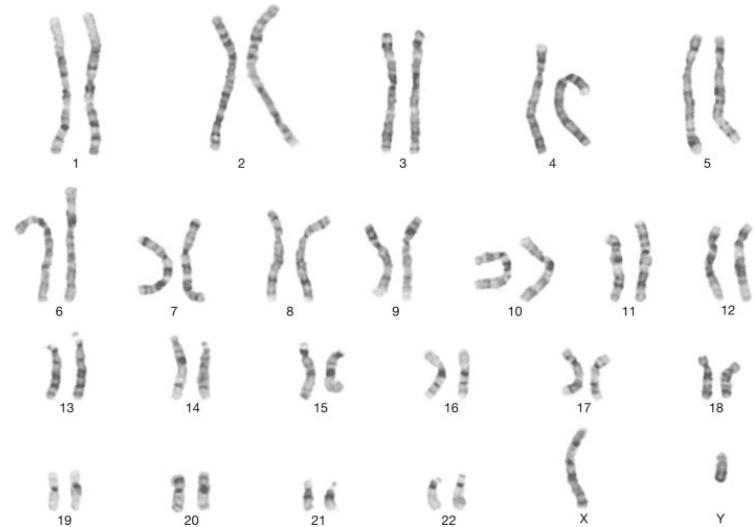
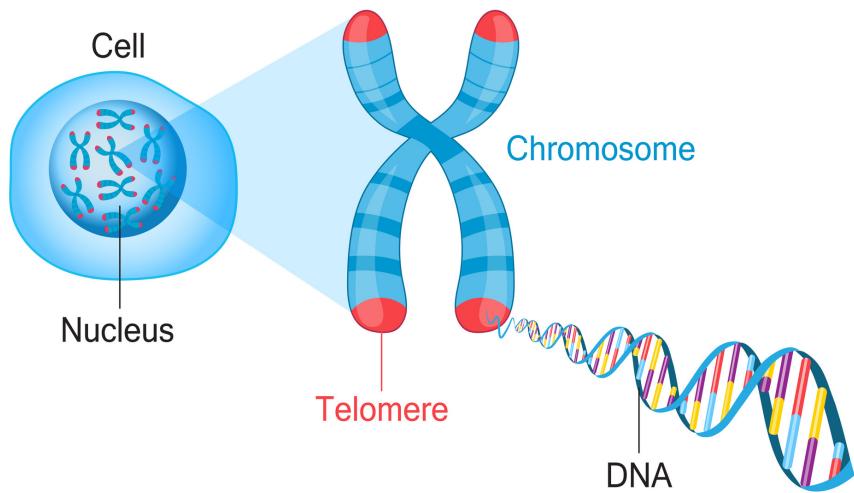
Genome and Genetics

- Genome: The genome refers to the complete set of genetic material (DNA) in an organism's cells.
- Genetics: Genetics is the scientific study of genes, heredity, and variation in living organisms.

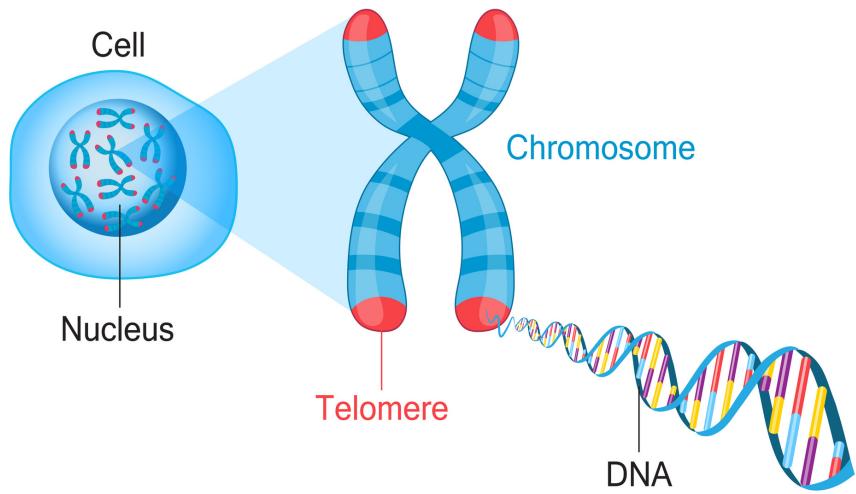
Haploid and Diploid cells

<https://www.youtube.com/watch?v=NR9zTvMg-pE>

Chromosomes

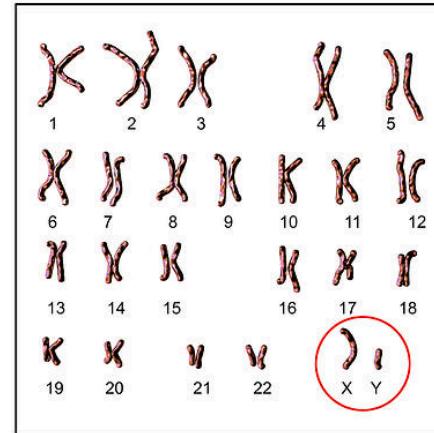


Chromosomes

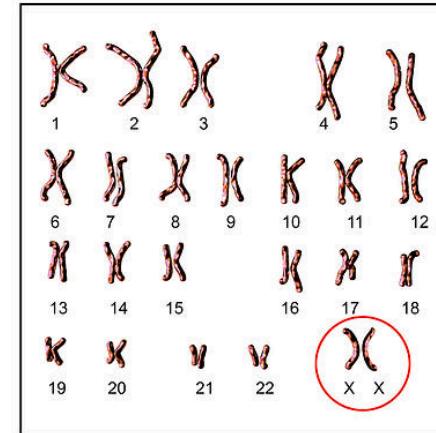


Human karyotype

Male

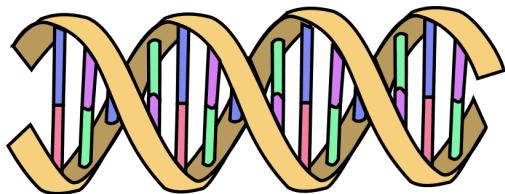


Female



Human Genetics

DNA



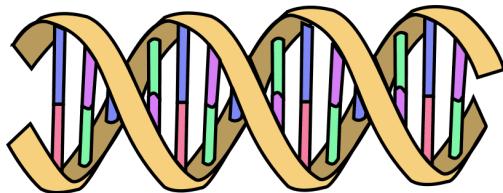
3 billions base pair long

....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

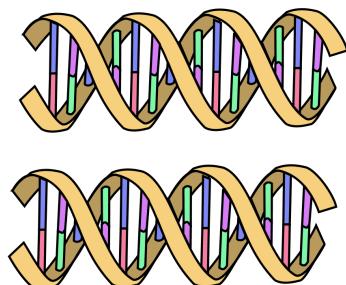
....TAAACGGTCAGTCATGGGT CCTACGACCTTGCCTA....

Human Genetics

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



....ATTGCCAGTCAGTACCCAGGATGCTGGA....

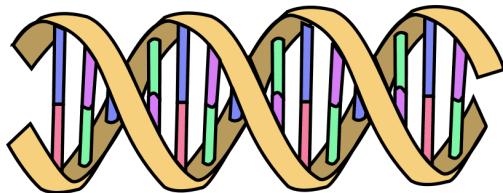
....TAAACGGTCAGTCATGGGTCTACGACCT....

....ATTGCCAGTCAGTACCCAGGATGCTGGA....

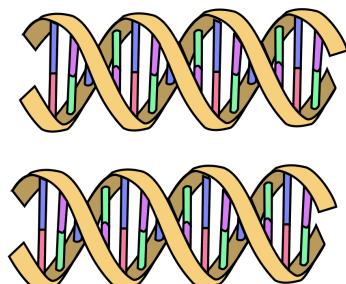
....TAAACGGTCAGTCATGGGTCTACGACCT....

Human Genetics

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



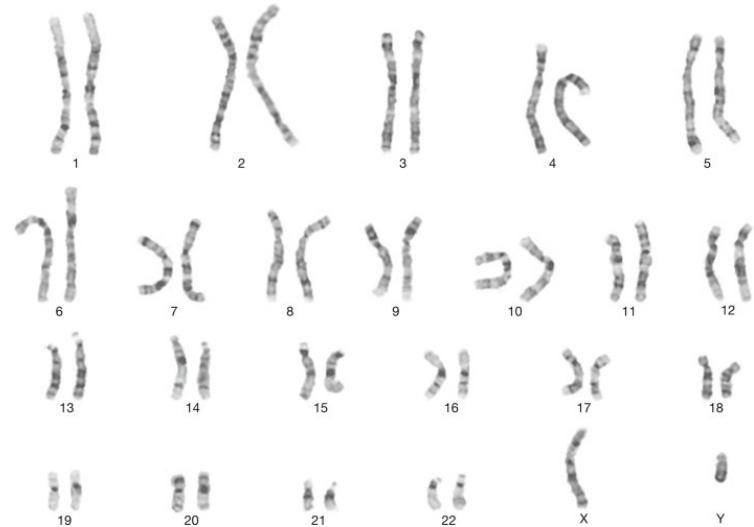
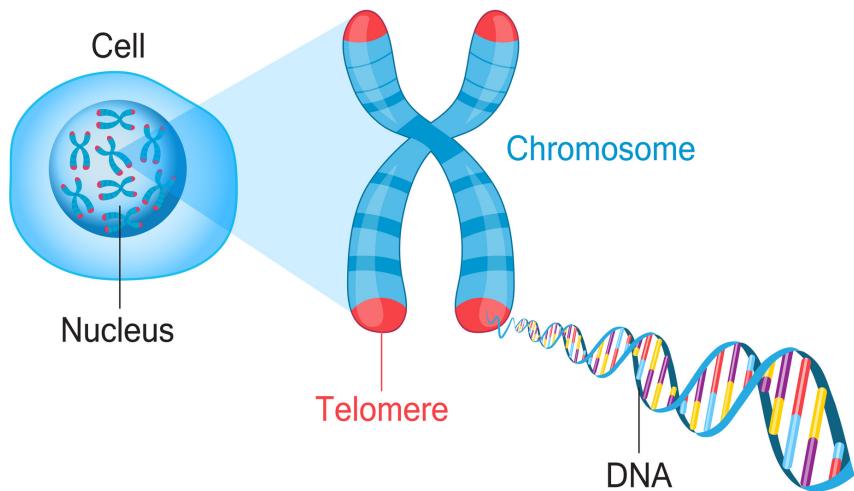
....ATTGCCAGTCAGTACCCAGGATGCTGGA....

....TAAACGGTCAGTCATGGGTCTACGACCT....

....ATTGCCAGTCAGTACCCAGGATGCTGGA....

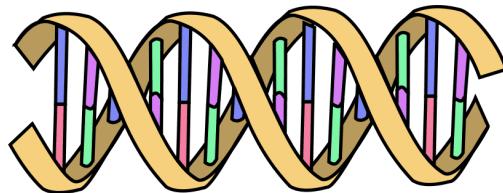
....TAAACGGTCAGTCATGGGTCTACGACCT....

Chromosomes

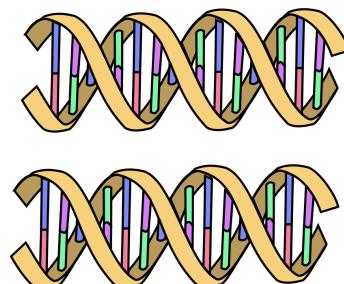


Human Genetics

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



Copy 1, Chromosome 1 (Mother)

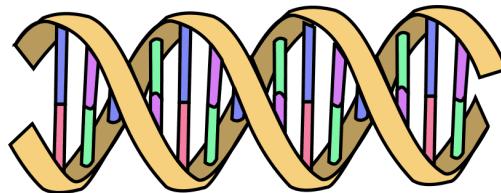
....ATTGCAAGTCAGTACCGAGGATGCTGGA....

....ATTGCCAGTCAGTACCCAGGATGCTGGA....

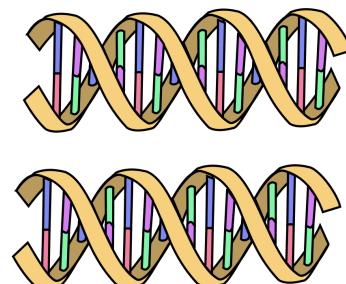
Copy 2, Chromosome 1 (Father)

Human Genetics

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

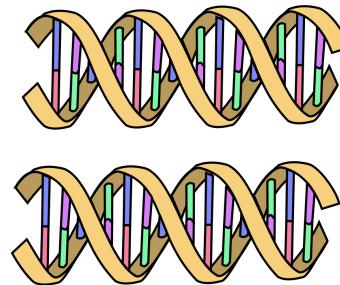


....ATTGCAAGTCAGTACC**G**AGGATGCTGGA....

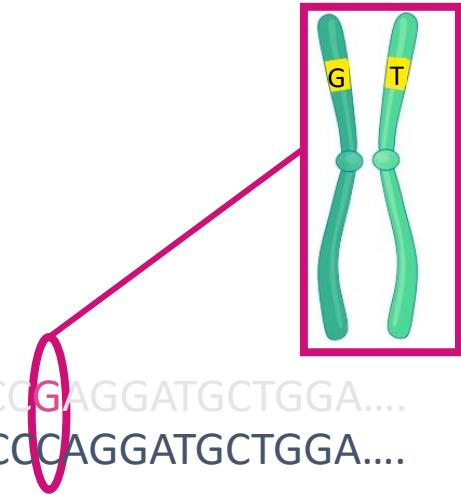
....AT**C**TGCCAG**A**CAGTACCCAGGATGCTGGA....

Human Alleles

- An allele is different versions of the same variant in a specific locus in a chromosome.
 - For example, a SNP may have two alternative bases, or alleles, C and T

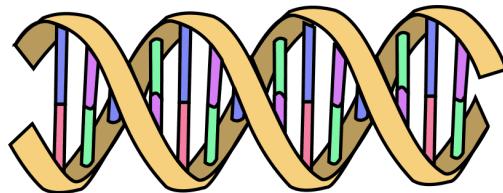


....ATTGCAAGTCAGTACCGAGGATGCTGGA....
....ATCTGCCAGACAGTACCCAGGATGCTGGA....

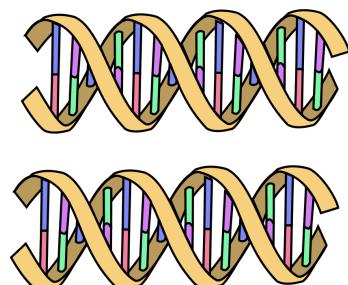


Human Genetics

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



....ATTGCAAGTCAGTACC**G**AGGATGCTGGA....

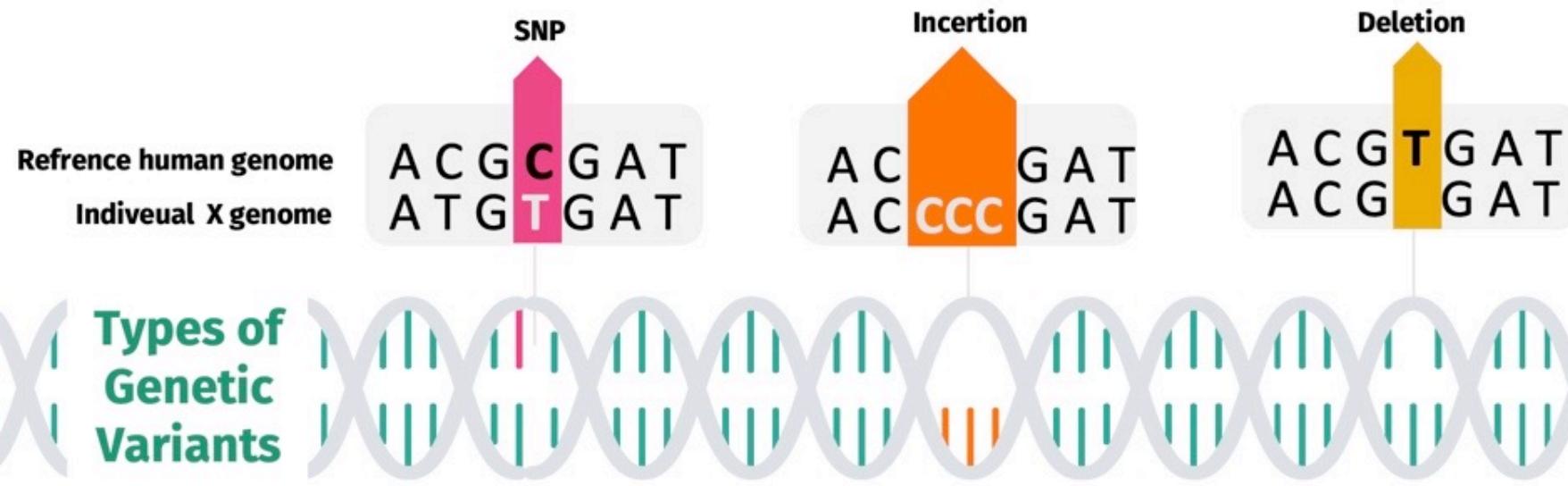
....AT**C**TGCCAG**A**CAGTACCCAGGATGCTGGA....

Single Nucleotide Polymorphism
(SNP)

Genetic Variants Forms



An alteration in the most common DNA nucleotide sequence



Sequence Variants

SNV (Single Nucleotide Variant)

Ref	A	A	G	G	G	C	T	G
Query	A	A	G	G	A	C	T	G

_____ | _____

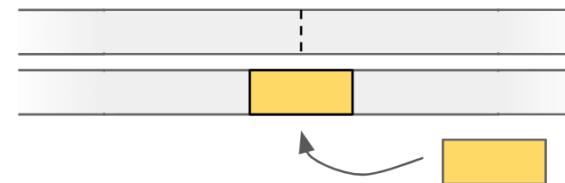
INDEL (Insertion or Deletion)

Ref	A	A	G	G	G	C	T	G
Query	A	A	G	-----	C	T	G	

_____ | _____

Structural Variants

Insertion

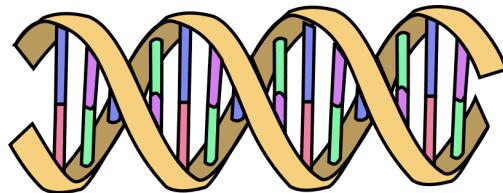


Inversion

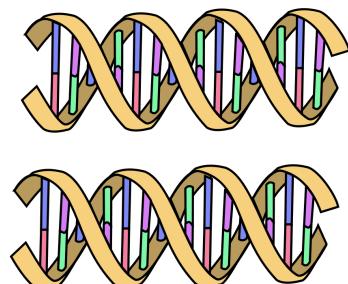


Human Genotype

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



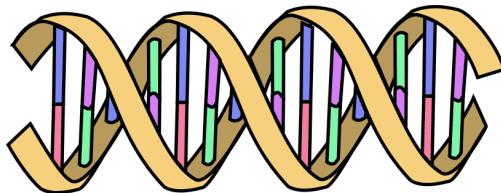
....ATTGCAAGTCAGTACC**G**AGGATGCTGGA....

....AT**C**TGCCAG**A**CAGTACCCAGGATGCTGGA....

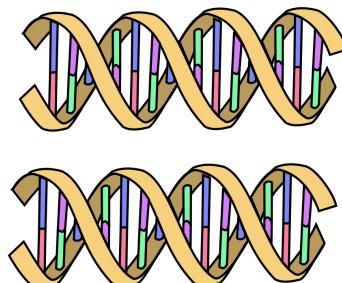
What is a genotype? (A/T)

Human Phenotype

DNA



....ATTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



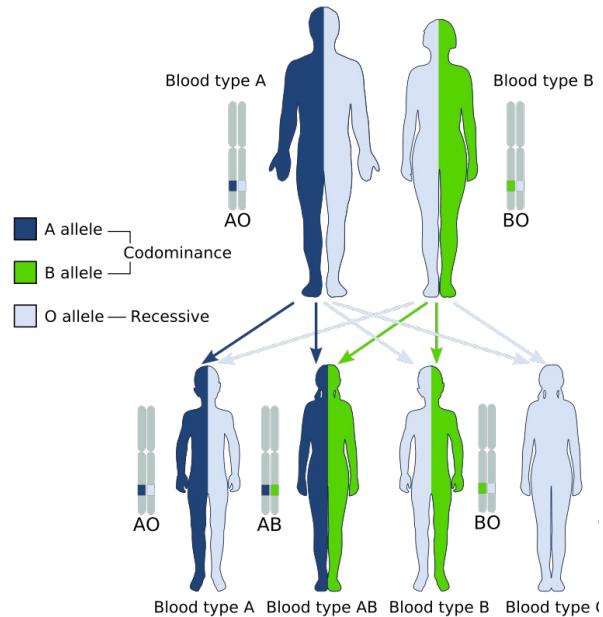
....ATTGCAAGTCAGTACCAGGGATGCTGGA....

....ATCTGCCAGACAGTACCCAGGATGCTGGA....

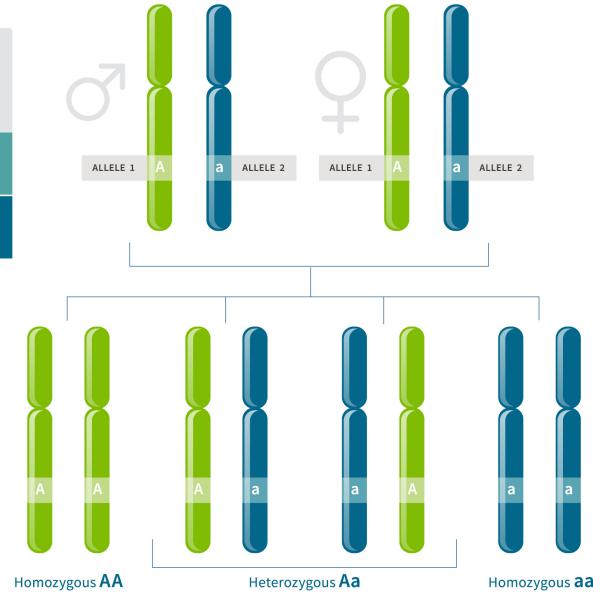


What is a phenotype? (Dark brown eyes)

Genotype and Phenotype



	A	AA	Aa
	a	Aa	aa



In a pair of homologous chromosomes, one is inherited from the male parent, and the other from the female parent.

Paternal homologue

Maternal homologue

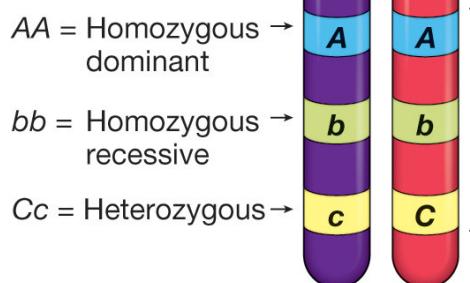
A genetic locus is the location of a particular gene on a chromosome.

At each genetic locus, an individual has two alleles, one on each homologous chromosome.

AA = Homozygous dominant

bb = Homozygous recessive

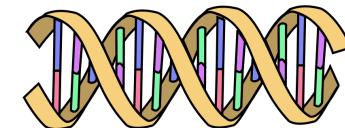
Cc = Heterozygous



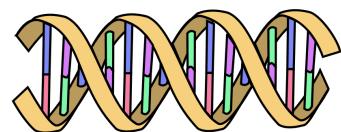
Three gene pairs at three different loci

Human Genetics

DNA

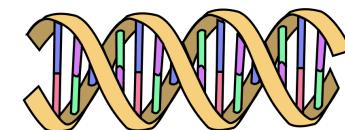


....ATTGCCAGTCAGTACCG**G**AGGGATGCTGGA....

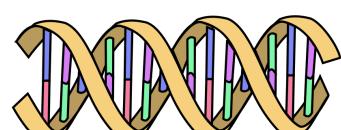


....TAAACGGTC**A**GTCATGGGTCC**T**GACCT....

DNA



....ATTGCCAGTCAGTACCG**G**AGGGATGCTGGA....



....TAAACGGTC**A**GTCATGGGTCC**T**GACCT....

Human Genetics

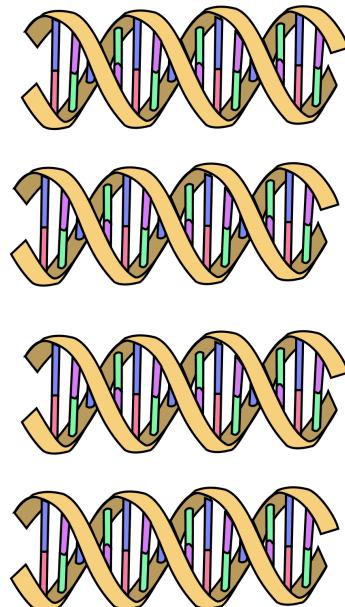


DNA

?



DNA



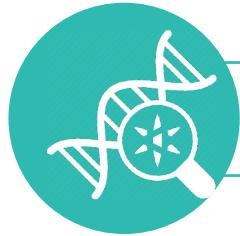
....ATTGCCAGTCAGTACCG**G**AGGGATGCTGGA....

....TAAACGGTC**A**GTCATGGGTCC**T**GACCT....

....ATTGCCAGTCAGTACCG**G**AGGGATGCTGGA....

....TAAACGGTC**A**GTCATGGGTCC**T**GACCT....

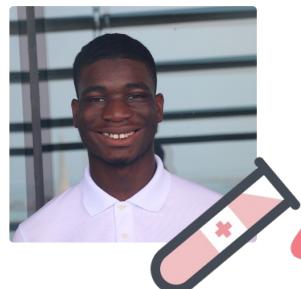
DNA Sequencing

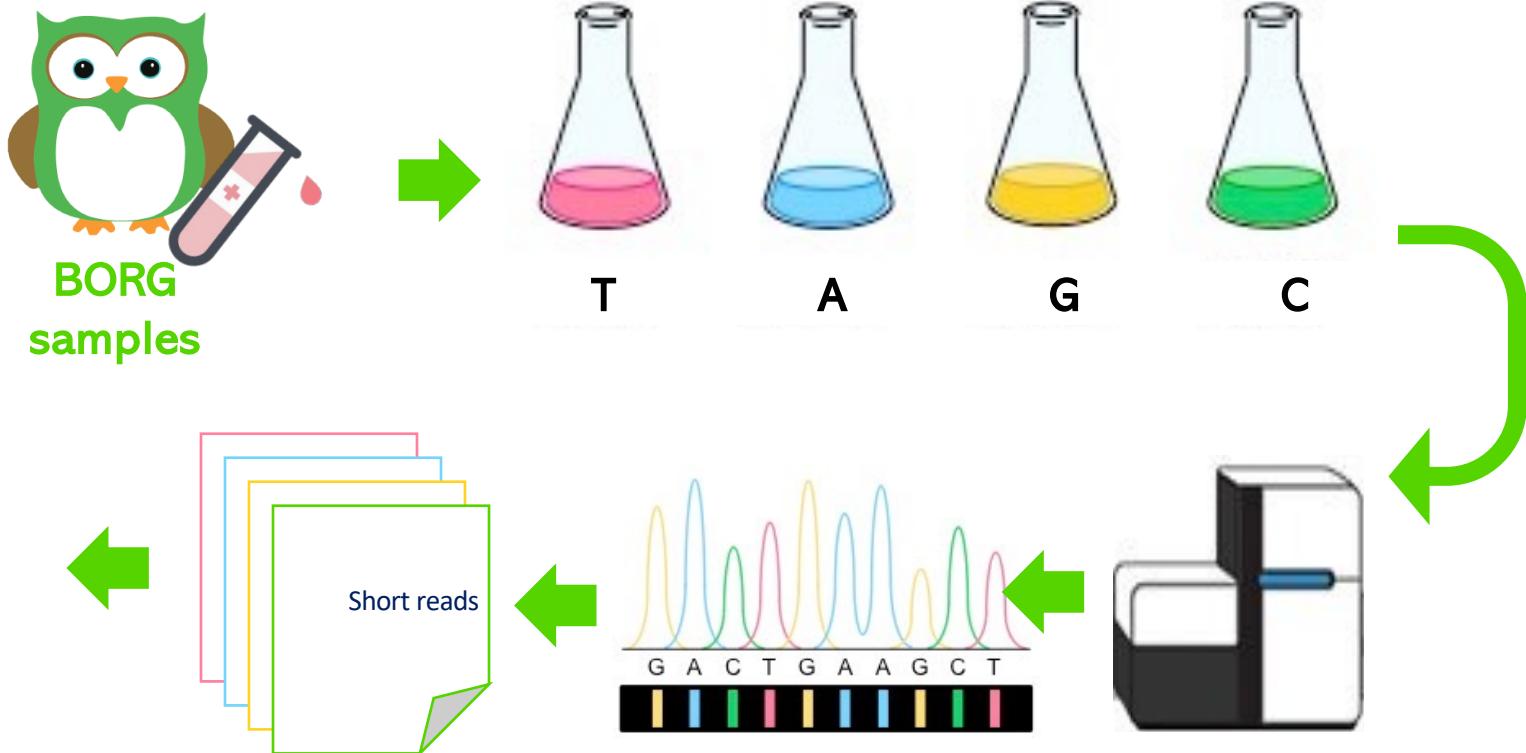


DNA sequencing is the process of determining the order of nucleotides in DNA.

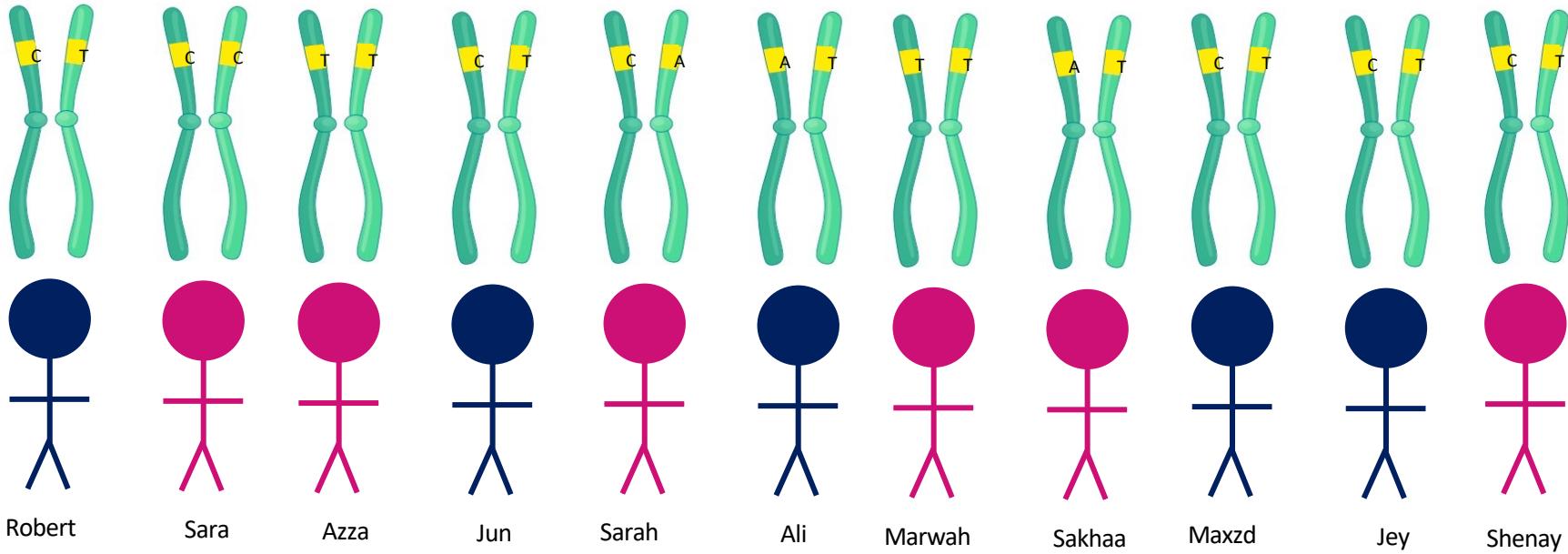


KAUST





How we calculate AF?



BORG TEAM Cohort example: NOT RELATED TO SCIENCES ... BUT IT IS TRUE example

How We Calculate AF?

AA

$$P(A) = \frac{0+1+2}{11*2} = 0.136$$

AC

AT

$$q(C) = \frac{1+5+(2*1)}{11*2} = 0.364$$

TC

TT

$$z(T) = \frac{2+5+(2*2)}{11*2} = 0.5$$

CC

$$P(A) + q(C) + Z(T)$$

Allele Frequency = $\frac{\text{Number of copies of a particular allele in the population Total}}{\text{number of gene copies in the population}}$

How We Calculate AF?

AA

$$P(A) = \frac{0+1+2}{11*2} = 0.136$$

AC

AT

$$q(C) = \frac{1+5+(2*1)}{11*2} = 0.364$$

TC

TT

CC

MAF

$$z(T) = \frac{2+5+(2*2)}{11*2} = 0.5$$

$$P(A) + q(C) + Z(T)$$

Allele Frequency = $\frac{\text{Number of copies of a particular allele in the population Total}}{\text{number of gene copies in the population}}$

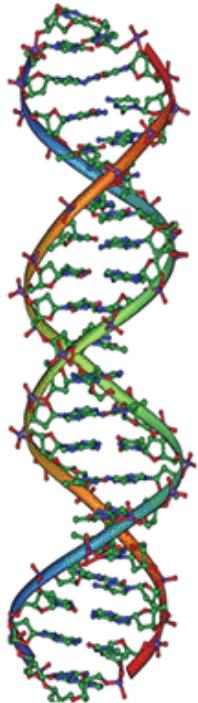
Introduction to Bioinformatics

Calling variants in diploid systems

Part 2: Variant Calling

Basic DNA Sequencing Workflow

Human genome



Short reads

.....

```
GGTCTGGATGC  
CGGTCTGGATGC  
GCGGTCTGGATG  
GCGGTCTGGAT  
GGCGGTCTGGAT  
GGCGGTCTGGA  
TCTATGCAGGGCCCT  
TCTATGCAGGGCCC  
ATCTATGCAGGGC  
TATCTATGCAGGGC  
TTATCTATGCAGGG  
CTTATCTATGCAGGG
```



Alignment of reads to
the reference genome
and SNP calling

SNP:A->G

GTCTGGATGCT	TCTATGCAGGGCCCT
GGTCTGGATGC	TCTATGCAGGGCCC
CAGGTCTGGATGC	ATCTATGCAGGGCC
GCGGTCTGGATG	TATCTATGCAGGGC
GCGGTCTGGAT	TTATCTATGCAGGG
GCGGTCTGGAT	CTTATCTATGCAGGG
GCGGTCTGGAT	CTTATCTATGCAGGG
GCGGTCTGGA	CTTATCTATGCAGGG

GGCGGTCTAGATGCTTATCTATGCAGGGCCCT

Reference genome sequence

Variant Identification and Analysis



Variant and Genotype Calling

- **Variant calling** - identification of positions where the sequenced sample is different from the reference

Aspect	Variant Calling	Genotype Calling
Definition	Identifying differences between DNA sequences and a reference genome	Determining specific alleles an individual carries at genomic positions based on sequencing data
Process	Analyzes sequenced DNA reads to detect variations (SNPs, insertions, deletions)	Assigns genotypes (allele combinations) to variants in each sample
Objective	List of genomic positions with identified variants	Genotype calls for each variant in each sample (e.g., AA, AG, GG for a SNP)
Application	Crucial for understanding genetic variation in populations, diseases, or evolution	Essential for studying heritability, population genetics, and disease association studies

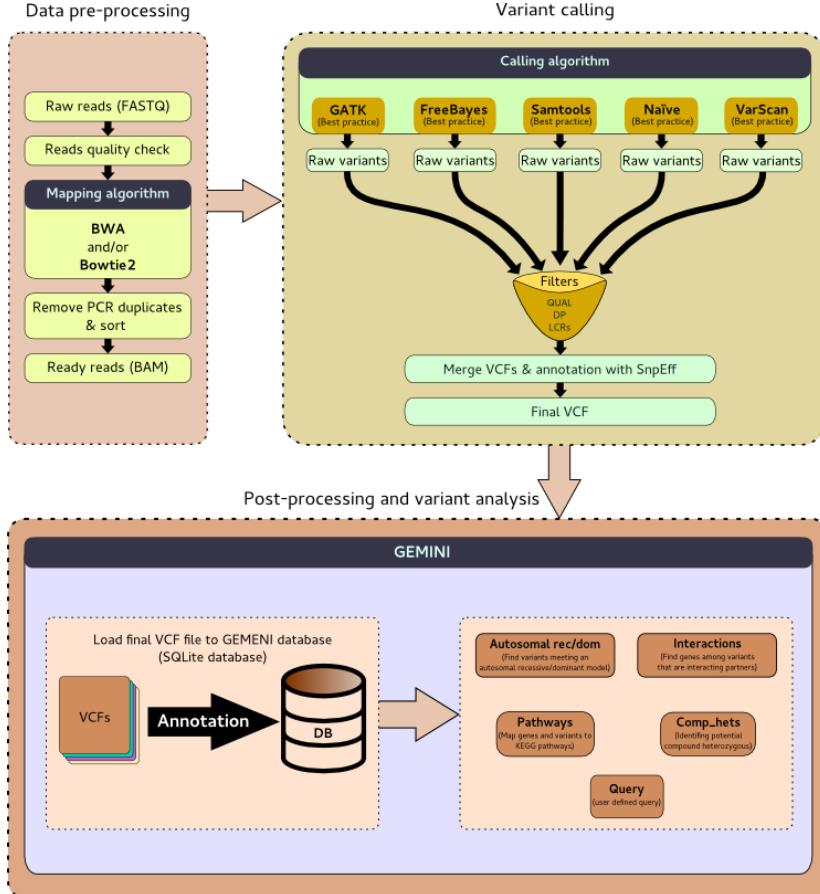
Advantages and features of variant calling



- **Abundance:** In-depth analysis of all aspects of genetic variation, including SNP, InDel, SNV, novel gene, et al.
- **Flexibility:** with or without reference is suitable
- **Accuracy:** different sequencing methods can be applied based on different material



Variant Calling Workflow



Calling with FreeBayes



أكاديمية كاوهست
KAUST ACADEMY

	Variant Region		Variant Region	
Ref	TACCGAT	CATTGGATCA	CGATTCC...GCATTGC	AAAAAAA-
Reads	TACCGAT	CATTGGATCA	CGATTCC...GCATTGC	-AAAAAA-
	ACCGAT	TATTGCATCG	CGATTCC...GCATTGC	-AAAAAA-
	ACCGAT	CATTGGATCA	CGATTCC...GCATTGC	AAAAAA-A
	ACCGAT	TATTGGATCG	CGATTCC...GCATTGC	-AAAAAAA
	CCGAT	C-TTGGATCA	CGATTCC...GCATTGC	AAAAAAA-
	CCGAT	CATGGGATCA	CGATTCC...GCATTGC	AAAAAAA-A
• • •	• • •	• • •	• • •	• • •
Observed Haplotypes	CATTGGATCA x8		(A)₇ x10	
	TATTGGATCG x9		(A)₆ x7	
	CTTGGATCA x1		(A)₅ x1	
	CATGGGATCA x1		(A)₈ x1	
• • •	• • •		• • •	



Hands-on and Practical Part



Part 1: Variant analysis
➤ Calling variants

Introduction to Bioinformatics

Molecular identification of phenotype-causing mutations

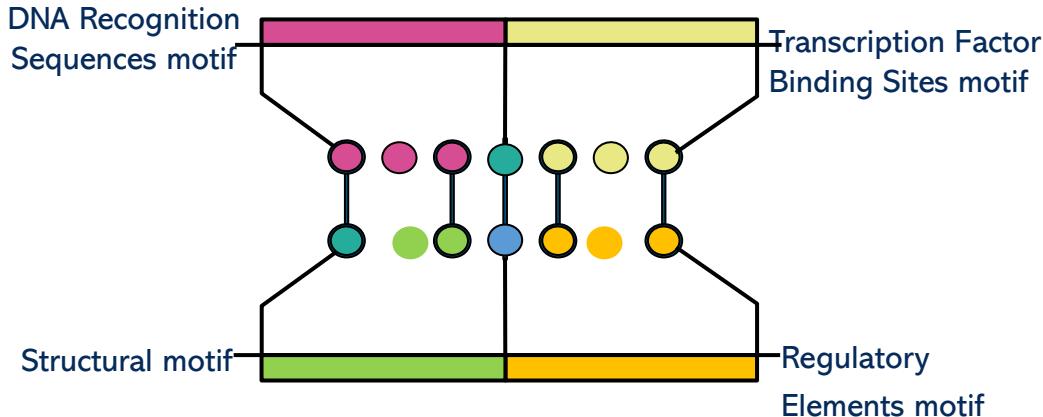
Part 3: Variant annotations

Finding Hidden Messages in DNA



DNA Motifs: Decoding the Language of Genes

DNA motifs is a specific recurring sequence patterns in DNA



GTG_{green}CATCTGA_{yellow}CTC_{yellow}CTGAGGAGAAG
CAC_{yellow}GTAG_{green}ACTGAGGACTC_{magenta}CTCTTC
GTG_{green}CATCTGA_{yellow}CTC_{yellow}CTGAGGAGAAG
CAC_{yellow}GTAG_{green}ACTGAGG_{yellow}TG_{green}CATCTGAC
CCT_{yellow}GAGGAGAAGCAC_{green}GTAG_{yellow}ACTGG
GACTC_{magenta}CTCTTC_{green}GACTC_{magenta}CTCTTC_{green}GTG_{yellow}C
GA_{green}CTC_{yellow}CTGAGGAGAAGCAC_{green}GTAGA
CTGAGGACTC_{magenta}CTCTTCATTGCCTT

Understanding the Significance of Sequence Patterns

Example of DNA Motifs

1. Transcription Factor Binding Sites (TFBS):

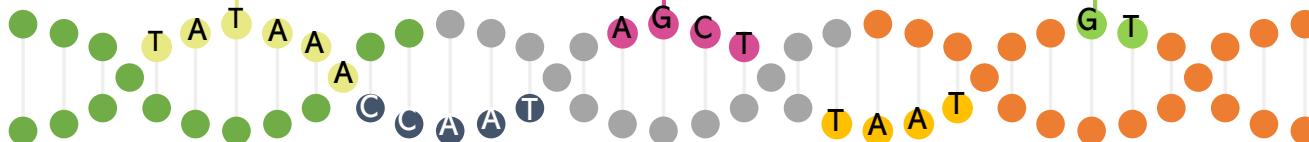
- Motif: TATA Box (TATAAA)
- Group: Promoter Motif
- Function: It's a promoter motif that helps initiate transcription by binding with transcription factors.

3. Repetitive Elements:

- Motif: Alu Sequence (AGCT)
- Group: Repetitive Motif
- Function: These sequences are repeated many times throughout the genome and may play a role in genetic recombination.

5. Splicing Sites:

- Motif: 5' Splice Site (GT)
- Group: Splicing Motif
- Function: This motif marks the beginning of an intron and is crucial for mRNA splicing.



2. Enhancer Motifs:

- Motif: CAAT Box (CCAAT)
- Group: Enhancer Motif
- Function: Enhancer motifs increase the rate of transcription by binding to enhancer regions and facilitating the binding of transcription factors.

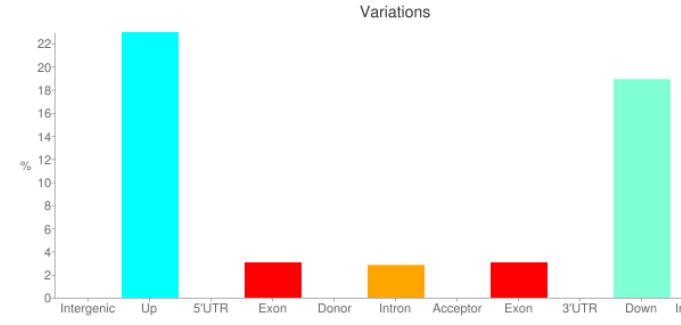
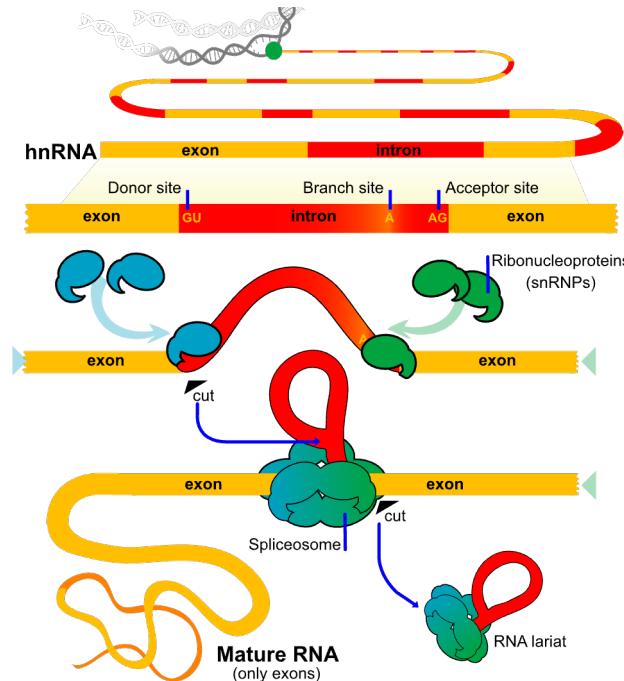
4. Protein Binding Sites:

- Motif: Homeodomain (TAAT)
- Group: Protein Interaction Motif
- Function: This motif is recognized by specific proteins (homeodomain proteins) involved in gene regulation and development.

Genetic Variants Annotation



أكاديمية كاوهست
KAUST ACADEMY



	ATC	ATT	CAA	CAC	CAT	CCC	CCG	CTG	GAA	GAC	GAG	GAT	GCA	GCC	GCT	GGA	GGG	GTG	TAC	TCC	TGG	TTG
ATC	2																					
ATT																						
CAA			1																			
CAC				2	1																	
CAT																						
CCC							1															
CCG																						
CTG																						
GAA																1						
GAC																						
GAG																	2					
GAT																						
GCA																						
GCC																						
GCT															1	1						
GGA																						
GGG																						
GTG																	1					
TAC																			1			
TCC																				1		
TGG																					1	
TTG																						1



Hands-on and Practical Part



Part 1: Variant analysis

- Annotating variants
- Manipulating variation data



Introduction to Bioinformatics

Molecular identification of phenotype-causing mutations

Part 3: Mapping-by-Sequencing

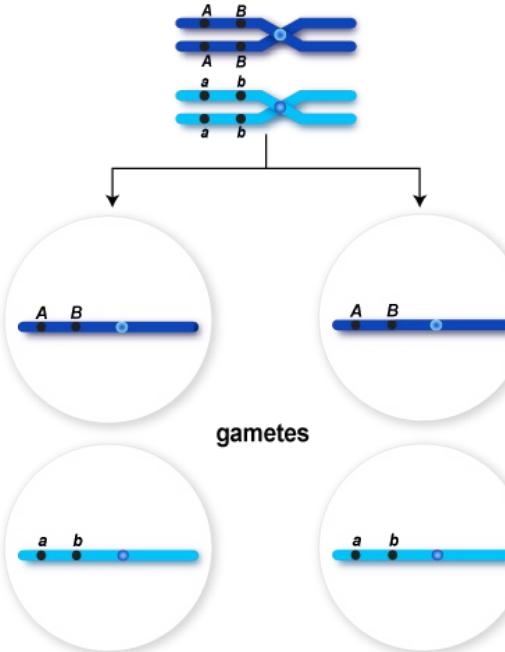


Will start with basic genetics concepts...

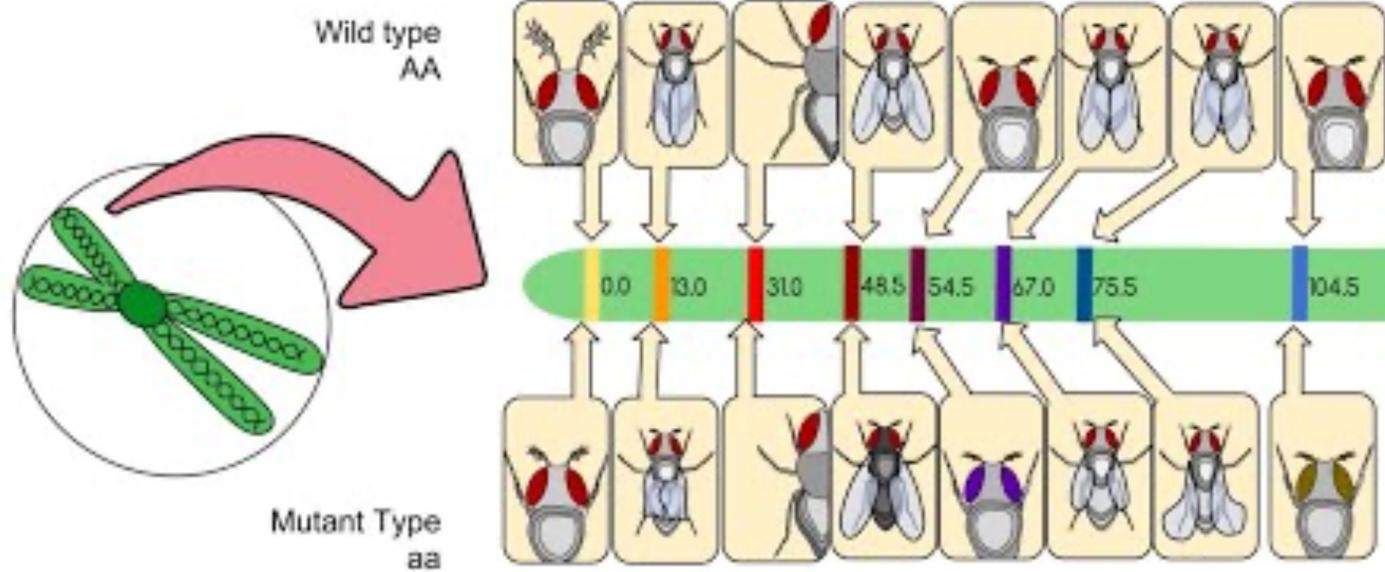
Genetic Linkage

The tendency of genes to stay together on a chromosome.

Linkage of two genes on one chromosome without change

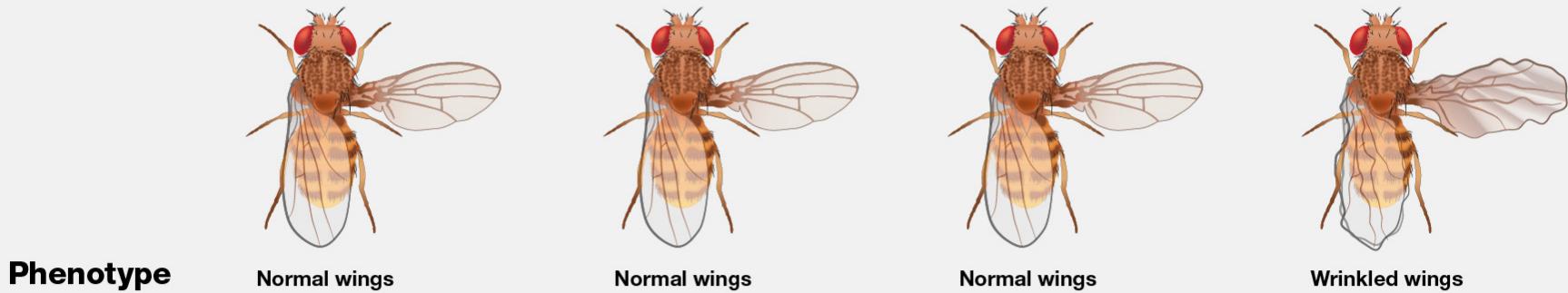
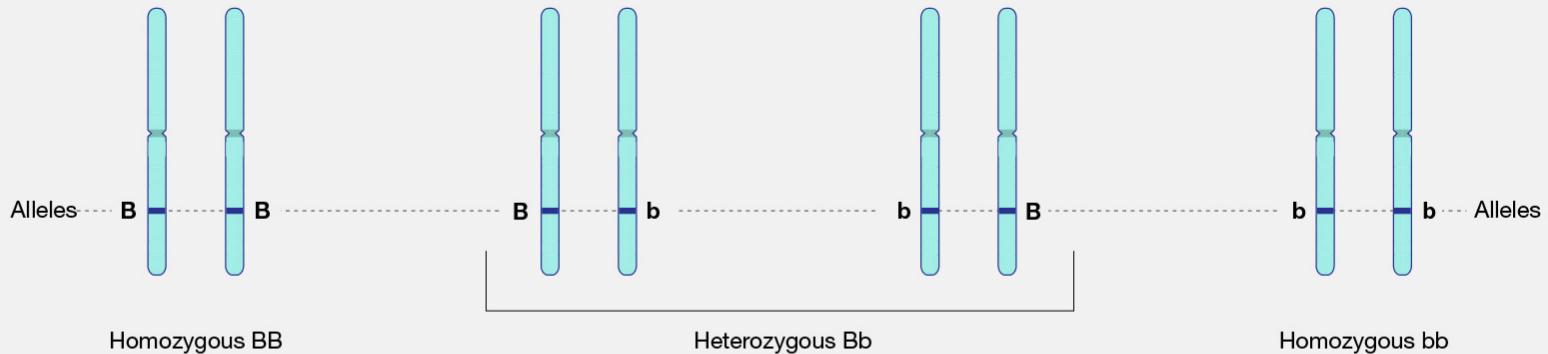


Sturtevant and his colleagues were able to map many of the fruit fly genes in this way



Linkage	Crossing Over
It keeps the gene together.	It leads to separation of linked genes.
DNA sequences that are close together on a chromosome has a ability to be inherited together during meiosis phase of sexual reproduction.	Exchange of genes between two chromosomes
It involves individual chromosomes	It involves non - sister chromatids of homologous chromosomes
Degree of linkage is inversely proportional to distance between two genes	The probability of two genes crossing over is directly proportional to distance between two genes

The paternal characteristics are passed down to offsprings	Changes in the parental characteristics occurs
Reduces with age	Doesn't reduce with age
Reduces variability	Increases variability.



Phenotype-causing mutations

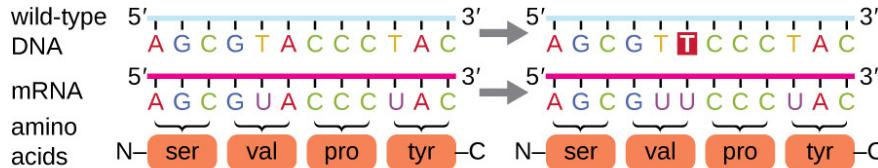


Phenotypic mutations are errors that occur during protein synthesis. These errors can lead to amino acid substitutions that create abnormal proteins.

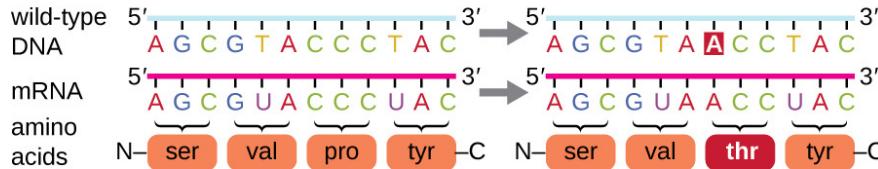


point mutation: substitution of a single base

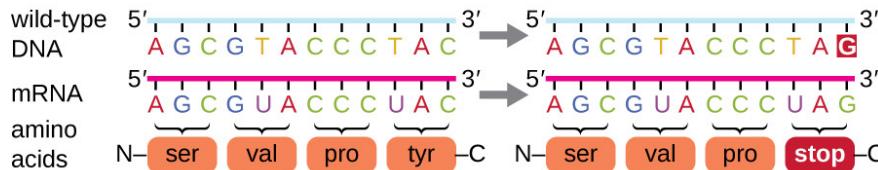
silent: has no effect on the protein sequence



missense: results in an amino acid substitution

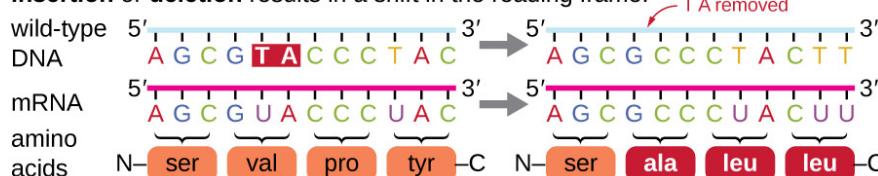


nonsense: substitutes a stop codon for an amino acid



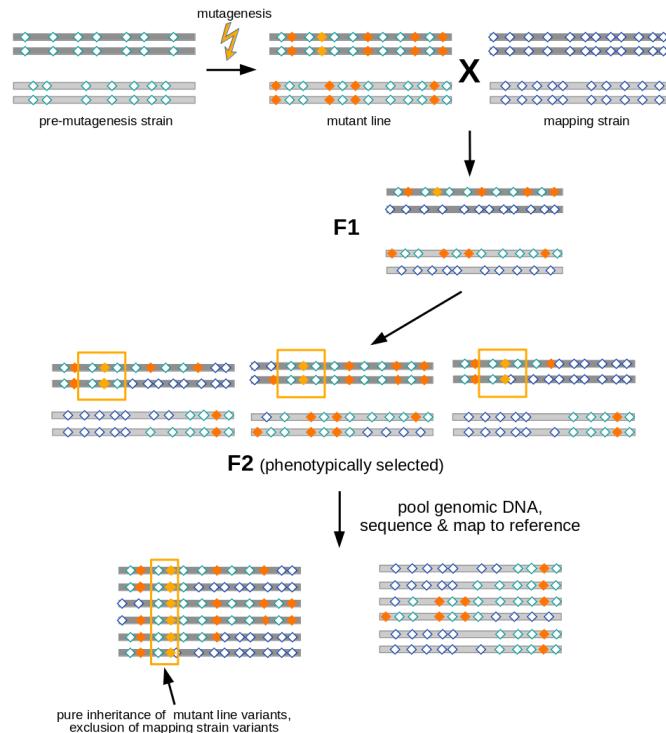
frameshift mutation: insertion or deletion of one or more bases

Insertion or **deletion** results in a shift in the reading frame.



Mapping-by-sequencing

Mapping-by-sequencing combines genetic mapping with whole-genome sequencing in order to accelerate mutant identification.



Hence, when the recombined DNA from phenotypic F2 progeny gets pooled and sequenced, the causative variant can be found by looking for a region for which all sequenced reads support variant alleles found in the original mutant line, rather than mapping strain variant alleles.

By following these steps, researchers can narrow down the location of the causative mutation and identify the specific genetic variant responsible for a particular phenotype of interest

Hands-on and Practical Part



Part 1: Mapping by sequencing

- [Manipulating variation data](#)
- [Data Preparation](#)
- [Joint Variant Calling and Extraction](#)
- [Linkage Analysis](#)
- [Identifying Candidate Mutations](#)



Done with Day 2, Heyyyyy!

Thank You !