أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Instructor :Sakhaa Alsaedi
TA: Ebtihal Hani
Sakhaa.Alsaedi@kaust.edu.sa

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

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Exploring the Central Dogma of Molecular Biology

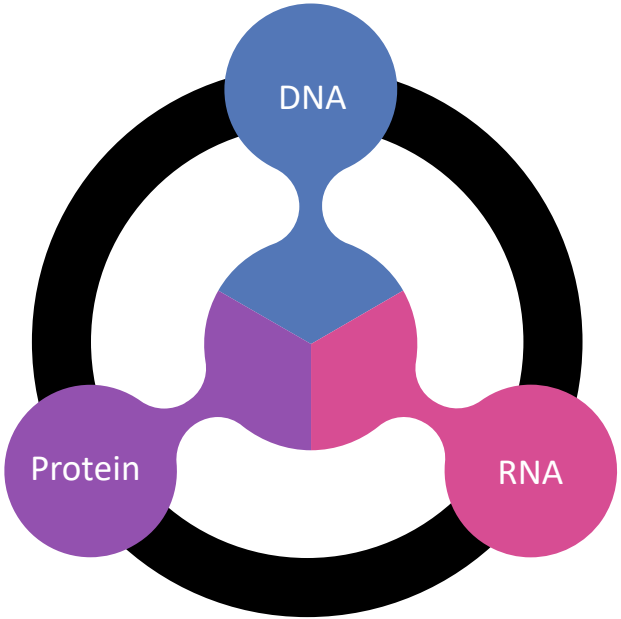
Part 1: Introduction

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

What is Molecular Biology?

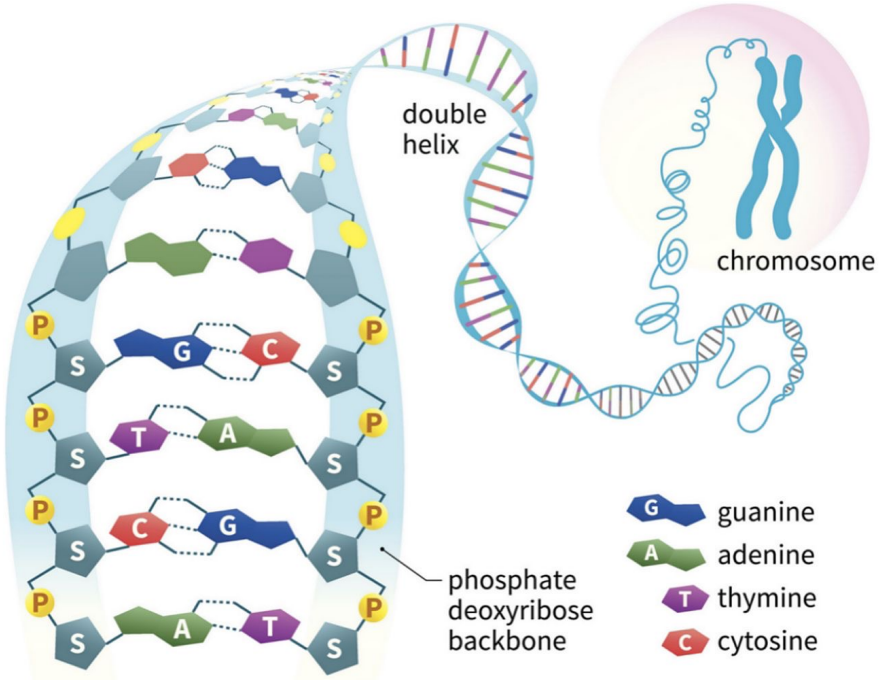
The study of structure and function of nucleic acids and proteins of biological molecules to understand the molecular mechanisms underlying various biological processes.

The Central Dogma

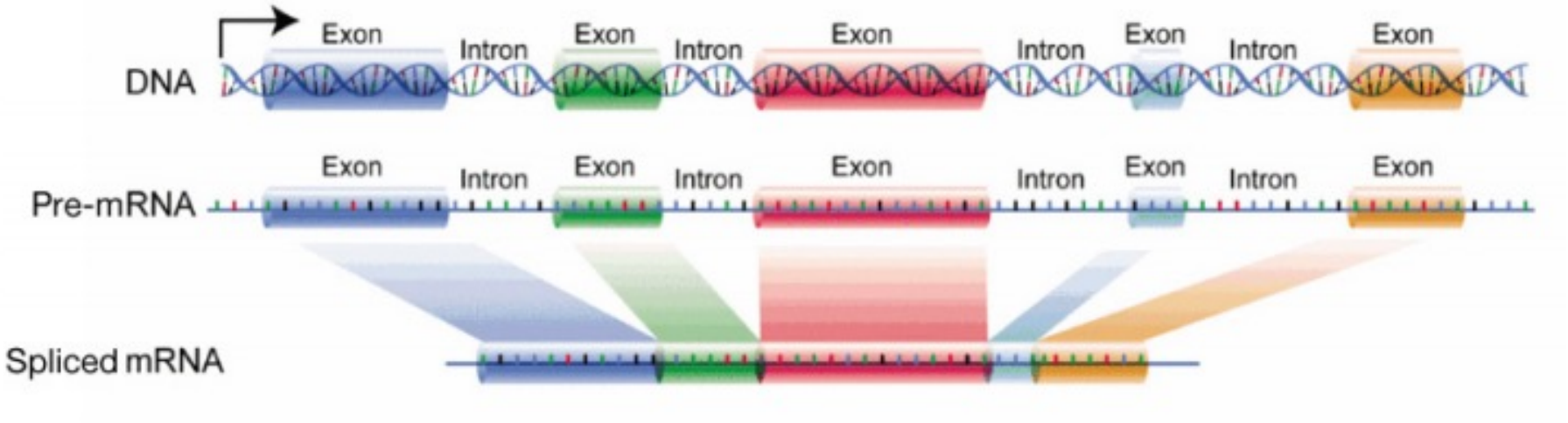


Human Genome

- 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

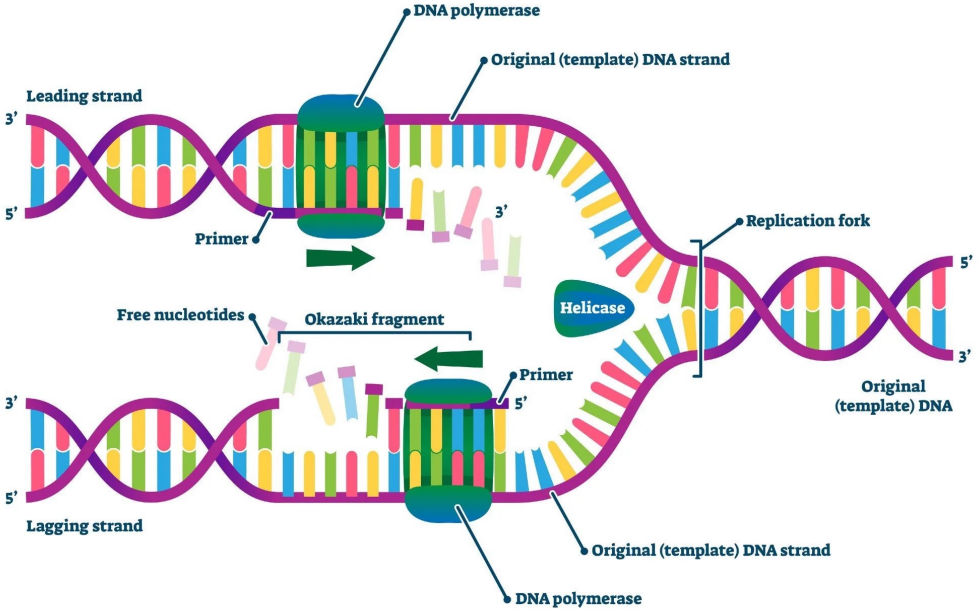
DNA Replication

RNA Transcription

RNA Translation

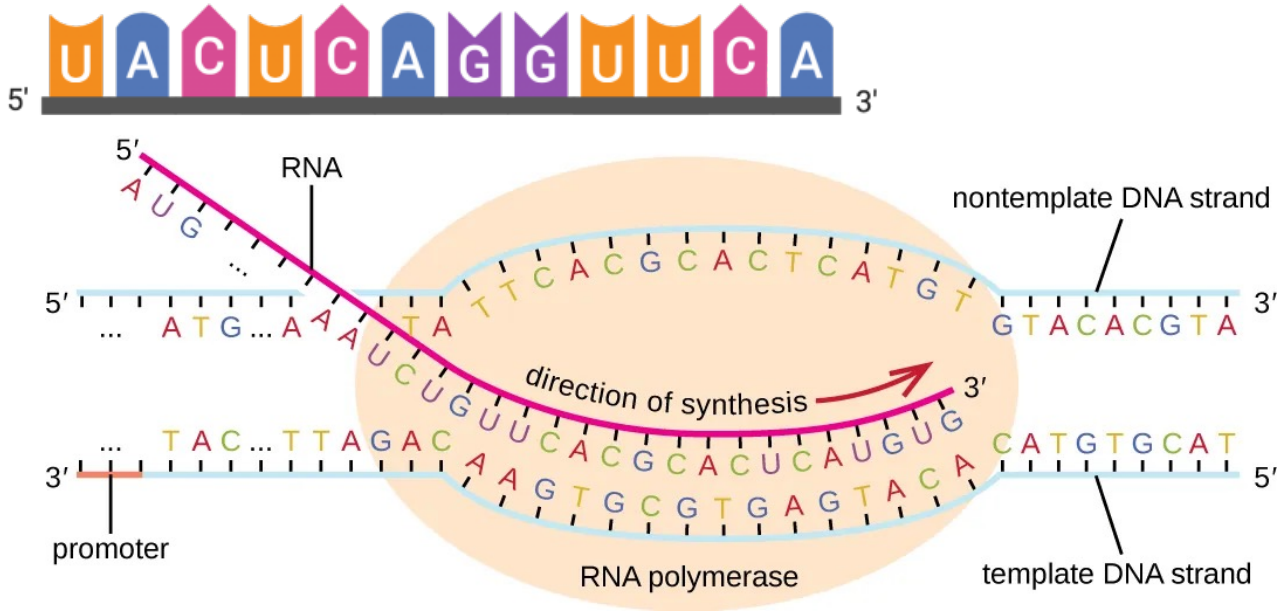
DNA Replication and Repair: Understanding how DNA is copied (replication) and maintained (repair) is crucial for maintaining genetic stability

- Helicase
- DNA polymerase



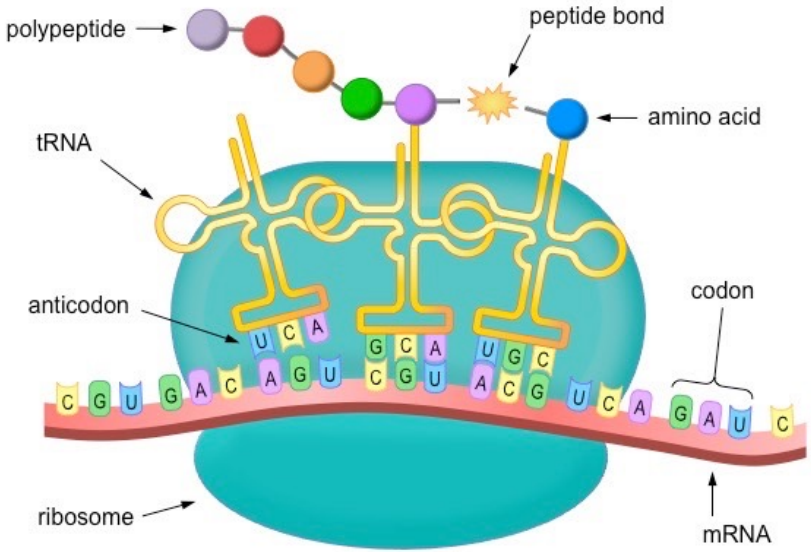
Transcription: Molecular biologists study the process by which genetic information encoded in DNA is transcribed into RNA molecules, particularly messenger RNA (mRNA), which serves as a template for protein synthesis

- RNA polymerase



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.

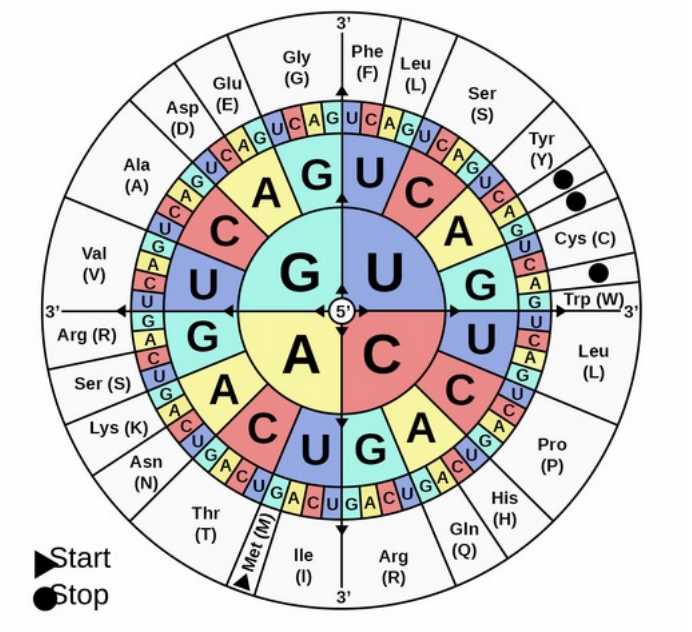
- Ribosome
- tRNA



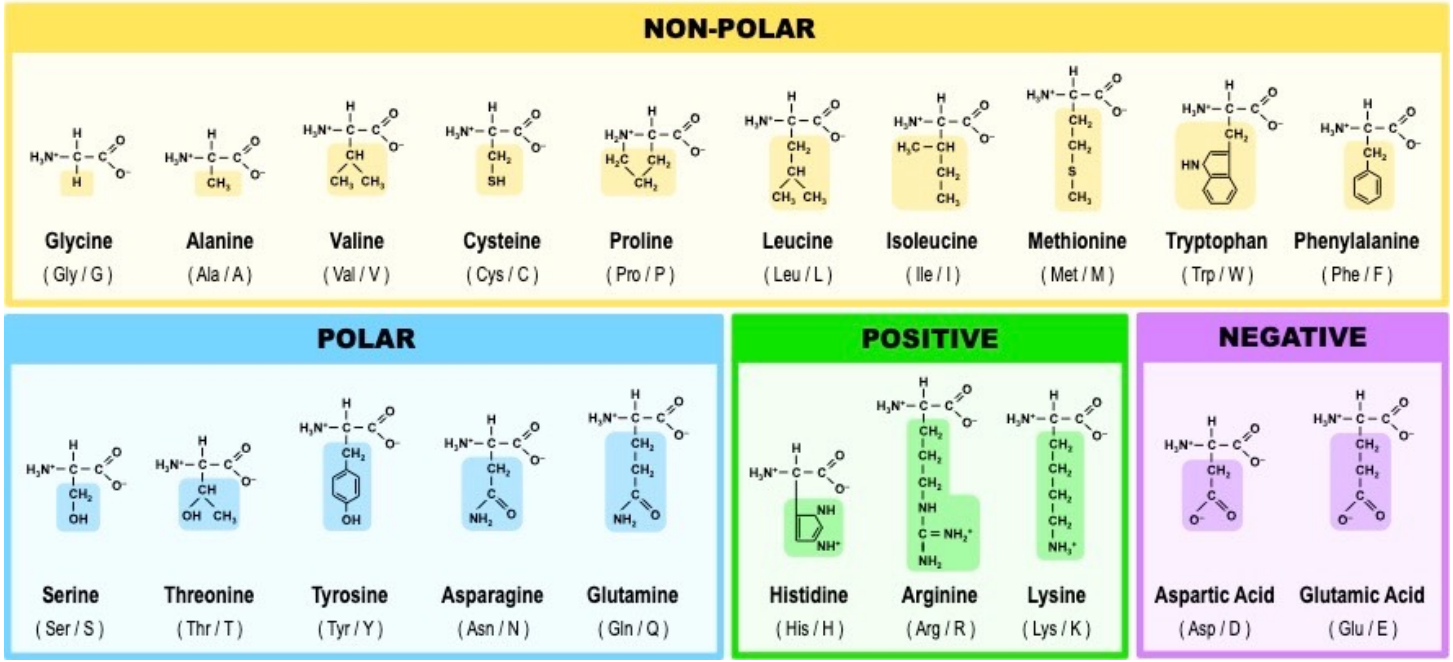
DNA CODE

SECOND POSITION OF CODON

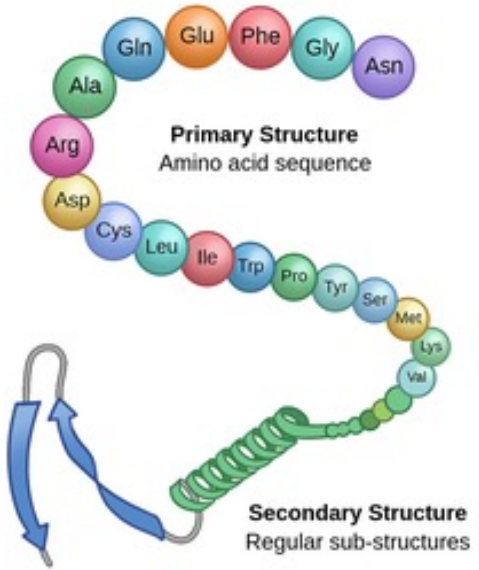
FIRST POSITION OF CODON	T	C	A	G
T	ttt	tct	tat	tgt
	ttc	tcc	tac	tgc
	tta	tca	taa	tga
	ttg	tcg	tag	tgg
C	ctt	cct	cat	cgt
	ctc	ccc	cac	cgc
	cta	cca	caa	cga
	ctg	ccg	cag	cgg
A	att	act	aat	agt
	atc	acc	aac	agc
	ata	aca	aaa	aga
	atg	acg	aag	agg
G	gtt	gct	gat	ggt
	gtc	gcc	gac	ggc
	gta	gca	gaa	gga
	gtg	gcg	gag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

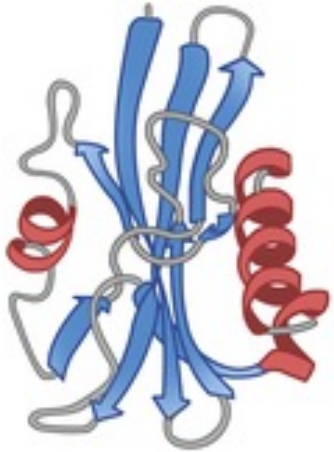


Amino Acids Types



Protein Folding and Structure

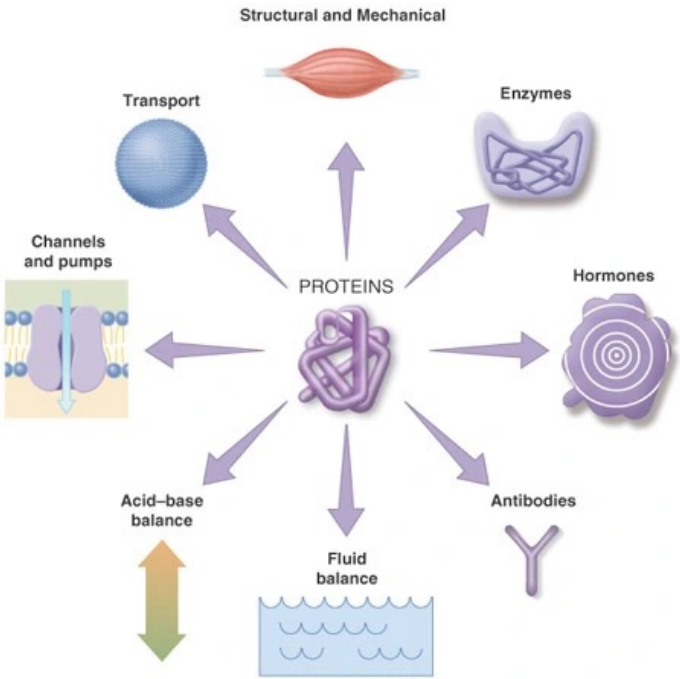




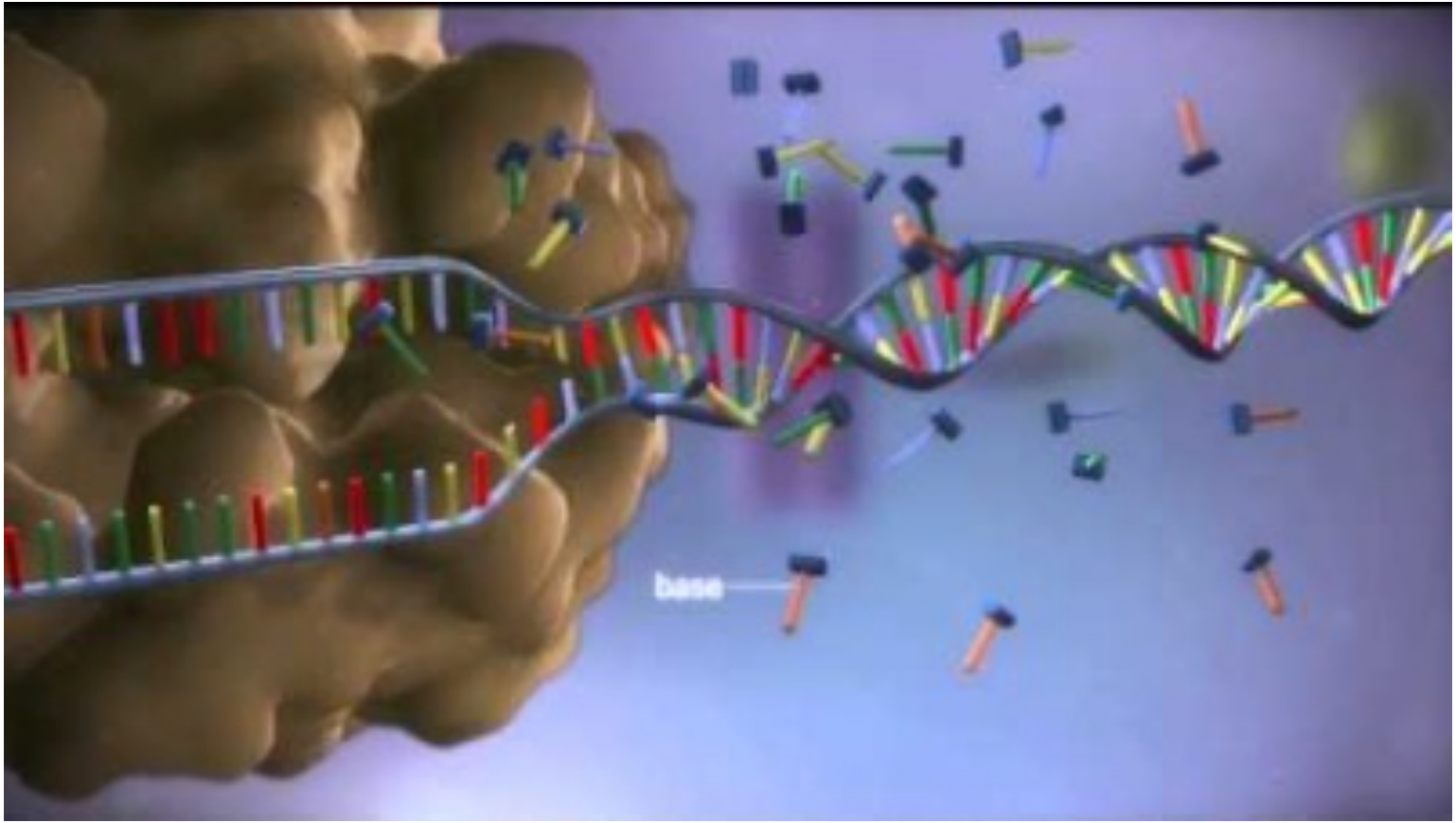
Tertiary Structure
Three-dimensional structure

Quaternary Structure
complex of protein molecules

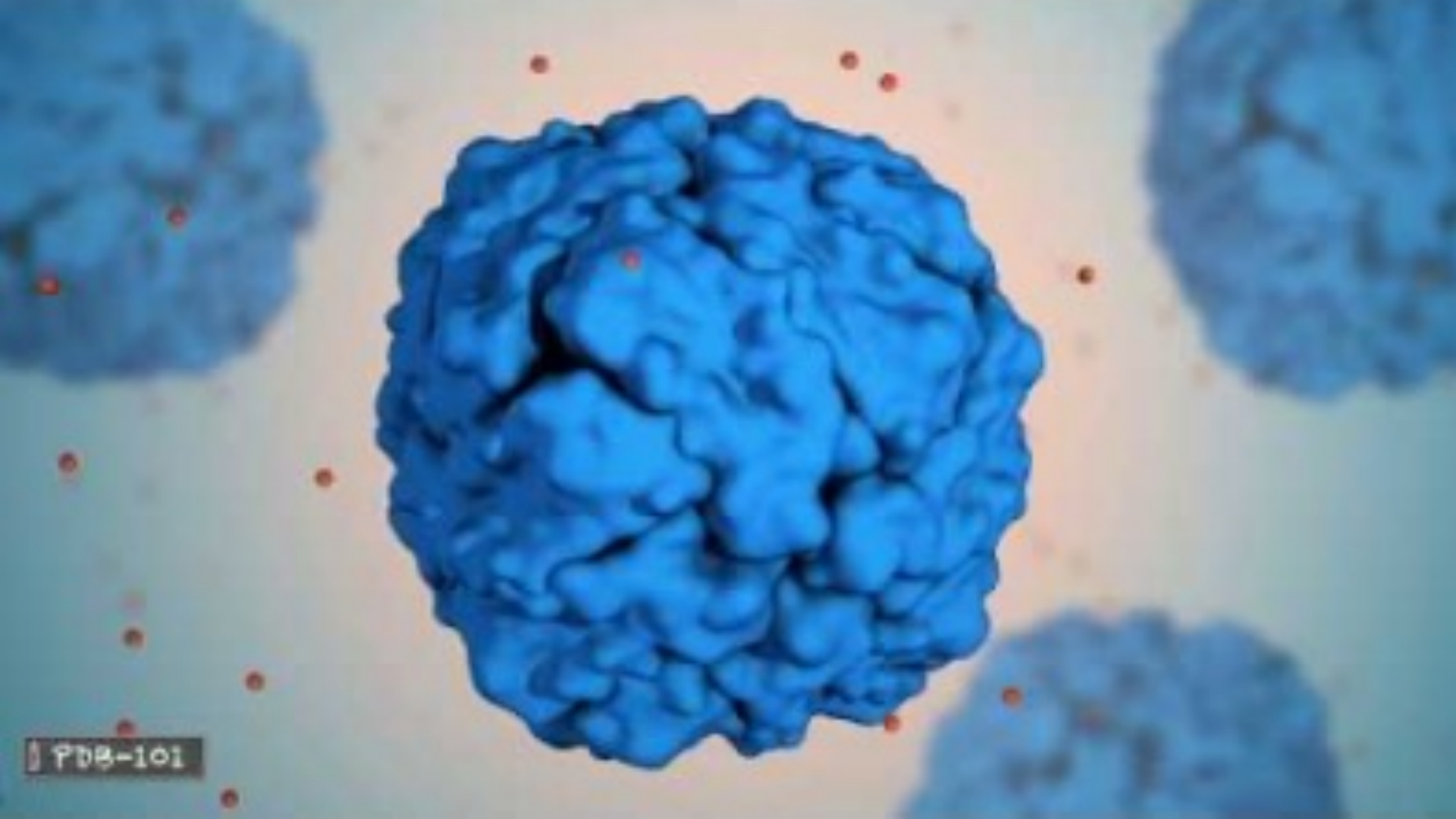
Protein Functions



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


PDB-101

أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Unraveling the Human Genome: Basics and Beyond

Part 2: Basic Genome

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

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

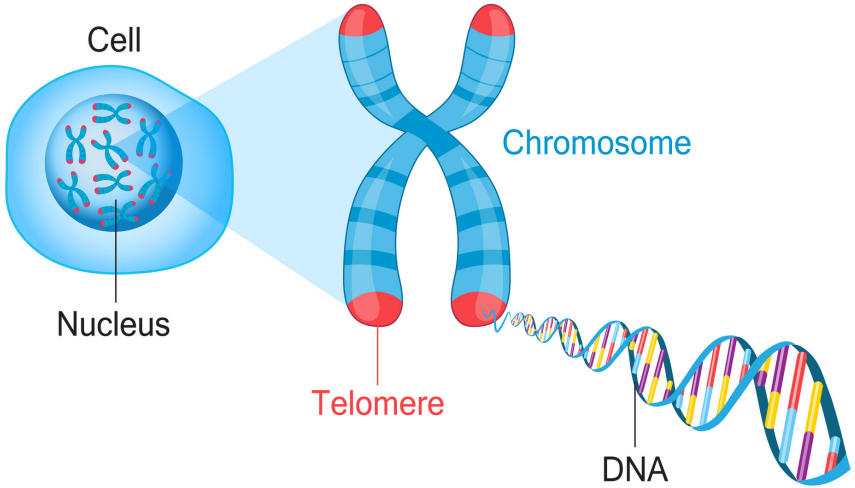
- Genome: The genome refers to the complete set of genetic material (DNA) in an organism's cells.

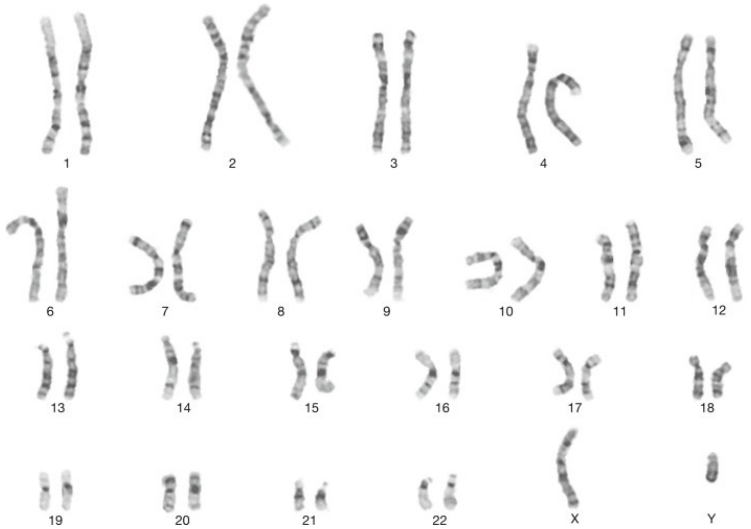
- Genetics: Genetics is the scientific study of genes, heredity, and variation in living organisms.

Haploid and Diploid cells

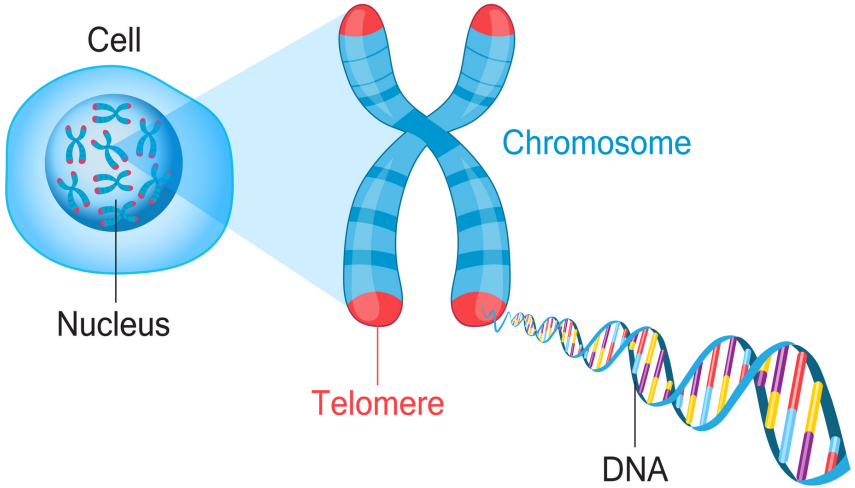
https://www.youtube.com/watch?v=NR9zTvMg-pE

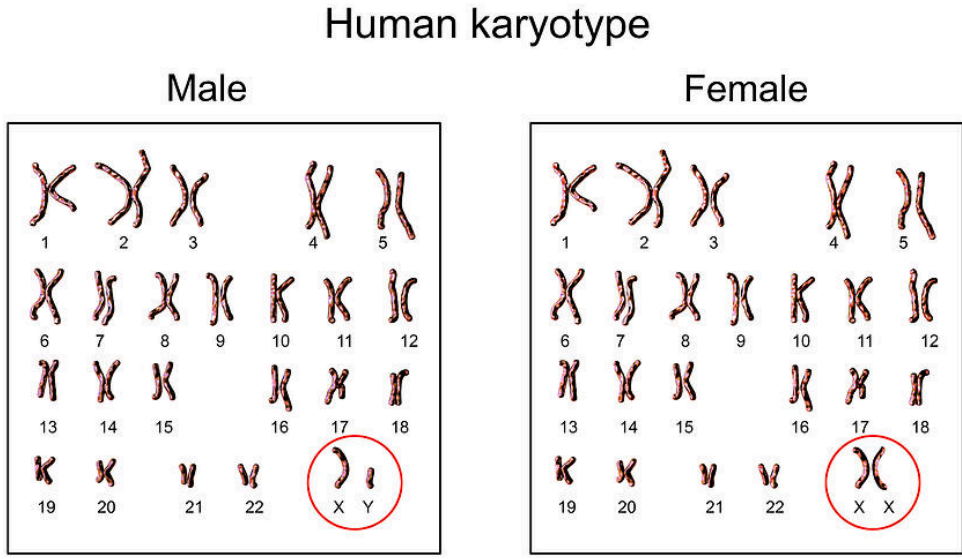
Chromosomes





Chromosomes





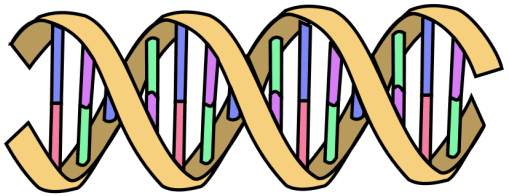
SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Human Genetics

DNA

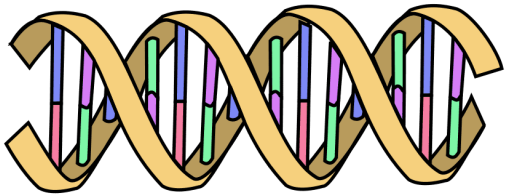


3 billions base pair long

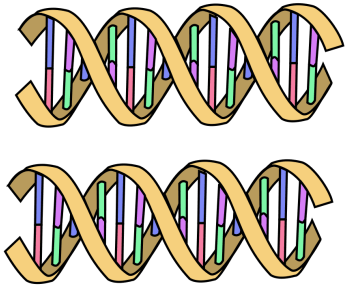
….ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT….

….TAAACGGTCAGTCATGGGTCCTACGACCTTGCCTA….

Human Genetics

DNA 

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



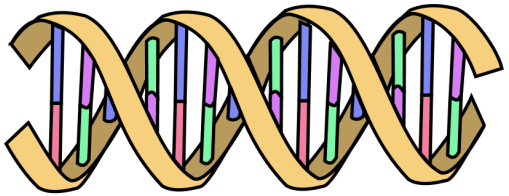
....ATTTGCCAGTCAGTACCCAGGATGCTGGA....

....TAAACGGTCAGTCATGGGTCCTACGACCT....

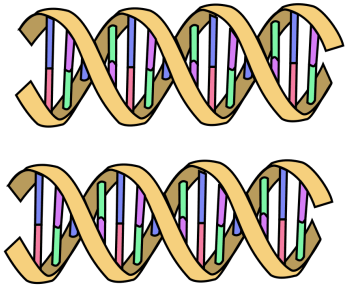
....ATTTGCCAGTCAGTACCCAGGATGCTGGA....

....TAAACGGTCAGTCATGGGTCCTACGACCT....

Human Genetics

DNA 

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....



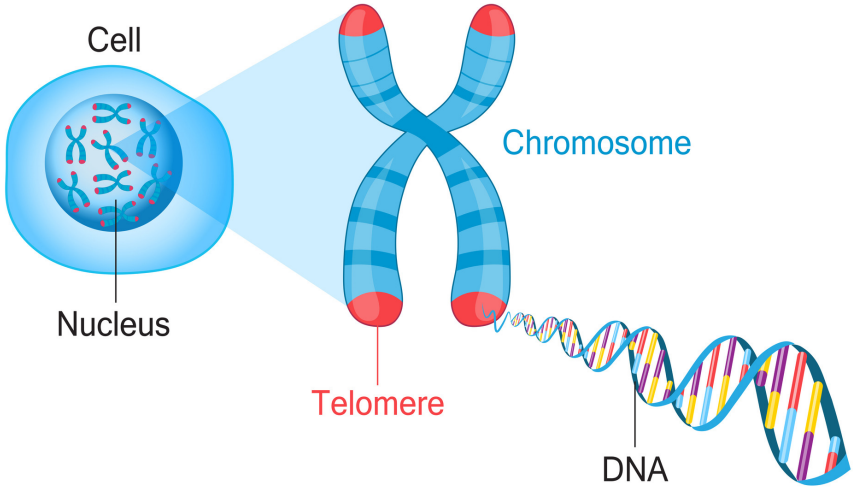
....ATTTGCCAGTCAGTACCCAGGATGCTGGA....

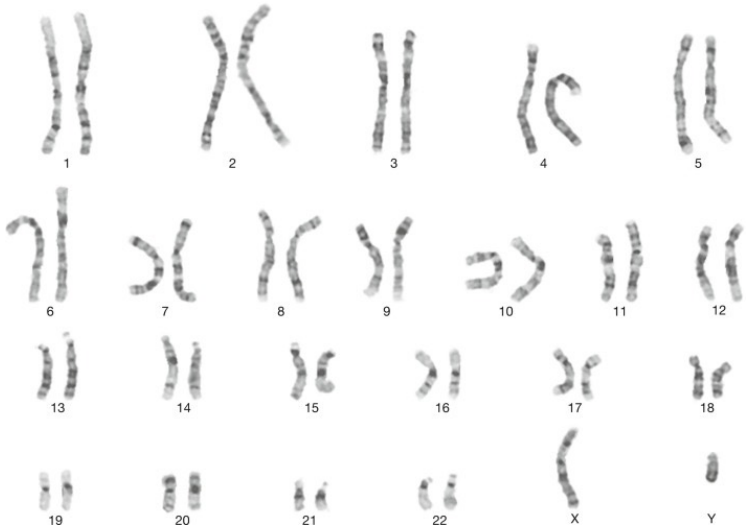
....TAAACGGTCAGTCATGGGTCCTACGACCT....

....ATTTGCCAGTCAGTACCCAGGATGCTGGA....

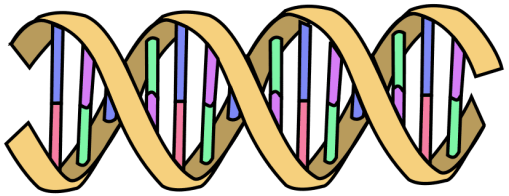
....TAAACGGTCAGTCATGGGTCCTACGACCT....

Chromosomes

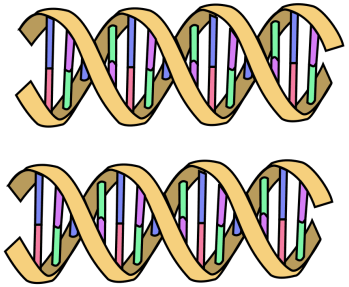




Human Genetics

DNA 

….ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT….



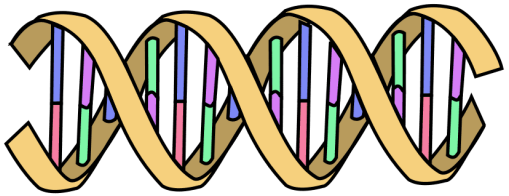
Copy 1, Chromosome 1 (Mother)

….ATTTGCAAGTCAGTACCGAGGATGCTGGA….

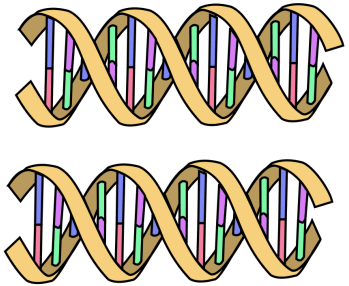
….ATTTGCCAGTCAGTACCCAGGATGCTGGA….

Copy 2, Chromosome 1 (Father)

Human Genetics

DNA 

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

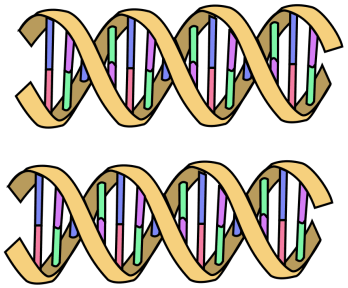


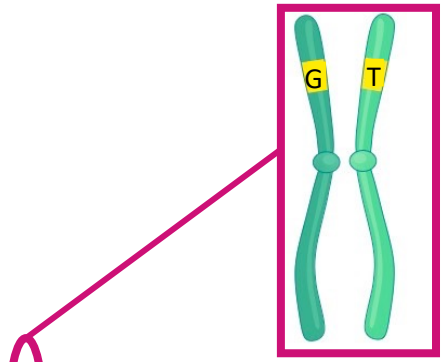
....ATTTGCAAGTCAGTACCGAGGATGCTGGA....

....ATCTGCCAGACAGTACCCAGGATGCTGGA....

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

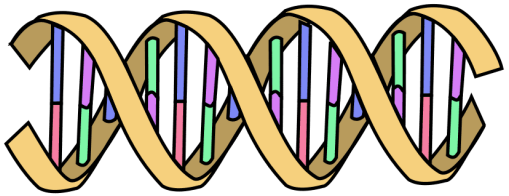




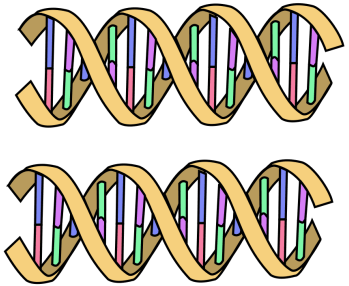
....ATTTGCAAGTCAGTACCGAGGATGCTGGA....

....ATCTGCCAGACAGTACCCAGGATGCTGGA....

Human Genetics

DNA 

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

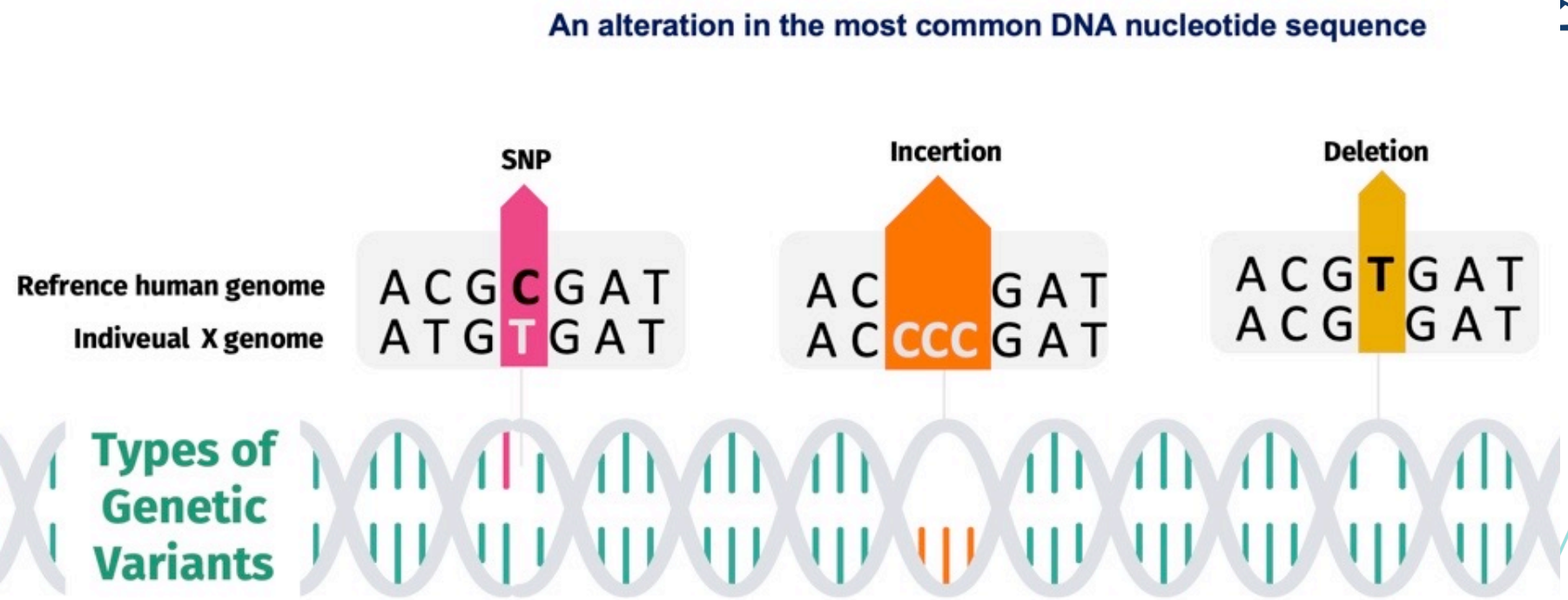


....ATTTGCAAGTCAGTACCGAGGATGCTGGA....

....ATCTGCCAGACAGTACCCAGGATGCTGGA....

Single Nucleotide Polymorphisme (SNP)

Genetic Variants Forms



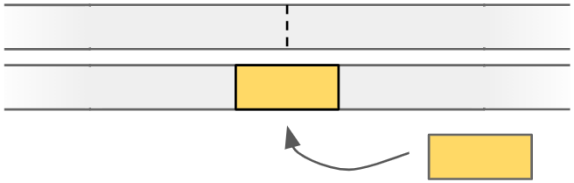
Sequence Variants

SNV (**S**ingle **N**ucleotide **V**ariant)

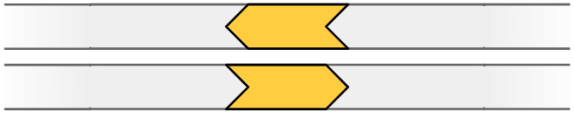
INDEL (**In**sertion or **Del**etion)

Structural Variants

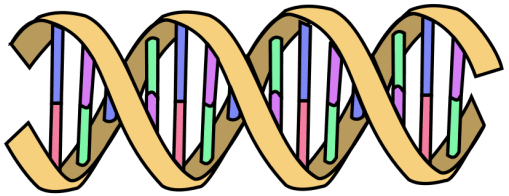
Insertion



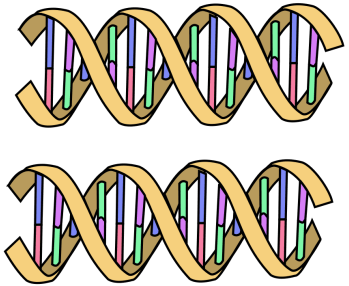
Inversion



Human Genotype

DNA 

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

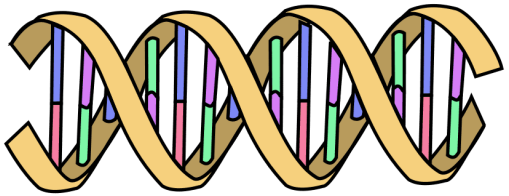


....ATTTGCAAGTCAGTACCGAGGATGCTGGA....

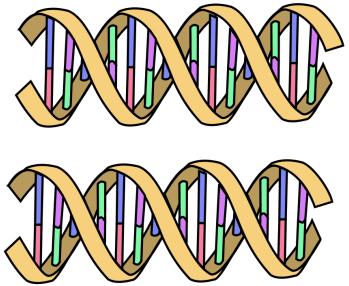
....ATCTGCCAGACAGTACCCAGGATGCTGGA....

What is a genotype? (A/T)

Human Phenotype

DNA 

....ATTTGCCAGTCAGTACCCAGGATGCTGGAACGGAT....

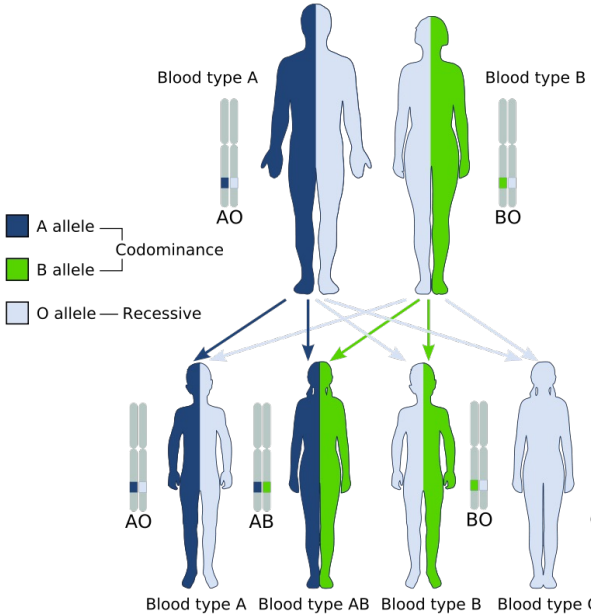


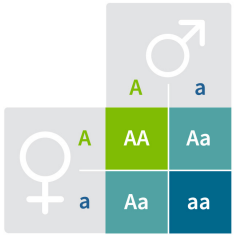
....ATTTGCAAGTCAGTACCGAGGATGCTGGA....

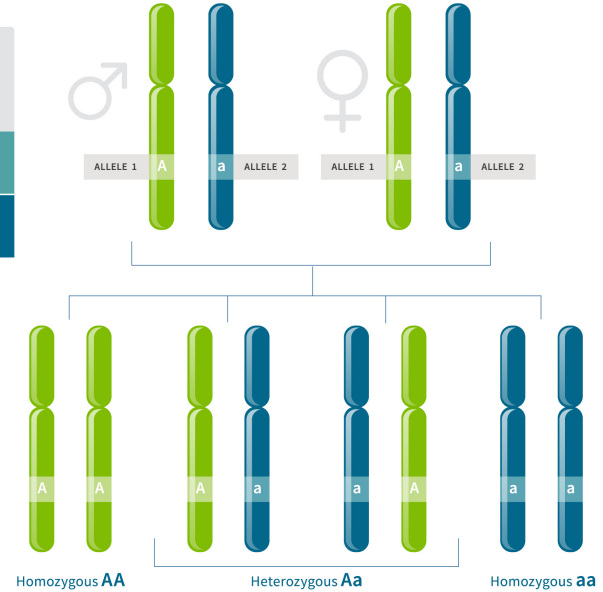
....ATCTGCCAGACAGTACCCAGGATGCTGGA....

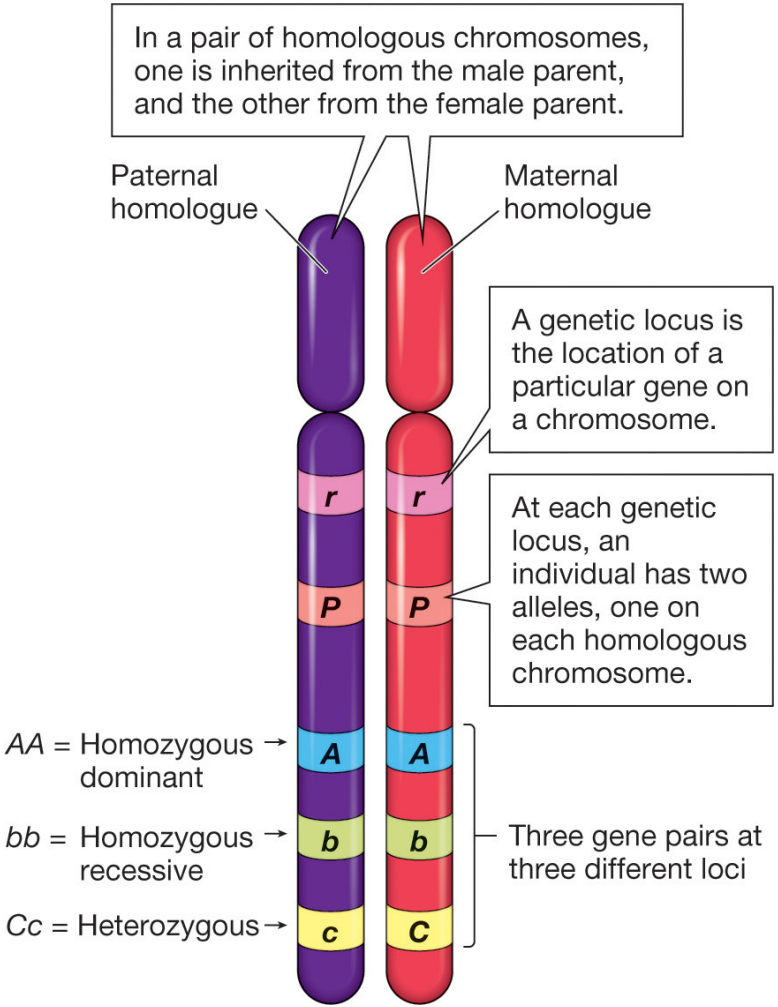
What is a phenotype? (Dark brown eyes)

Genotype and Phenotype



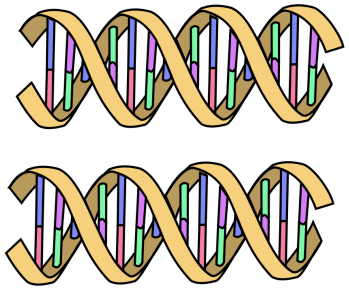





In a pair of homologous chromosomes, one is inherited from the male parent, and the other from the female parent.
Paternal homologue
Maternal homologue
A genetic locus is the location of a particular gene on a chromosome.
At each genetic locus, an individual has two alleles, one on each homologous chromosome.
r
r
P
P
AA = Homozygous dominant
A
A
bb = Homozygous recessive
b
b
Cc = Heterozygous
c
C
Three gene pairs at three different loci

Human Genetics

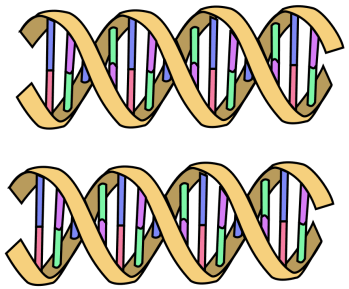
DNA



....ATTTGCCAGTCAGTACCGAGGATGCTGGA....

....TAAACGGTCAGTCATGGGTCCTATGACCT....

DNA

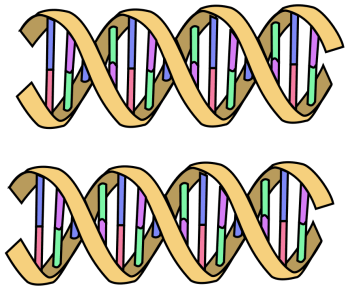


....ATTTGCCAGTCAGTACCGAGGATGCTGGA....

....TAAACGGTCAGTCATGGGTCCTATGACCT....

Human Genetics

DNA

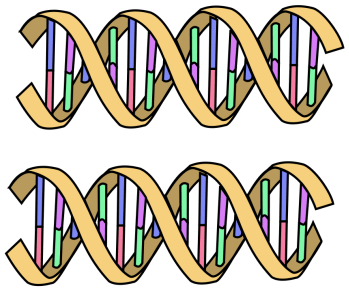


….ATTTGCCAGTCAGTACCGAGGATGCTGGA….

….TAAACGGTCAGTCATGGGTCCTATGACCT….

?

DNA



….ATTTGCCAGTCAGTACCGAGGATGCTGGA….

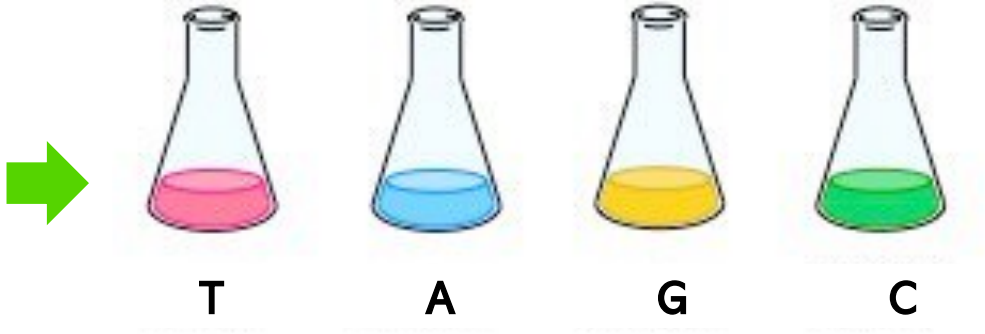
….TAAACGGTCAGTCATGGGTCCTATGACCT….

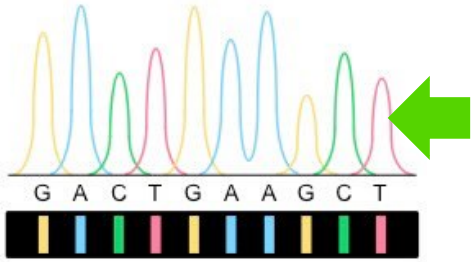
DNA Sequencing

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

KAUST

BORG samples



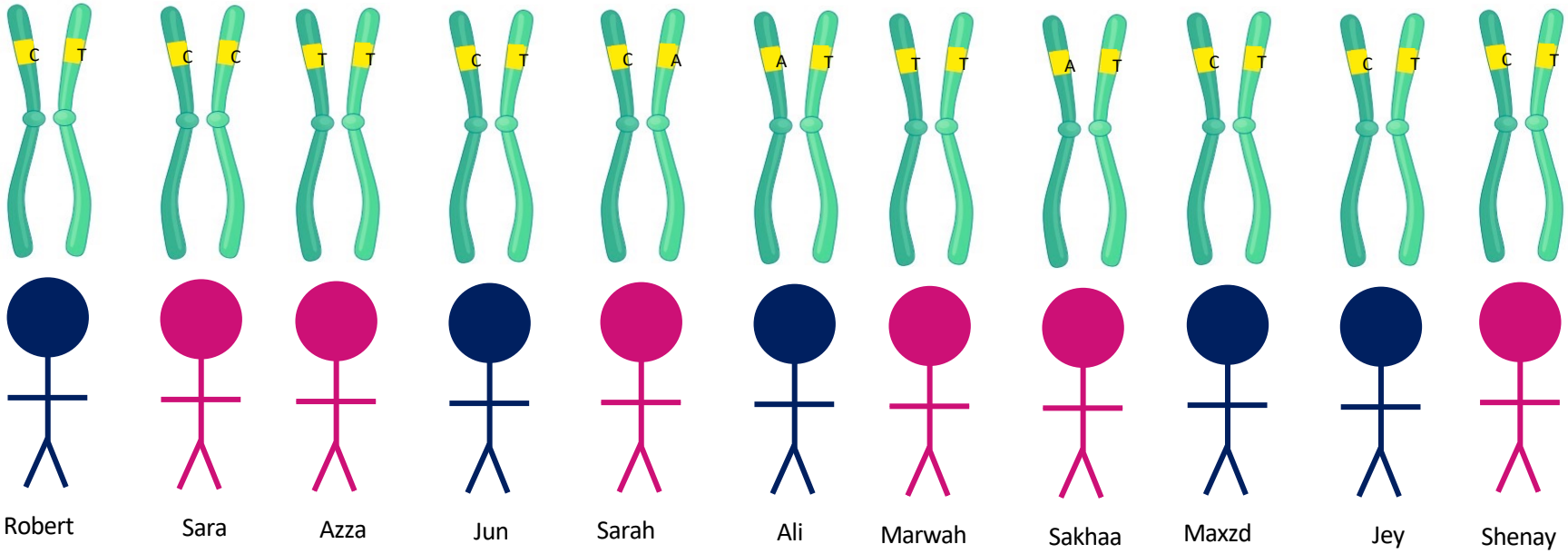


SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

How we calculate AF?



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

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

How We Calculate AF?

AA

AC

AT

TC

TT

CC

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

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

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

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

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

How We Calculate AF?

AA

AC

AT

TC

TT

CC

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

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

MAF

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

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

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

أكاديمية كاوست
KAUST ACADEMY

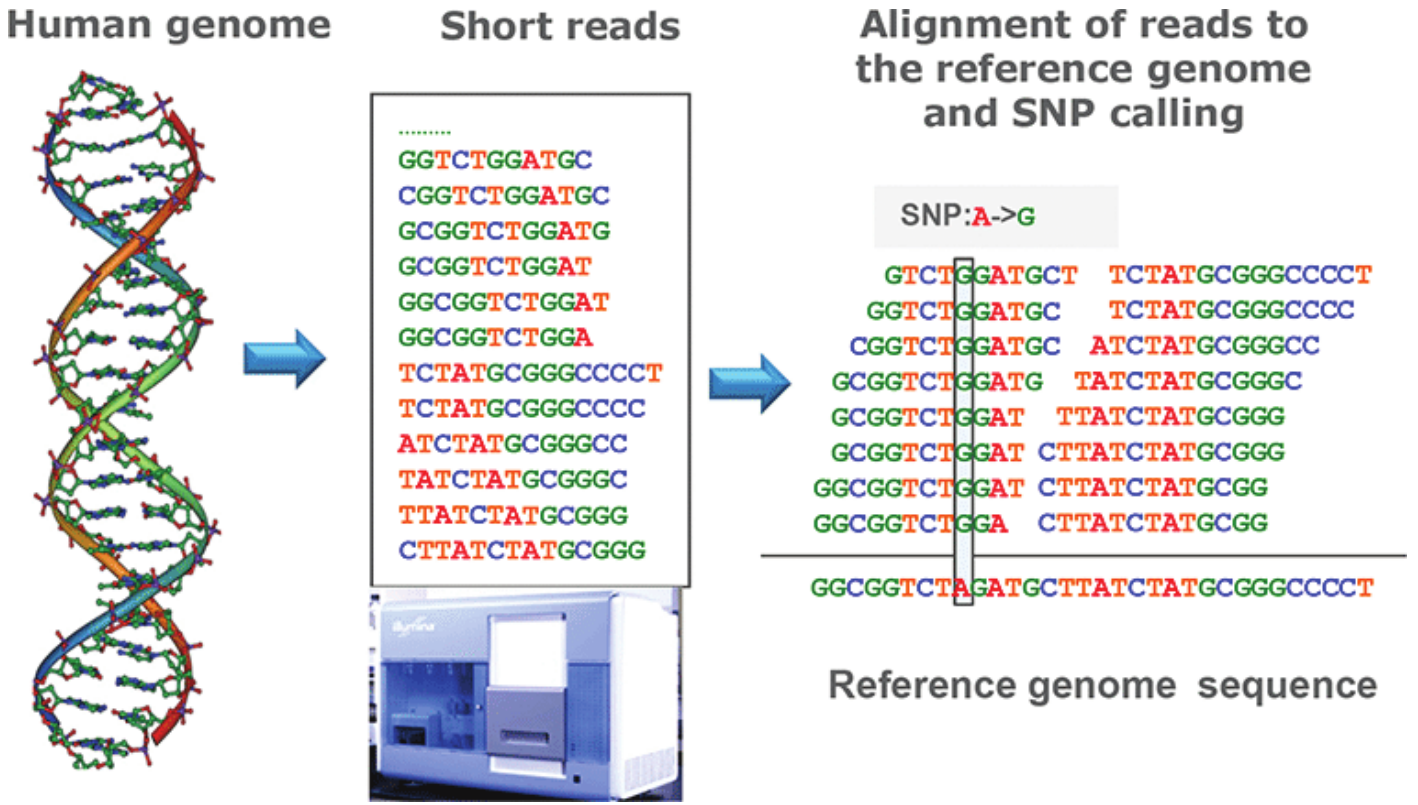
Introduction to Bioinformatics

Calling variants in diploid systems

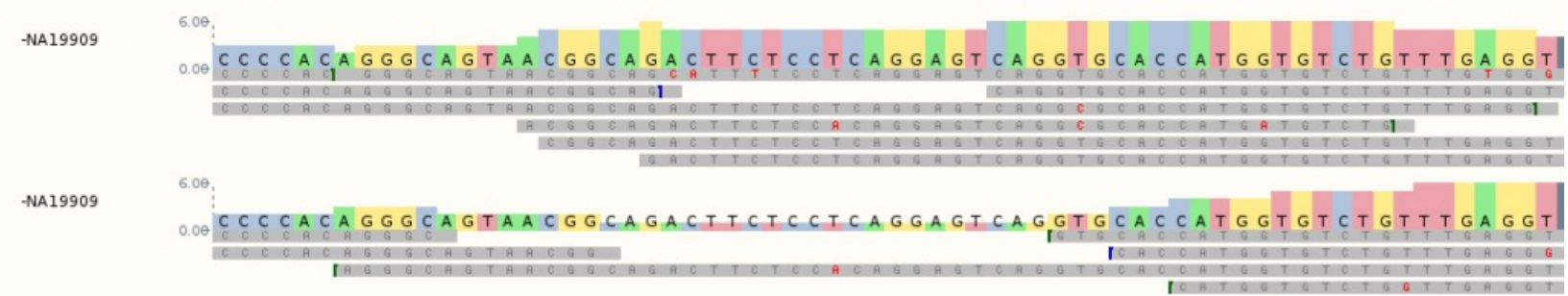
Part 2: Variant Calling

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Basic DNA Sequencing Workflow



Variant Identification and Analysis




```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

Variant and Genetype Calling

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

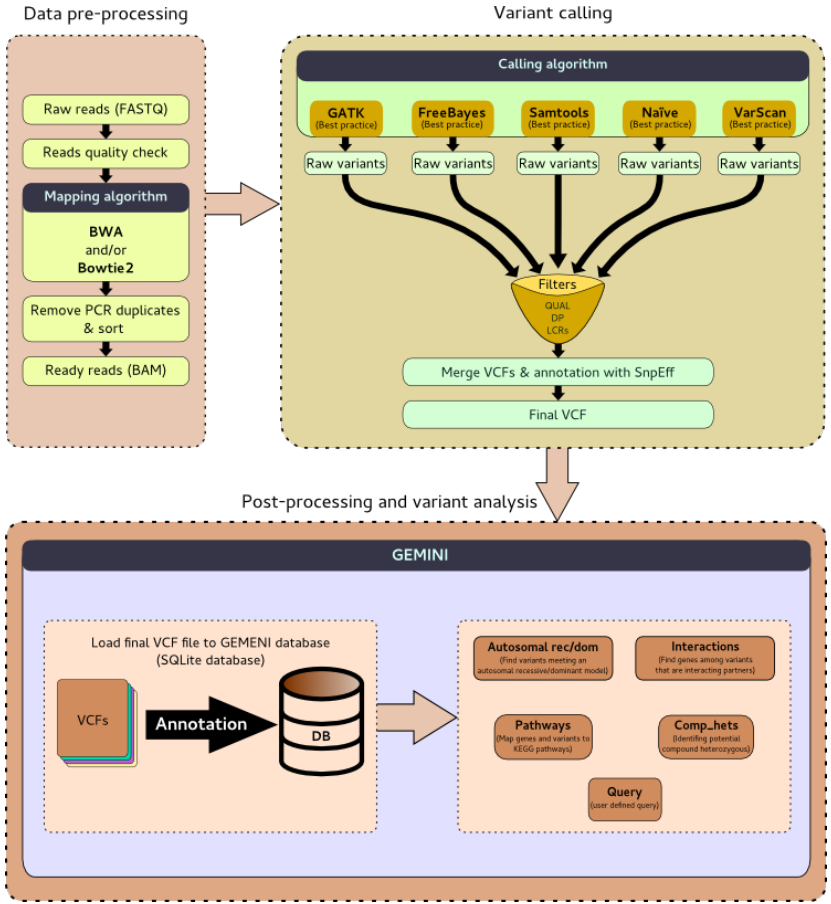
Aspect	Variant Calling	Genotype Calling
Definition	Identifying differences between DNA sequences and a reference genome	Determining specific alleles an individual carries at genomic positions based on sequencing data
Process	Analyzes sequenced DNA reads to detect variations (SNPs, insertions, deletions)	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

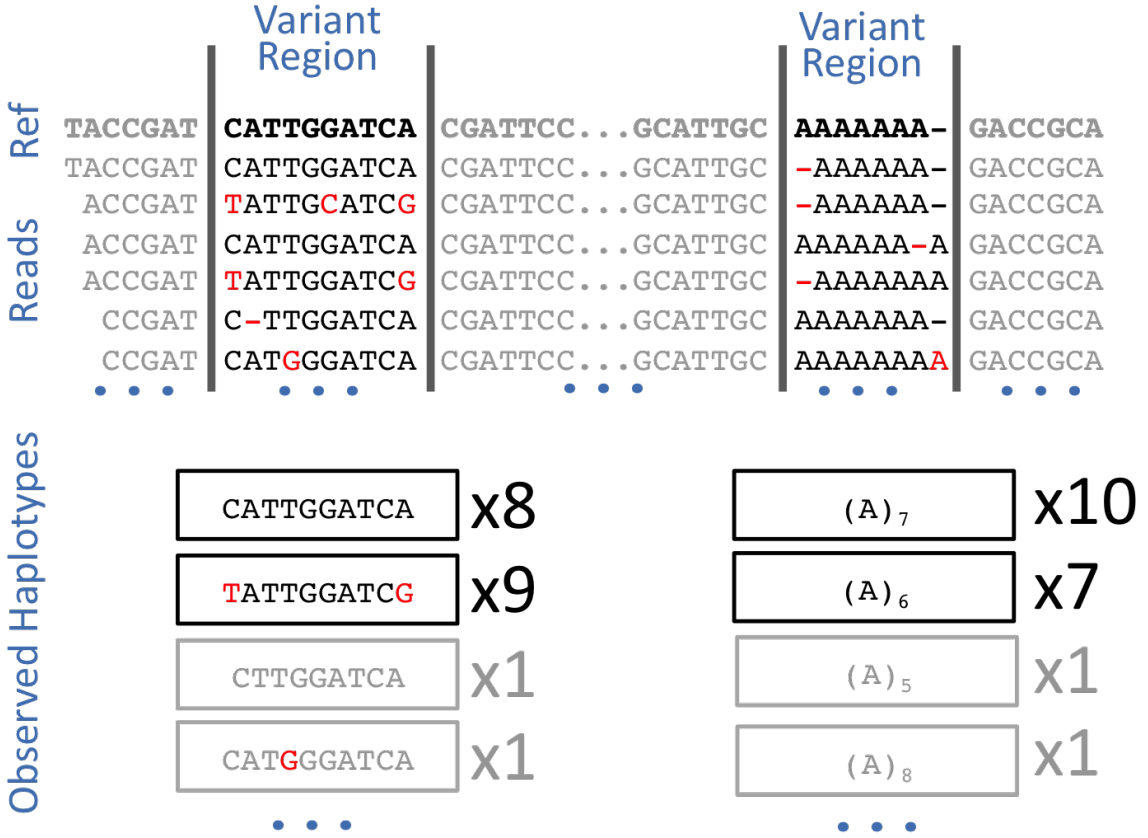
أكاديمية كاوست
KAUST ACADEMY

- **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



Hands-on and Practical Part

Galaxy PROJECT

GitHub

Part 1: Variant analysis

- Calling variants

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Molecular identification of phenotype-causing mutations

Part 3: Variant annotations

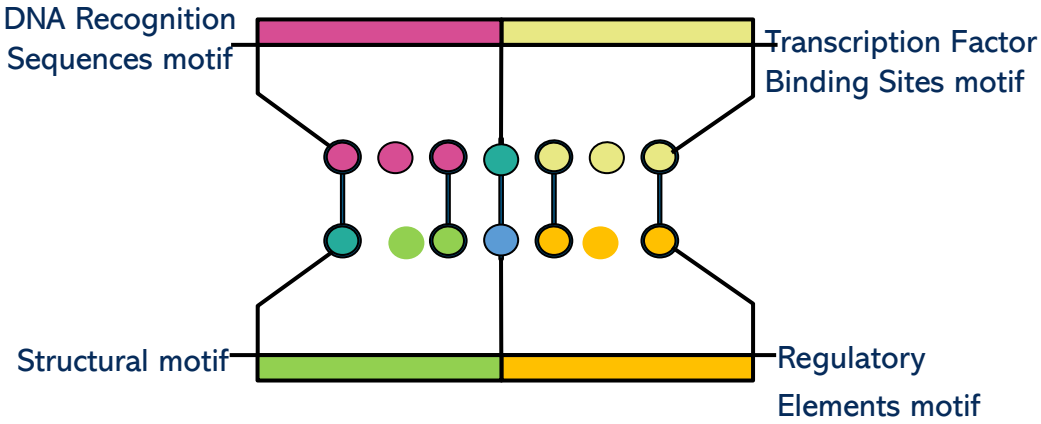
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Finding Hidden Messages in DNA

DNA sequence reads

DNA Motifs: Decoding the Language of Genes

DNA motifs is a specific recurring sequence patterns in DNA



GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTC
GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGTGCATCTGAC
CCTGAGGAGAAGCACGTAGACTGG
GACTCCTCTTCGACTCCTCTTCGTGC
GACTCCTGAGGAGAAGCACGTAGA
CTGAGGACTCCTCTTCATTGCCTTT

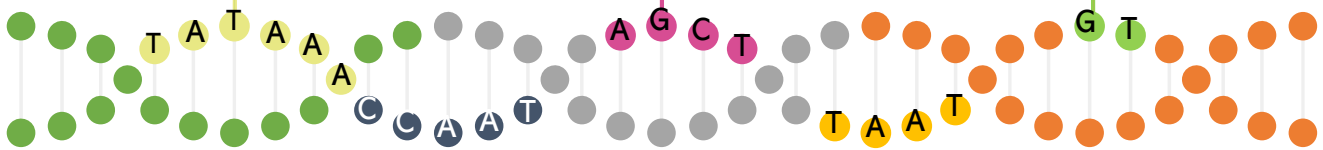
Understanding the Significance of Sequence Patterns

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.

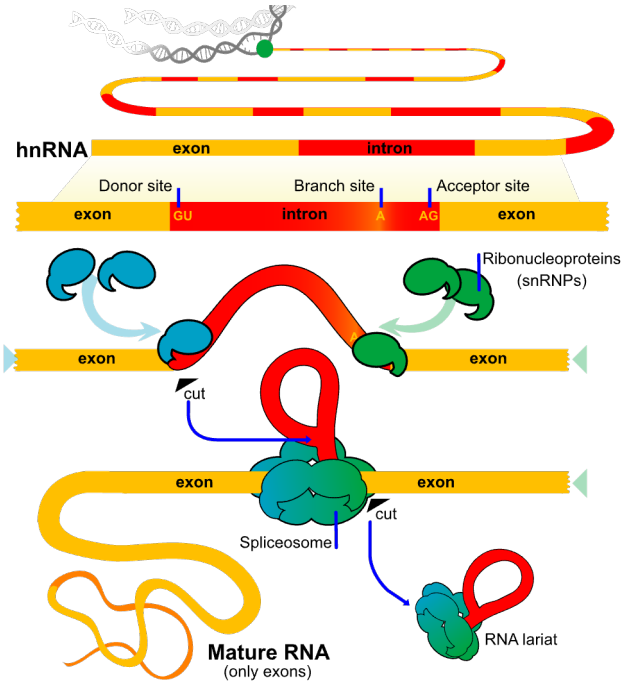
5. Splicing Sites:
- **Motif:** 5' Splice Site (GT)
- **Group:** Splicing Motif
- **Function:** This motif marks the beginning of an intron and is crucial for mRNA splicing.

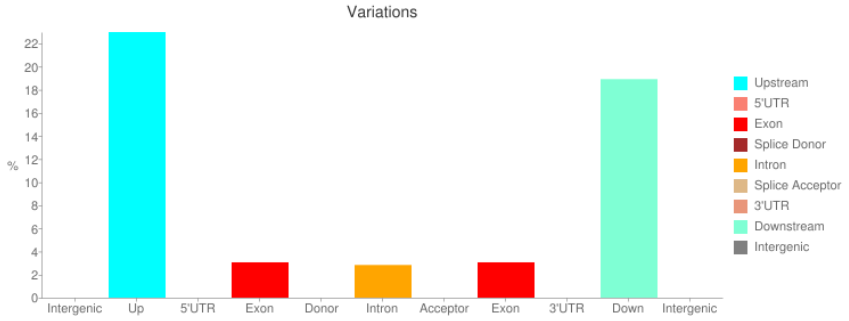


2. Enhancer Motifs:
- **Motif:** CAAT Box (CCAAT)
- **Group:** Enhancer Motif
- **Function:** Enhancer motifs increase the rate of transcription by binding to enhancer regions and facilitating the binding of transcription factors.

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





	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												1										
GAG																	2					
GAT																						
GCA																						
GCC																						
GCT													1	1								
GGA																						
GGG																						
GTG																		1				
TAC																				1		
TCC																						
TGG																		1				
TTG								1														

Hands-on and Practical Part

Galaxy PROJECT

GitHub

Part 1: Variant analysis

- Annotating variants
- Manipulating variation data

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Molecular identification of phenotype-causing mutations

Part 3: Mapping-by-Sequencing

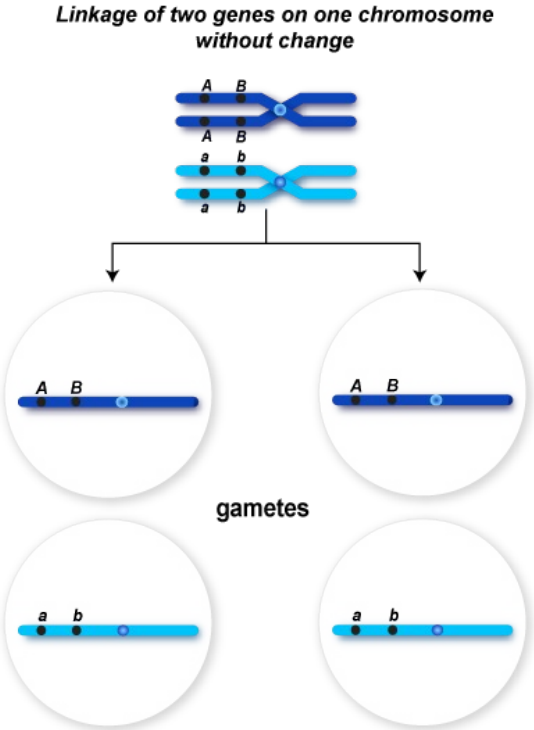
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

أكاديمية كاوست
KAUST ACADEMY

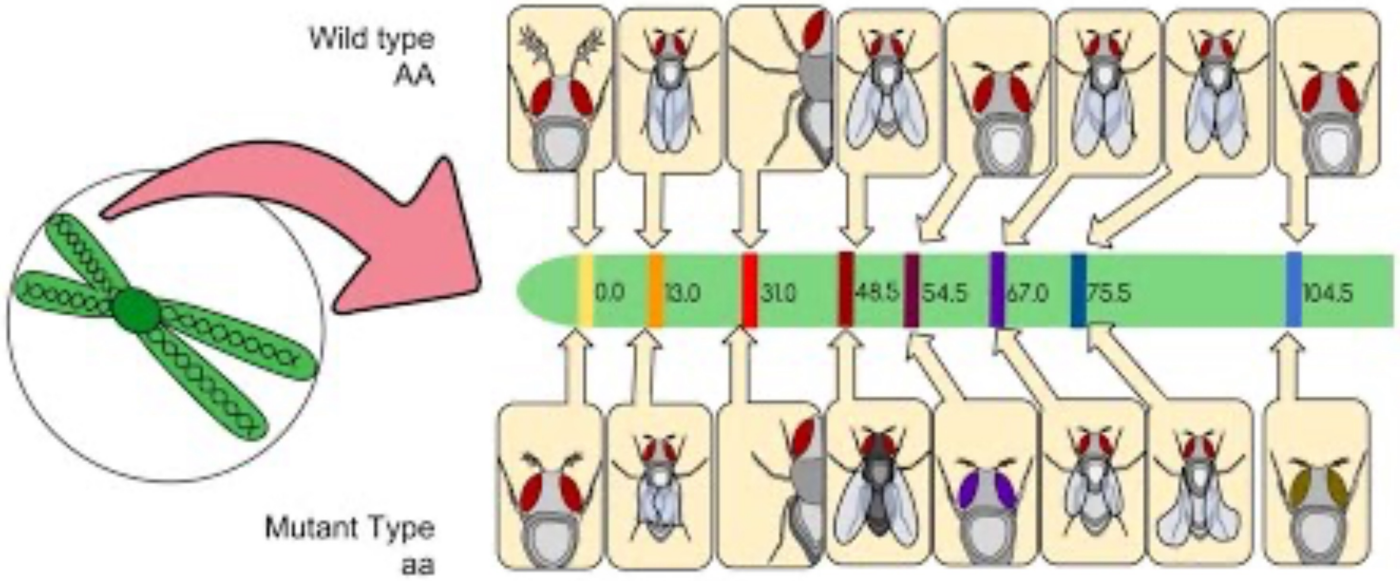
Will start with basic genetics concepts…

Genetic Linkage

The tendency of genes to stay together on a chromosome.

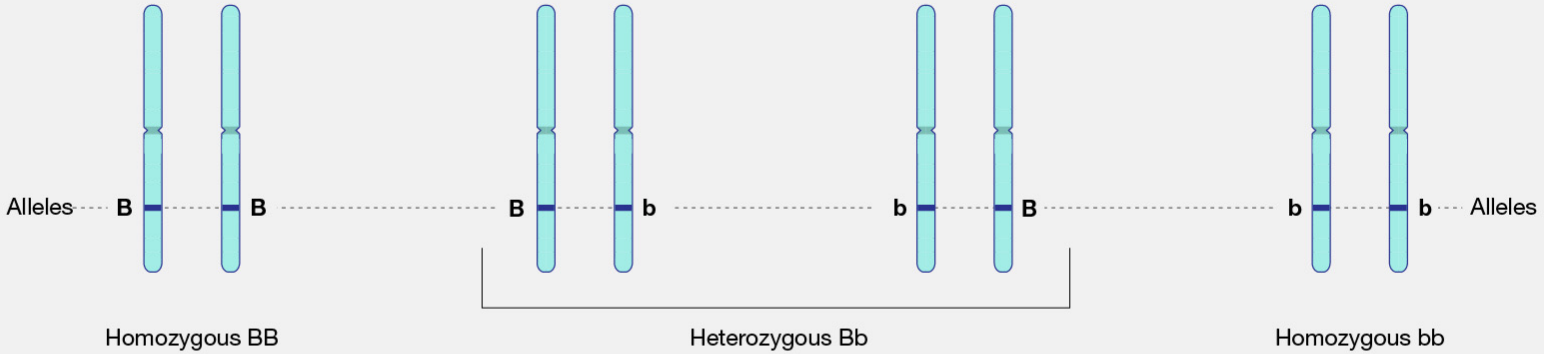


Sturtevant and his colleagues were able to map many of the fruit fly genes in this way



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 probabilityof 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.

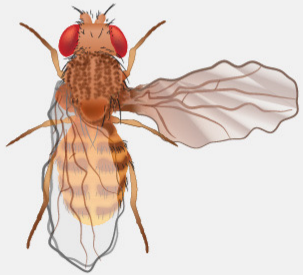


Genotype

Homozygous BB

Heterozygous Bb

Homozygous bb



Phenotype

Normal wings

Normal wings

Normal wings

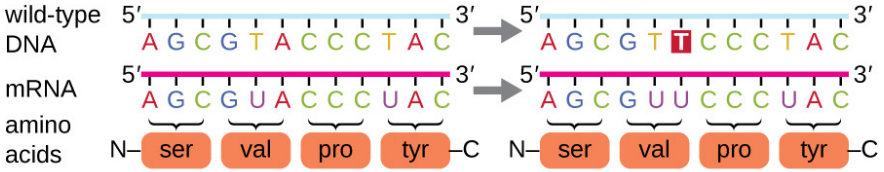
Wrinkled wings

Phenotype-causing mutations

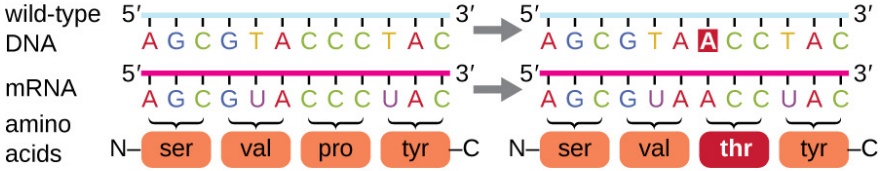
Phenotypic mutations are errors that occur during protein synthesis. These errors can lead to amino acid substitutions that create abnormal proteins.

point mutation: substitution of a single base

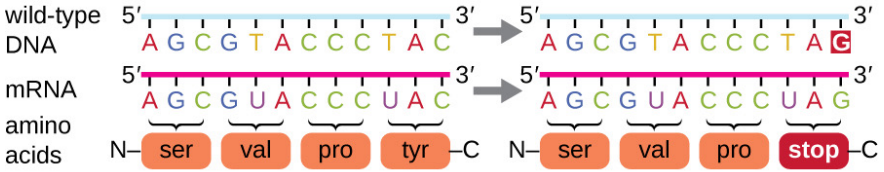
silent: has no effect on the protein sequence



missense: results in an amino acid substitution

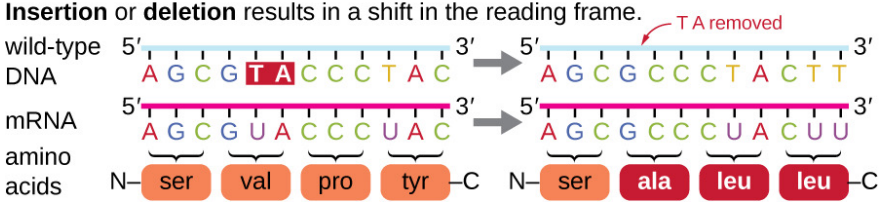


nonsense: substitutes a stop codon for an amino acid



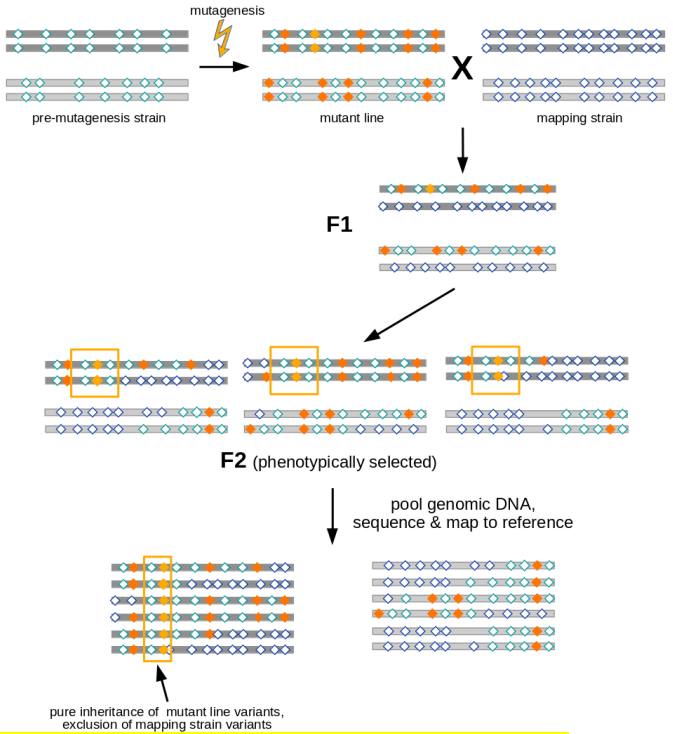
frameshift mutation: insertion or deletion of one or more bases

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

Galaxy PROJECT

GitHub

Part 1: Mapping by sequencing

- Manipulating variation data
- Data Preparation
- Joint Variant Calling and Extraction
- Linkage Analysis
- dentifying Candidate Mutations

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Done with Day 2, Heyyyyy!

Thank You !

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology