Big Data Analysis

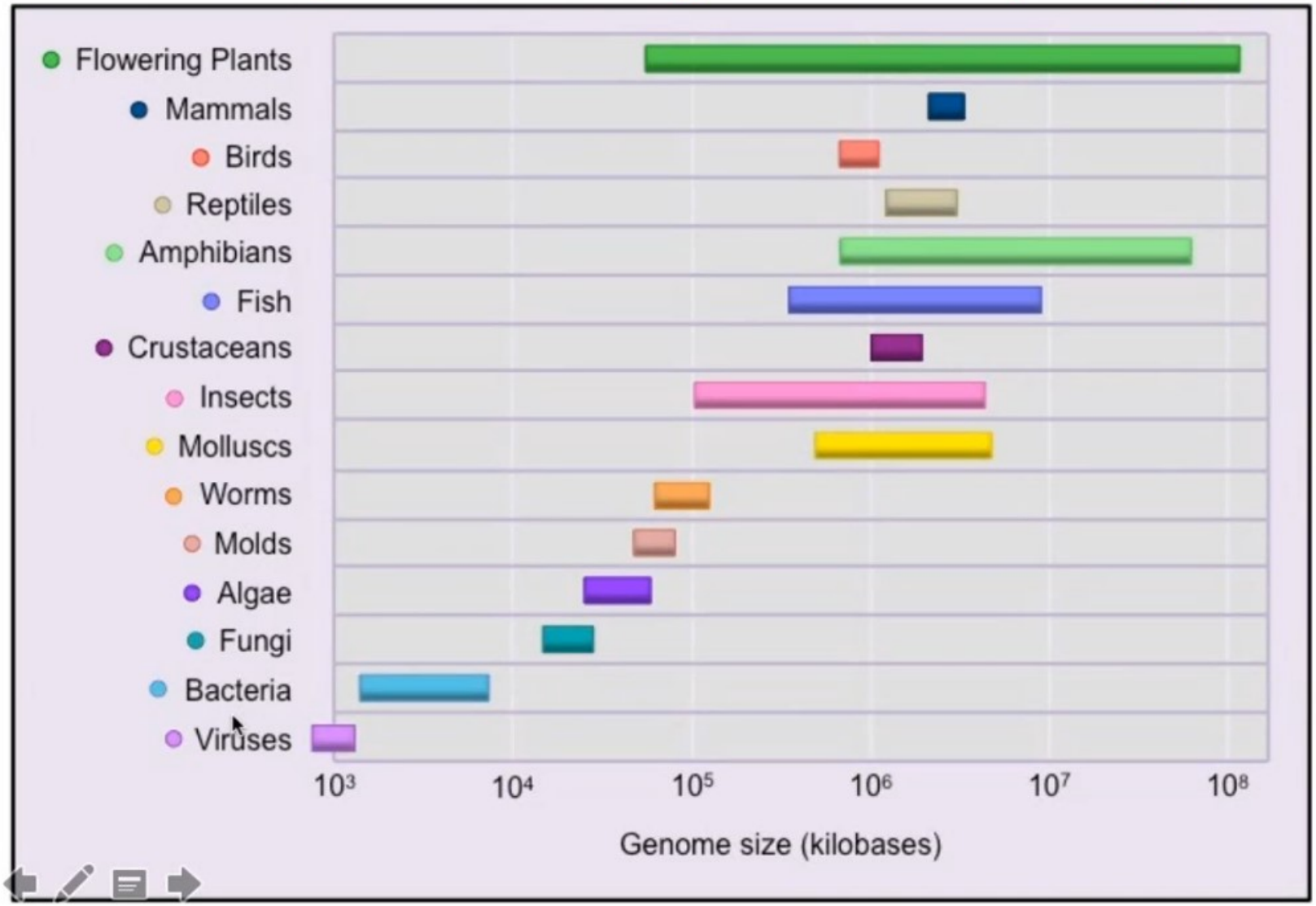
Introduction to Next-Generation Sequencing

Genome Size Does Matter!

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Largest known genome (Paris japonica) – 150 billion bp
Smallest known genome (*Carsonella ruddi*) – 160 kb

Are there any correlation with genome size and phenotypic complexity?



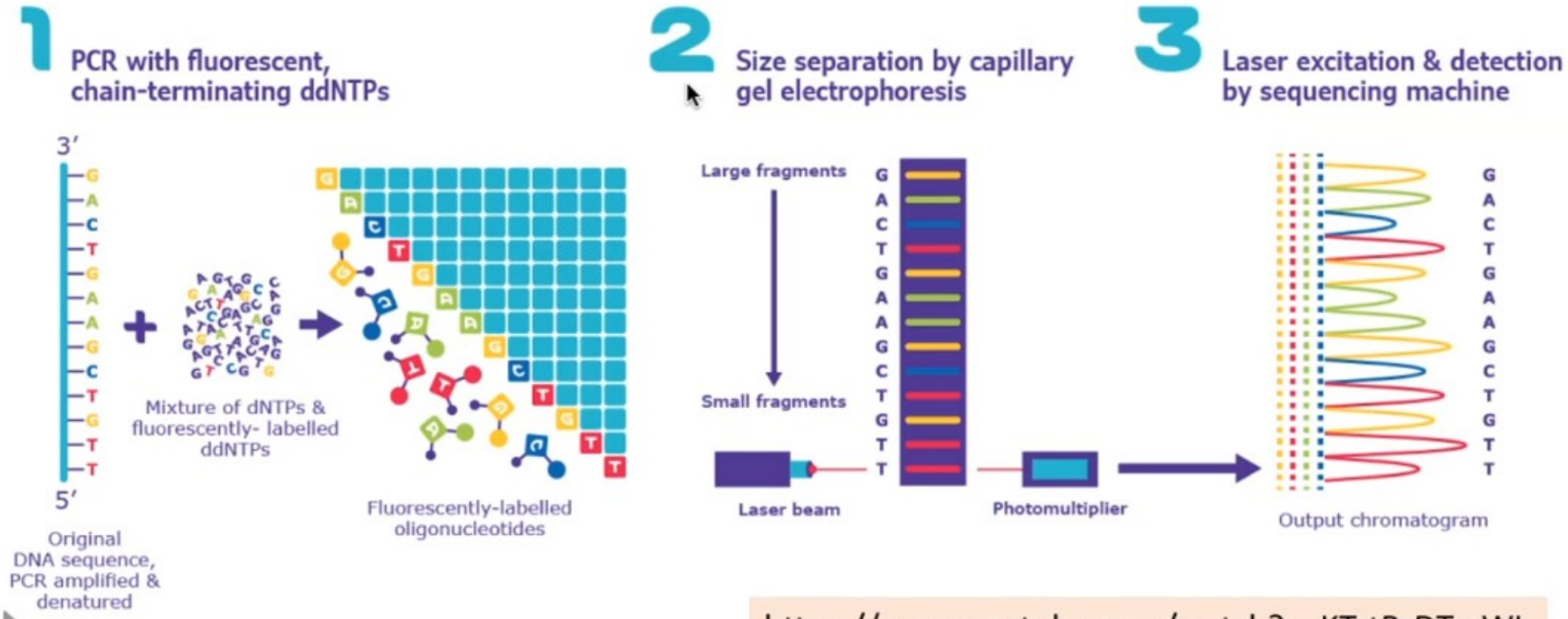
https://ib.bioninja.com.au

Frederick Sanger

Sanger Sequencing Overview (1st Generation)

The Nobel Prize in Chemistry 1958 - Protein Sequencing

The Nobel Prize in Chemistry 1980 - Nucleic acid sequencing



https://www.youtube.com/watch?v=KTstRrDTmWI

High-through put sequencing
(Next-Generation Sequencing)

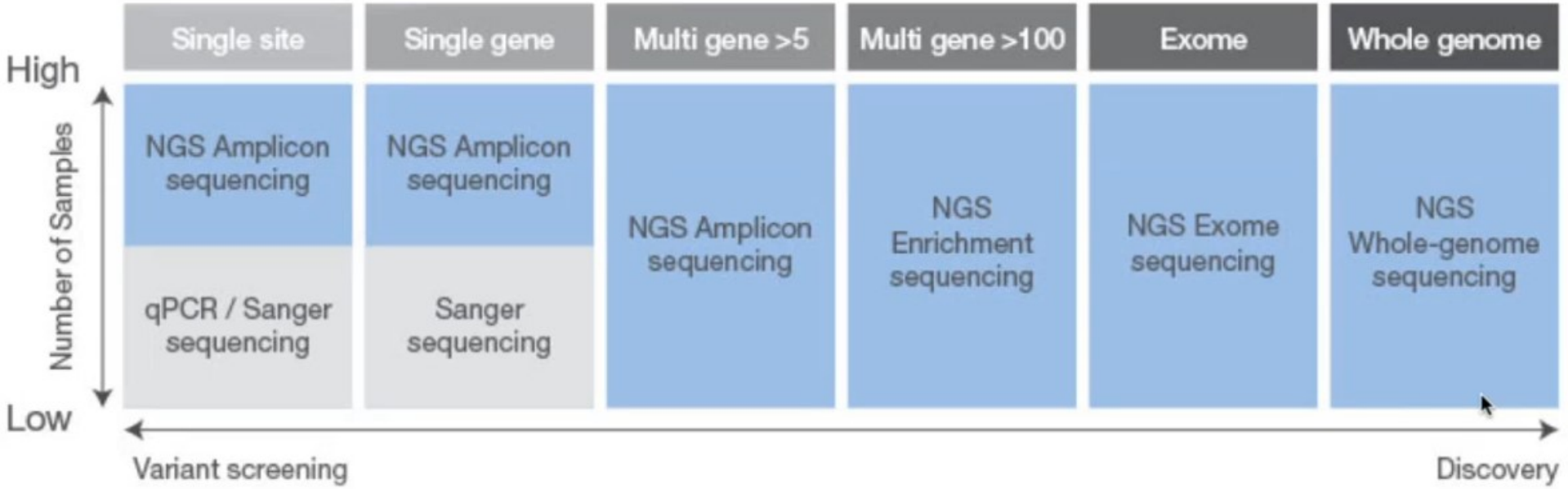
The critical difference between Sanger (dideoxy) sequencing and NGS is sequencing volume.

While the Sanger method only sequences a single DNA fragment at a time, NGS is massively parallel, sequencing millions of fragments simultaneously per run.

This high-throughput process translates into sequencing hundreds to thousands of genes at one time.

NGS also offers greater discovery power to detect novel or rare variants with deep sequencing.

High throughput sequencing (Next-Generation Sequencing)



High-through put sequencing (Next-Generation Sequencing)

Advantages

- Higher sensitivity to detect low-frequency variants
- Faster turnaround time for high sample volumes
- Comprehensive genomic coverage
- Lower limit of detection
- Higher throughput with sample multiplexing
- Ability to sequence hundreds to thousands of genes or gene regions simultaneously

High-through put sequencing
(Next-Generation Sequencing)

"It is remarkable to reflect on the fact that the first human genome, famously co-published in Science and Nature in 2001, required 15 years to sequence and cost nearly three billion dollars.

In contrast, the HiSeq X® Ten System, released in 2014, can sequence over 45 human genomes in a single day for approximately $1000 each."

Illumina.com

High-through put sequencing
(Next-Generation Sequencing)

Illumina sequencing (2nd Generation)

The “sequencing-by-synthesis” technology is used by Illumina. It was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge.

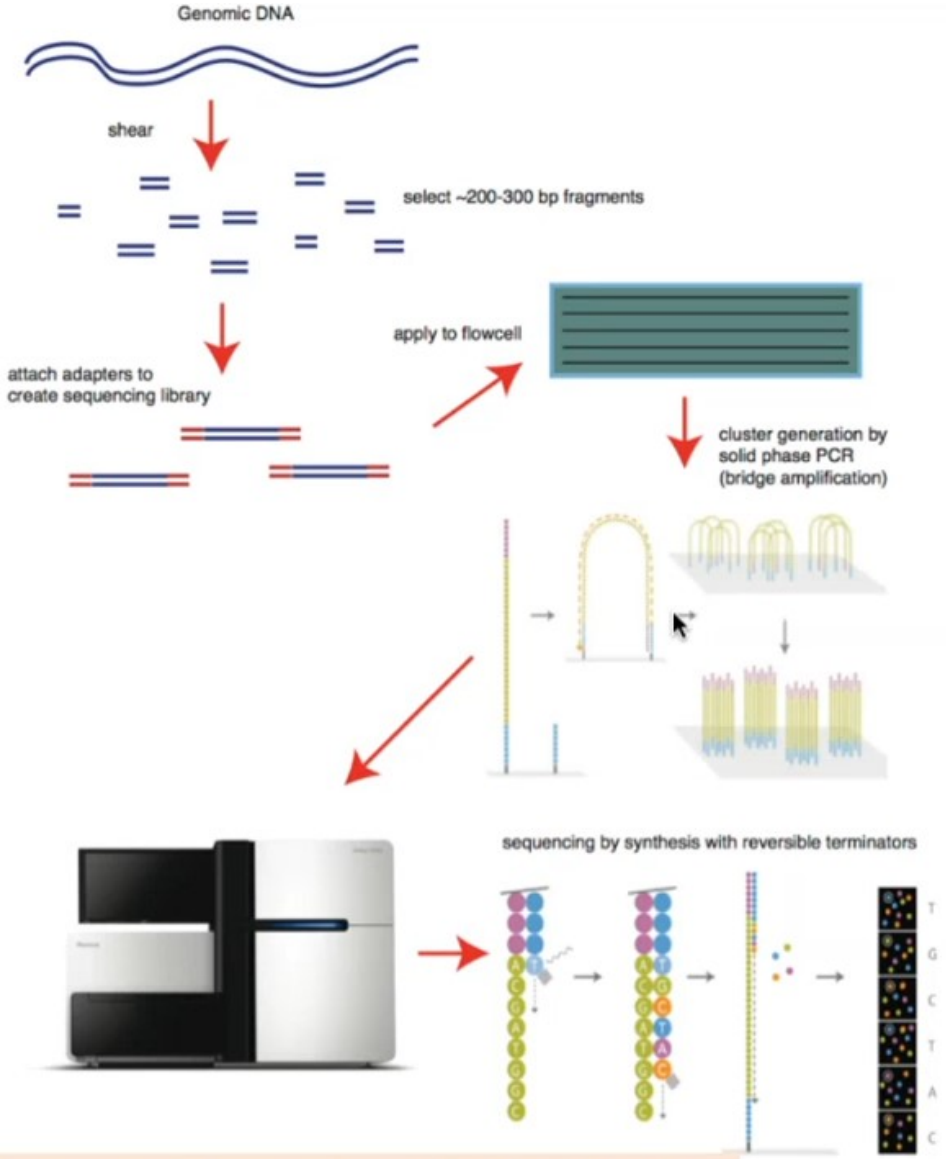
DNA or cDNA samples are randomly fragmented, usually into segments of 200 to 600 base pairs. These fragments are then ligated to adaptors and made single-stranded

Each fragment is amplified on the flow cell, and unlabeled nucleotides and polymerization enzymes are added. These additions, called “Bridge amplification,” connect and lengthen the fragments of DNA on the flow cell

Illumina’s “sequencing by synthesis” involves a proprietary method whereby four labeled reversible dNTP terminators, primers and DNA polymerase are added to the templates on the flow cell.

When excited by a laser, fluorescence from each cluster can be detected, which identifies the first base.

Base calls are made from signal intensity measurements during each cycle, reducing error rates further



https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina sequencing (2nd Generation)

The "sequencing-by-synthesis" technology is used by Illumina. It was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge.

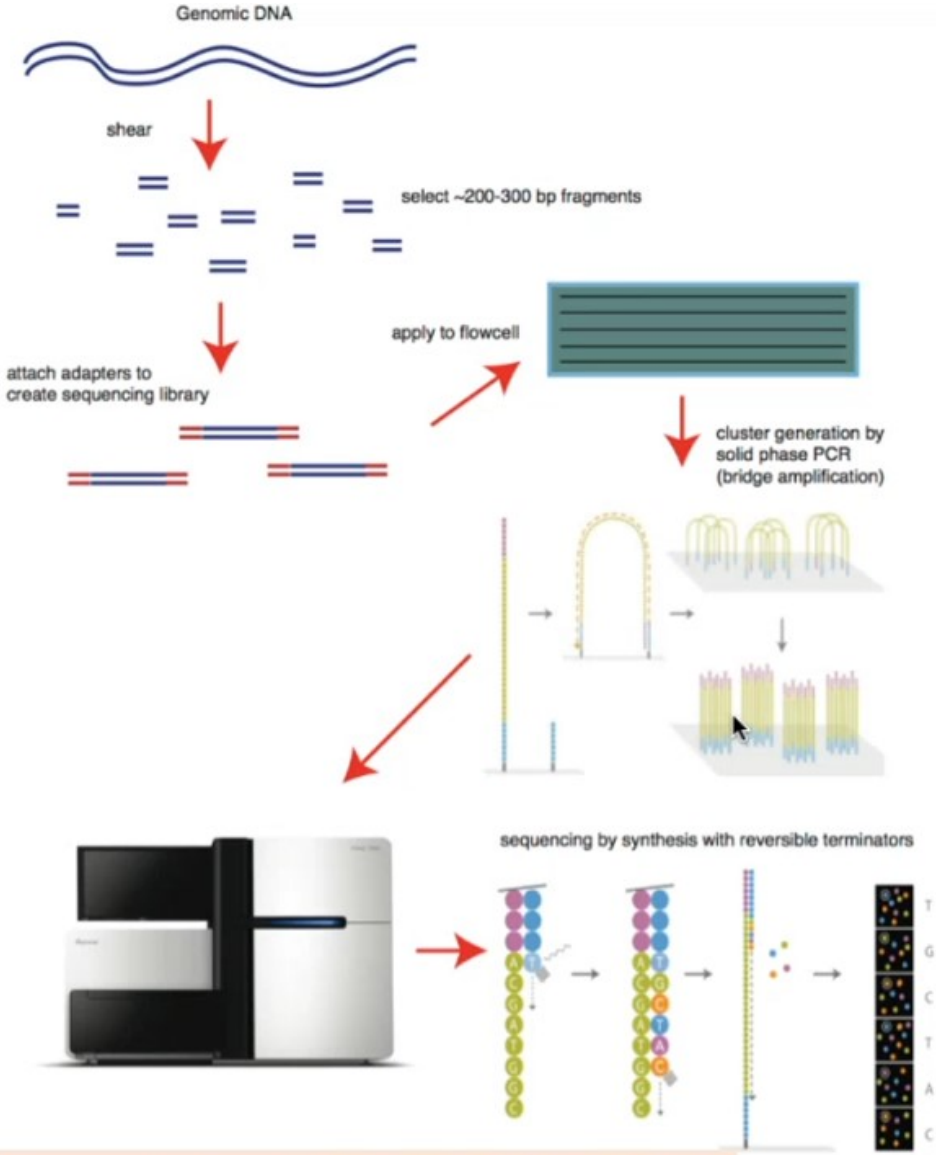
DNA or cDNA samples are randomly fragmented, usually into segments of 200 to 600 base pairs. These fragments are then ligated to adaptors and made single-stranded

Each fragment is amplified on the flow cell, and unlabeled nucleotides and polymerization enzymes are added. These additions, called "Bridge amplification," connect and lengthen the fragments of DNA on the flow cell

Illumina's "sequencing by synthesis" involves a proprietary method whereby four labeled reversible dNTP terminators, primers and DNA polymerase are added to the templates on the flow cell.

When excited by a laser, fluorescence from each cluster can be detected, which identifies the first base.

Base calls are made from signal intensity measurements during each cycle, reducing error rates further



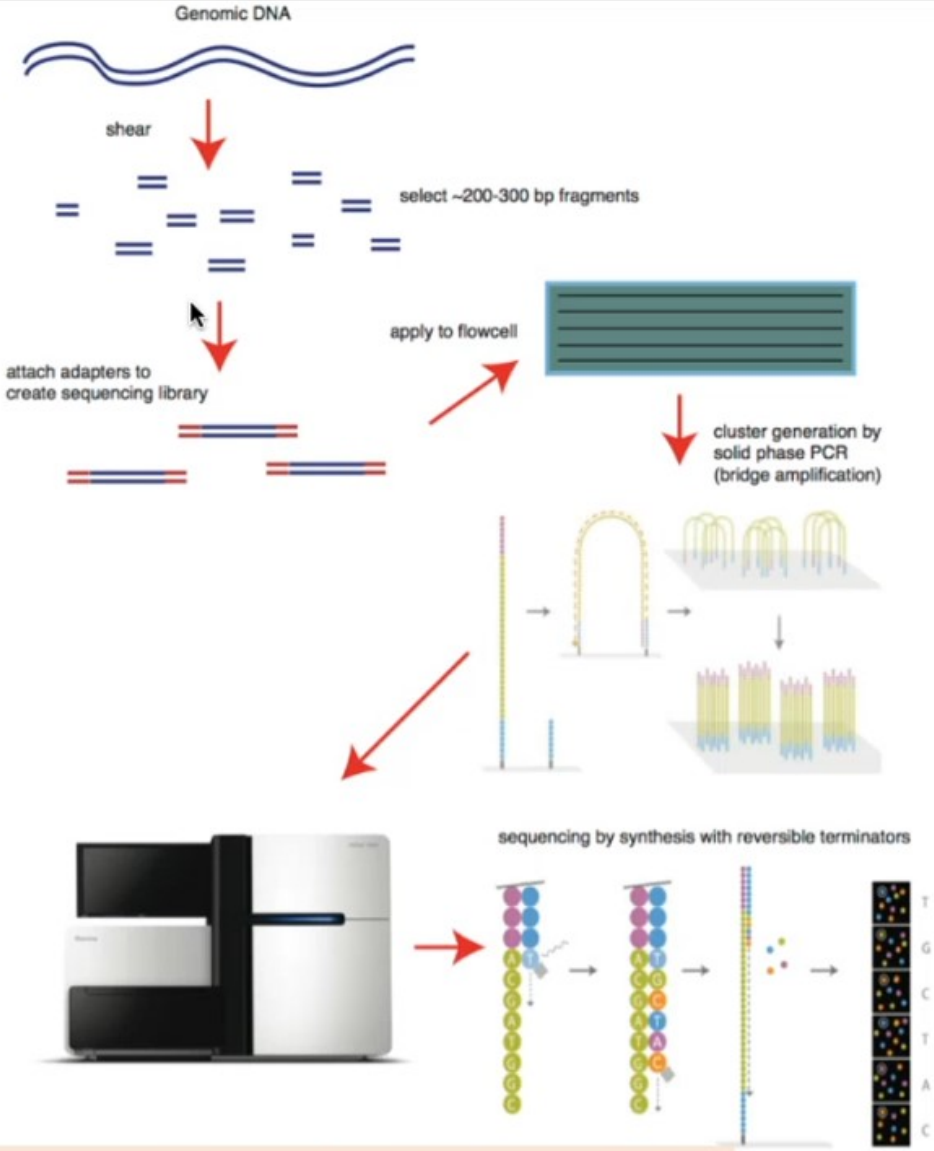
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina sequencing

Disadvantages

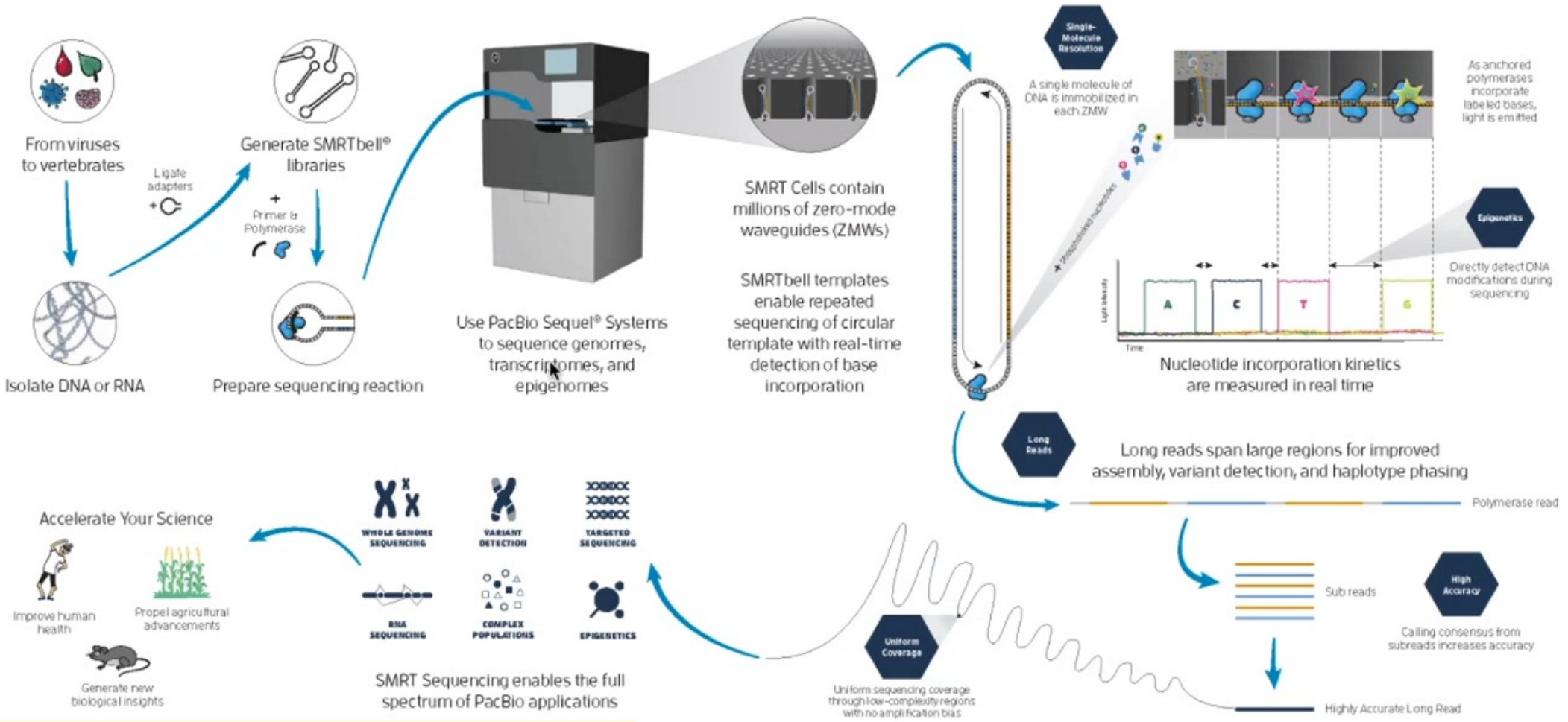
The short read lengths hinder

1. assigning reads to complex parts of the genome
2. phasing of variants
3. resolving repeat regions
4. introduce gaps and ambiguous regions in *de novo* assemblies

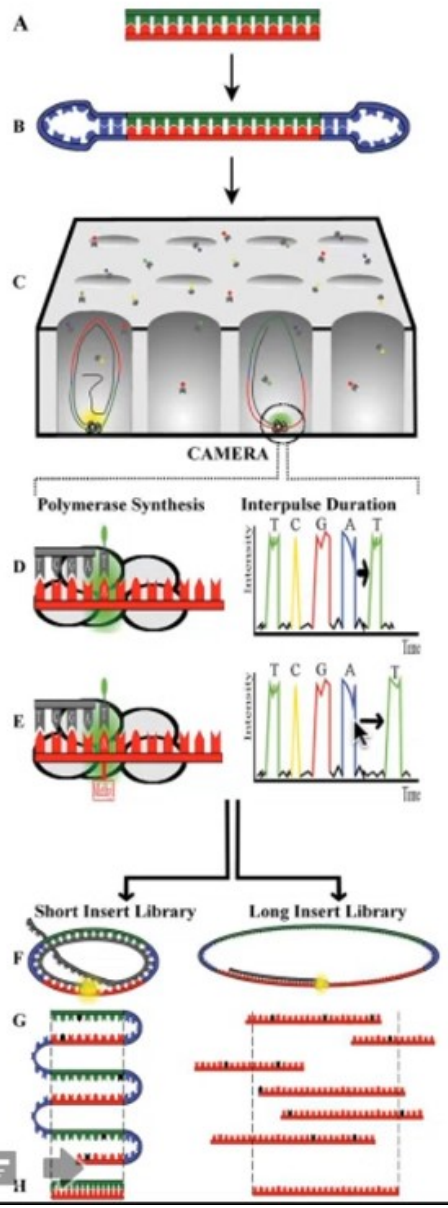


https://www.youtube.com/watch?v=fCd6B5HRaZ8

Pacbio sequencing (**3rd Generation**)



https://www.youtube.com/watch?v=v8p4ph2MAvI



Pacbio sequencing

Comparison of PacBio sequencing platforms to two current industry standards

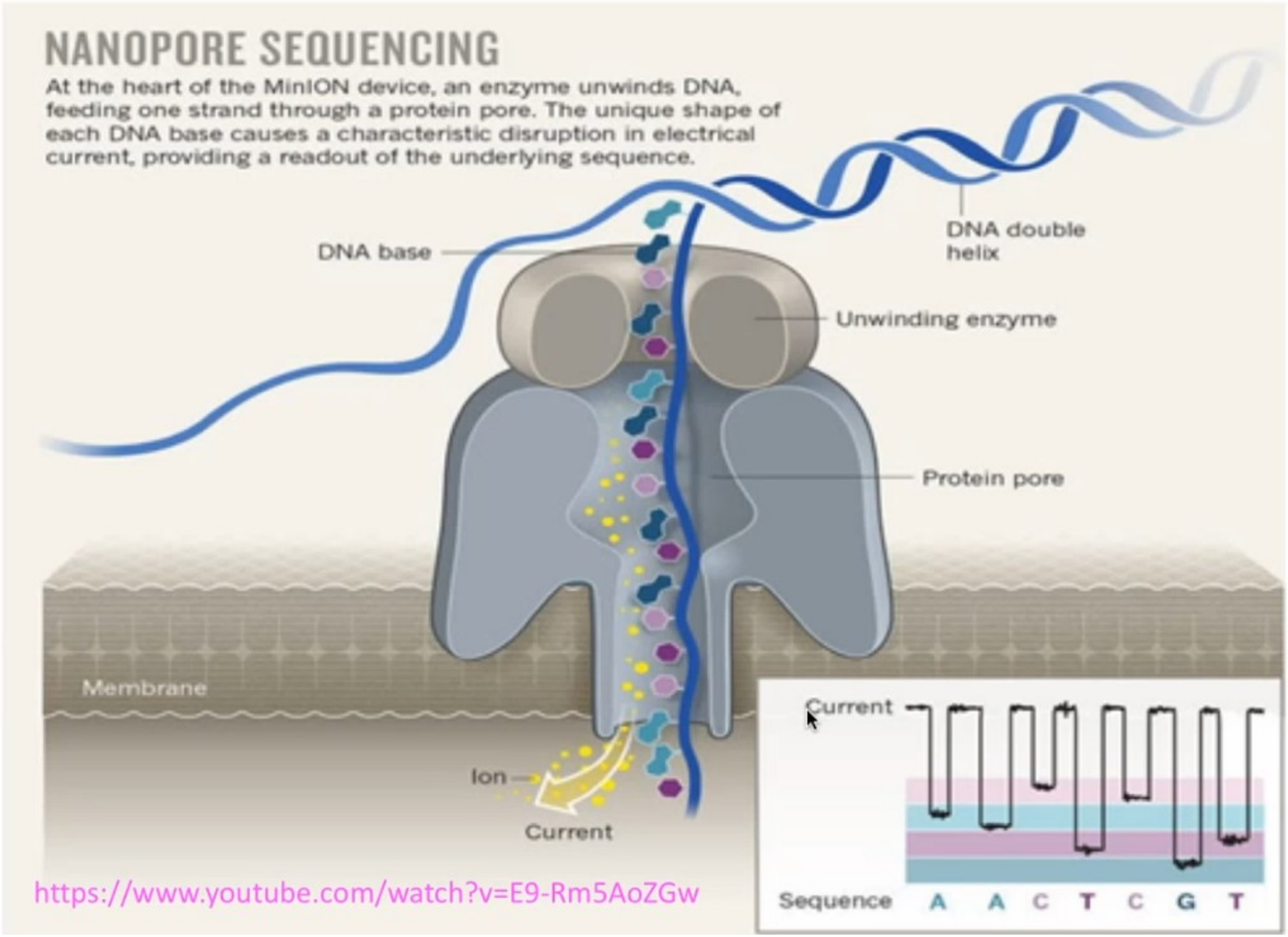
Platform	Read length	Number reads	Error rate	Run rime
PacBio RSII (per SMRT cell)	Average 10–16 kb	~55 000	13–15%	0.5–6 hours
PacBio Sequel (per SMRT cell)	Average 10–14 kb	~365 000	13–15%	0.5–10 hours
Illumina HiSeq 4000	2 × 150 bp	5 billion	~0.1%	<1–3.5 days
Illumina MiSeq	2 × 300 bp	25 million	~0.1%	4–55 hours

Advantages

1. Long Reads
2. Uniform Coverage
3. Epigenetics
4. Single-molecule resolution


Oxford
NANOPORE™
Technologies


00:01
Nanopore DNA sequencing(720p).mp4
04:29



https://www.youtube.com/watch?v=E9-Rm5AoZGw

Oxford Nanopore Sequencing (4th Generation)

Advantages

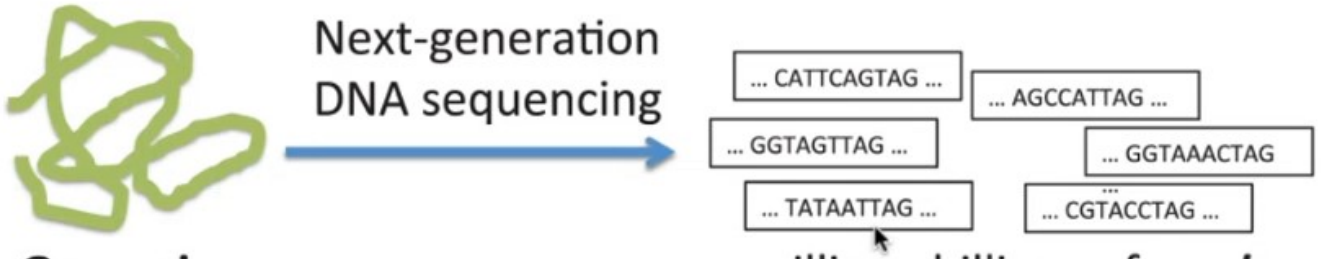
Long Reads (> kb to Mb)

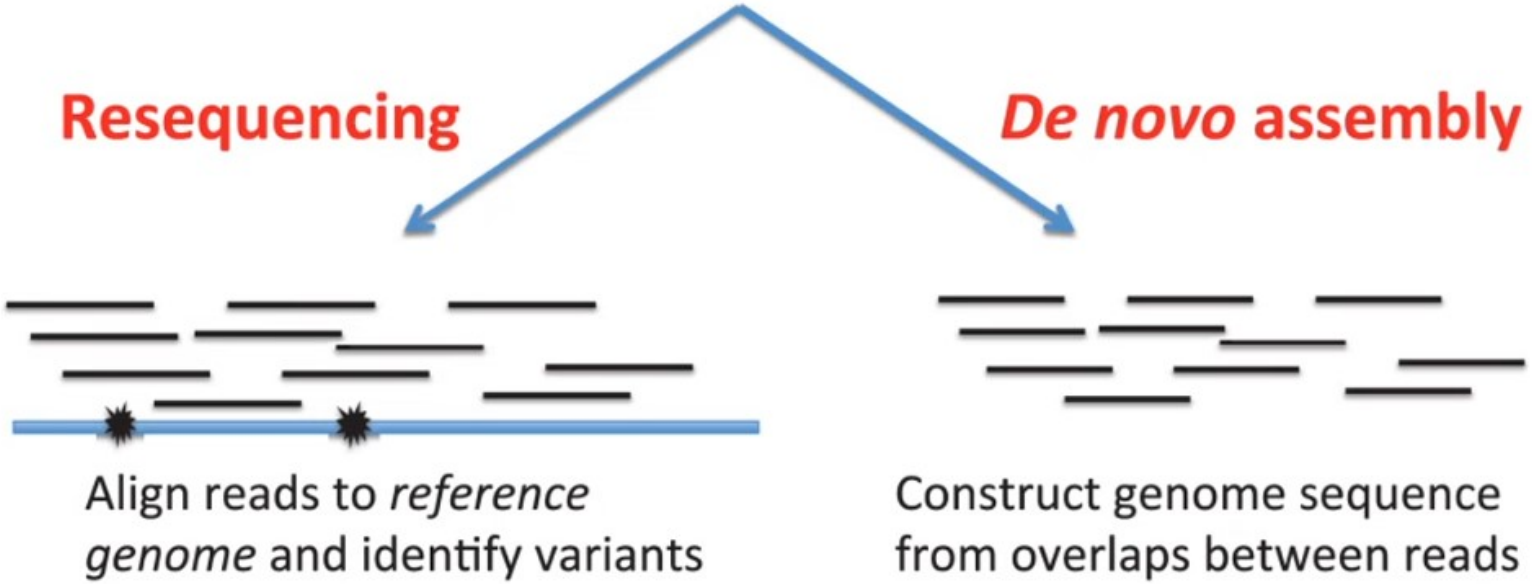
Direct DNA/RNA Sequencing

Epigenetics

Economical

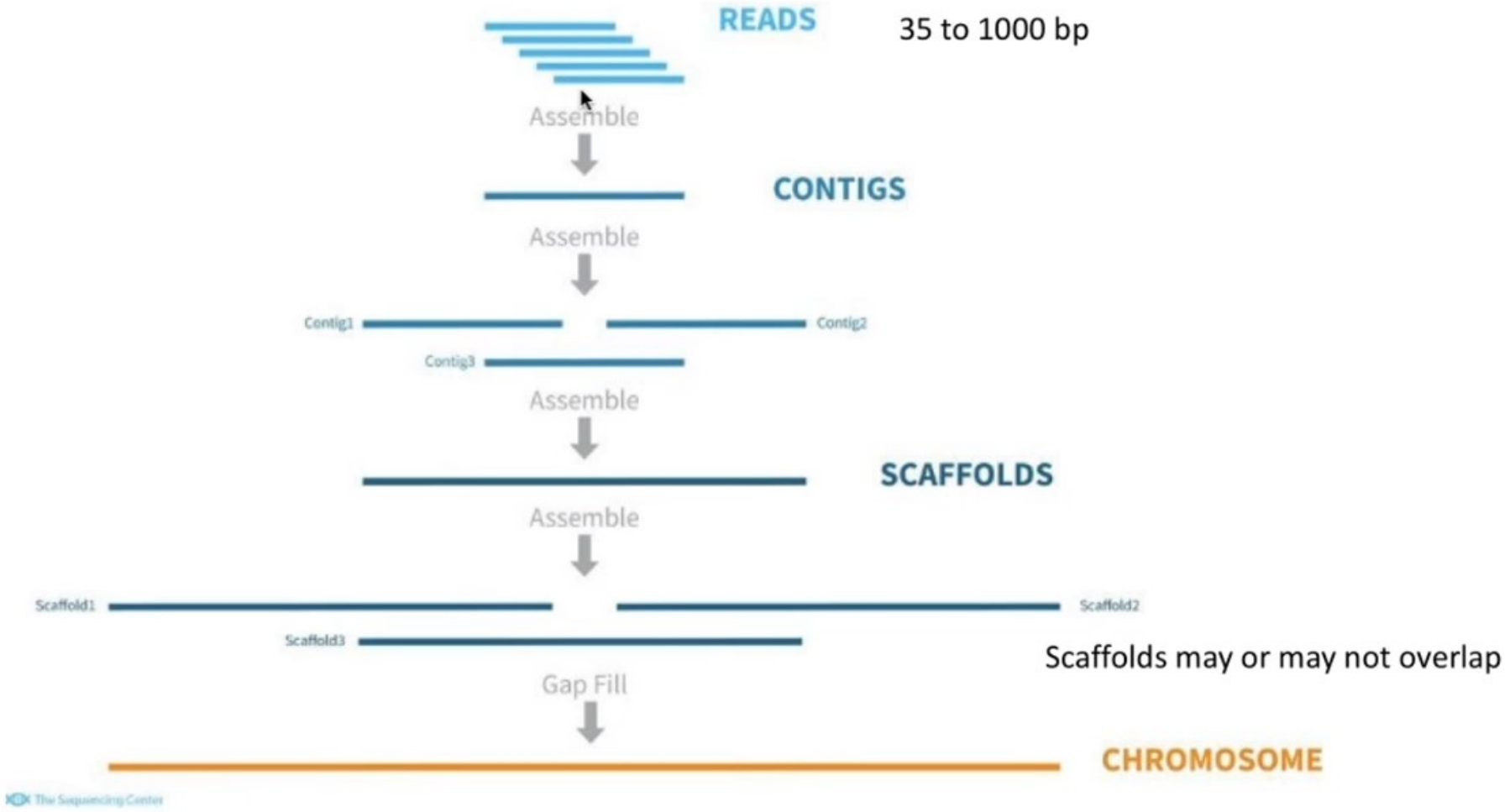
Genome Assembly





doi: https://doi.org/10.1371/journal.pcbi.1002821.g002

De novo Genome Assembly



Big Data Analysis

Transcriptomics

Why the normal cell turns into a tumour cell?

Normal Cell

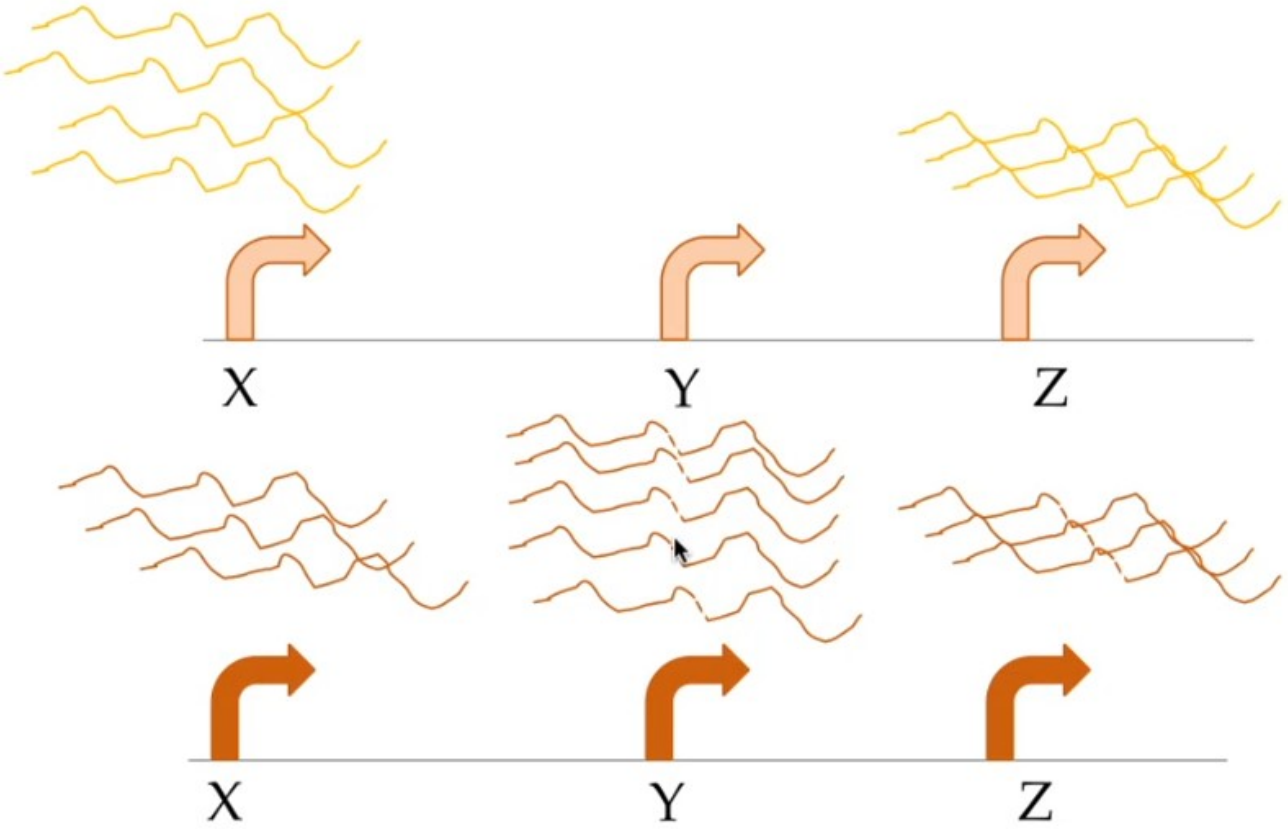
There is an underlying difference in the genetic mechanism

Tumour Cell

What is the difference in the expression of genes between normal and tumour cells?

Normal Cell

Tumour Cell



What are the genes that are upregulated (active) or downregulated (inactive) ?

How much the genes are expressed?

Next-Gen Sequencing - RNA-seq

Steps in RNA-seq

Steps in RNA-seq

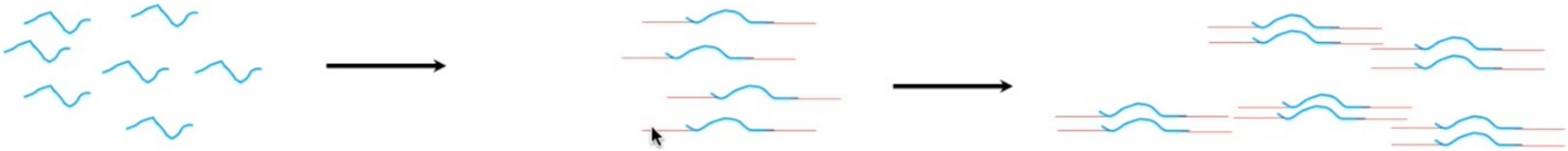
Sample (Library) Preparation

mRNA

Fragmentation
(200 – 300 nt)

cDNA Generation

Sample (Library) Preparation



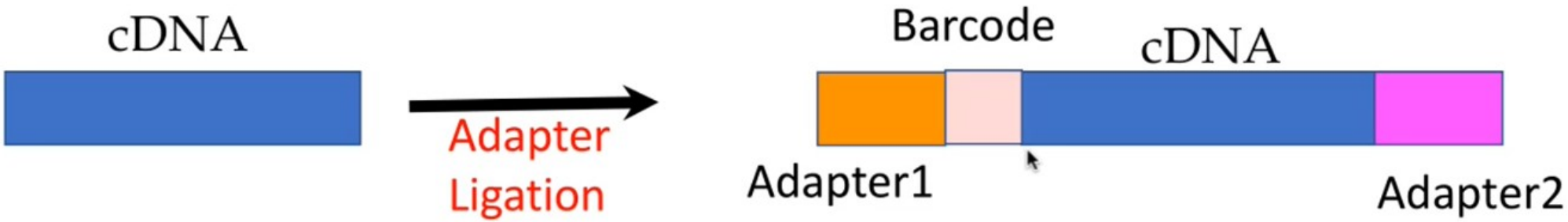
cDNA → Adapter Ligation → PCR Amplification

Quality Check

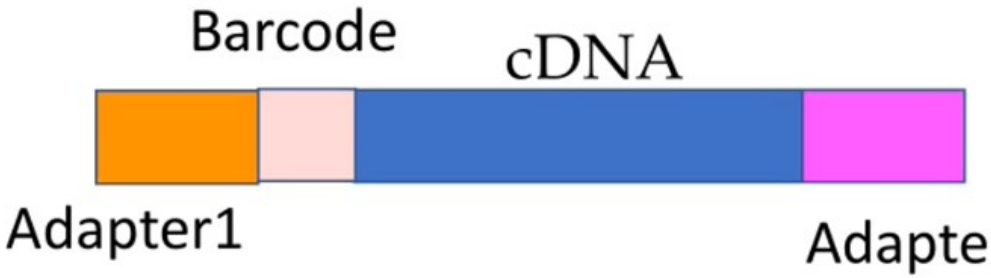
→ Check for optimum concentration of sample

Check for uniform length of the fragments (don't want too short and too long fragments)

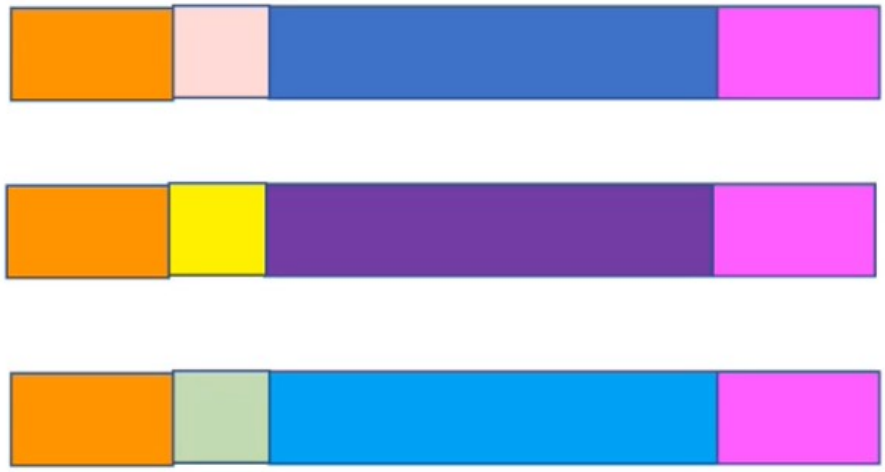
Sample (Library) Preparation



Sample (Library) Preparation

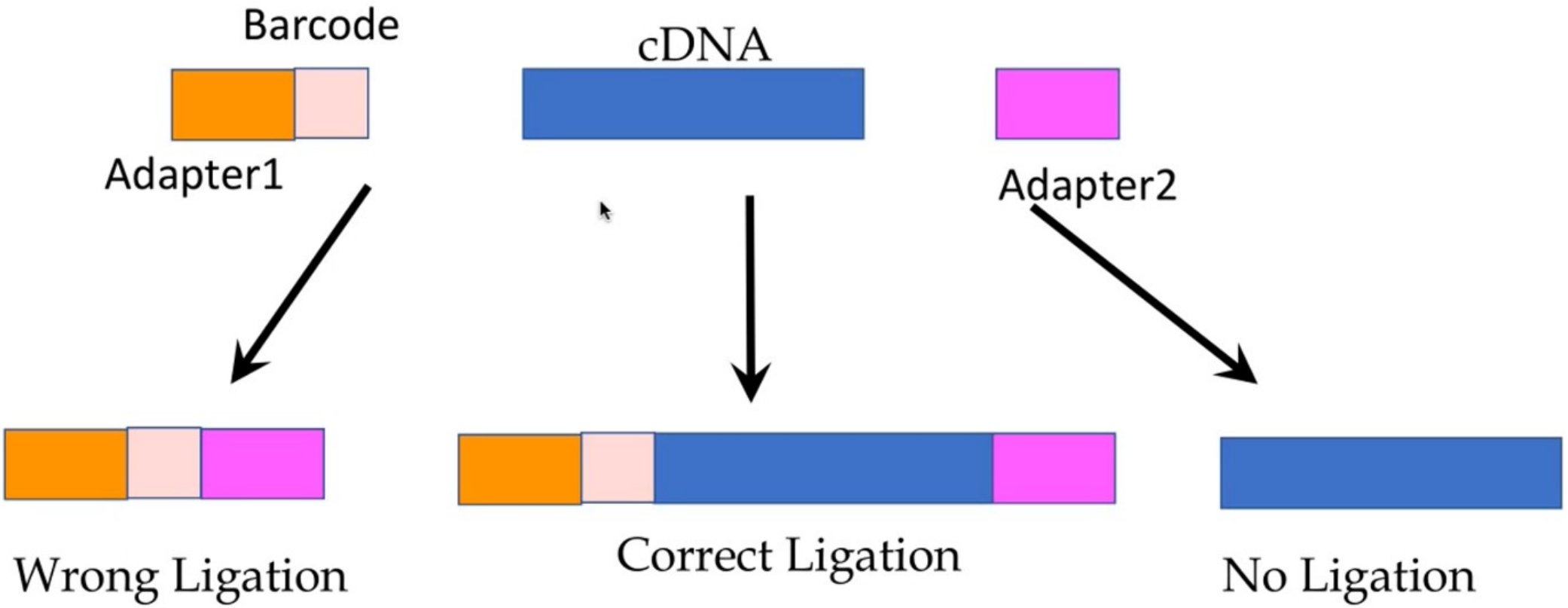


Barcode - A short unique sequence



All samples are sequenced together - Multiplexing

Sample (Library) Preparation Artefact(s)



Stranded vs Unstranded Library Preparation

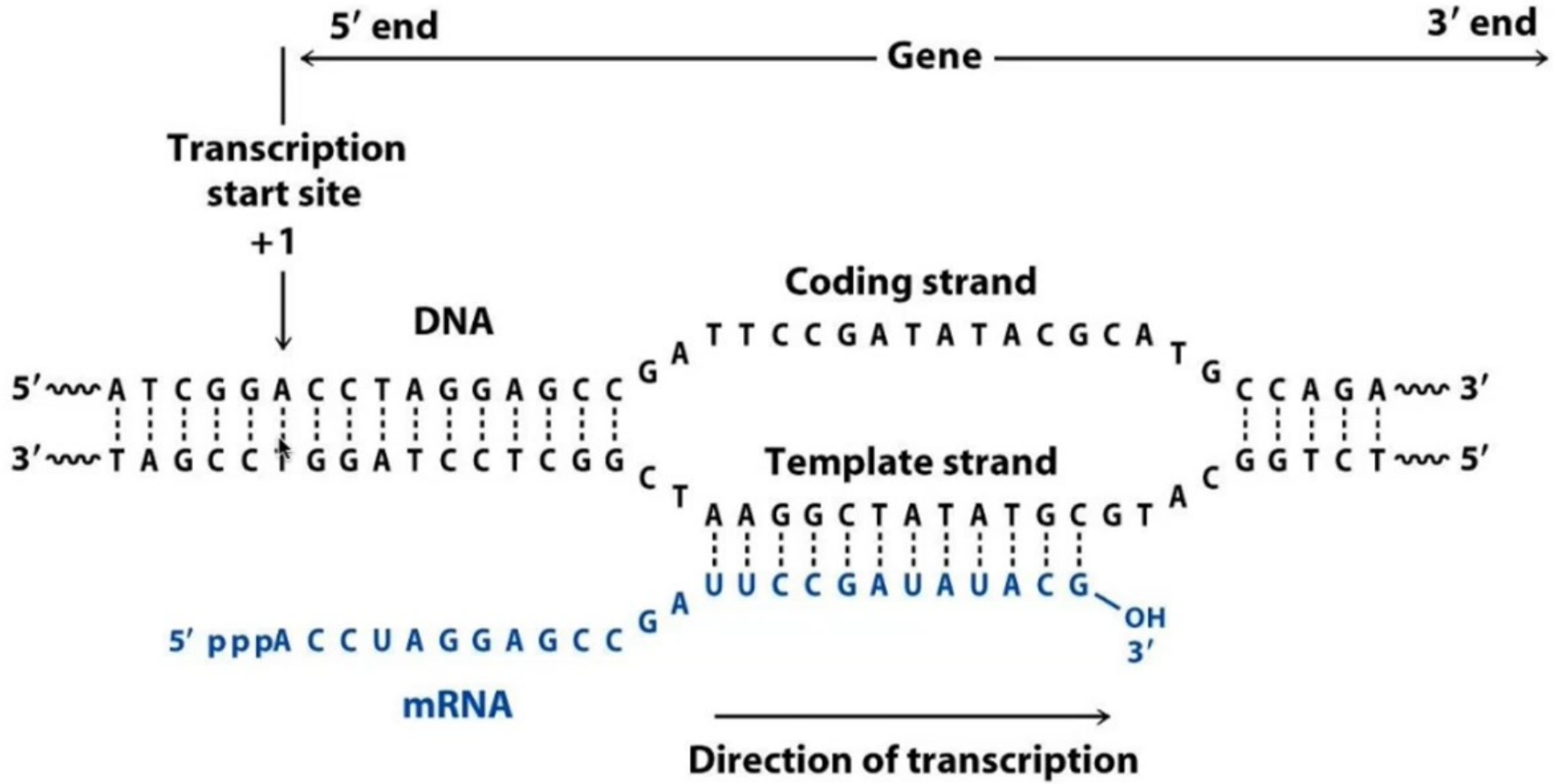
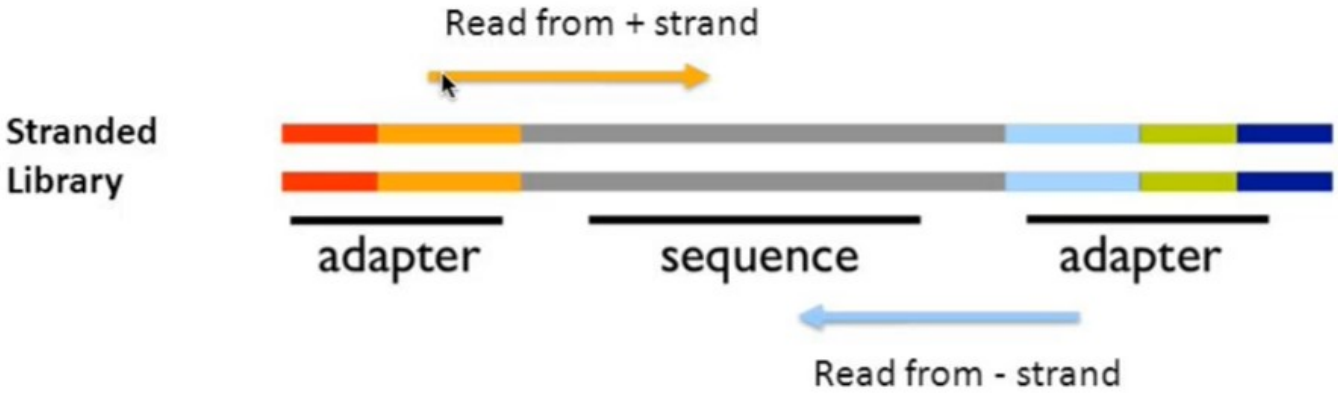
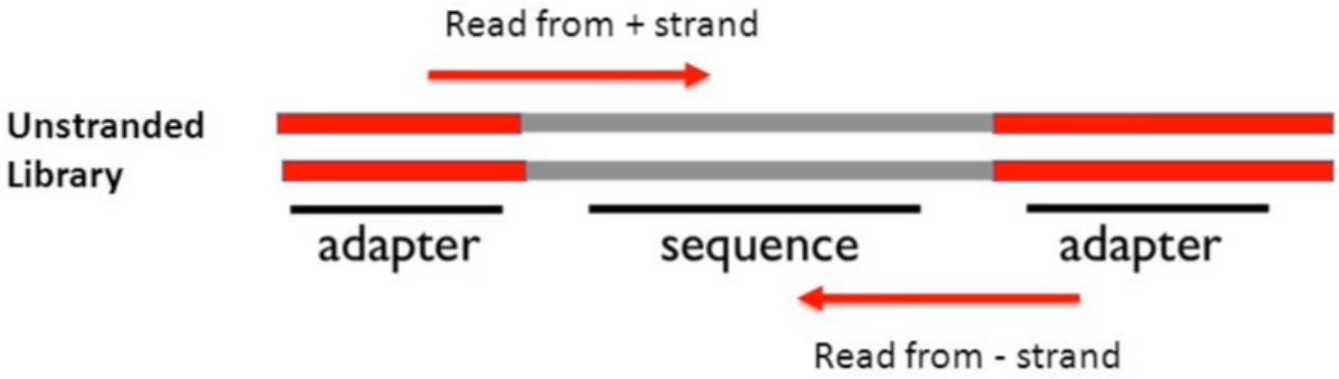


Figure 21-5 Principles of Biochemistry, 4/e

Stranded vs Unstranded Library Preparation

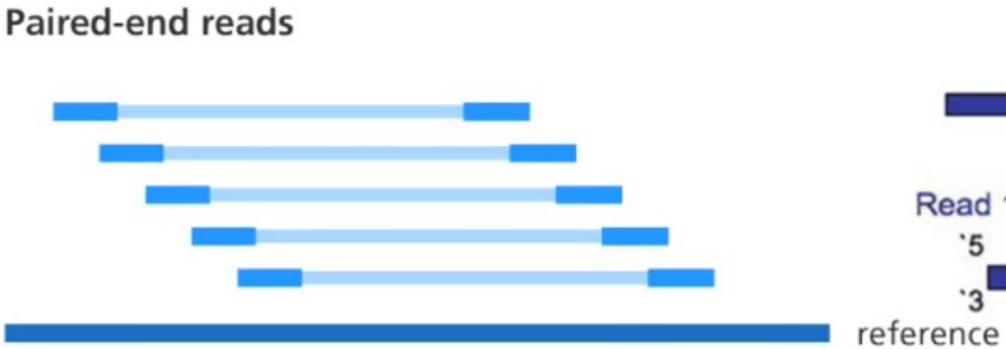
Stranded libraries provide information on the strand that codes for the RNA. This allows to disentangle transcription of overlapping genes.

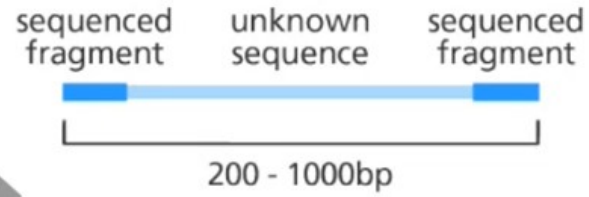


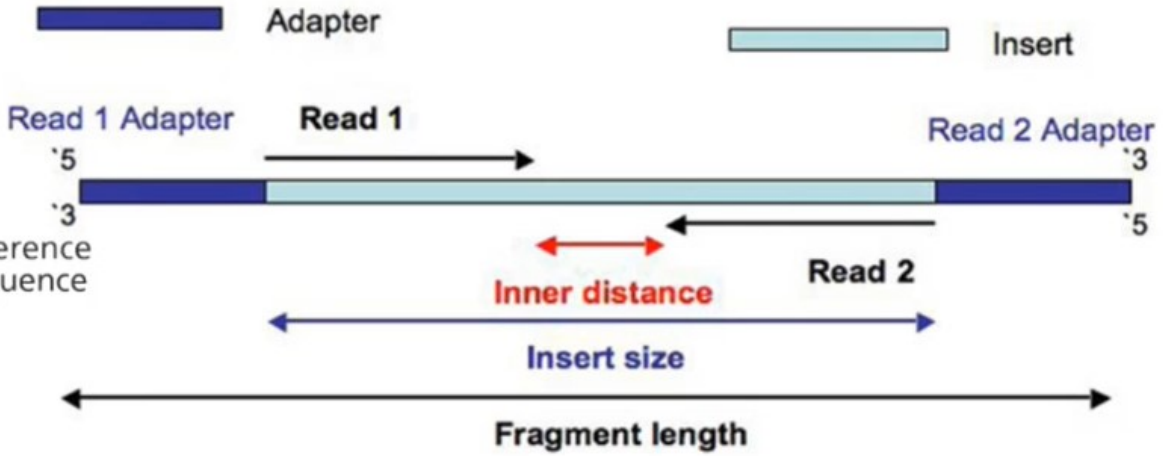


Single Vs. Paired End Read

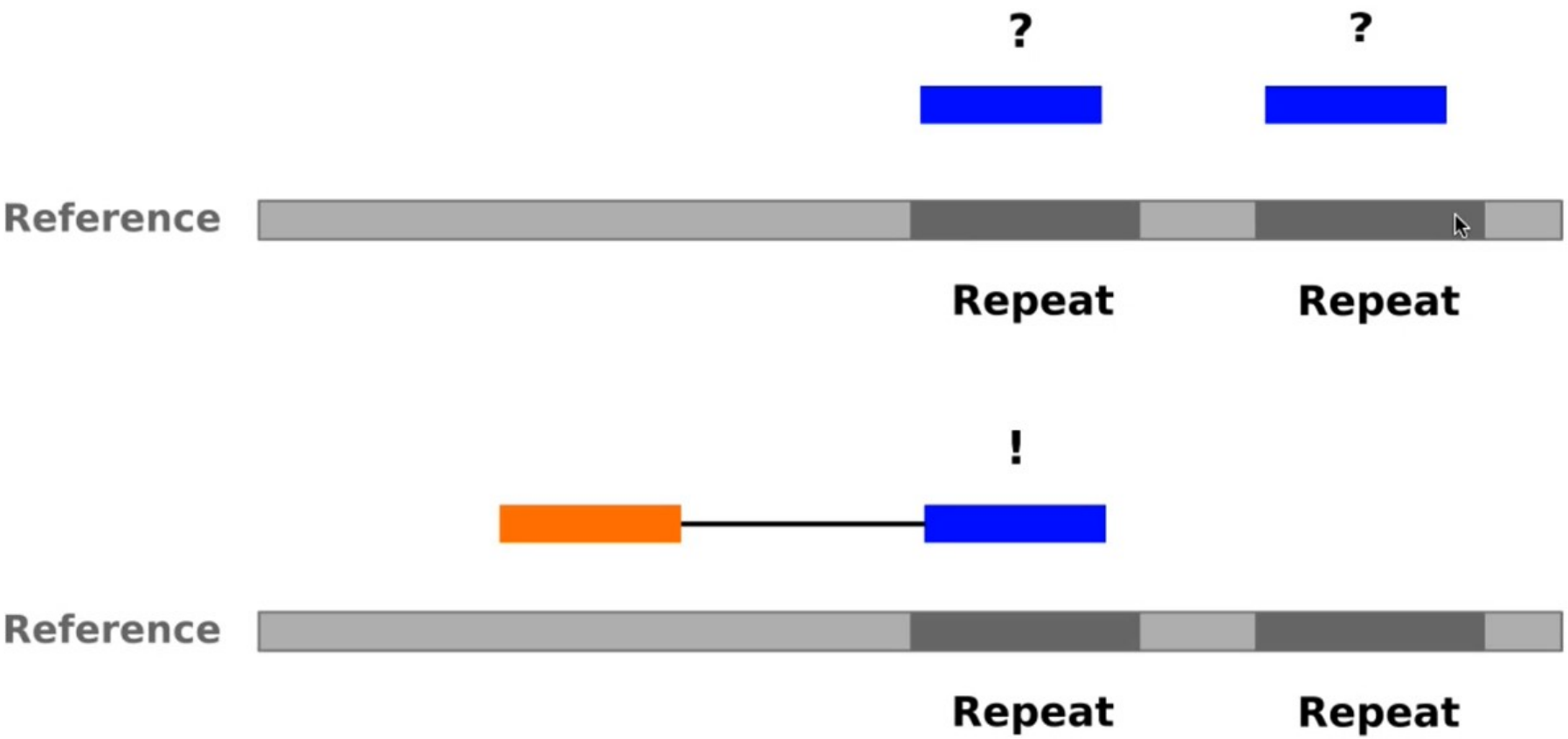
- Paired-end (mate-pair) sequencing enables sequencing both ends of the same fragment.
- This increases the coverage.
- Helps to resolve ambiguous alignment to the reference sequence







Single Vs. Paired End Read



Illumina sequencing (2nd Generation)

The "sequencing-by-synthesis" technology is used by Illumina. It was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge.

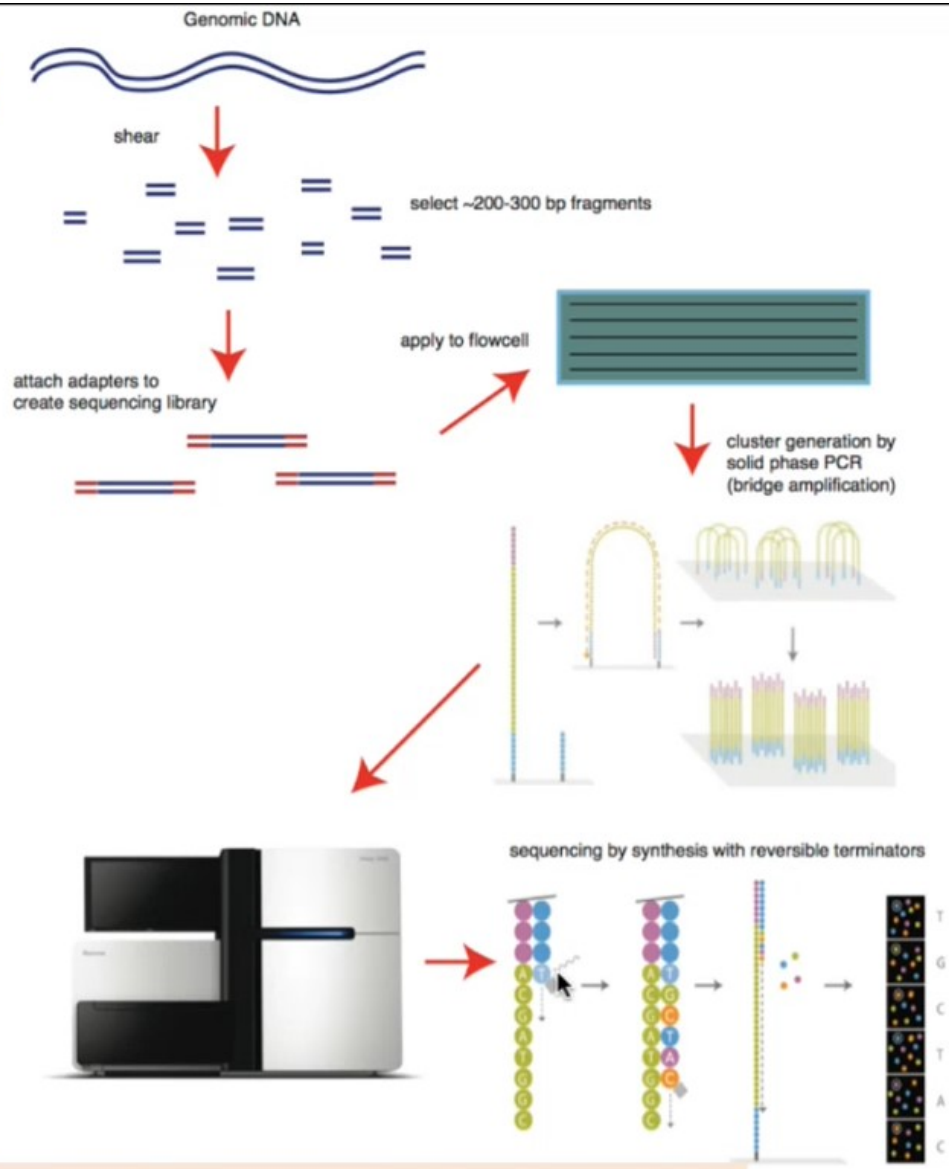
DNA or cDNA samples are randomly fragmented, usually into segments of 200 to 600 base pairs. These fragments are then ligated to adaptors and made single-stranded

Each fragment is amplified on the flow cell, and unlabeled nucleotides and polymerization enzymes are added. These additions, called "Bridge amplification," connect and lengthen the fragments of DNA on the flow cell

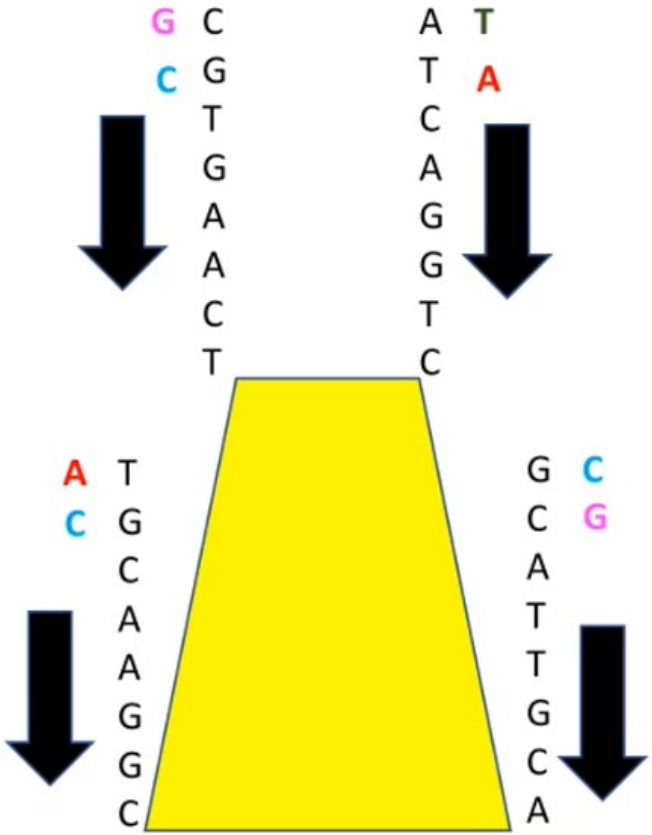
Illumina's "sequencing by synthesis" involves a proprietary method whereby four labeled reversible dNTP terminators, primers and DNA polymerase are added to the templates on the flow cell.

When excited by a laser, fluorescence from each cluster can be detected, which identifies the first base.

Base calls are made from signal intensity measurements during each cycle, reducing error rates further

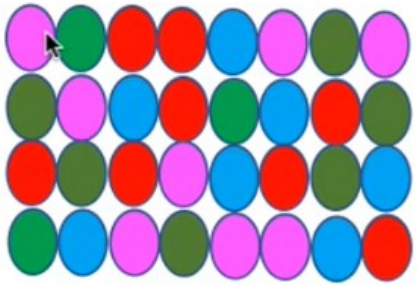


https://www.youtube.com/watch?v=fCd6B5HRaZ8

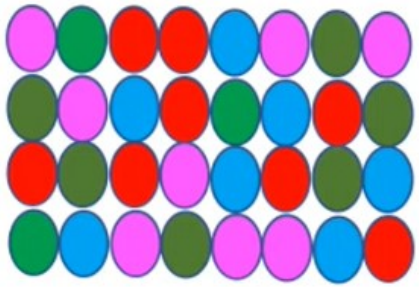

G C G T G A A C T
C G T G A A C T
A T A T C A G G T C
T A T C A G G T C
A C T G C A A G G C
C T G C A A G G C
G C G C A T T G C A
C G C A T T G C A

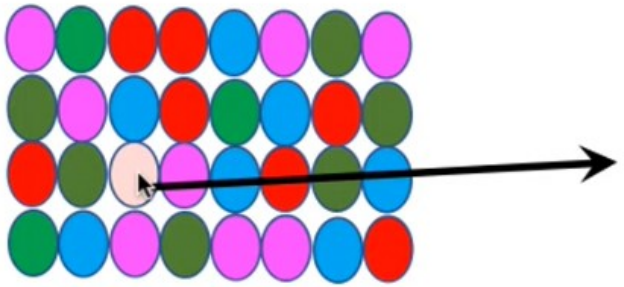
View from top of the flow cell

G A T C



G A T C

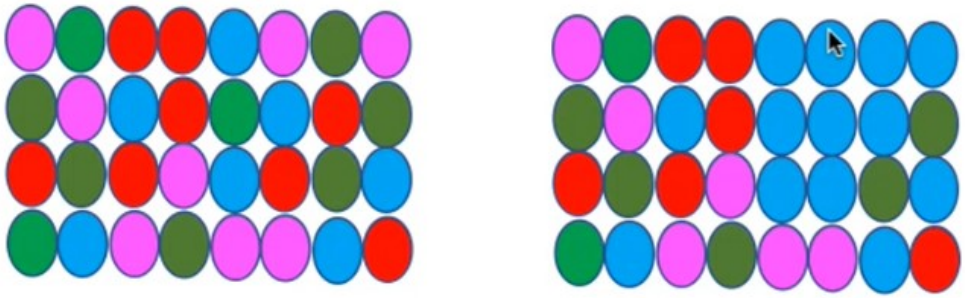




Probe is dull. Machine finds it difficult to call the correct base

Such bases will be called out with low quality scores

G A T C



Same bases occur very closely – “Low complexity”.
This leads to ambiguity in finding the correct position of bases due to blurring of colours.

Such bases will be called out with low quality scores

Steps in NGS Data Analysis

FASTQ Format

@SEQ_ID [Header Information Line starts with @ and followed by sequence identifier]

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT [Sequence]

+ [+ followed by sequence identifier or description]

!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65 [Quality score identifier]

Symbols = Low quality
Numbers = better
Alphabets = Best

Data Processing and Analysis Steps

1. Remove the low quality reads and adapters
2. Align the reads to the genome of interest
3. Count the number of reads per gene

Data Processing and Analysis Steps

1. Remove the low quality reads and adapters
2. Align the reads to the genome of interest
3. Count the number of reads per gene

Quality Scores

$$Q = -10 \log_{10}(P)$$

Q -Phred Quality Score
P- Probability of calling an incorrect base

Q	P	Base Call Accuracy
10	0.1	90%
20	0.01	99%
30	0.001	99.9%
40	0.0001	99.99%
50	0.00001	99.999%

Genome Alignment

ggttaaatagtctaaggcccggaa....

ggttaa aggcct
atagtc taaggc
ttcgaa ccggaa

Fragment genome

Index fragments and their respective locations

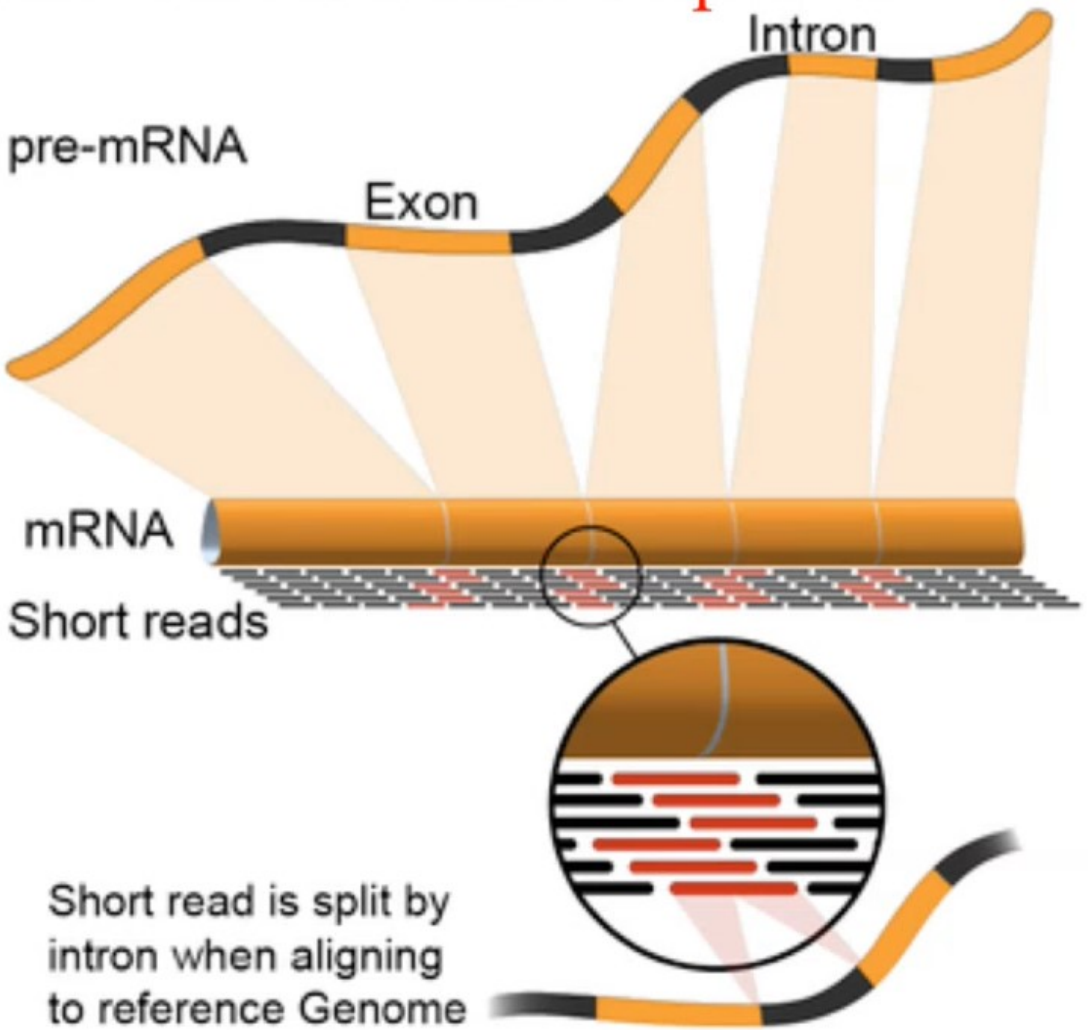
AGCTAAGCCGA

AGCTAA
AAGCCA
AGCCGA

Fragment Read

Matched fragment map the chromosome and its position within the chromosome

Alignment to Reference Sequence



Alignment involves the knowledge of splice junctions

Credit: Rgocs

Sequenced reads have to be normalised for

1. Sequencing depth → no. of times a particular gene gets sequenced

2. Length of genes

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Sample 1, 2 and 3 are replicates – experiments repeated multiple times under similar experimental conditions.

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Sample 3 has more reads than others irrespective of gene sizes.

It represents that Sample 3 has more depth than others.

1. Sample 1 and 2 may have more low quality reads
2. Sample 3 has slightly higher concentration than sample 1 & 2 on the flow cell

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Gene C has twice as many as reads as Gene B regardless of samples.

It represents that Sample 3 has more depth than others.

Reads per kilo base of transcript per Million mapped Reads (RPkM)

$$\text{RPkM} = \frac{\text{\# of Reads} \times 10^6}{\text{Total \# of Reads} \times \text{Gene size}}$$

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1
Total	**35**	**45**	**106**
Scaling Factor	**3.5**	**4.5**	**10.6**

$$\text{Scaling factor for Read counts} = \frac{\text{Total \# counts}}{10}$$

In Reality, we have to divide by 10^6

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	1.43	1.78	1.42
B (2 kb)	2.86	2.67	2.83
C (4 kb)	5.71	5.56	5.56
D (10 kb)	0	0	0.09

Scaled for Read count

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Raw Count

RPkM

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	1.43	1.78	1.42
B (2 kb)	1.43	1.39	1.42
C (4 kb)	1.43	1.78	1.42
D (10 kb)	0	0	0.009

Scaled for Read count and Gene length

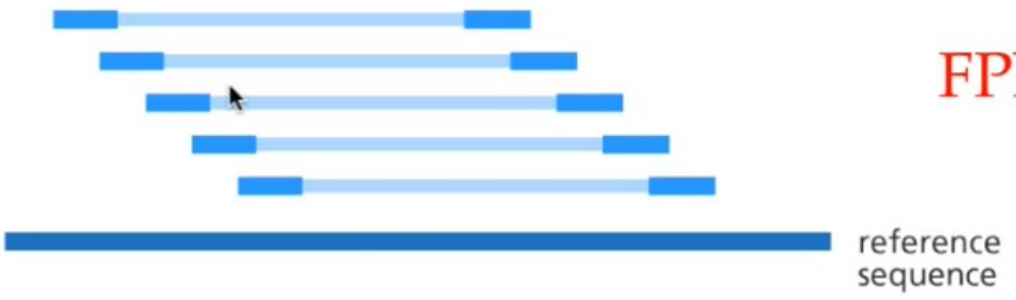
Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Raw Count

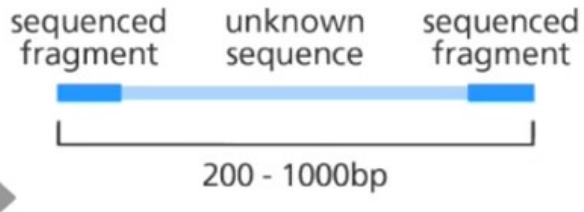
Single-end reads

RPkM = Reads Per kilobase Million

Paired-end reads



FPkM = Fragments Per kilobase Million



Both reads mapped to the same fragment are NOT counted twice

Transcripts Per Million (TPM)

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	5	6	15
C (4 kb)	5	6.25	15
D (10 kb)	0	0	0.1
Total	15	20.25	45.1

Scaled for Gene length – Divide each read count by respective gene length

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Raw Count

Transcripts Per Million (TPM)

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	5	6	15
C (4 kb)	5	6.25	15
D (10 kb)	0	0	0.1
Scaling Factor	1.5	2.025	4.51

Scaled for Gene length

$$\text{Scaling factor for Gene Size} = \frac{\text{Total \# counts after gene length scaling}}{10}$$

In Reality, we have to divide by 10^6

Transcripts Per Million (TPM)

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	3.33	3.95	3.326
B (2 kb)	3.33	2.96	3.326
C (4 kb)	3.33	3.09	3.326
D (10 kb)	0	0	0.02

Scaled for Gene length and Read count - Divide each read count by respective scaling factor

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	5	8	15
B (2 kb)	10	12	30
C (4 kb)	20	25	60
D (10 kb)	0	0	1

Raw Count

RPkM Vs. TPM

RPkM

Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	1.43	1.78	1.42
B (2 kb)	1.43	1.39	1.42
C (4 kb)	1.43	1.78	1.42
D (10 kb)	0	0	0.009
Total	4.29	4.5	4.25

Total Reads are different.

TPM

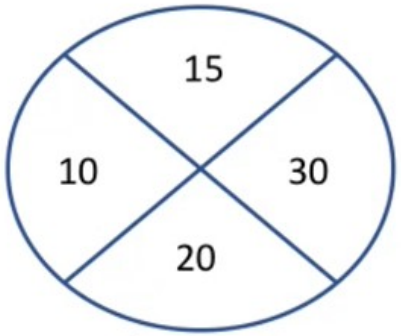
Gene name	Sample 1	Sample 2	Sample 3
A (1 kb)	3.33	3.95	3.326
B (2 kb)	3.33	2.96	3.326
C (4 kb)	3.33	3.09	3.326
D (10 kb)	0	0	0.02
Total	10	10	10

Total Reads are same. Comparisons across samples make sense.

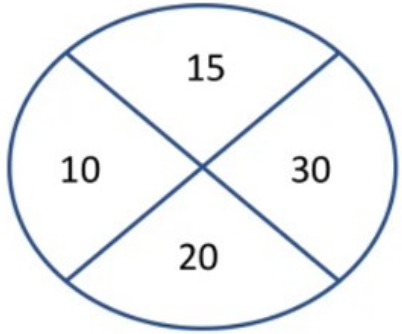
TPM

RPKM/FPKM

Sample 1

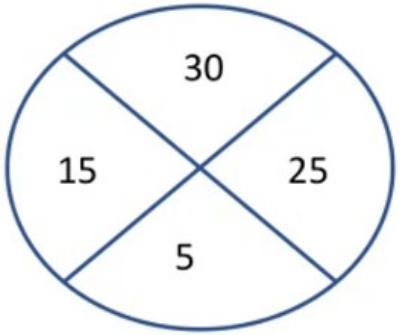


75

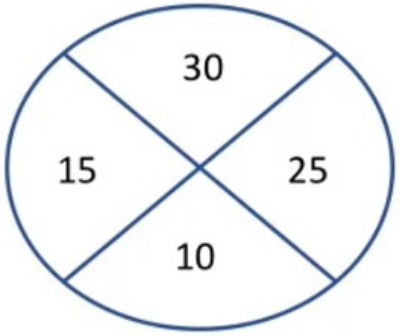


75

Sample 2



75



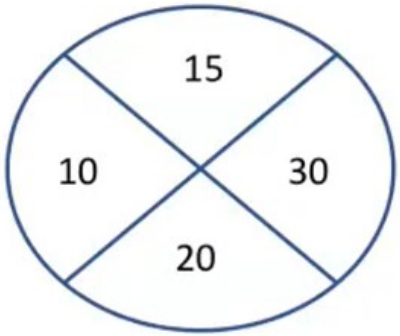
80

Total Reads are the same. Read counts can be compared across samples

Total Reads are not the same. Read counts can not be compared across samples

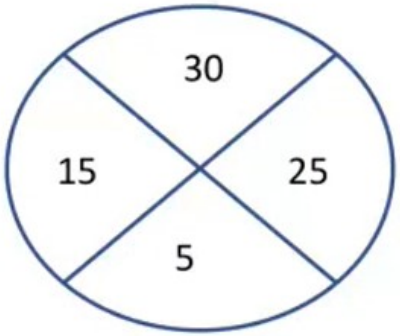
TPM

Sample 1



75

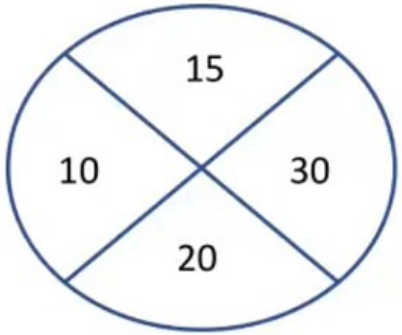
Sample 2



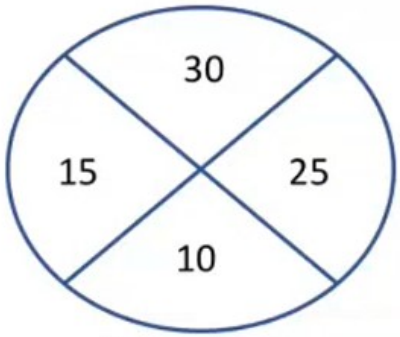
75

Total Reads are the same. Read counts can be compared across samples

RPKM/FPKM



75



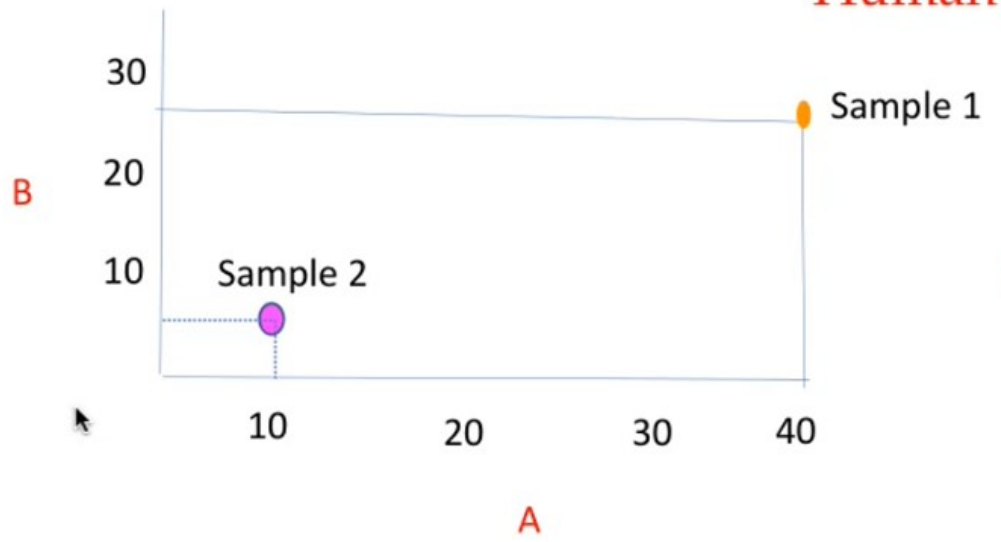
80

Total Reads are not the same. Read counts can not be compared across samples

Plot the Data

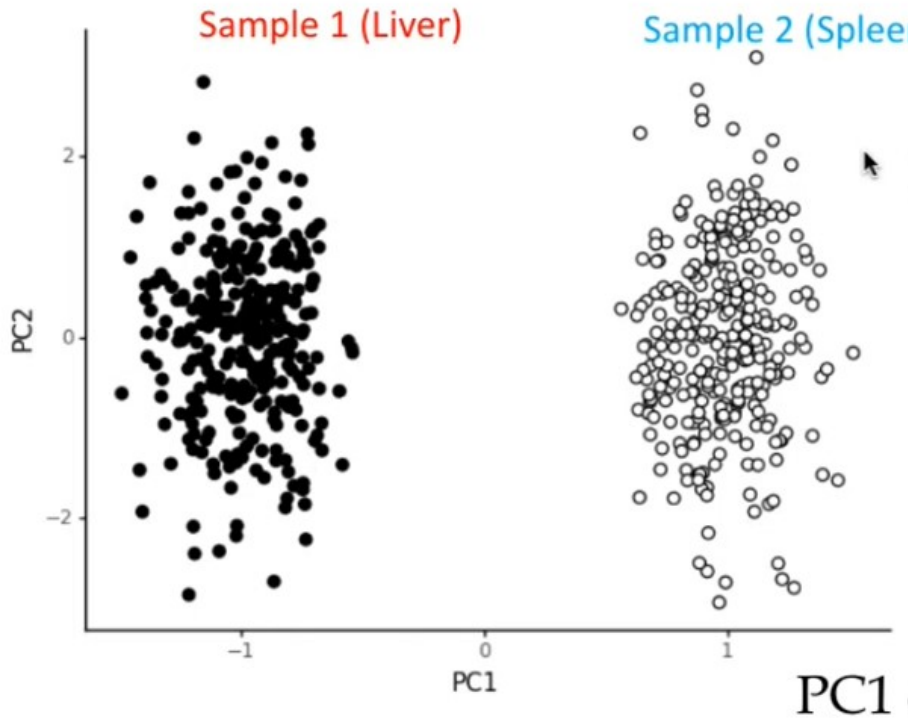
Gene	Sample 1	Sample 2
A	40	10
B	25	5

Human genome has ~ 25000 genes!



How to plot ~ 25000 genes?

Principal Component Analysis (PCA)



PCA reduces the dimensionality of the data

PC1 captures most of the variations in the data

PC2 captures second most variations in the data

Pic Source: nxn.se

Differential Gene Analysis

Problems in Library Normalisation

1. Differences in Library Size (Sequencing depth)

2. Differences in Library Composition

Problem 2 is not addressed by RPKM/FPKM/TPM

Differential Gene Analysis

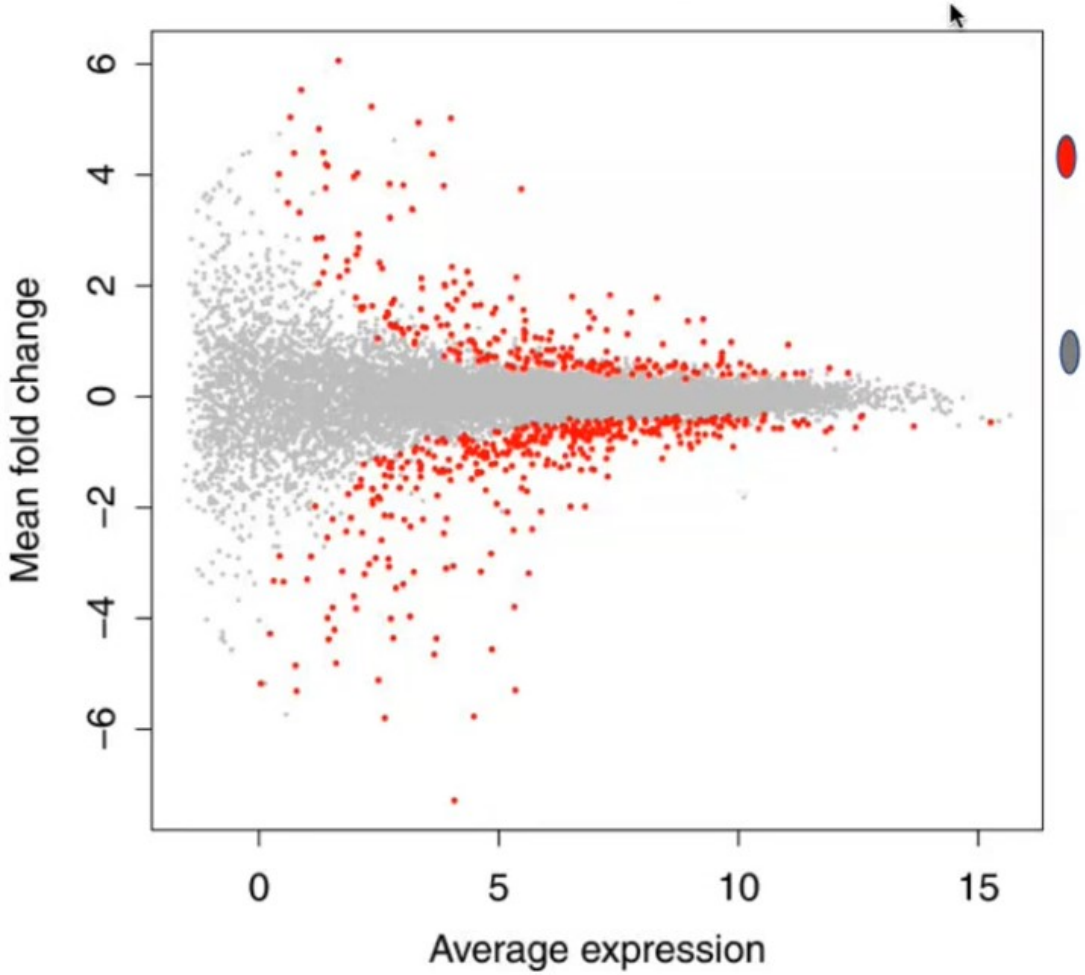
Differences in library composition

Gene name	Liver	Spleen
A	30	235
B	24	188
C	0	0
D	563	0
E	5	39
F	13	102
Total	635	635

D is the only differentially expressed gene

For other genes in Spleen sample, the reads got distributed

Differential Gene Analysis (MA plot)



Mean Average

Gene expressions that are different between WT and Mutant Sample

No significant change between WT and Mutant sample

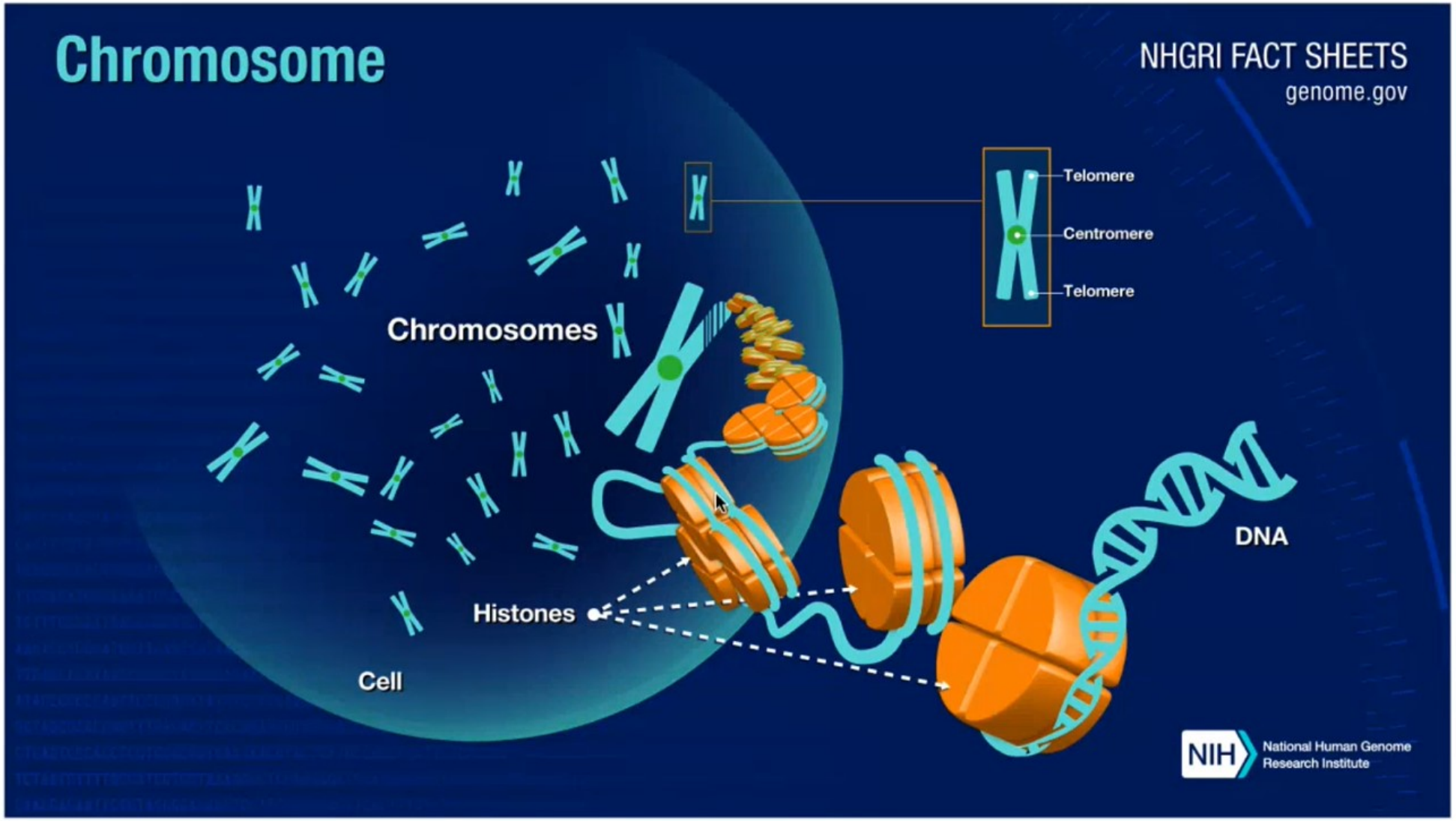
Horizontal : Left to Right

→ Low to High Expression

Pic Source: Fernandes et al, Microbiome, 2014

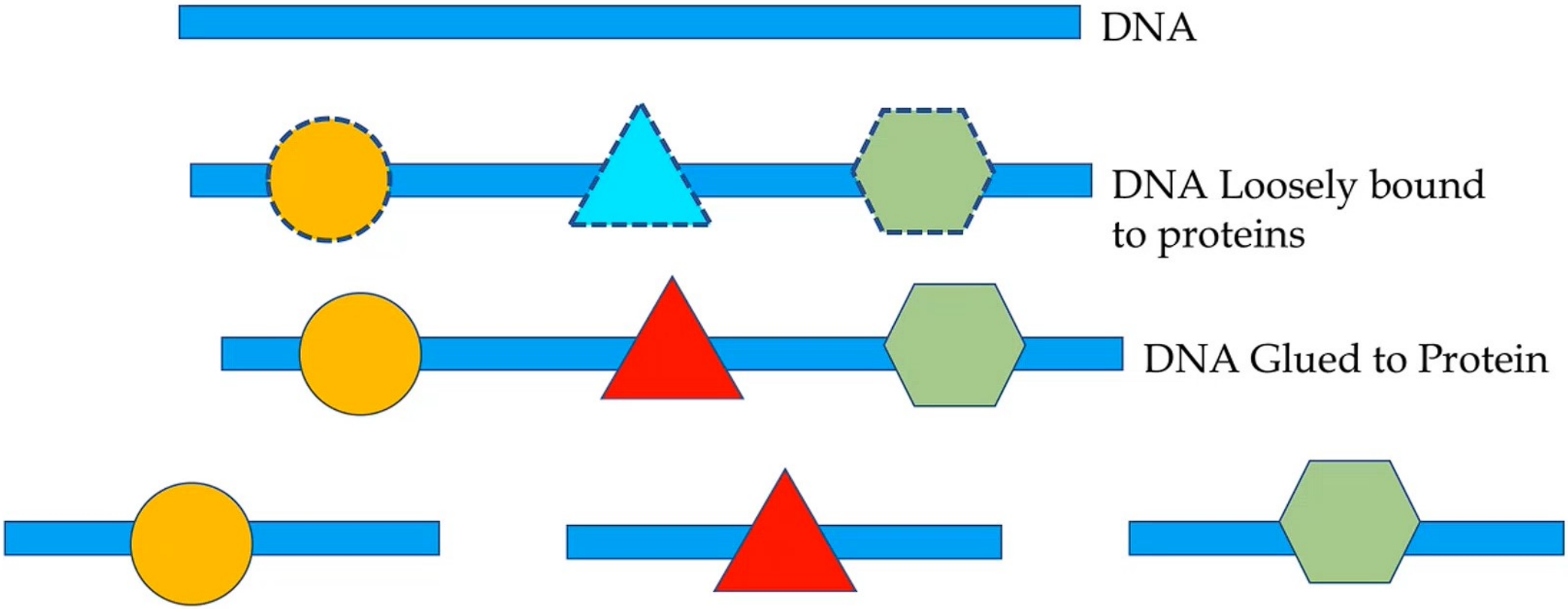
Big Data Analysis

CHIP-seq



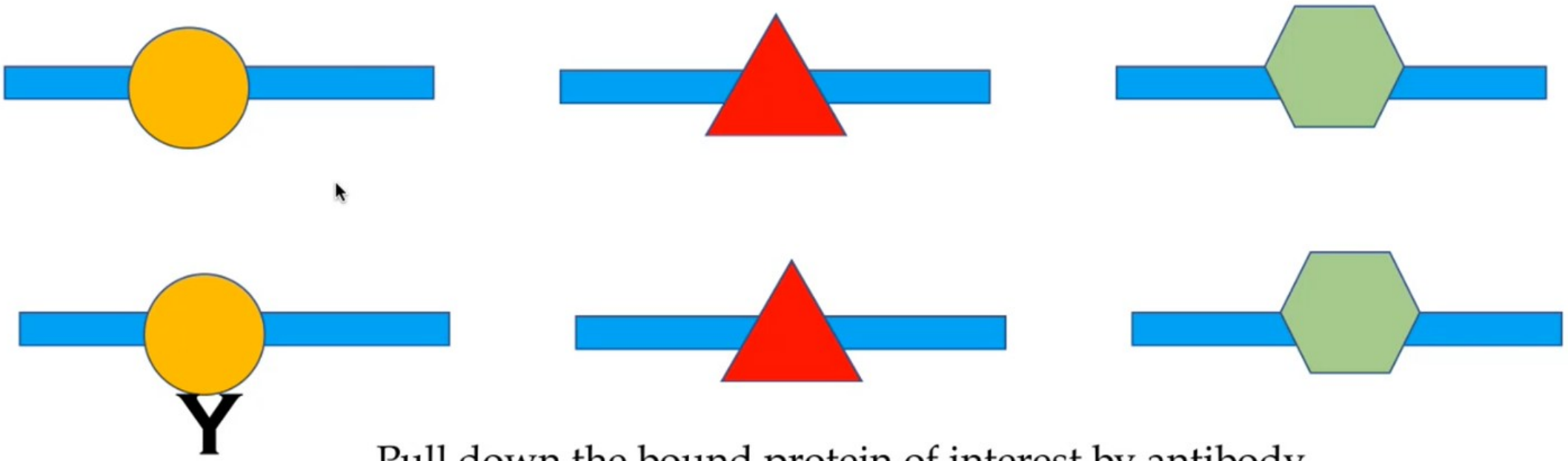
Chromatin = DNA + Histones + Auxiliary proteins

How do we find the region bound by protein of interest?



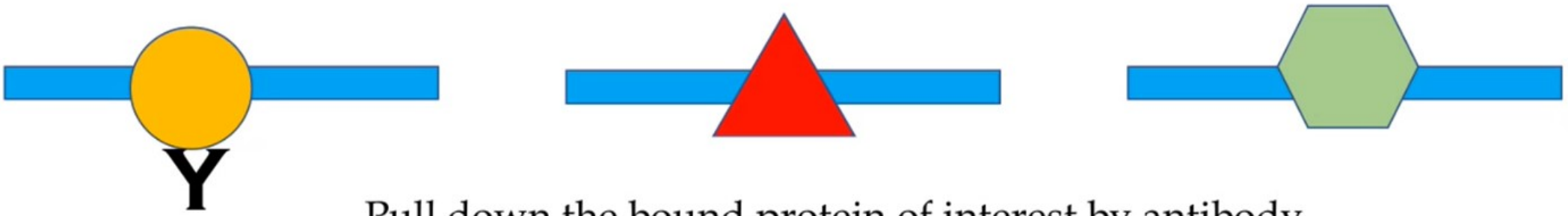
Fragmentation of DNA (300 bp)

CHIP-seq - Chromatin Immunoprecipitation followed by high throughput sequencing

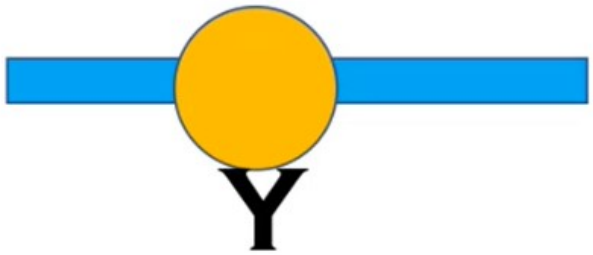


Pull down the bound protein of interest by antibody

CHIP-seq - Chromatin Immunoprecipitation followed by high throughput sequencing



Pull down the bound protein of interest by antibody



Remove the glue, the protein of interest and the antibody

DNA fragment bound by protein of interest

Adapter ligation at the end and PCR amplification (Library Preparation)

Illumina sequencing (2nd Generation)

The “sequencing-by-synthesis” technology is used by Illumina. It was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge.

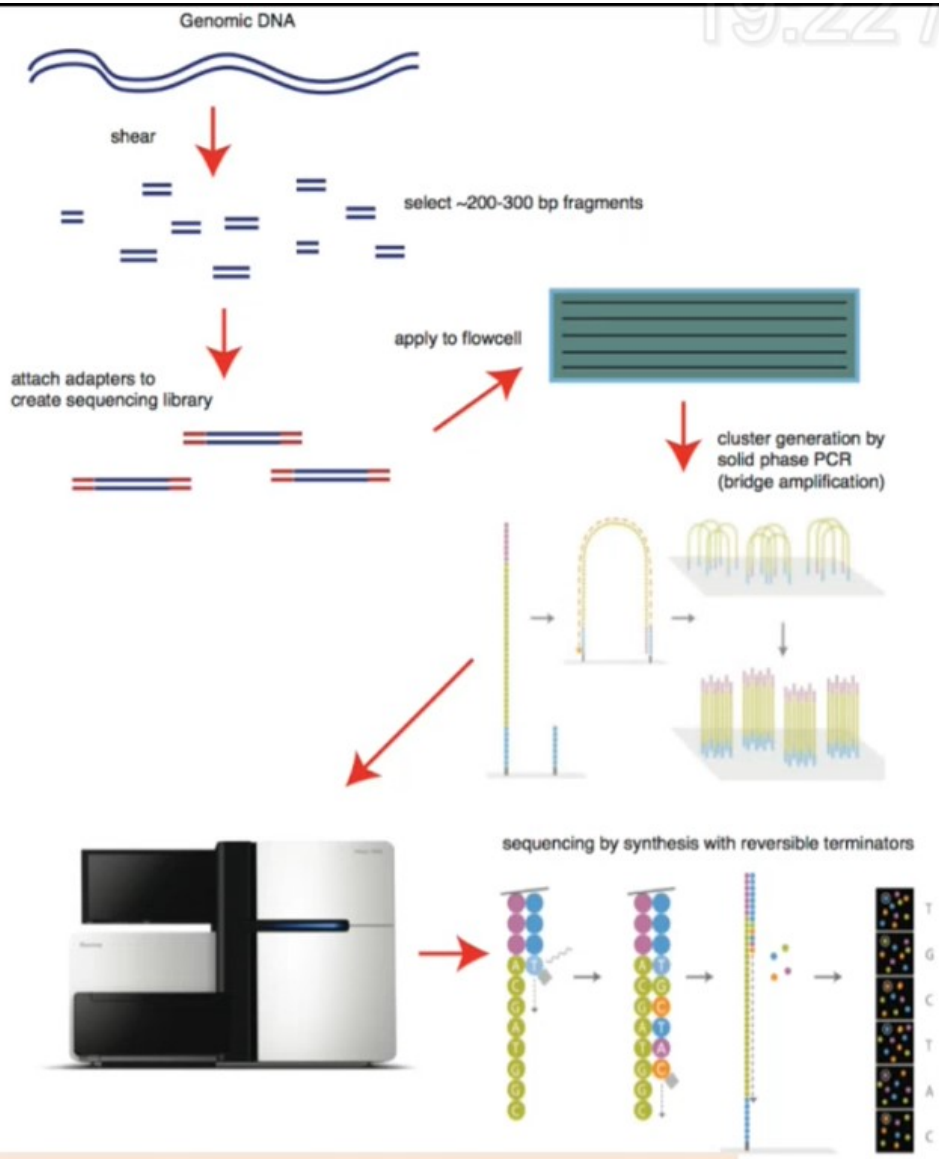
DNA or cDNA samples are randomly fragmented, usually into segments of 200 to 600 base pairs. These fragments are then ligated to adaptors and made single-stranded

Each fragment is amplified on the flow cell, and unlabeled nucleotides and polymerization enzymes are added. These additions, called “Bridge amplification,” connect and lengthen the fragments of DNA on the flow cell

Illumina’s “sequencing by synthesis” involves a proprietary method whereby four labeled reversible dNTP terminators, primers and DNA polymerase are added to the templates on the flow cell.

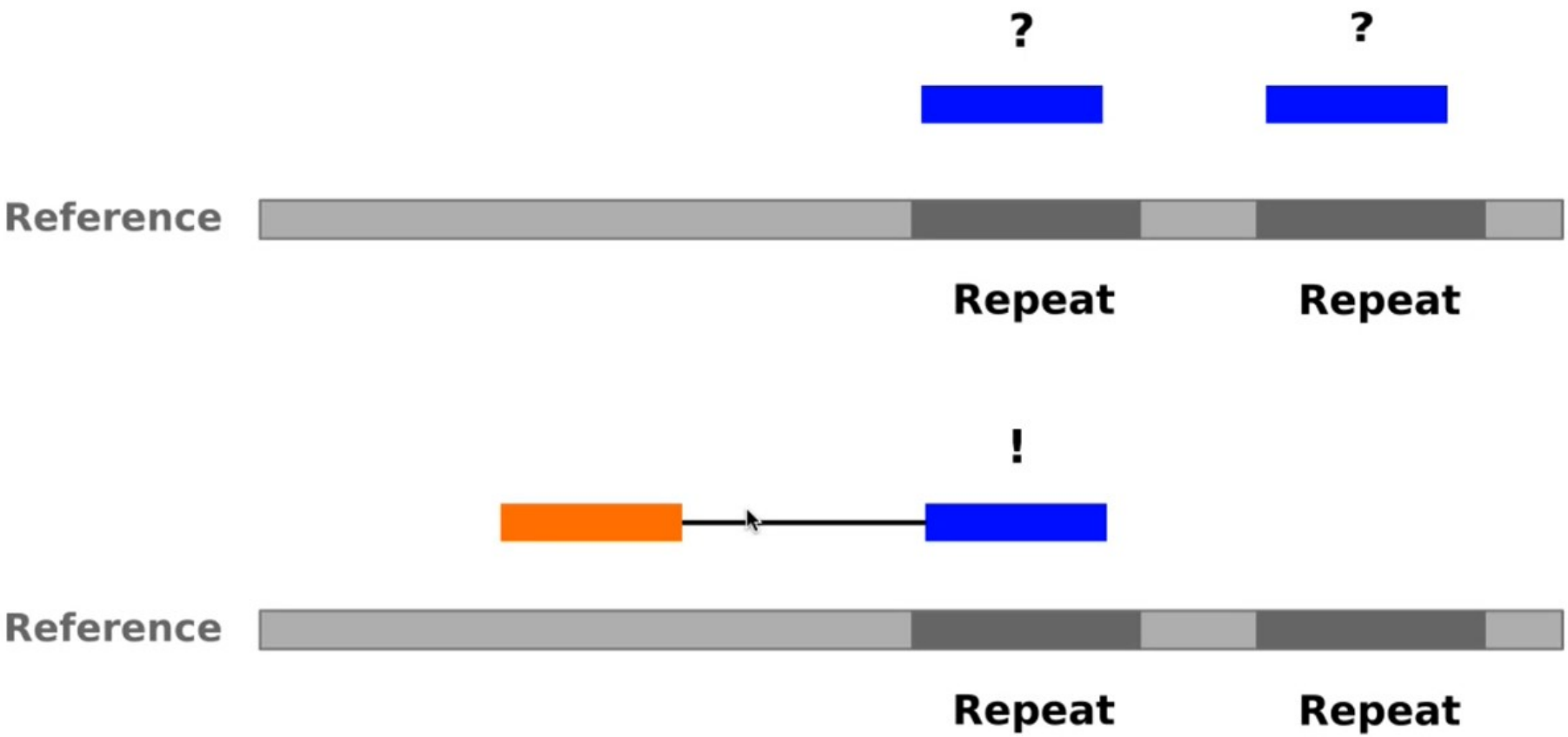
When excited by a laser, fluorescence from each cluster can be detected, which identifies the first base.

Base calls are made from signal intensity measurements during each cycle, reducing error rates further



https://www.youtube.com/watch?v=fCd6B5HRaZ8

Single Vs. Paired End Read



Galaxyproject.org

Steps in NGS Data Analysis

FASTQ Format

@SEQ_ID [Header Information Line starts with @ and followed by sequence identifier]

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT [Sequence]

+ [+ followed by sequence identifier or description]

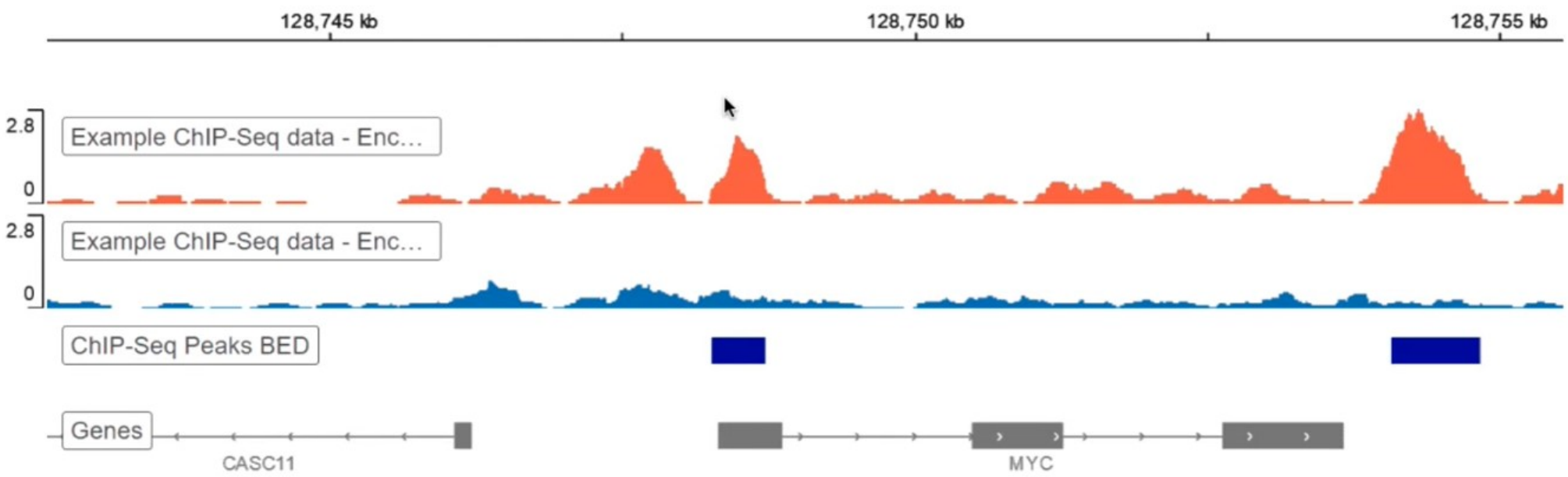
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65 [Quality score identifier]

Symbols = Low quality
Numbers = better
Alphabets = Best

Data Processing and Analysis Steps

1. Remove the low quality reads and adapters
2. Align the reads to the genome of interest (SAM/BAM format)
3. Count the number of reads per gene (Treated vs. Control)

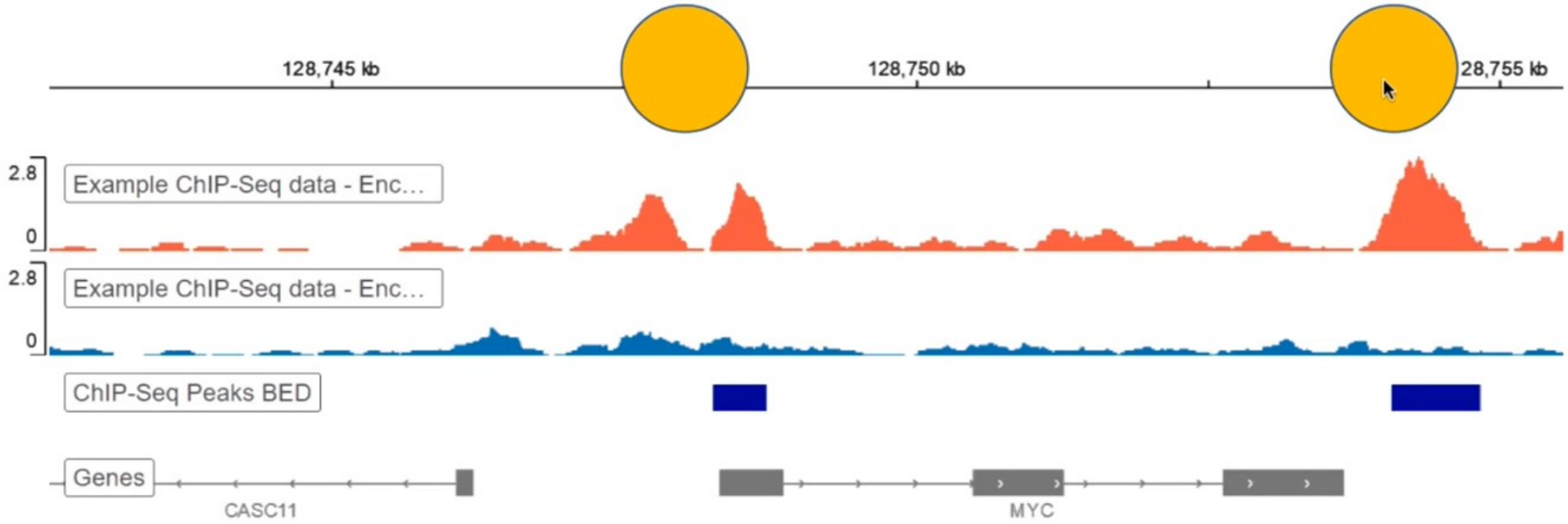
VISUALISING THE READS MAPPED ONTO THE REFERENCE GENOME



GACCT

Basepairtech.com

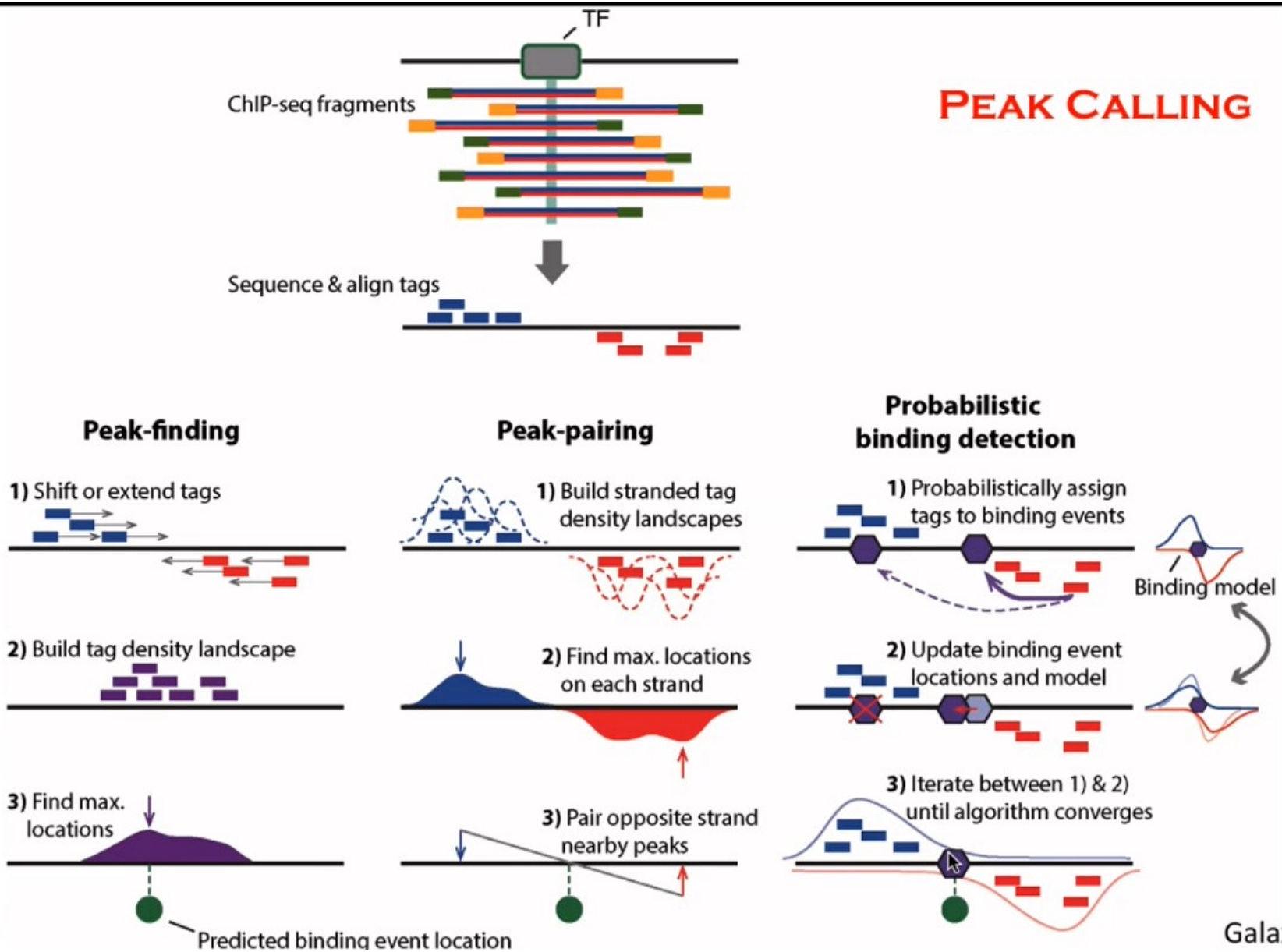
VISUALISING THE READS MAPPED ONTO THE REFERENCE GENOME



GACCT

Basepairtech.com

Peak Calling



Galaxyproject.org

Phylogenetic Analysis

What is Phylogenetics?

- The taxonomical classification of organisms on the basis of their degree of evolutionary relatedness.
  - originated by Willi Henning, a German zoologist. *Grundzüge einer Theorie der phylogenetischen Systematik* (1950),
  - *A Dictionary of Earth Sciences, © Oxford University Press 1999*

Tree of Life

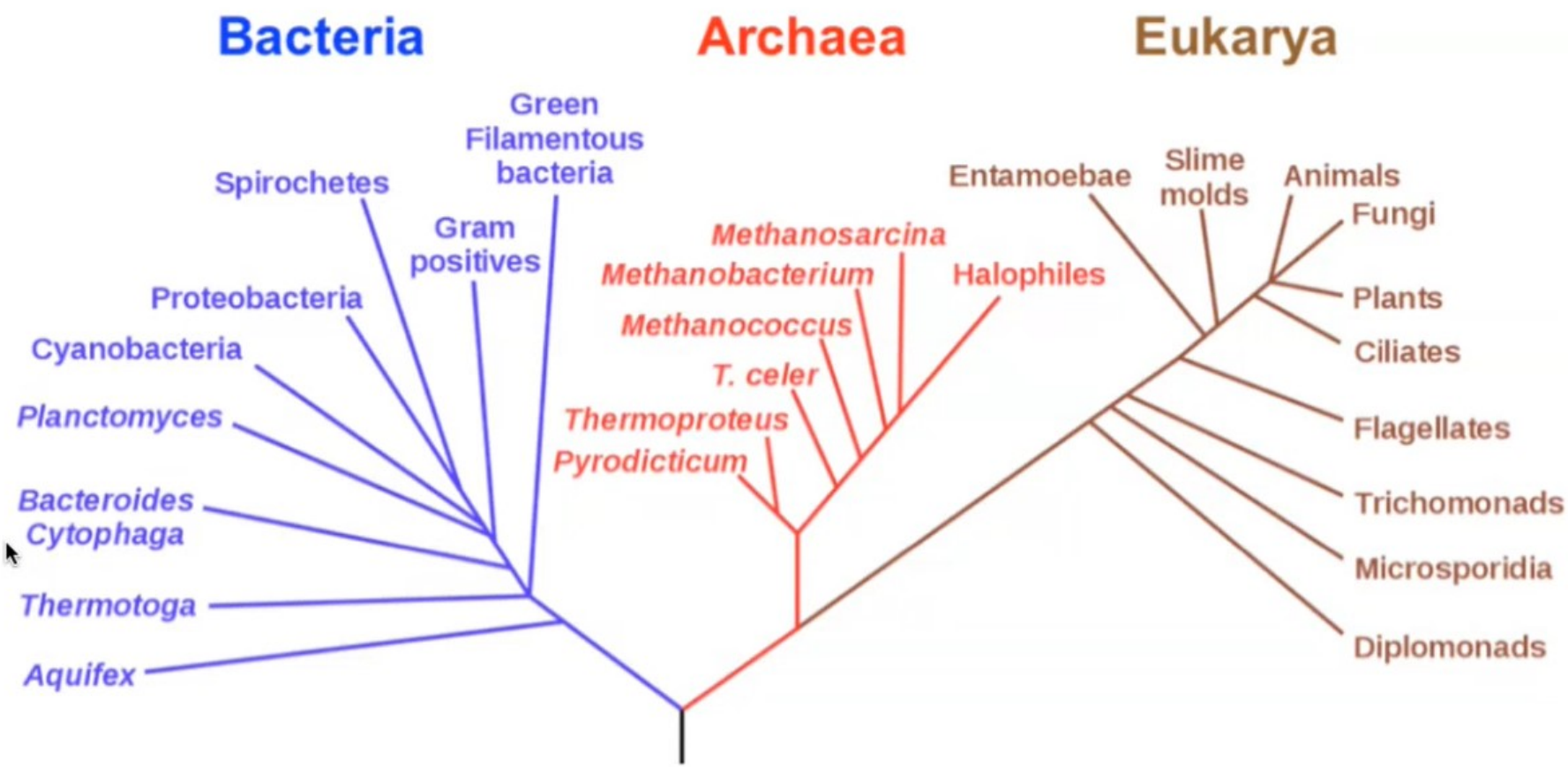
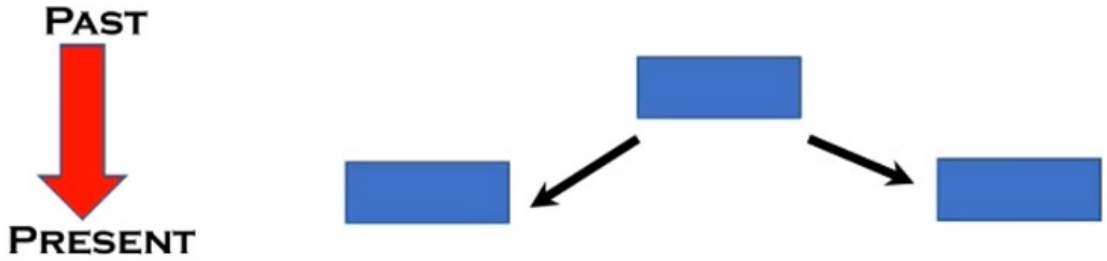
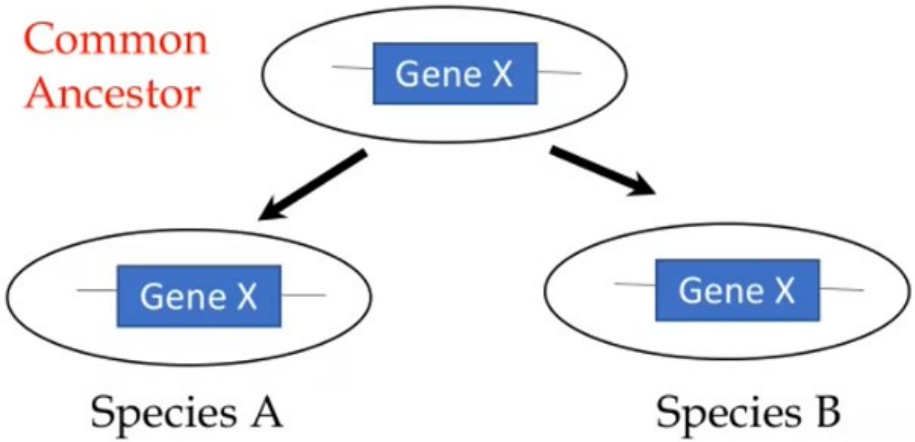


Image credit: NASA Astrobiology Institute & Eric Gaba

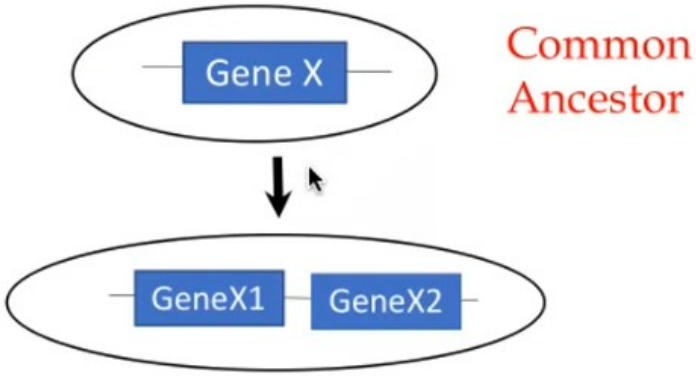
Evolutionary Relationships



HOMOLOGY:
Descent from a common ancestor

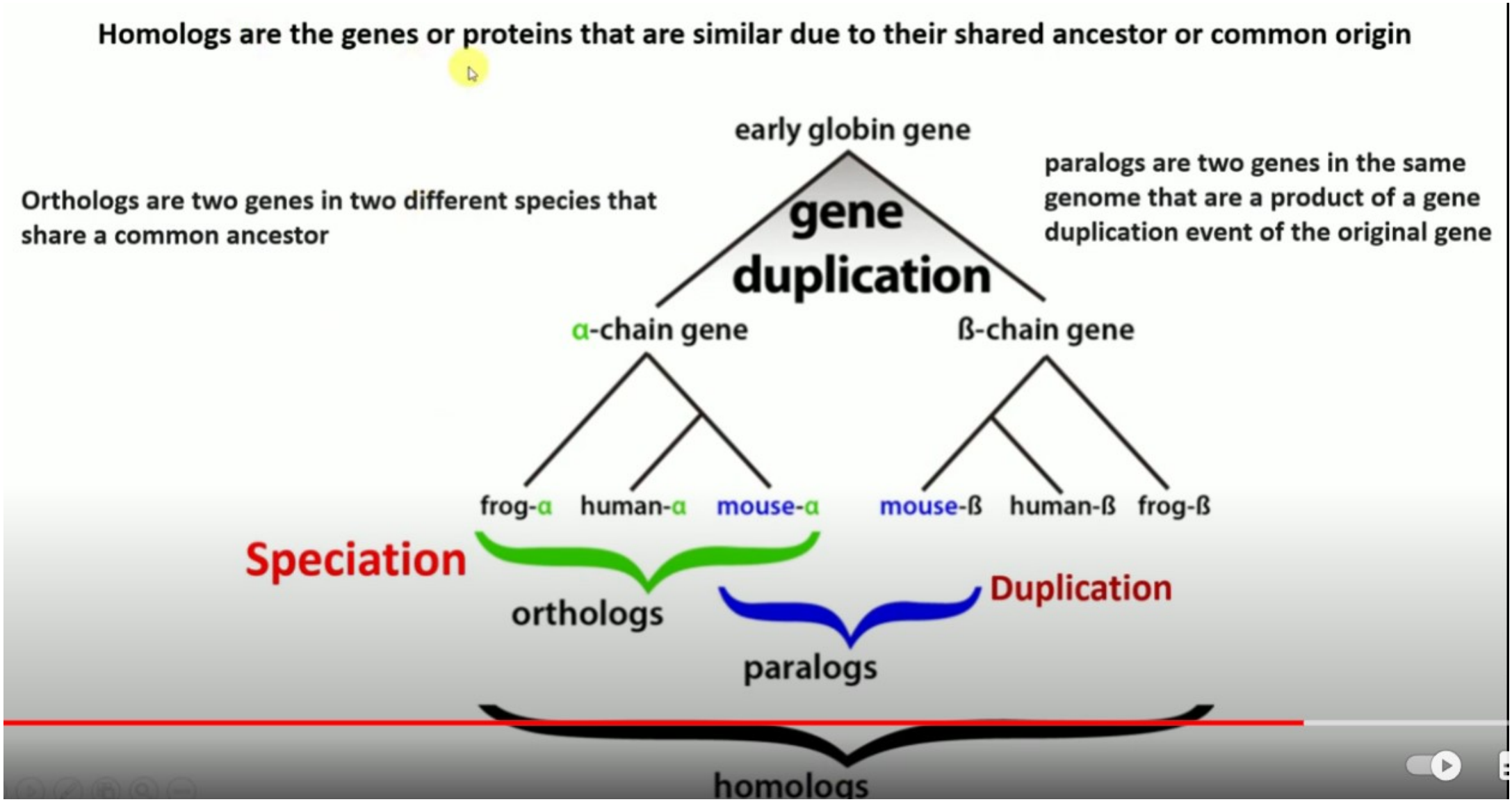


ORTHOLOGY: Descent from a common ancestor via speciation



PARALOGY: Descent from a common ancestor via gene duplication

Homologs are the genes or proteins that are similar due to their shared ancestor or common origin



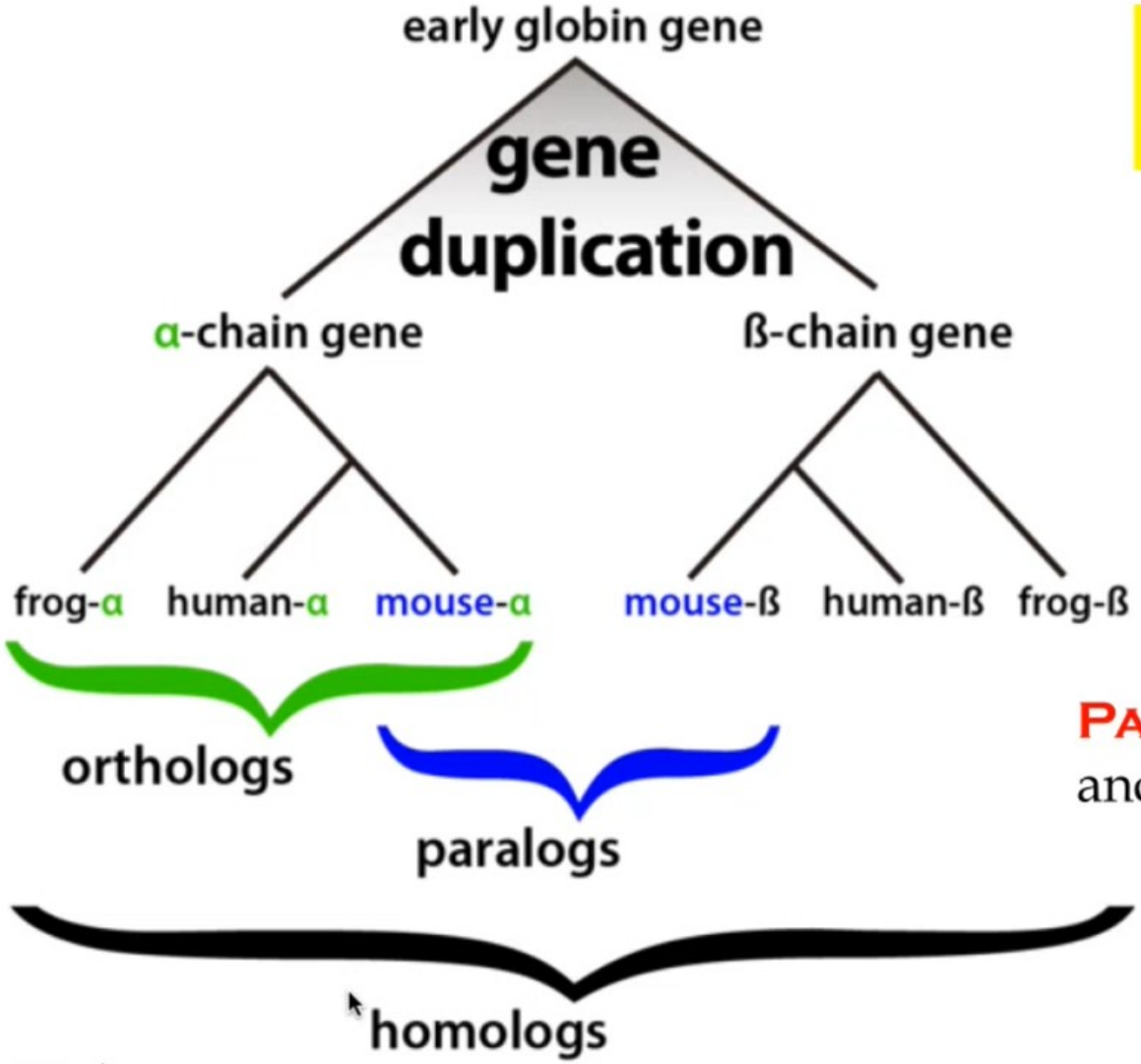
Similarity: Homology vs Analogy

Homology: Similarity in characteristics resulting from shared ancestry.

Analogy: The similarity of characteristics between two species that are not closely related; attributable to convergent evolution.

Two sisters: homologs

Two "Elvis": analogs



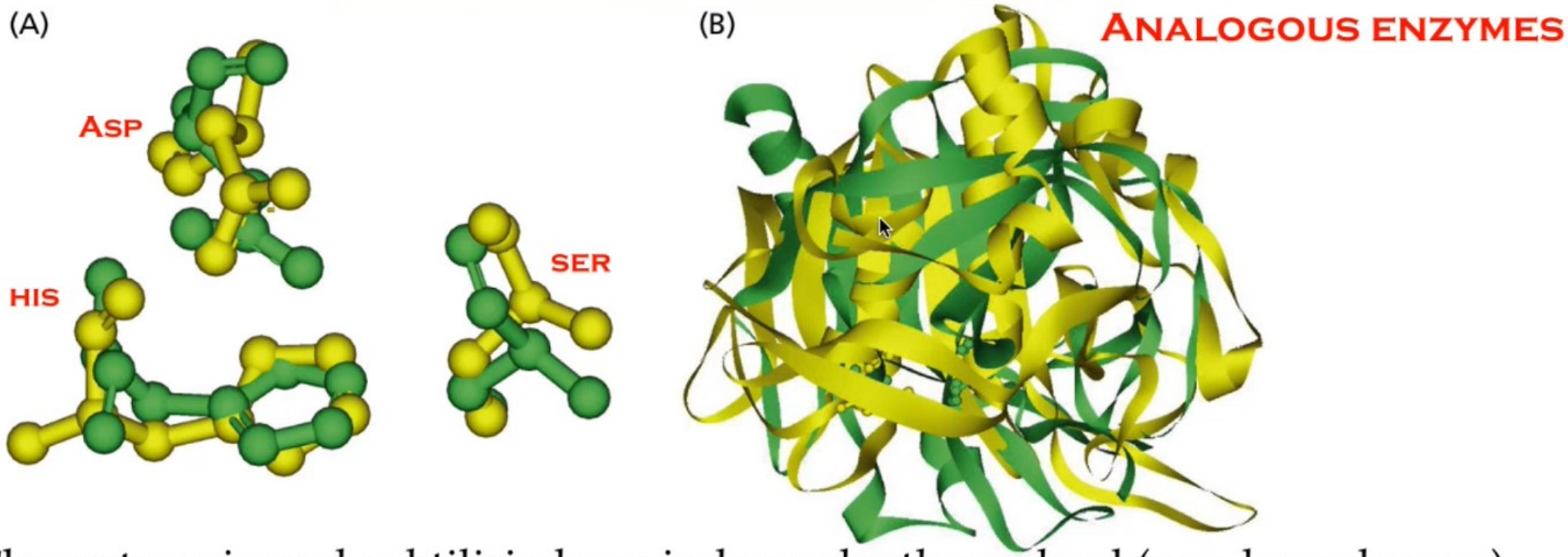
Homology (Divergent Evolution)

ORTHOLOG: Descent from a common ancestor via speciation

PARALOG: Descent from a common ancestor via gene duplication

Image Credit: Popo H. Liao

Homoplasy
(Convergent Evolution)



Chymotrypsin and subtilisin have independently evolved (non-homologous). They have different structure that doesn't overlay with each other. But, they have catalytic residues that independently evolved to retain the same catalytic triad.

Goal of Phylogenetic Analysis

To reconstruct the evolutionary history

A phylogenetic tree is a diagram that proposes an hypothesis for the reconstructed evolutionary relationships between a set of object.

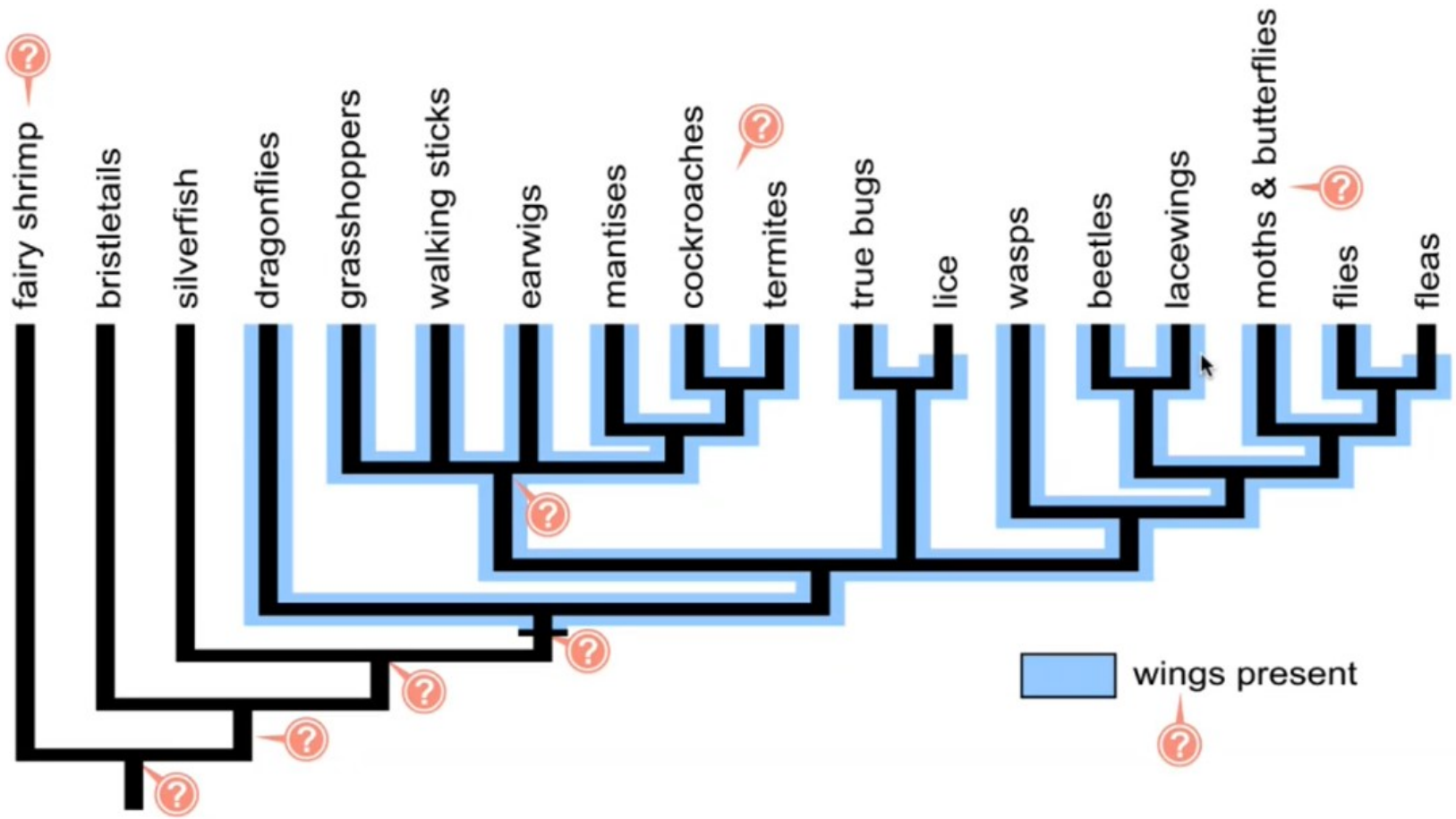


Image Credit: evolution.berkeley.edu

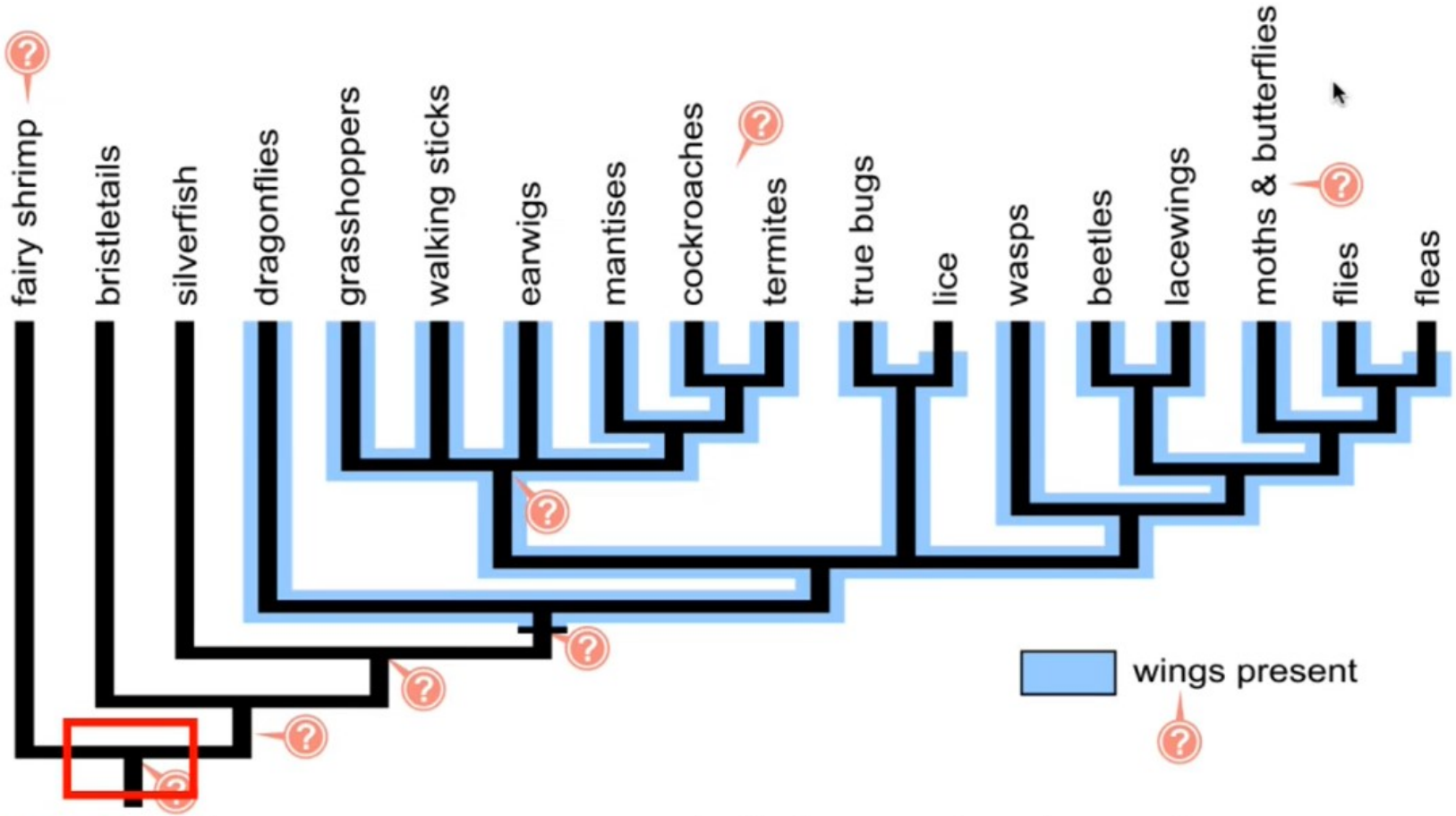
TREES

DEPICT

EVOLUTIONARY RELATIONSHIPS

NOT

EVOLUTIONARY PROGRESS



ROOT - the common ancestor of all the species shown here

Image Credit: evolution.berkeley.edu

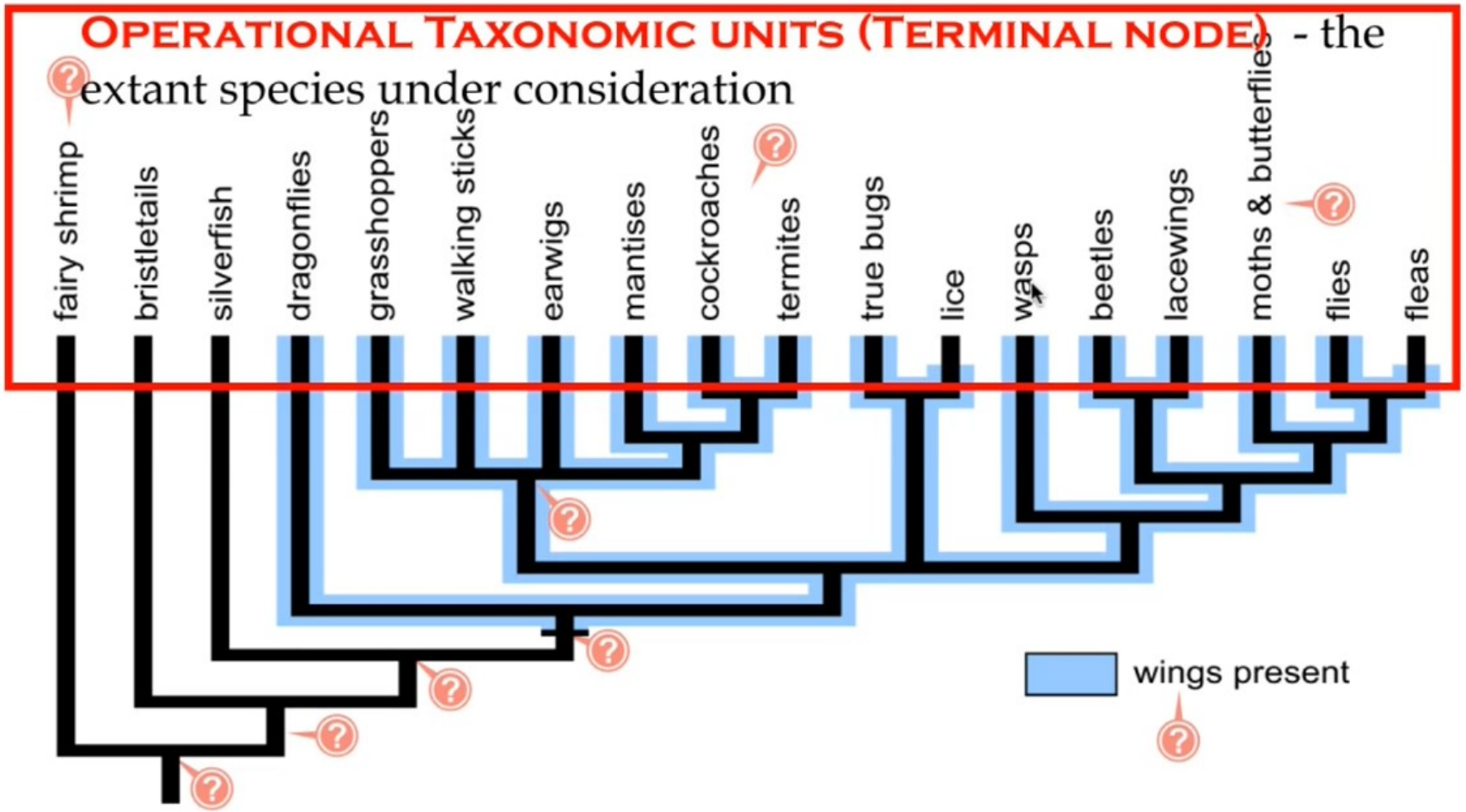


Image Credit: evolution.berkeley.edu

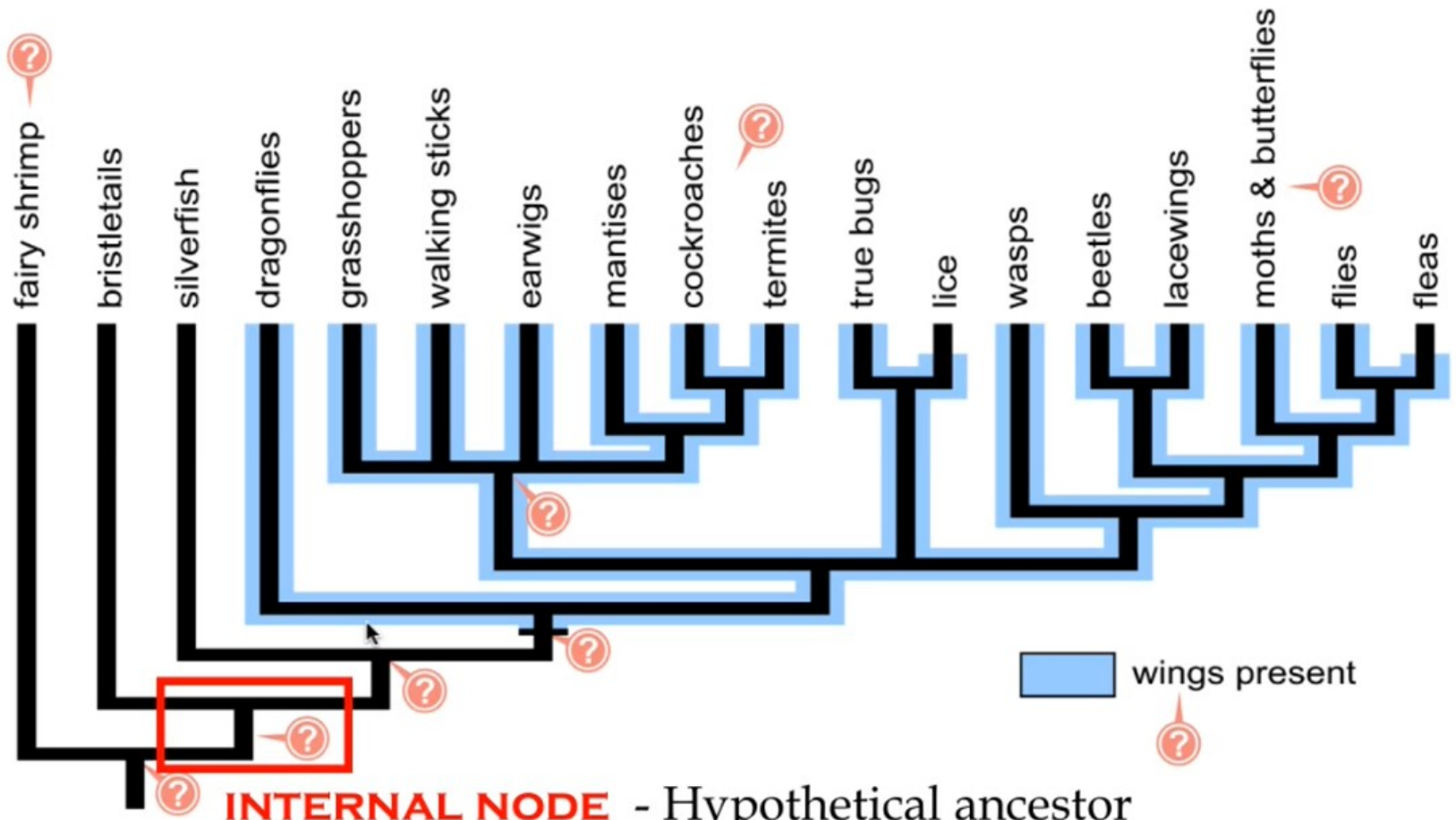
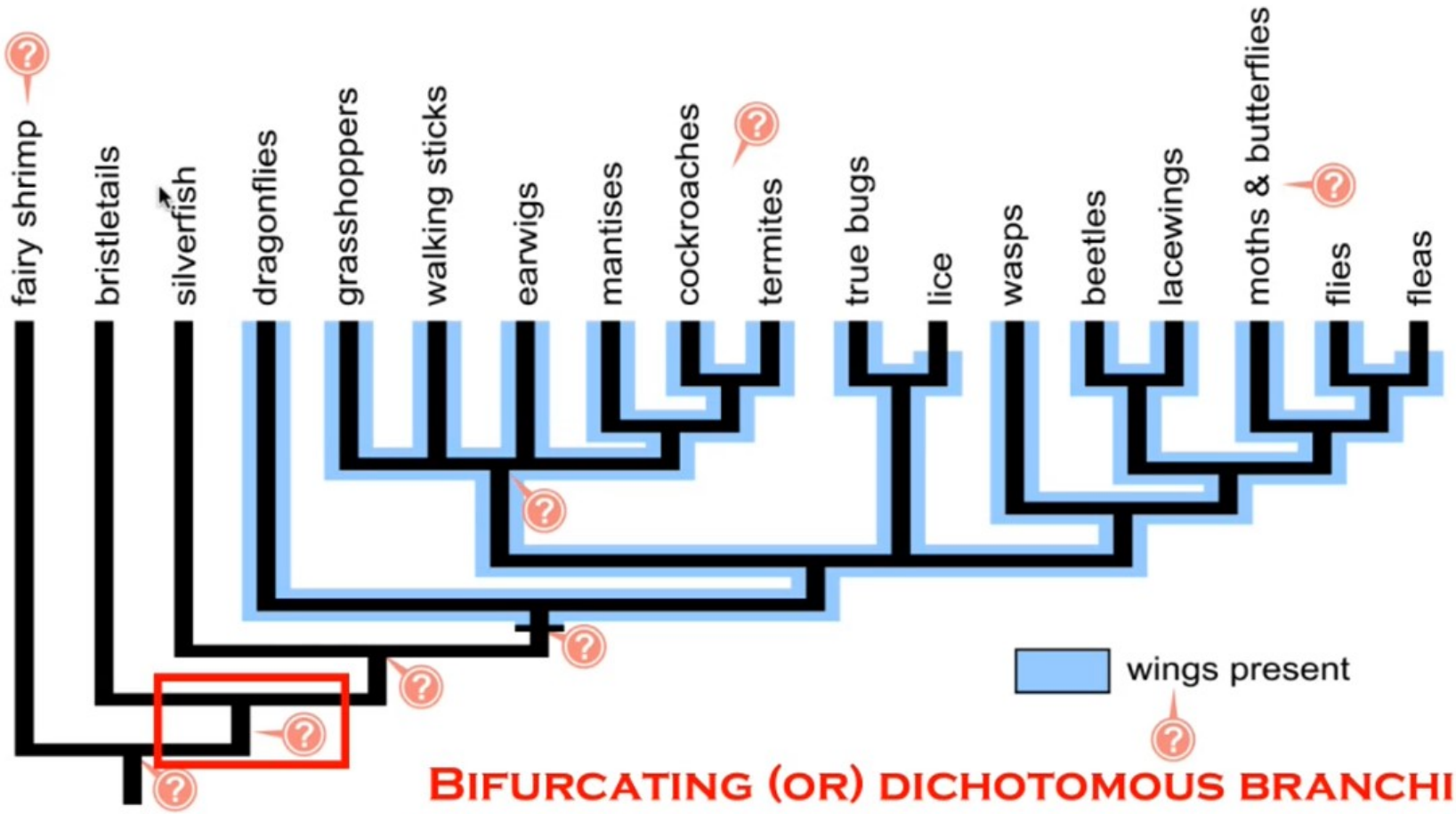
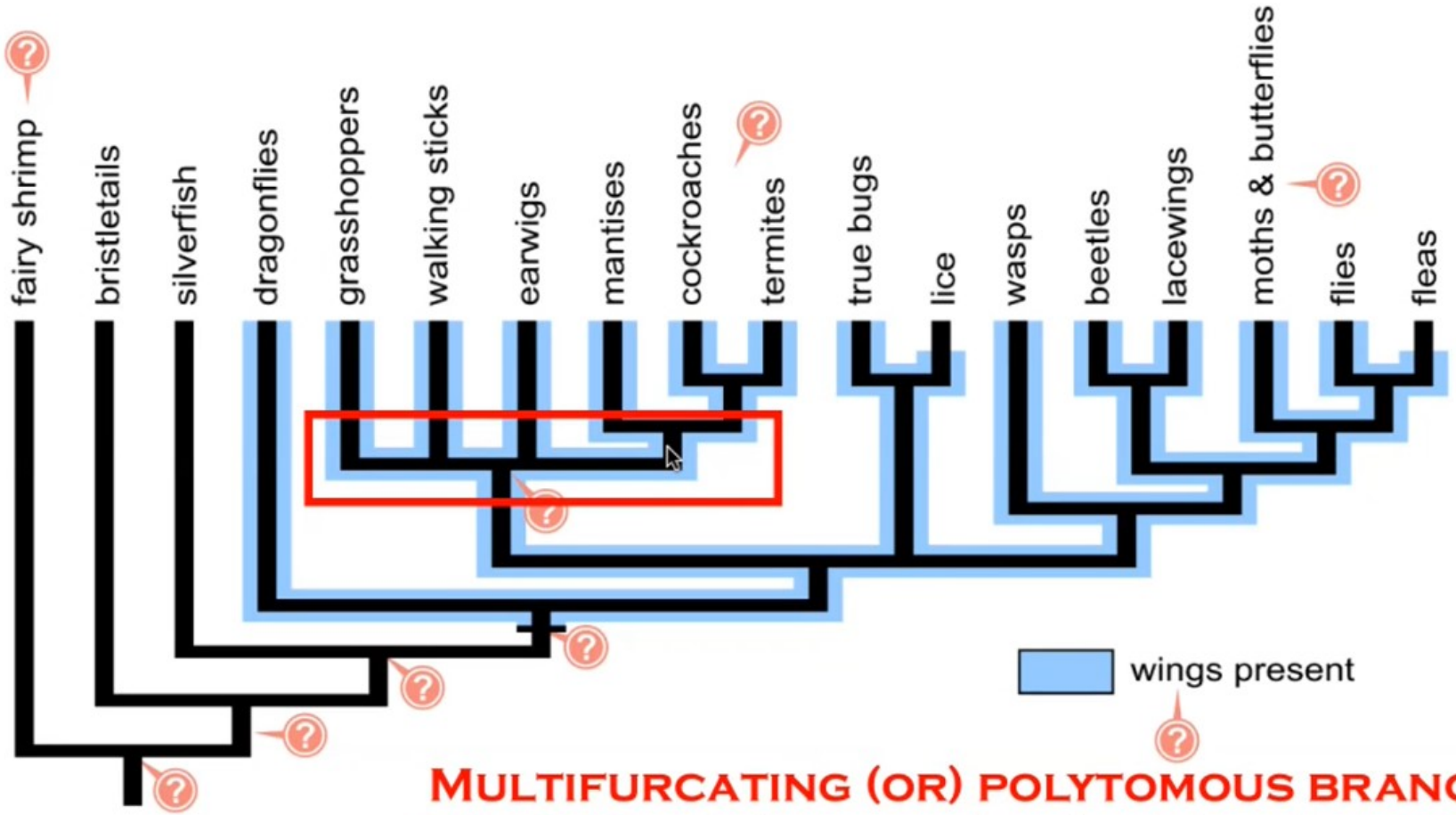


Image Credit: evolution.berkeley.edu



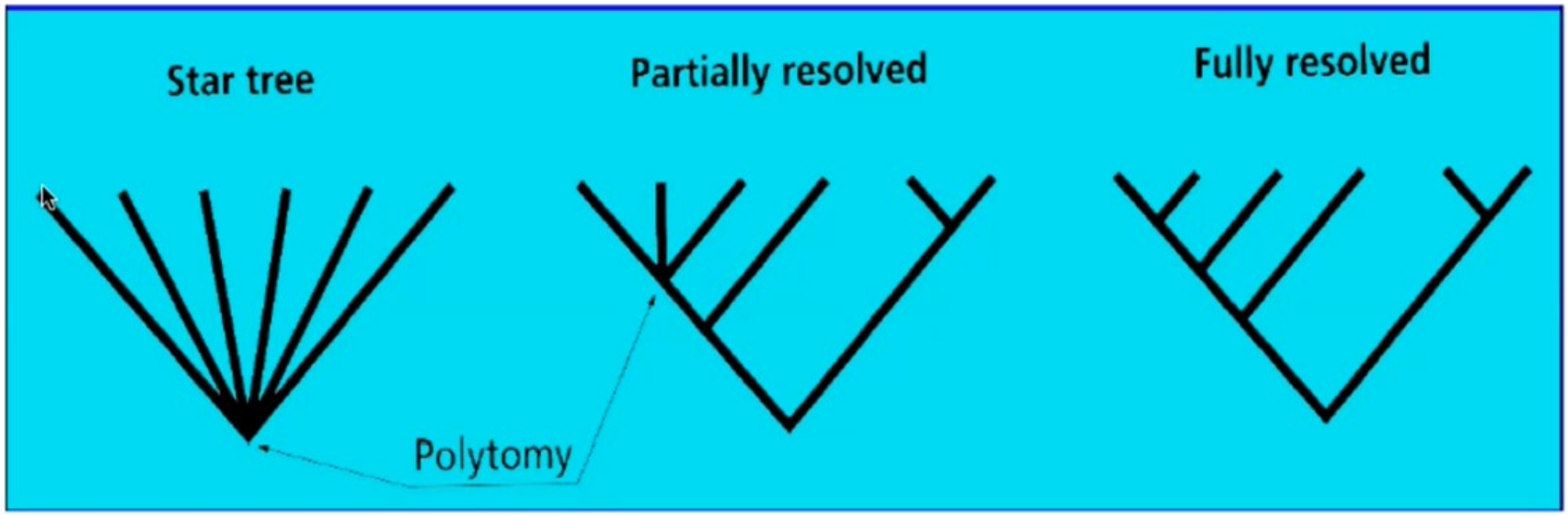
BIFURCATING (OR) DICHOTOMOUS BRANCHING PATTERN – Each ancestral node giving rise to **TWO** descendant nodes only

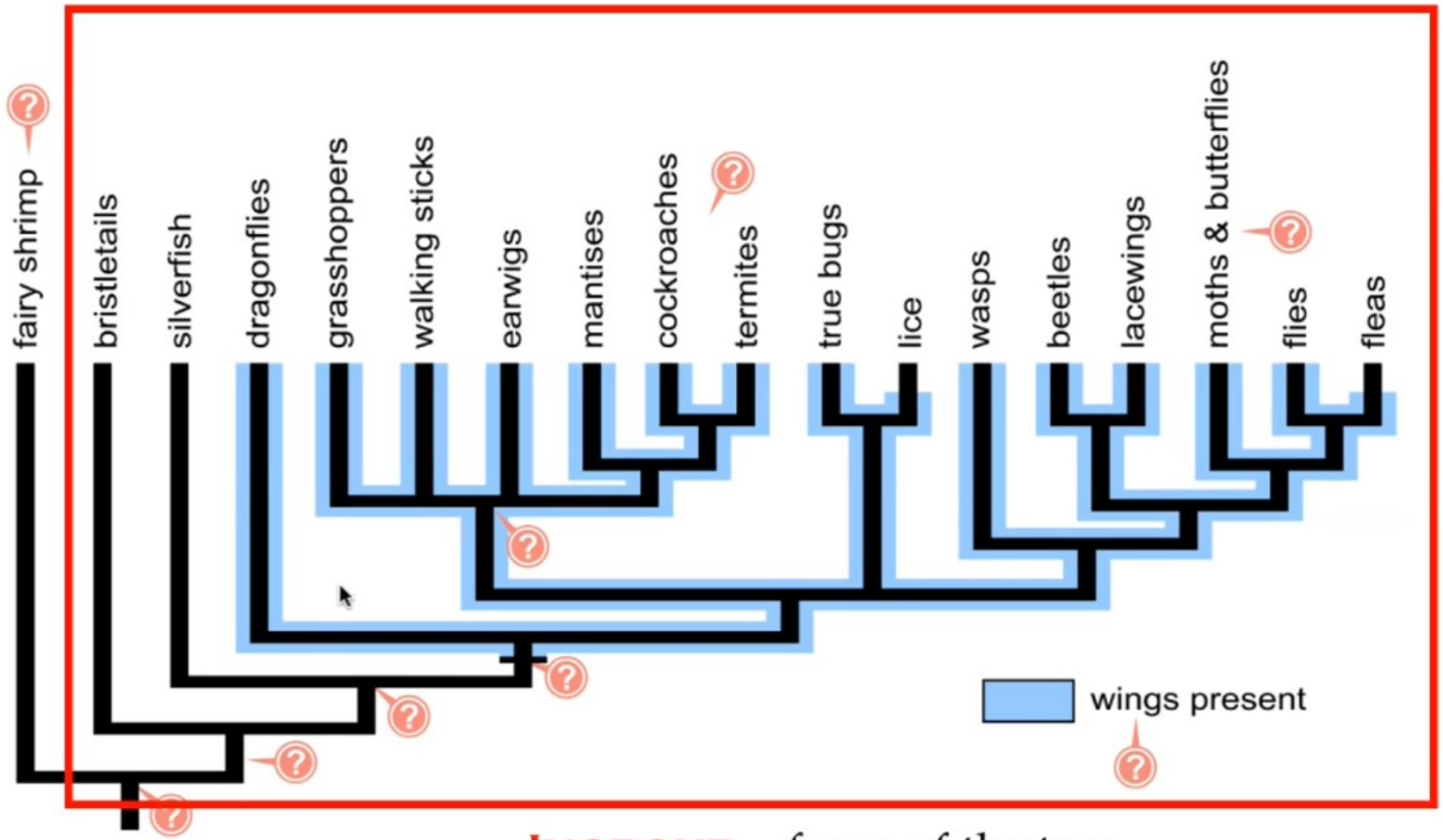
Image Credit: evolution.berkeley.edu



MULTIFURCATING (OR) POLYTOMOUS BRANCHING PATTERN – Each ancestral node giving rise to **MORE THAN TWO** descendant nodes

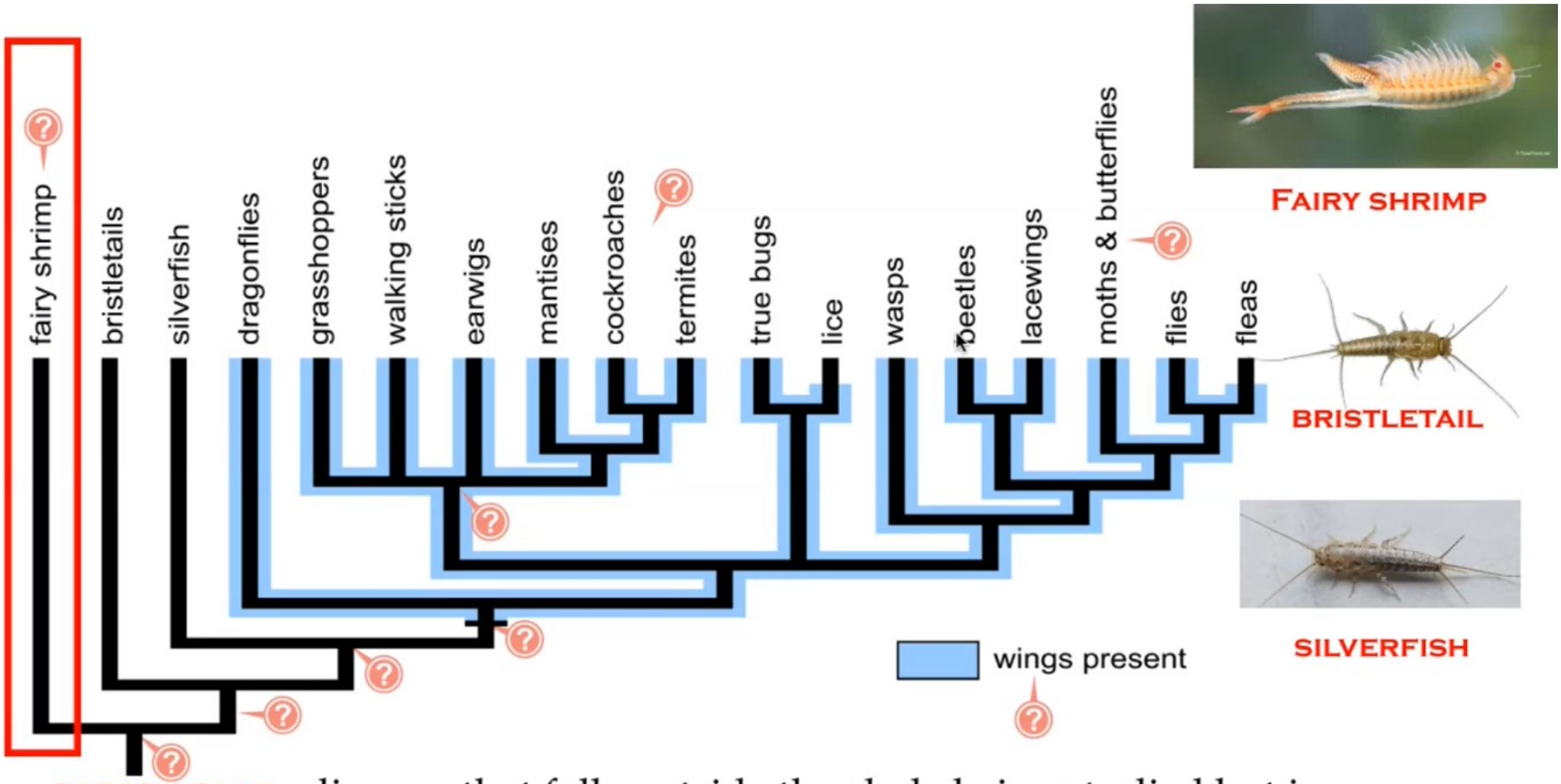
Image Credit: evolution.berkeley.edu


Star tree
Partially resolved
Fully resolved
Polytomy



INGROUP - focus of the tree

Image Credit: evolution.berkeley.edu

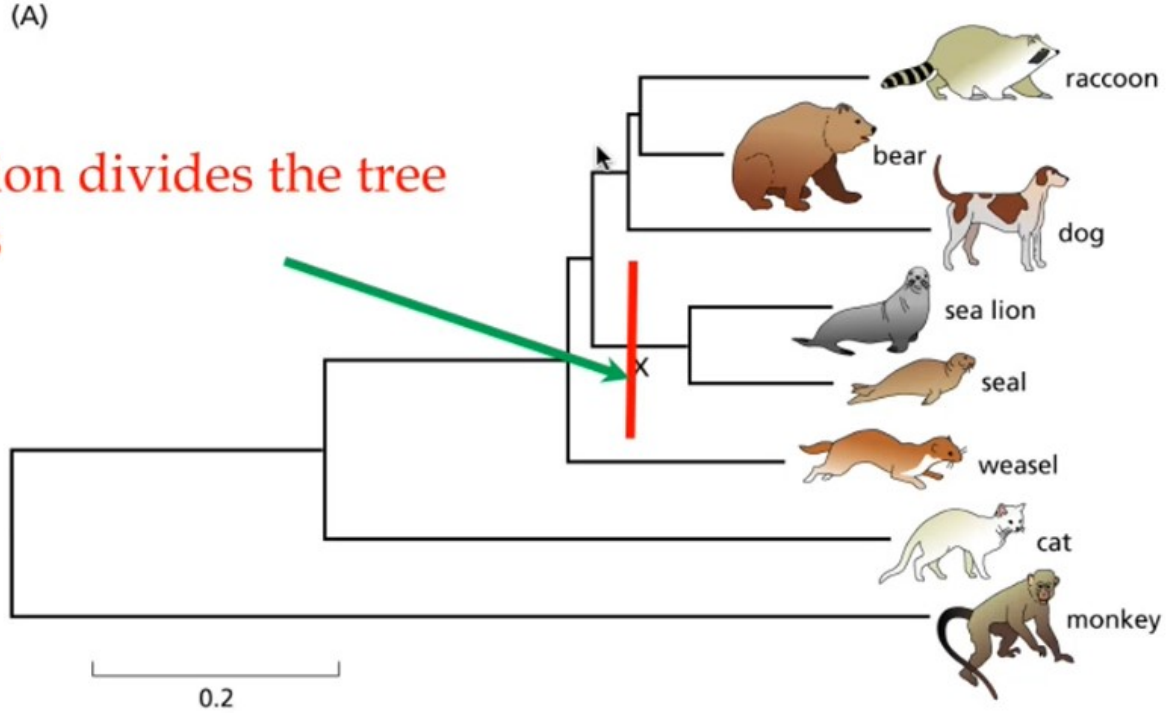


OUTGROUP - lineage that falls outside the clade being studied but is closely related to that clade

Image Credit: evolution.berkeley.edu

Split or Partition divides the tree into two parts

(A)



(B)

raccoon	bear	dog	sea lion	seal	weasel	cat	monkey
*	*						
*	*	*					
			*	*			
*	*	*	*	*			
*	*	*	*	*	*		

Every split produces two groups. Hence one group is suffice to be defined.

Rotations about the branches change the order of the taxa but NOT the evolutionary relationships

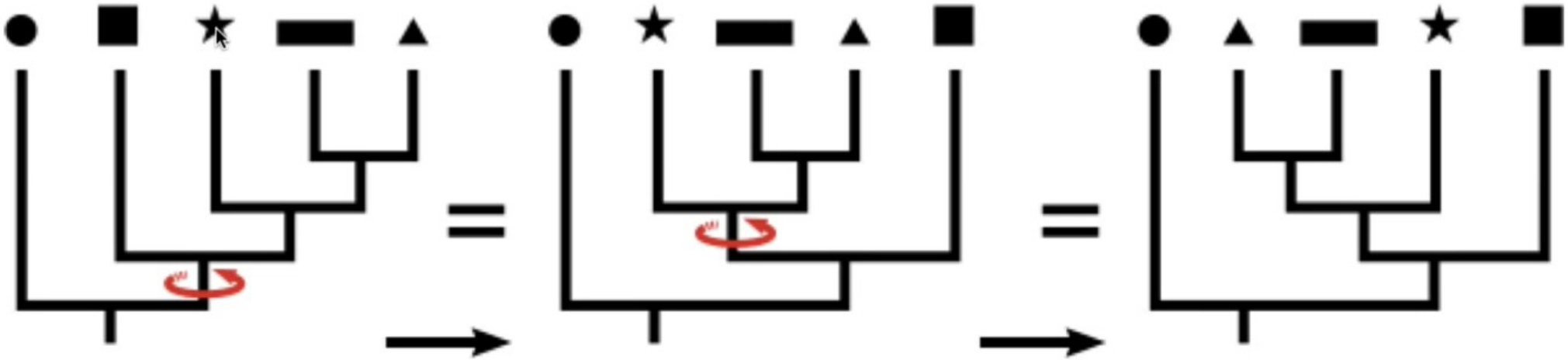


Image Credit: evolution.berkeley.edu

Rotations about the branches change the order of the taxa but NOT the evolutionary relationships

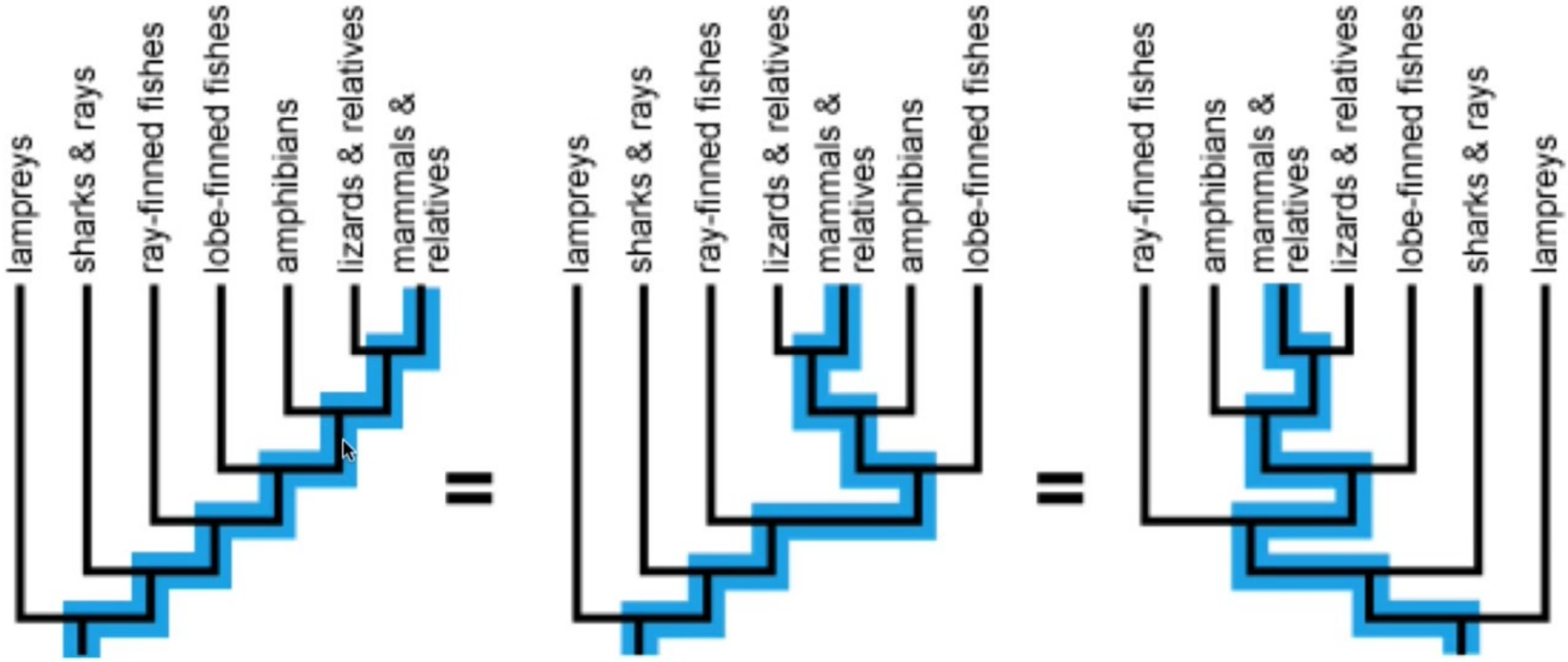
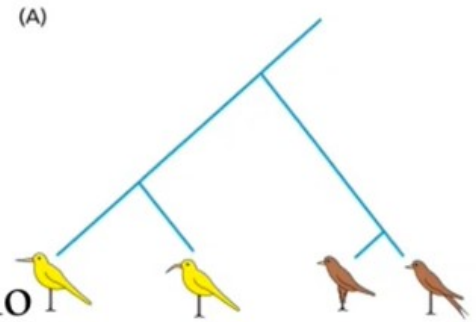
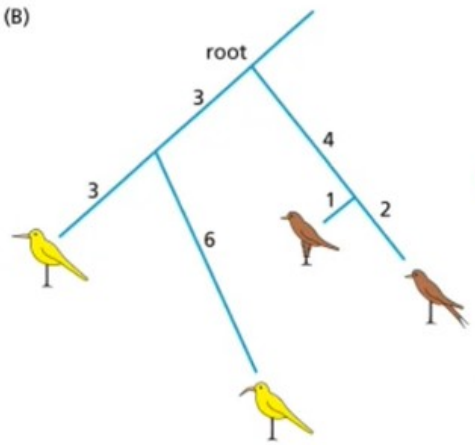


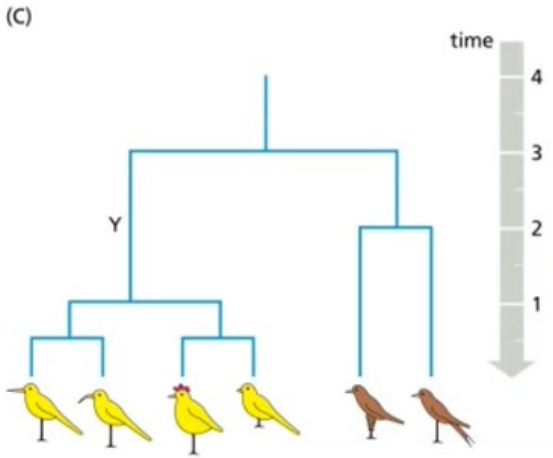
Image Credit: evolution.berkeley.edu

A ROOTED CLADOGRAM - Genealogy but no timing of divergence

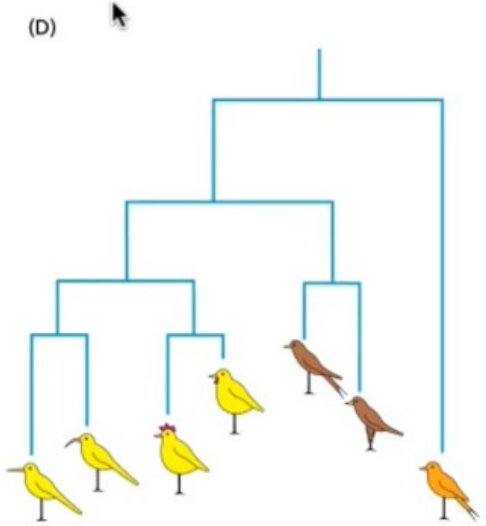




ADDITIVE TREE – branch length Proportional to number of mutations



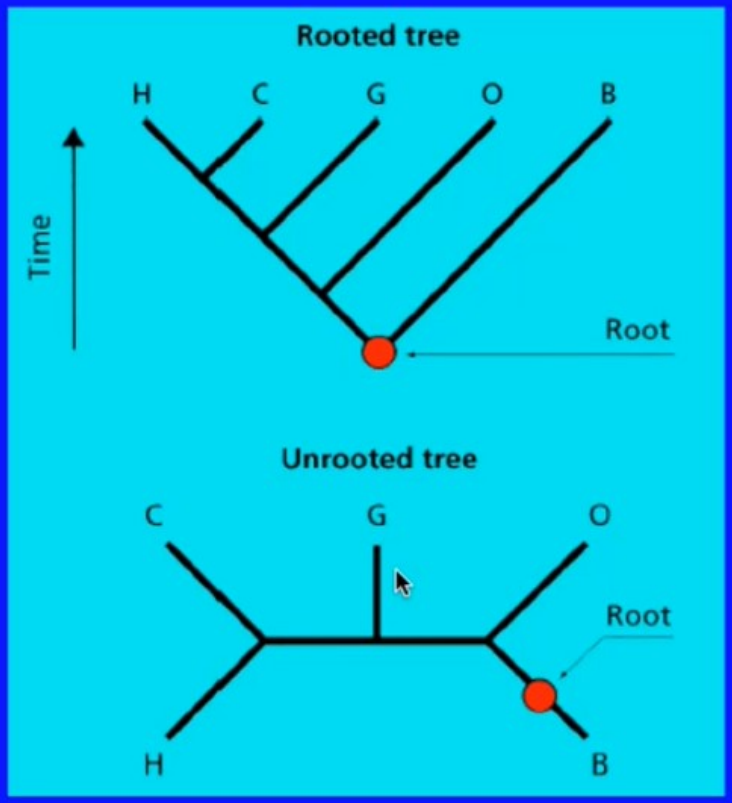
ULTRAMETRIC TREE – constant rate of Mutation assumed along all branches (molecular clock)



ADDITIVE TREE WITH OUTGROUPS

Rooted and Unrooted Trees

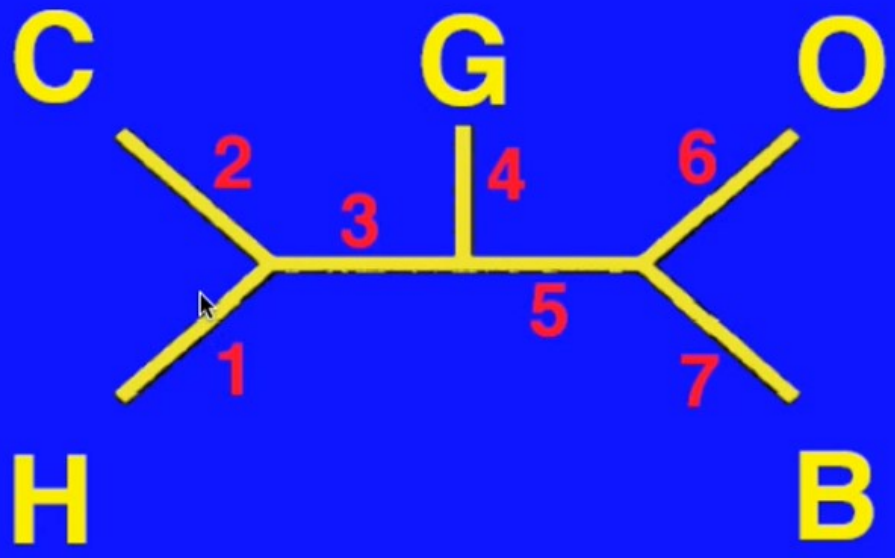
Rooted trees. Have direction corresponding to evolutionary time.



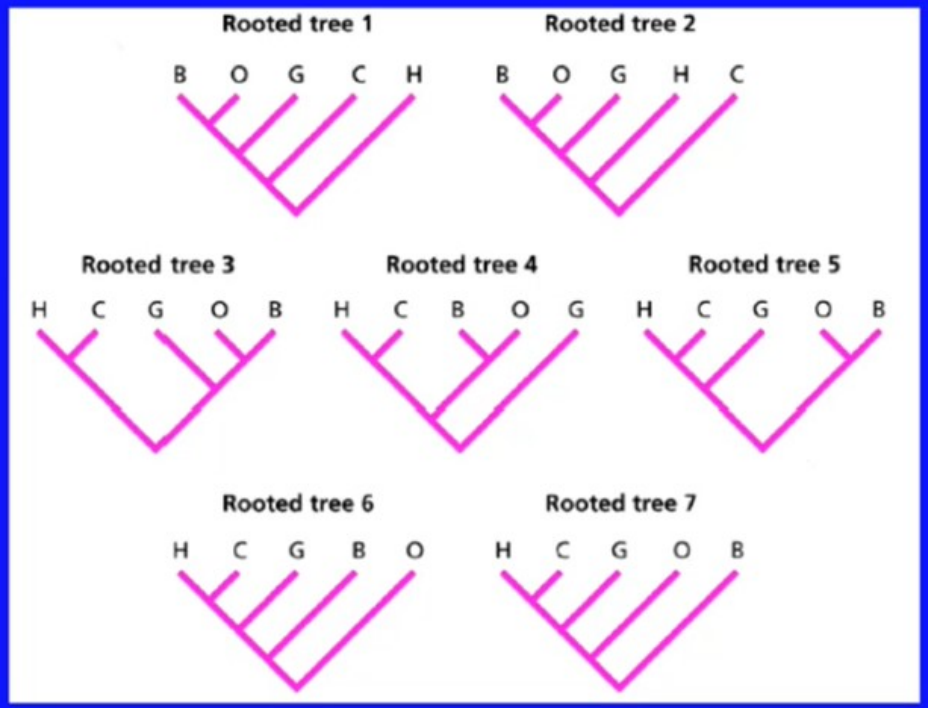
Unrooted trees. we cannot talk of ancestors and descendants

H = human
C = chimp
G = gorilla
O = orang
B = gibbon

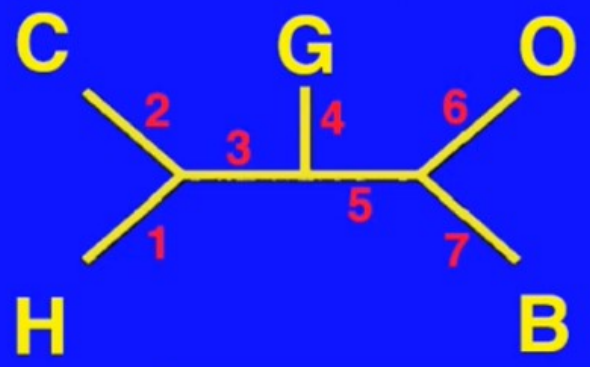
For this unrooted tree....



There are 7 different corresponding rooted trees



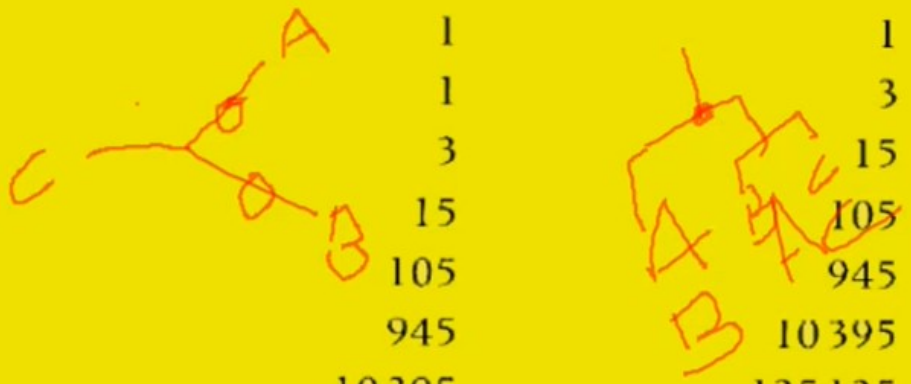
UNROOTED TREE



N.B. All correspond to same unrooted network shown previously but have the "root" placed on one of the 7 different branches of the original unrooted network

Number of possible trees increases exponentially with increasing number of sequences

Number of sequences	Number of unrooted trees	Number of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425



Rooted Trees = $\frac{(2n-3)!}{2^{n-2}\,(n-2)!}$

n> 2

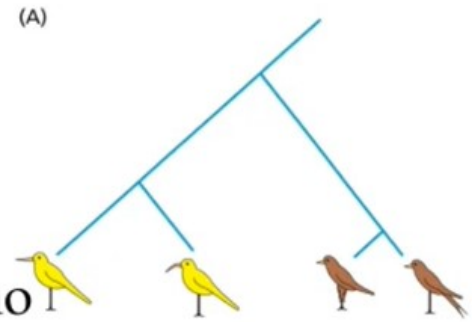
Unrooted Trees = $\frac{(2n-5)!}{2^{n-3}\,(n-3)!}$

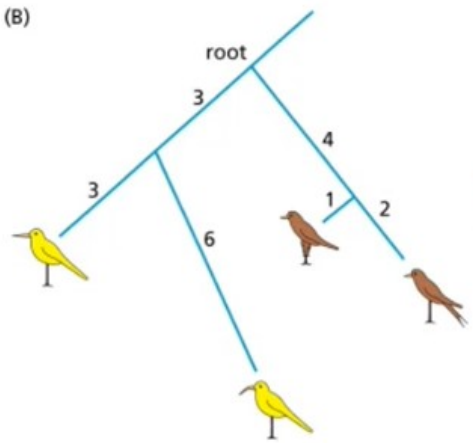
n> 3

For ten sequences there are more than 34 million rooted trees.

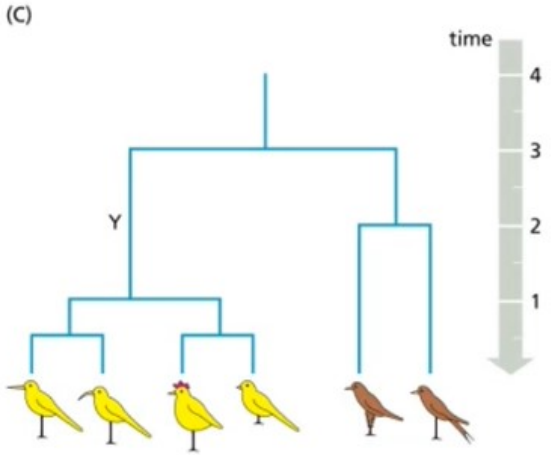
- For 20 sequences 8,200,794,532,637,891,559,000 trees.
- For a recent study of 135 human mtDNA sequences 2.113×10^{267} trees.
- Larger than number of particles in the known universe. (actually, larger than the volume of the universe measured in Plank Constant Units)

A ROOTED CLADOGRAM - Genealogy but no timing of divergence

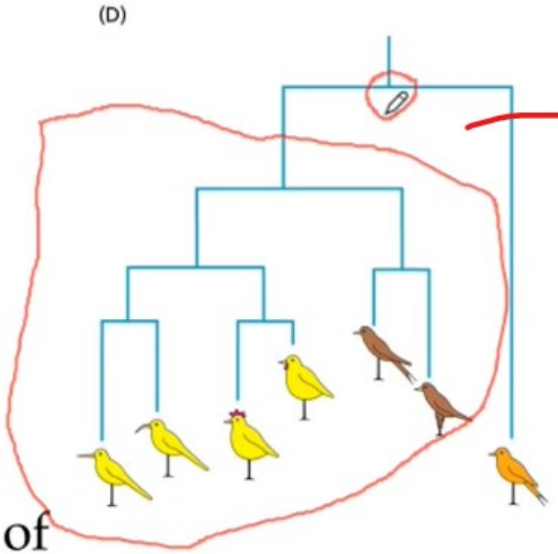




ADDITIVE TREE – branch length Proportional to number of mutations

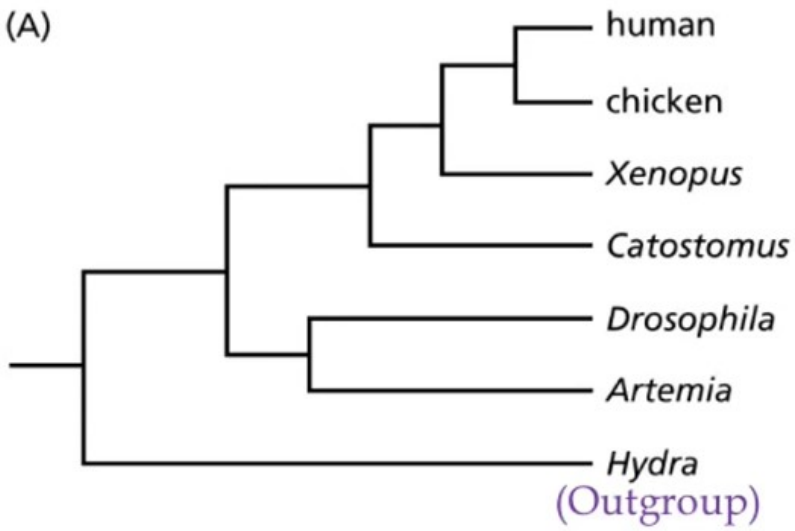


ULTRAMETRIC TREE – constant rate of Mutation assumed along all branches (molecular clock)

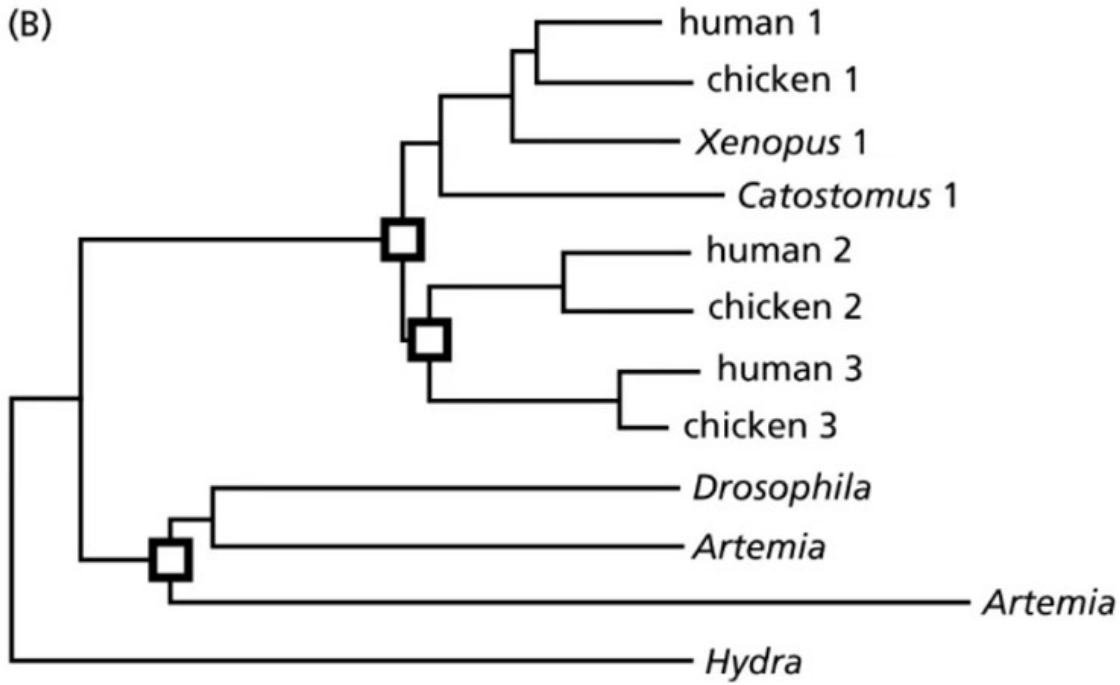


Rooting

ADDITIVE TREE WITH OUTGROUPS



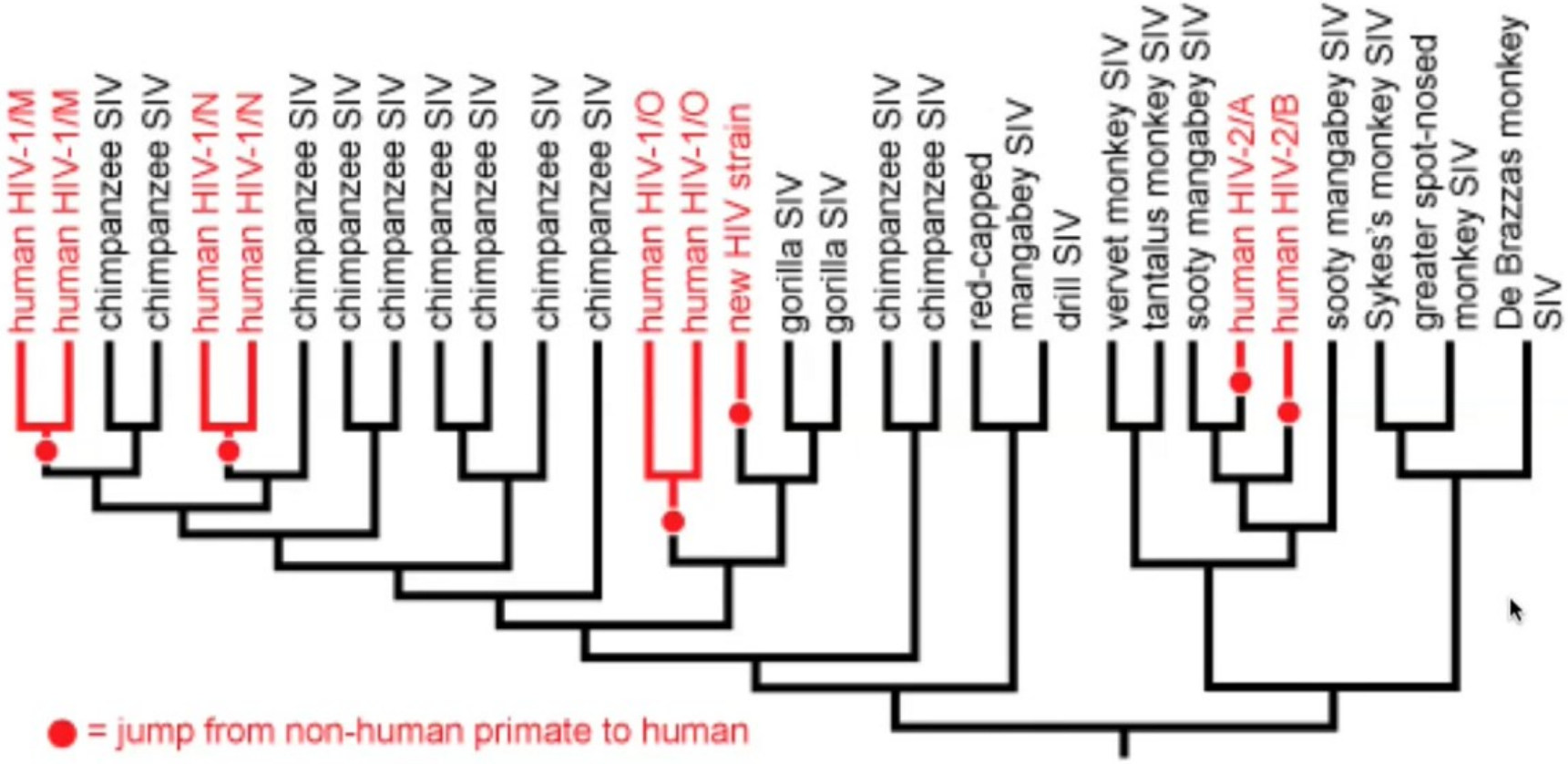
SPECIES TREE



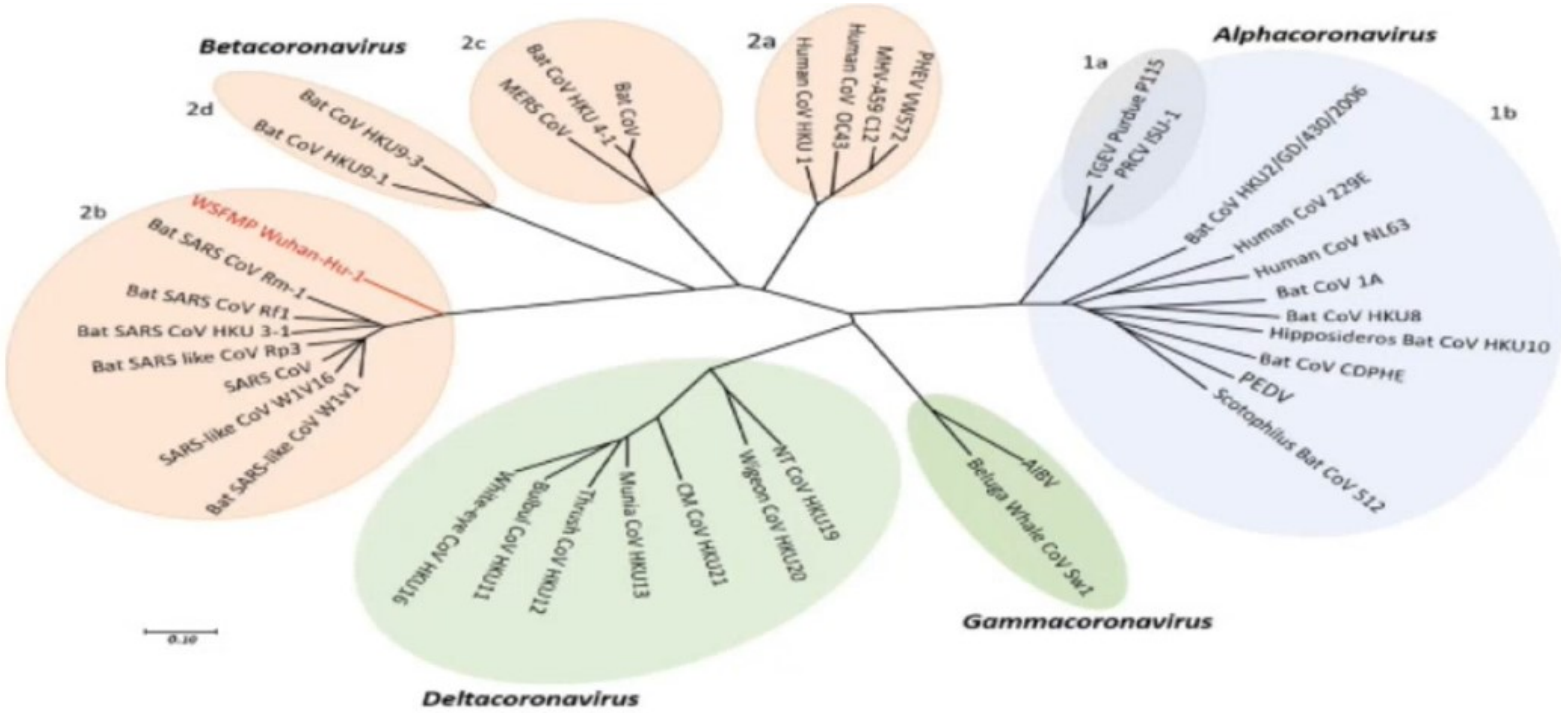
GENE TREE

Na^+-K^+ Pump Membrane Protein Family

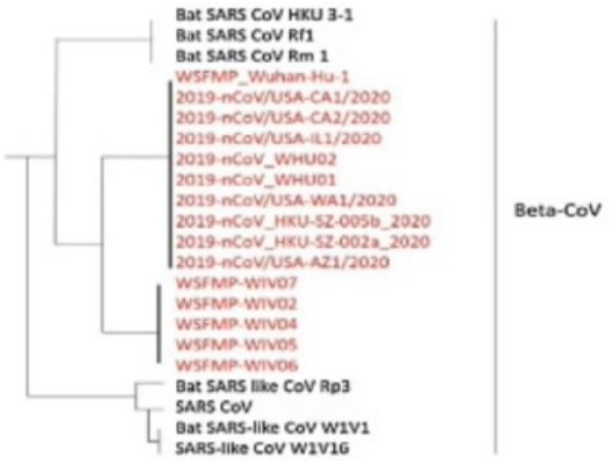
Phylogenetics reveals that HIV has evolved from Simian Immunodeficiency virus (SIV) several times



Keele *et al, Nature* 313:523-526 (2006)



Phylogenetic analysis predicted that the SARS-CoV-2 got transmitted to humans from bats

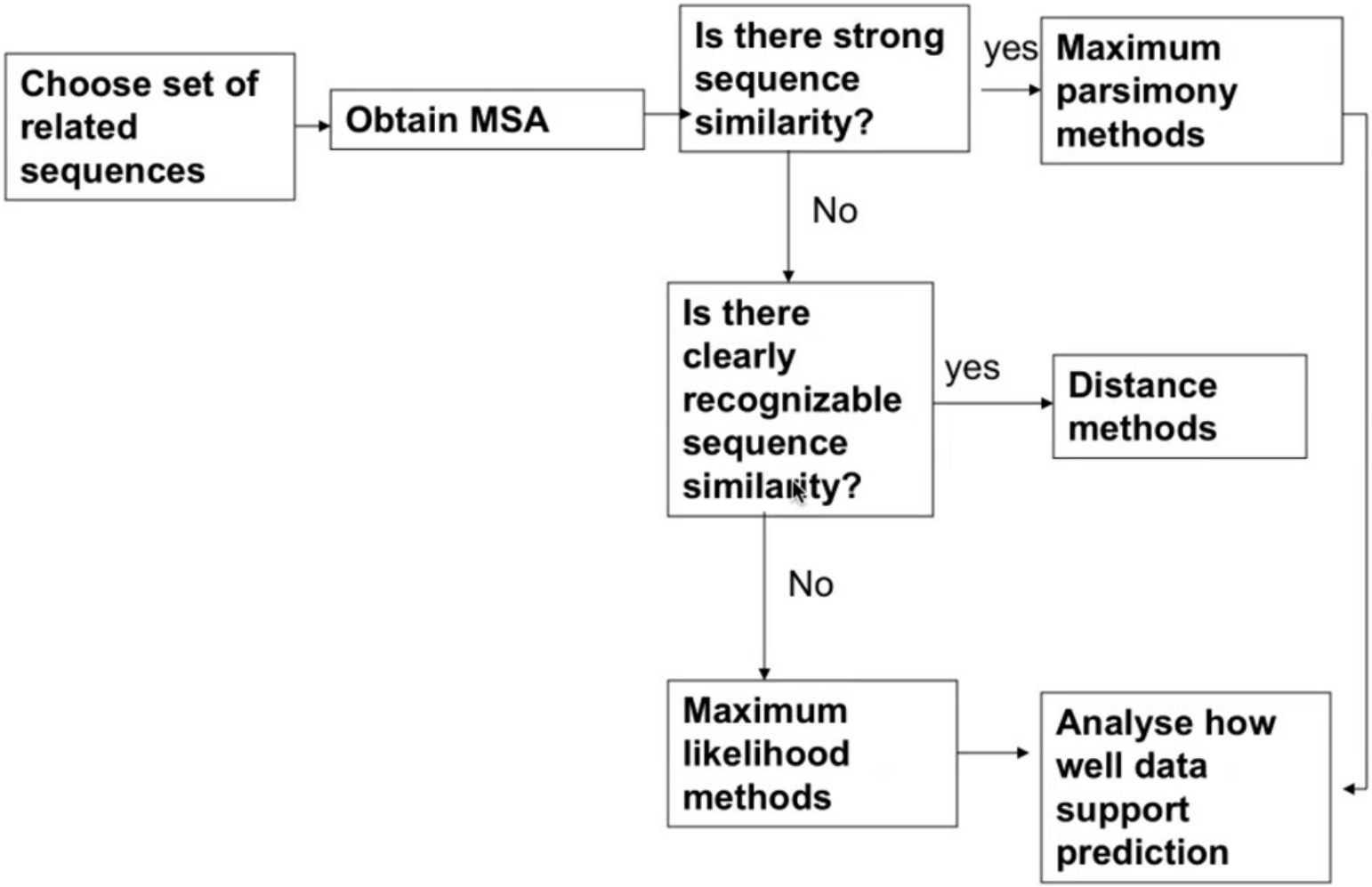


Shereen et al. Journal of Advanced Research 24 (2020) 91–98

Phylogenetic Analysis

Evolutionary Trees

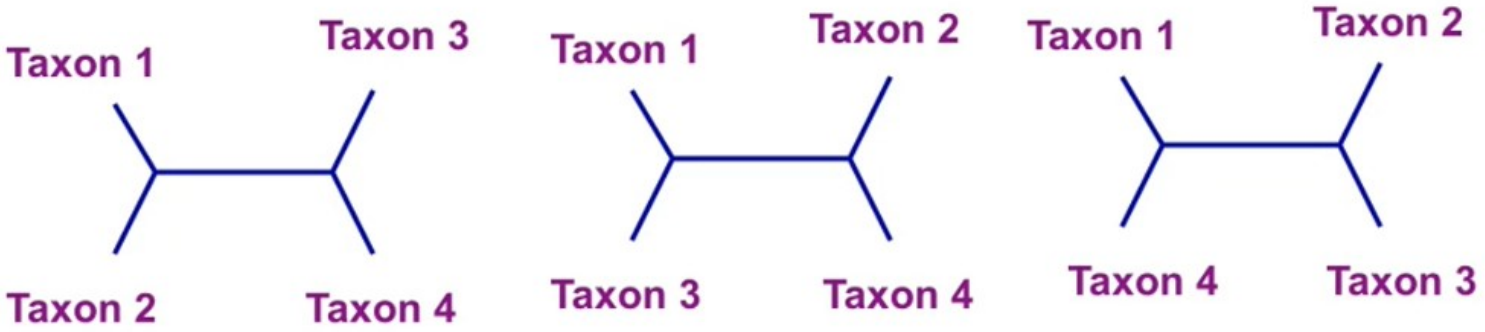
- Maximum Parsimony
- Distance based
- Maximum likelihood


Choose set of related sequences
Obtain MSA
Is there strong sequence similarity?
yes
Maximum parsimony methods
No
Is there clearly recognizable sequence similarity?
yes
Distance methods
No
Maximum likelihood methods
Analyse how well data support prediction

Maximum Parsimony Method

- Predicts the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences
- MSA is required to predict which sequence positions are likely to correspond
- Trees that produce the smallest number of changes overall for all sequence positions are identified
- All possible trees relating a group of sequences are examined

Sequence	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G



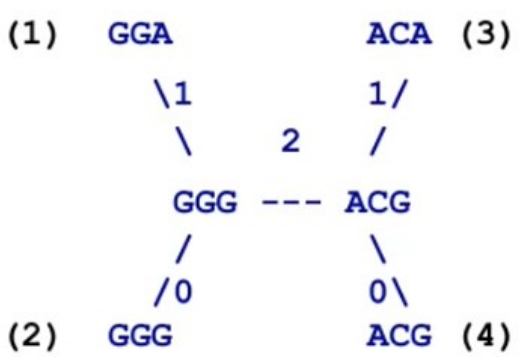
All possible unrooted trees are considered

Sequence	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Informative and non-informative sites

A site is informative only when there are at least two different kinds of nucleotides at the site, each of which is represented in at least two of the sequences under study.

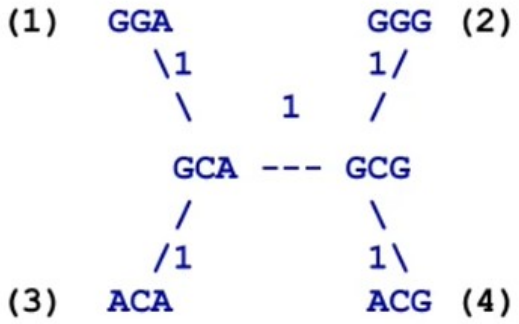
1 GGA
2 GGG
3 ACA
4 ACG
***



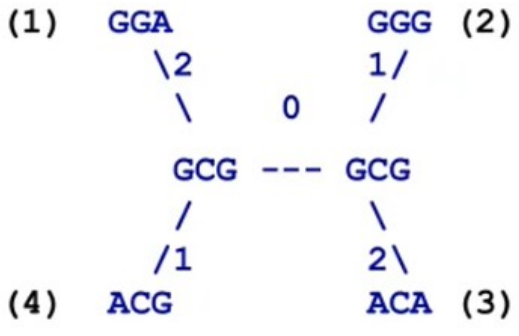
Number of mutations

Tree I: 4

Tree I is chosen because it requires the smallest number of changes (4) at the informative sites.



Tree II: 5



Tree III: 6

Maximum Parsimony Method

Disadvantages

Does not use all the information (only informative sites are used)

Does not provide information on branch length

Advantages

Evaluates different trees

Provides information about ancestral sequences

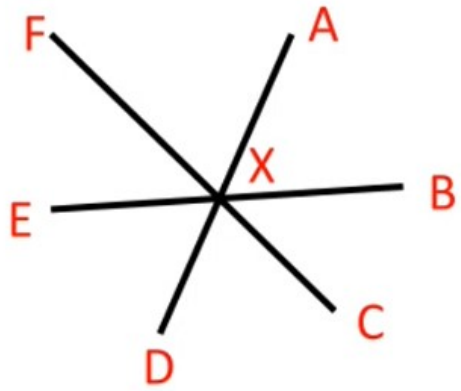
Distance Based Method

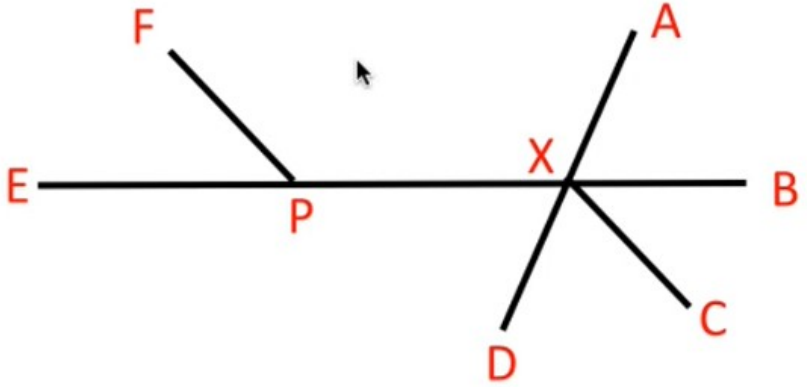
Sequence	1	2	3	4	5	6	7	8	9
P	A	A	G	A	G	T	G	C	A
Q	A	G	C	C	G	T	G	C	G
R	A	G	A	T	A	T	C	C	A
S	A	G	A	G	A	T	C	C	G

	P	Q	R	S
P		4	5	6
Q			5	4
R				2
S				

Distance Based Method

Neighbour Joining Method





	A	B	C	D	E
A		5	4	9	8
B			5	10	9
C				7	6
D					7
E					

$$U_i = \sum_{j=1}^{N} d_{ij} \qquad \text{.....(1)} \qquad U_A = 5 + 4 + 9 + 8 = 26$$

$$\delta_{ij} = d_{ij} - \left(\frac{U_i + U_j}{N-2}\right) \qquad \text{.....(2)} \qquad \delta_{AB} = 5 - \left(\frac{26+29}{3}\right) = \frac{-40}{3}$$

	A	B	C	D	E
A		5	4	9	8
B			5	10	9
C				7	6
D					7
E					

U
26
29
22
33
30

N =5

$3\ \delta_{ij}$

	A	B	C	D	E
A		-40	-36	-32	-32
B			-36	-32	-32
C				-34	-34
D					-42
E					

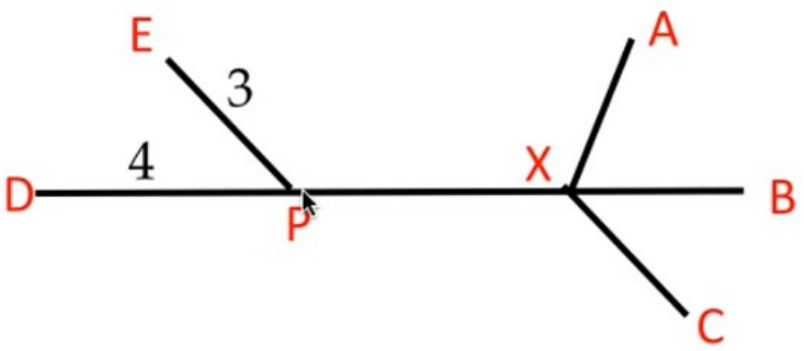
	A	B	C	D	E
A		5	4	9	8
B			5	10	9
C				7	6
D					7
E					

U
26
29
22
33
30

N =5

$3\,\delta_{ij}$

	A	B	C	D	E
A		-40	-36	-32	-32
B			-36	-32	-32
C				-34	-34
D					-42
E					



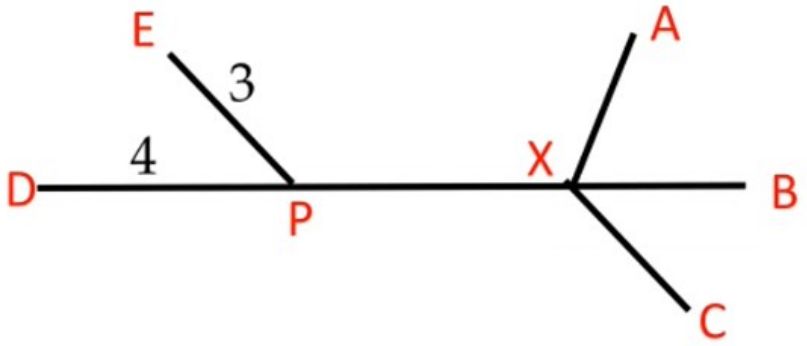
$$b_{DP} = \frac{1}{2}\left(d_{DE} + \left(\frac{U_D - UE}{N-2}\right)\right)$$

$$b_{DP} = \frac{1}{2}\left(7 + \left(\frac{33 - 30}{3}\right)\right) = 4$$

$$b_{iY} = \frac{1}{2}\left(d_{ij} + \left(\frac{U_i - Uj}{N-2}\right)\right) \text{.....(3)}$$

$b_{jY} = d_{ij} - b_{iY}$ .....(4)

$b_{EP} = 7 - 4 = 3$



$$b_{Yk} = \frac{1}{2}\left(d_{ik} + d_{jk} - dij\right) \text{.....(5)}$$

$$b_{PA} = \frac{1}{2}\left(d_{AD} + d_{AE} - dDE\right) = \frac{(9+8-7)}{2} = 5$$

	A	B	C	P
A		5	4	5
B			5	6
C				3
P				

U
14
16
12
14

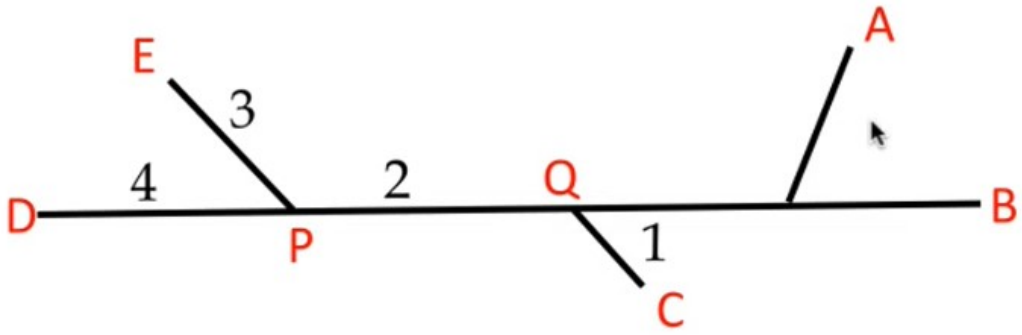
$N = 4$

$2\,\delta_{ij}$

	A	B	C	P
A		-20	-18	-18
B			-18	-18
C				-20
P				

$$U_i = \sum_{j=1}^{N} dij$$

$$\delta ij = dij - \left(\frac{Ui + Uj}{N - 2}\right)$$



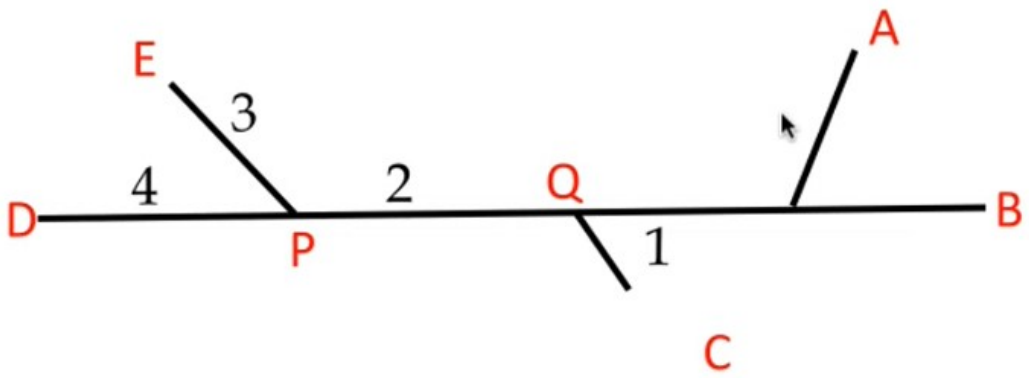
$$b_{PQ} = \frac{1}{2}\left(d_{PC} + \left(\frac{U_P - UC}{N-2}\right)\right)$$

$$b_{PQ} = \frac{1}{2}\left(3 + \left(\frac{14-12}{2}\right)\right) = 2$$

$$b_{iY} = \frac{1}{2}\left(d_{ij} + \left(\frac{U_i - Uj}{N-2}\right)\right)$$

$$b_{jY} = d_{ij} - b_{iY}$$

$$b_{QC} = b_{PQ} - b_{PC} = 3 - 2 = 1$$

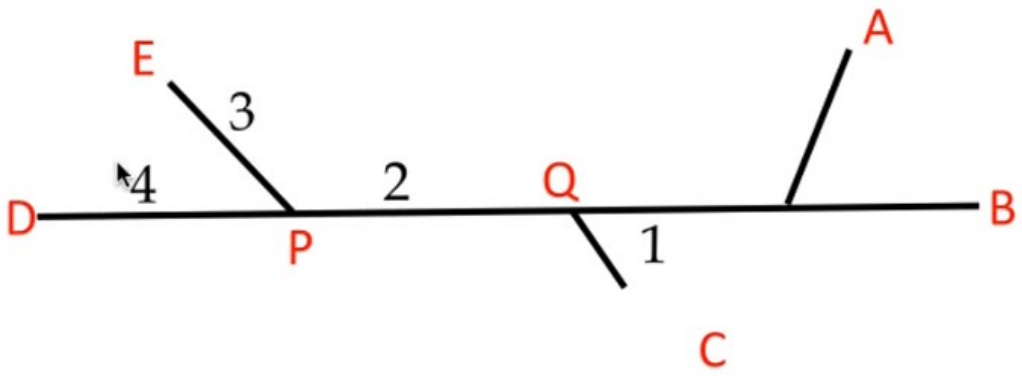


$$b_{Yk} = \frac{1}{2}\left(d_{ik} + d_{jk} - dij\right) \text{.....(5)}$$

	A	B	Q
A		5	3
B			4
Q			

U
8
9
7

$$U_i = \sum_{j=1}^{N} d_{ij}$$



$$b_{Yk} = \frac{1}{2}\left(d_{ik} + d_{jk} - dij\right) \text{.....(5)}$$

$$b_{QA} = \frac{1}{2}\left(d_{PA} + d_{CA} - dPC\right) = \frac{(5 + 4 - 3)}{2} = 3$$

	A	B	Q
A		5	3
B			4
Q			

U
8
9
7

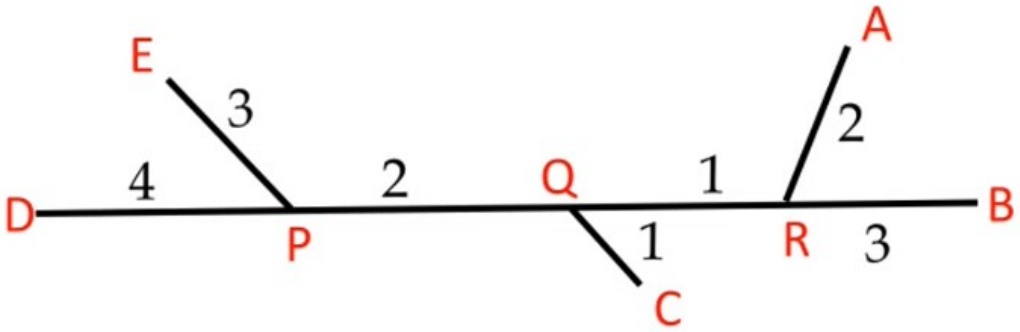
N = 3

δ_{ij}

	A	B	Q
A		-12	-12
B			-12
Q			

$$U_i = \sum_{j=1}^{N} d_{ij}$$

$$\delta_{ij} = dij - \left(\frac{U_i + Uj}{N-2}\right)$$



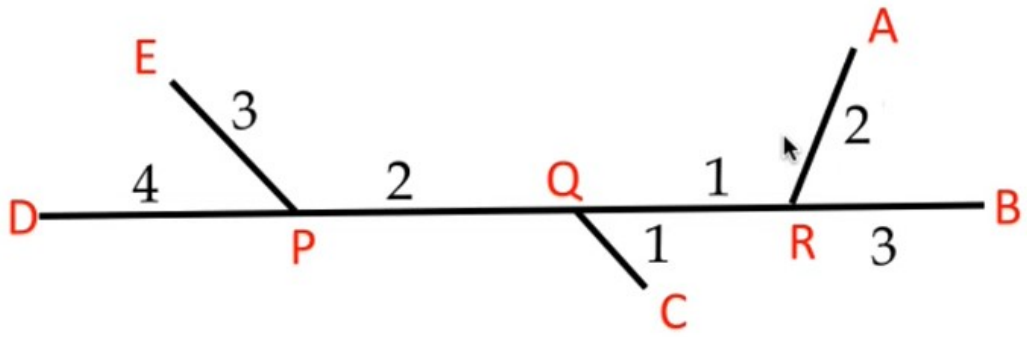
$$b_{QR} = \frac{1}{2}\left(d_{QA} + \left(\frac{U_Q - UA}{N-2}\right)\right)$$

$$b_{QR} = \frac{1}{2}\left(3 + \left(\frac{7-8}{1}\right)\right) = 1$$

$$b_{iY} = \frac{1}{2}\left(d_{ij} + \left(\frac{U_i - Uj}{N-2}\right)\right)$$

$$b_{jY} = d_{ij} - b_{iY}$$

$$b_{RB} = b_{QB} - b_{QR} = 4 - 1 = 3$$



- Unrooted Tree
- Additive Tree

Neighbour-joining Method

Advantages

Robust enough to recover the correct tree topology

Provides information on branch length – Avoids negative branch length

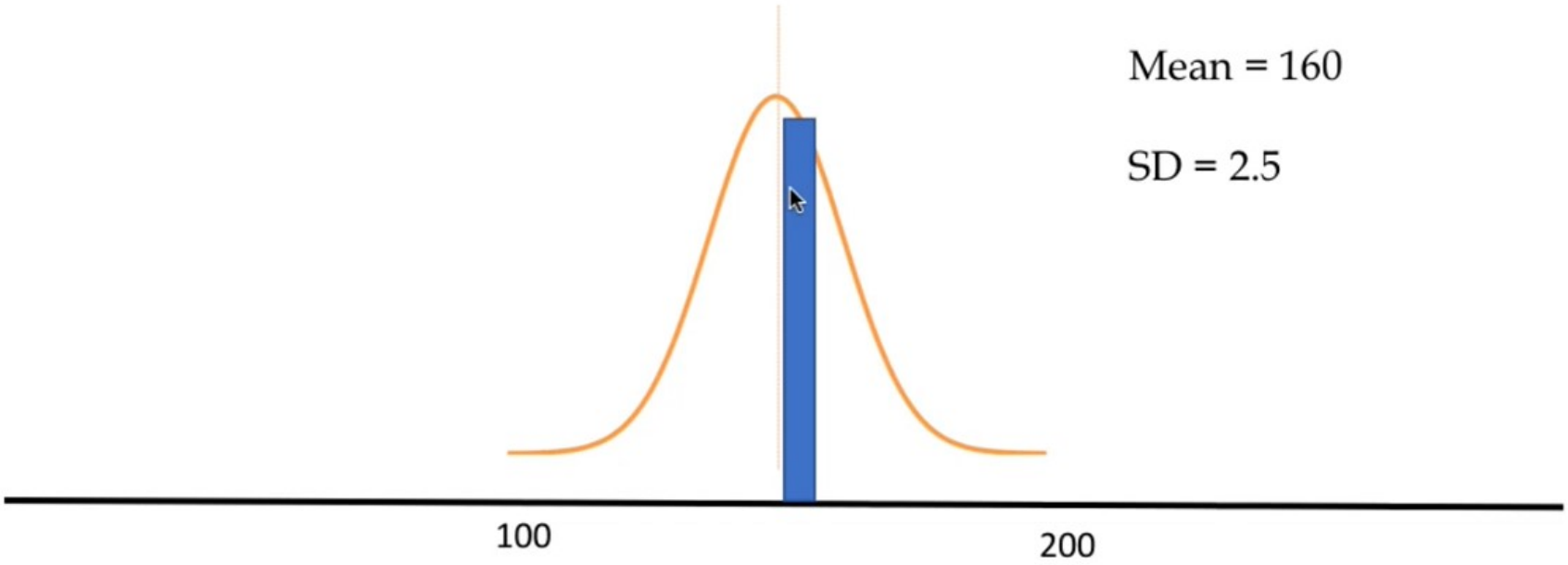
Disadvantages

Sequence information is lost

Does not provide information about ancestral sequences

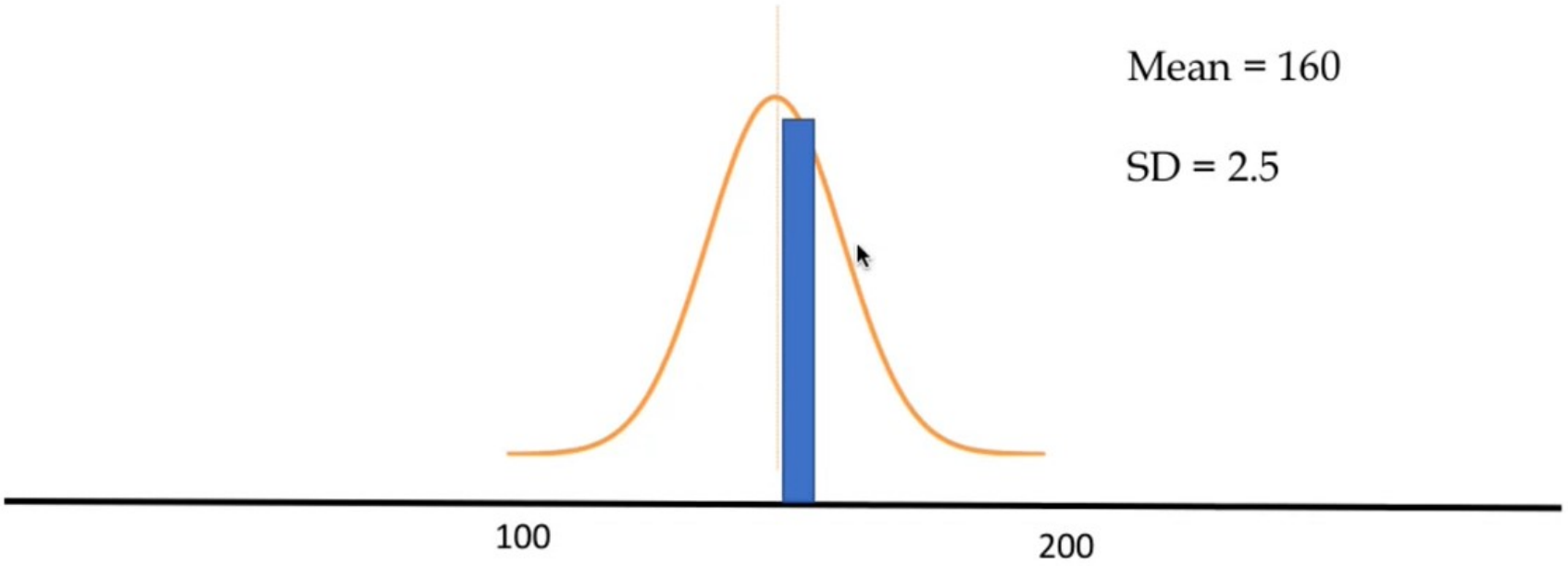
Maximum Likelihood

Probability versus Likelihood



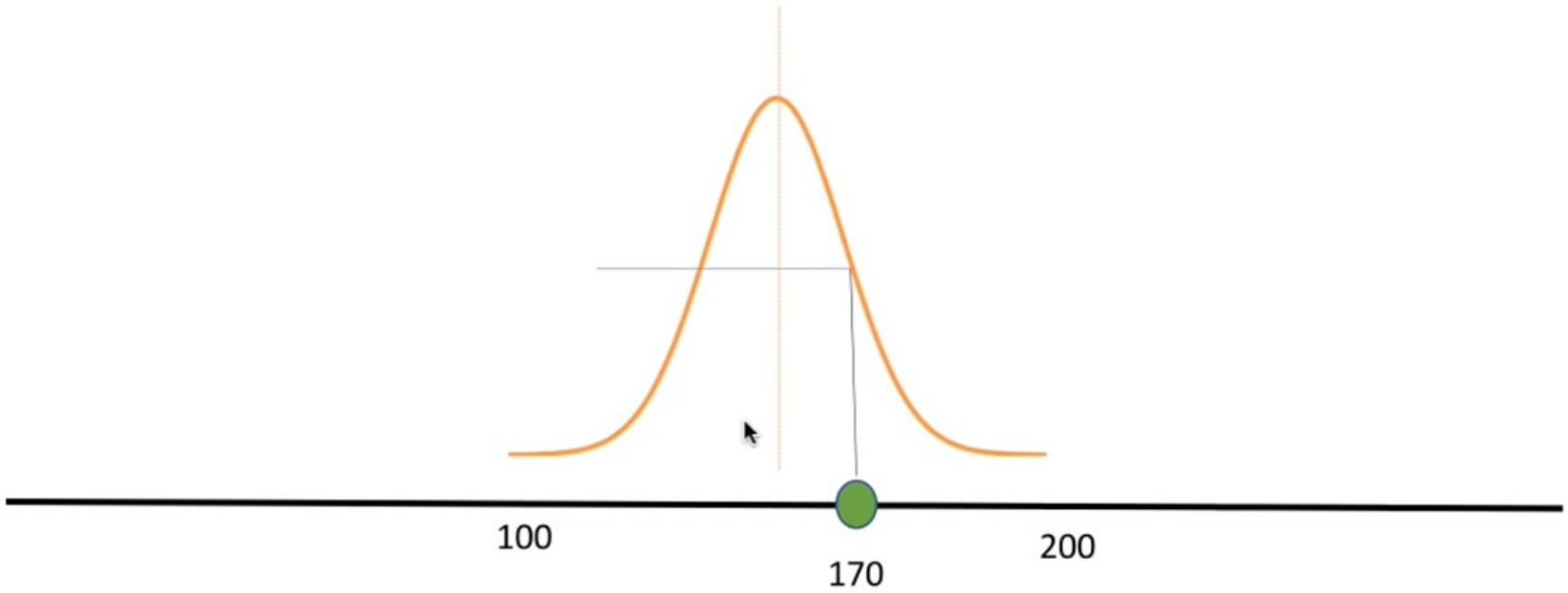
P (X>=150 and X < 160) = Area under the curve

Probability versus Likelihood



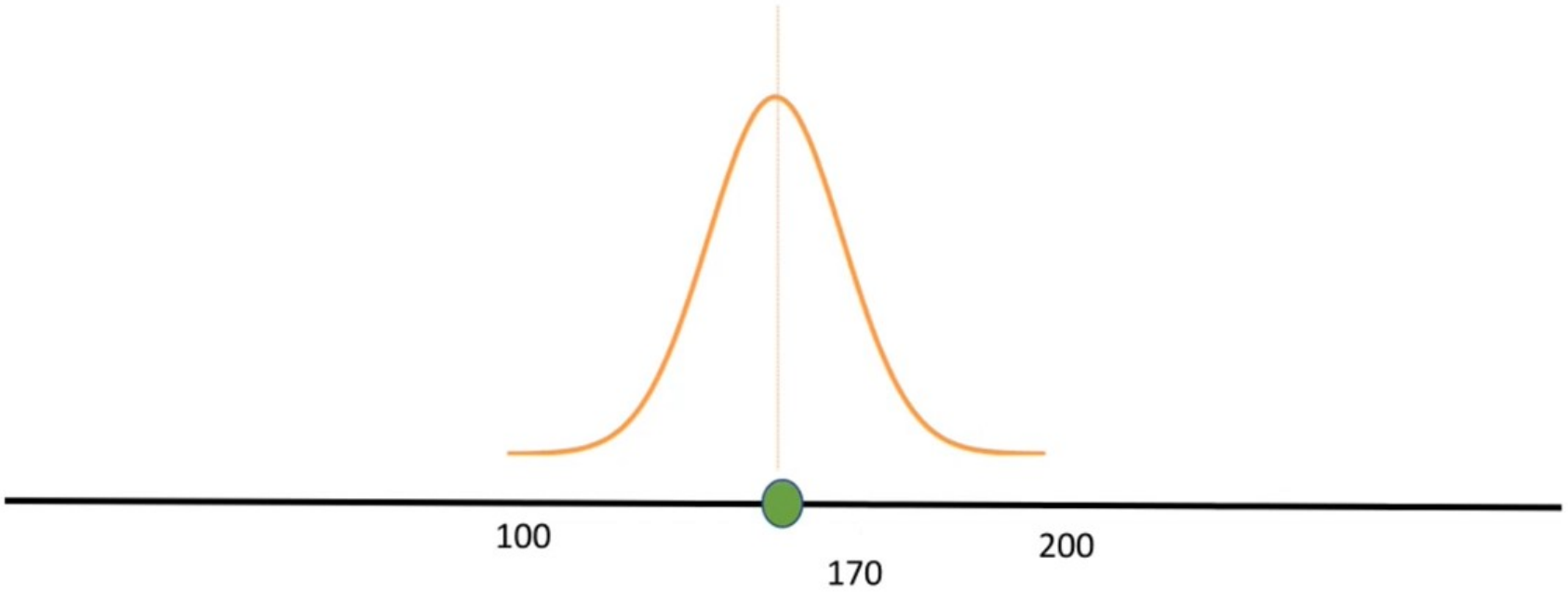
P (**X>=150 and X < 160 | mean =160 and SD =2.5**) = Area under the curve

Probability versus Likelihood



L (mean =160 and SD =2.5 | X = 170 |) = 0.15

Probability versus Likelihood

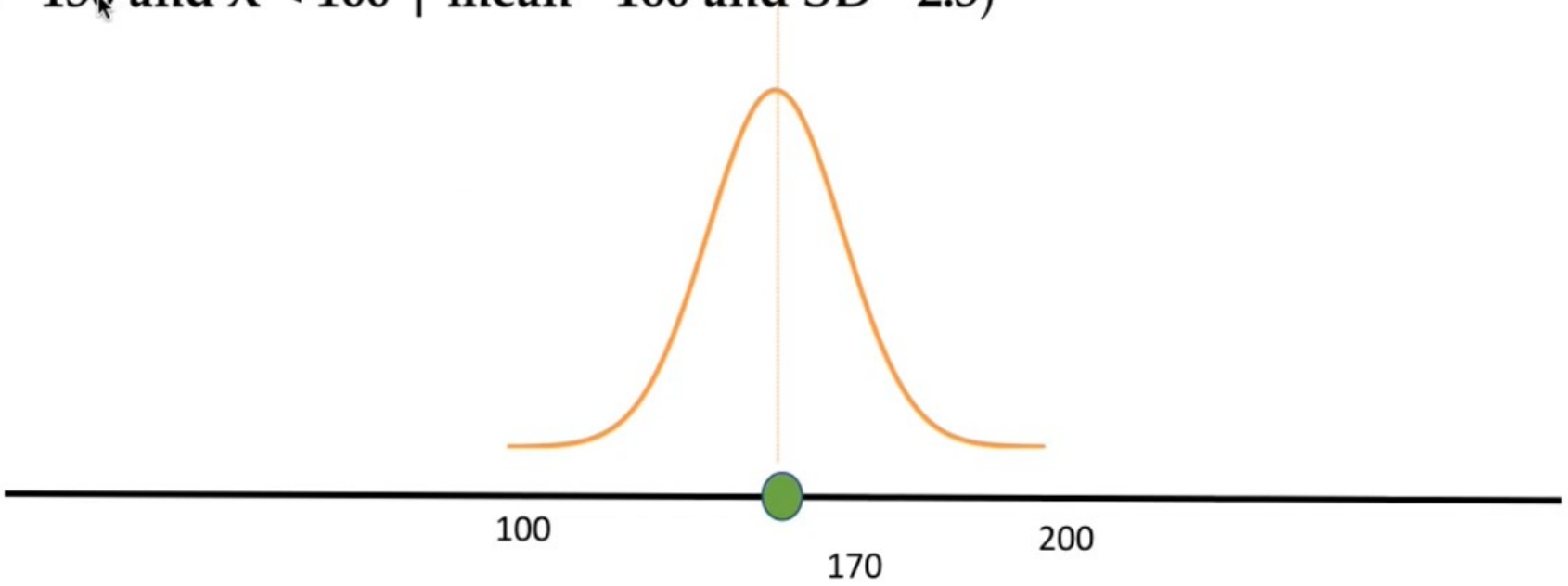


L (mean =160 and SD =2.5 | X = 160 |) = 0.25

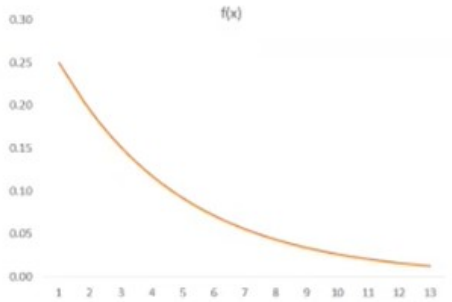
Probability versus Likelihood

L (**mean =160 and SD =2.5 | X = 160**)

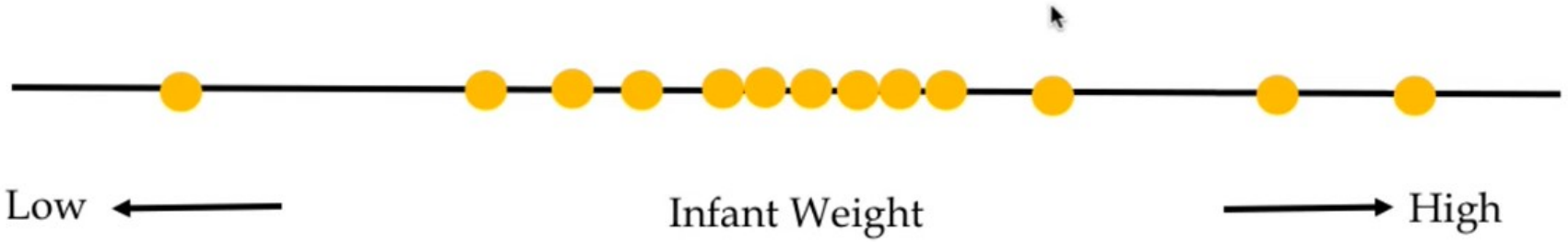
P (**X>=150 and X < 160 | mean =160 and SD =2.5**)



Normal Distribution

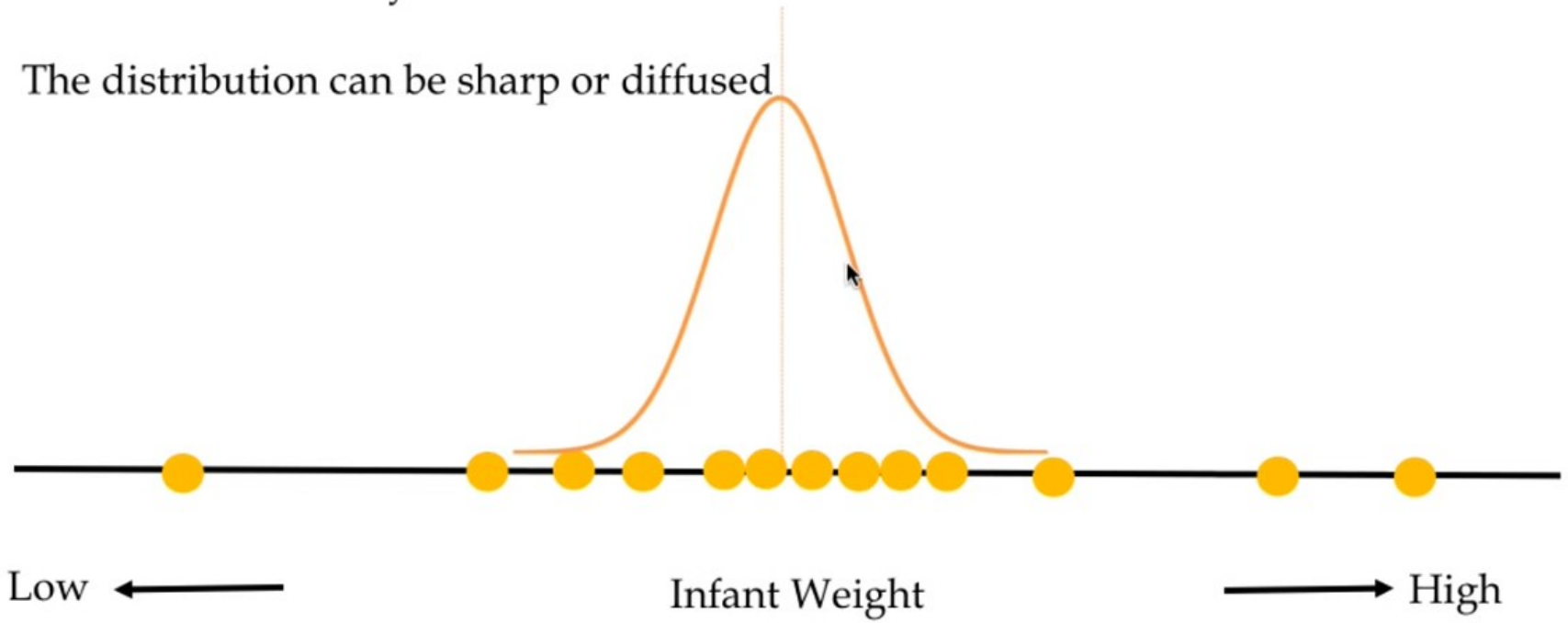


Exponential Distribution

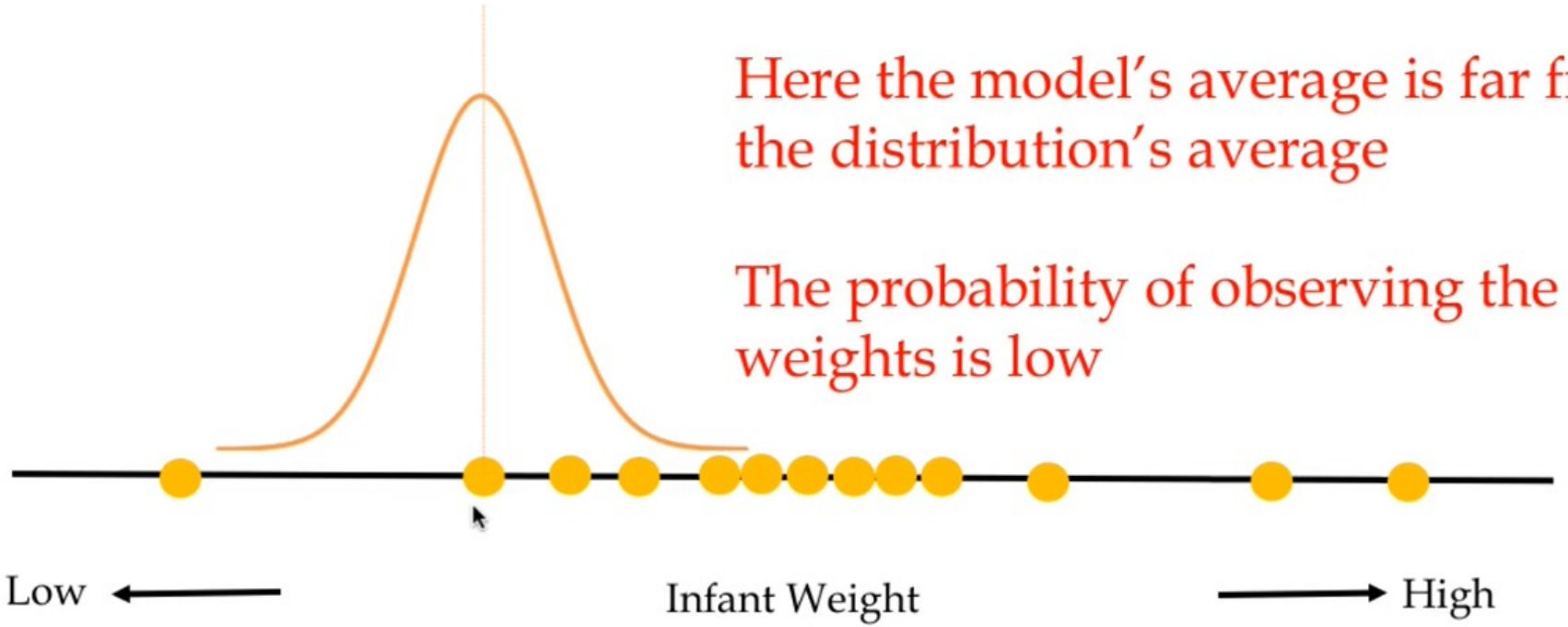


Normal Distribution

1. Most of the measurements are close to the mean of the distribution
2. Measurements are symmetrical about the mean
3. The distribution can be sharp or diffused



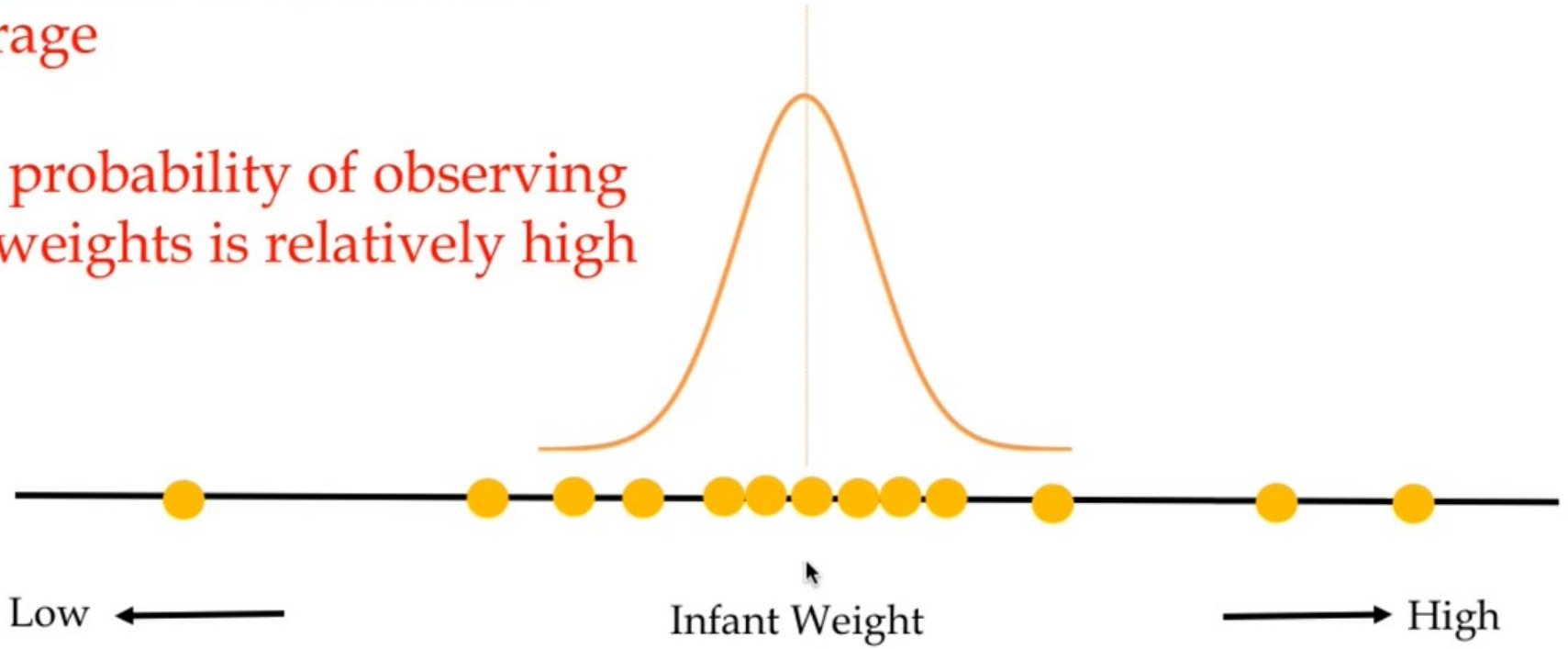
Normal Distribution



Normal Distribution

Here the model's average is close to the distribution's average

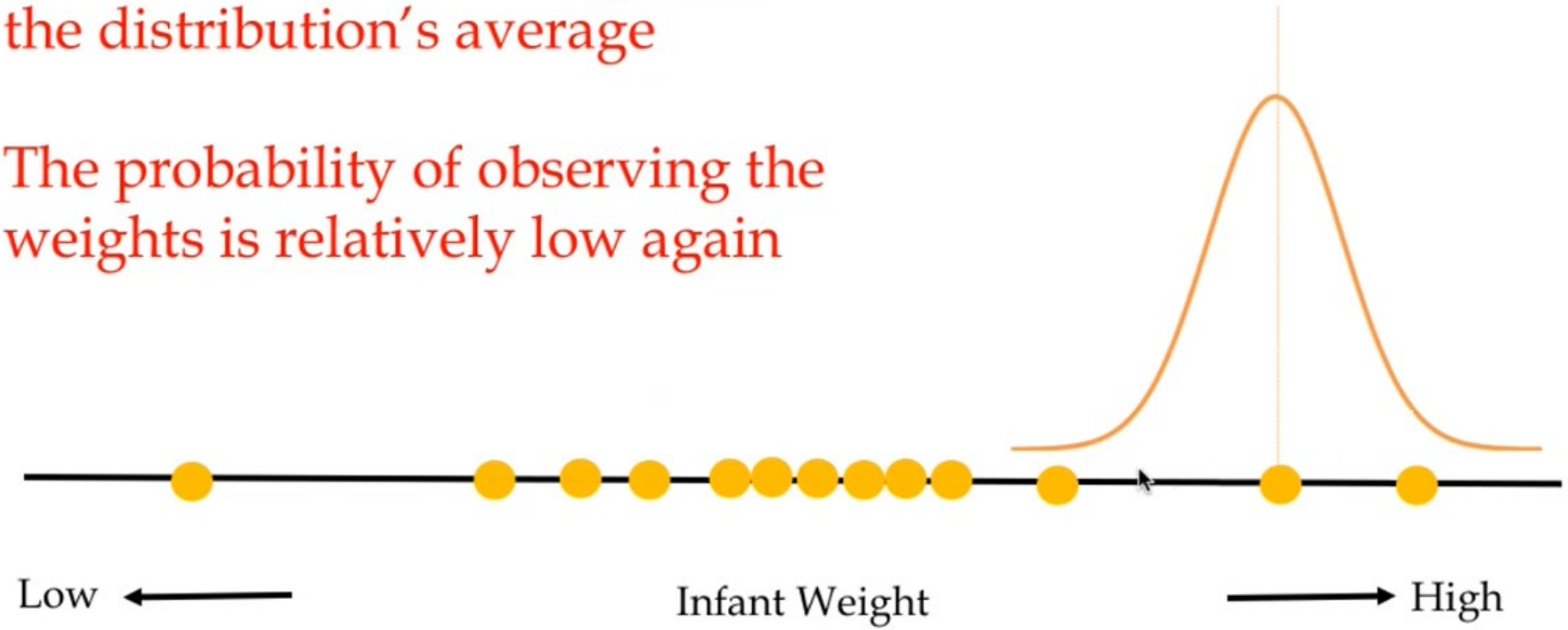
The probability of observing the weights is relatively high

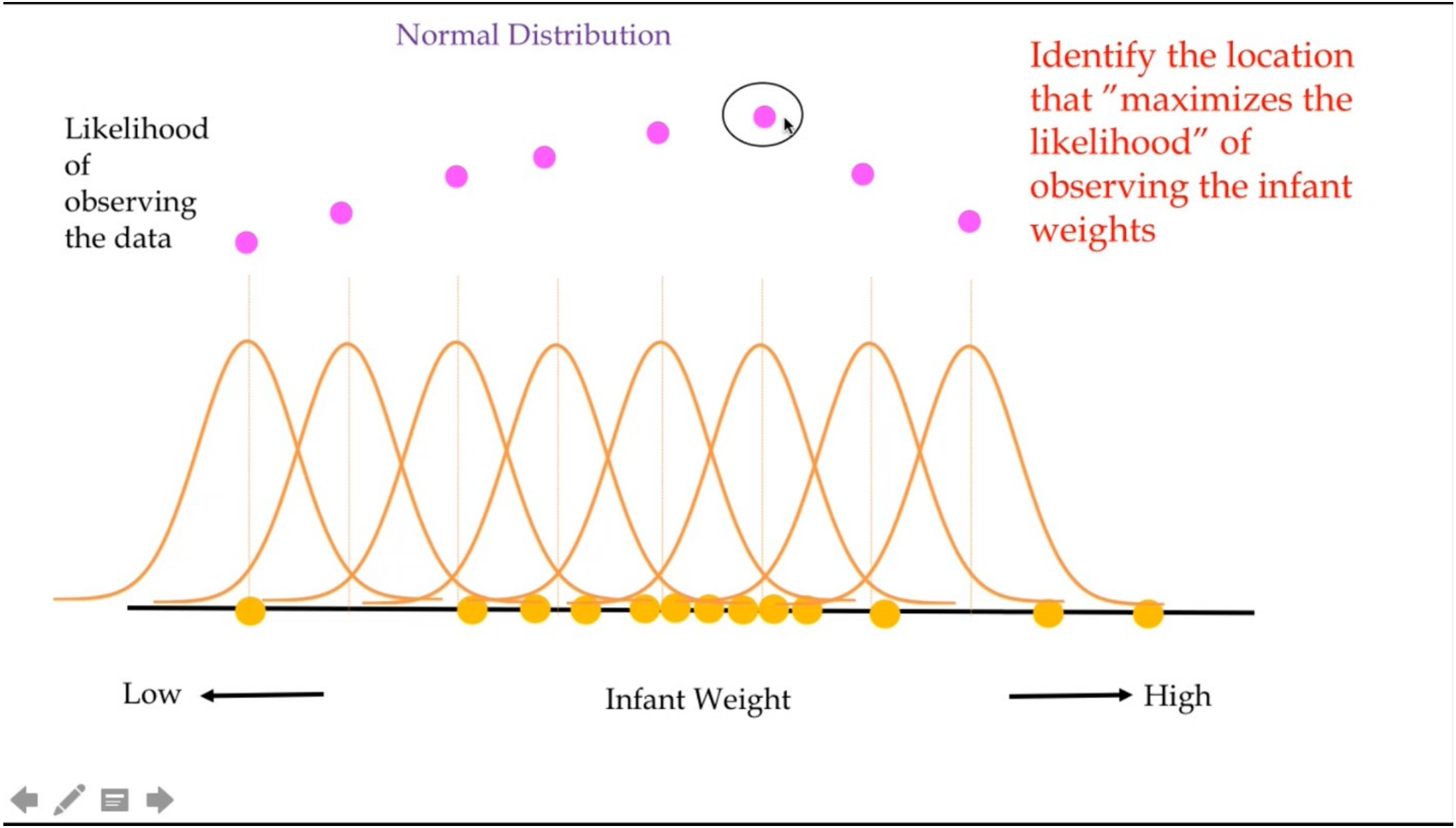


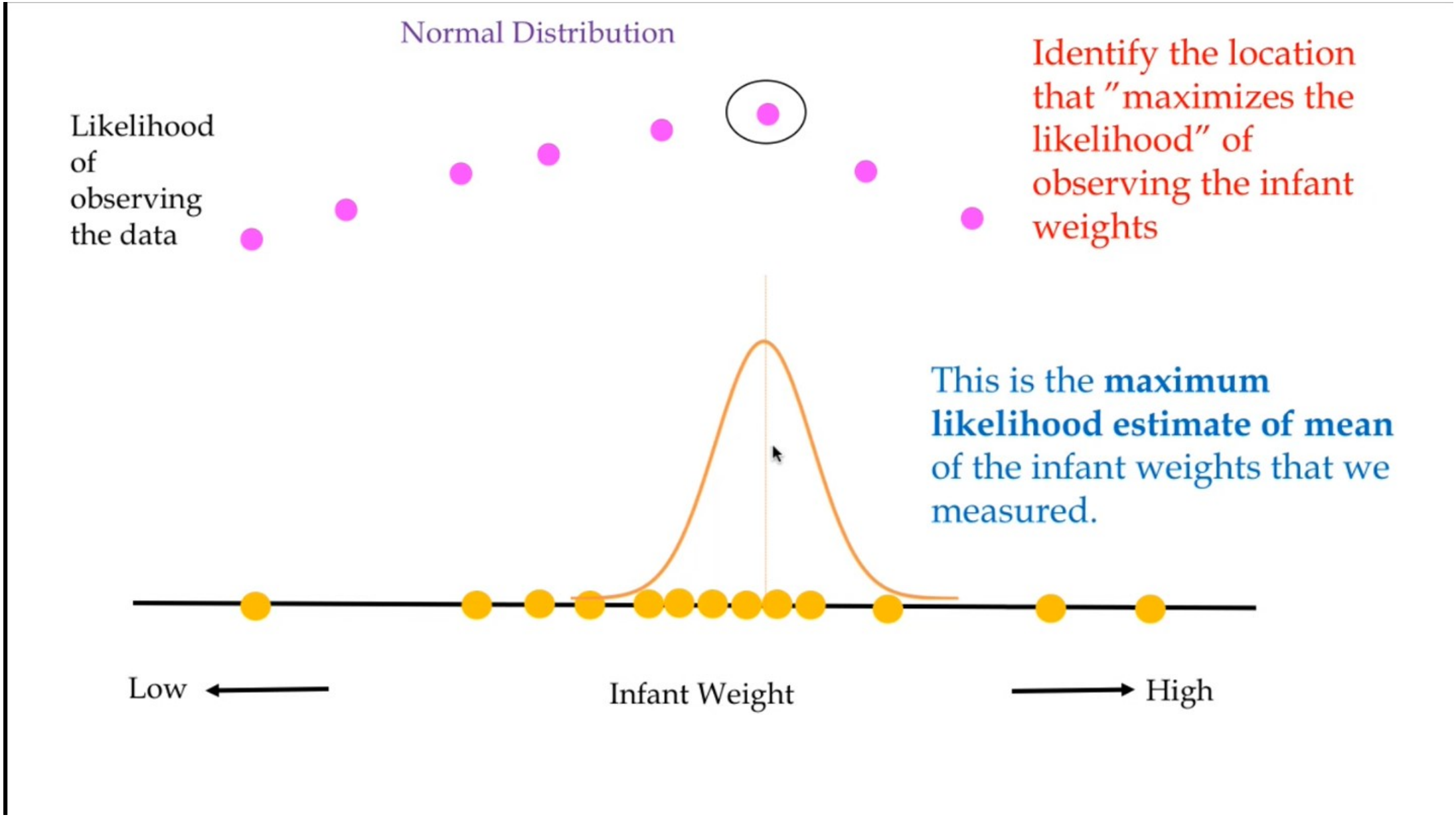
Normal Distribution

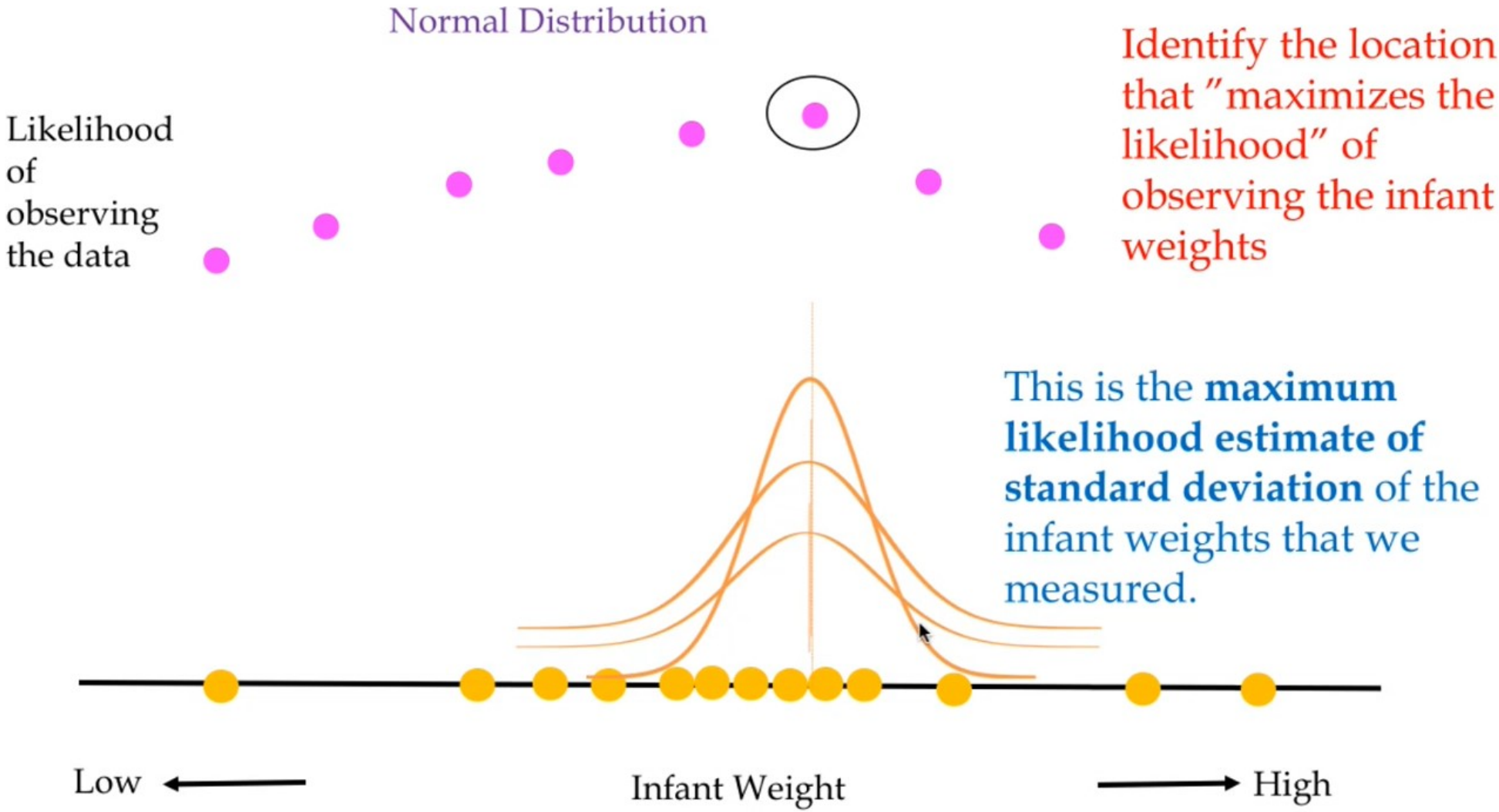
Here the model's average is far from the distribution's average

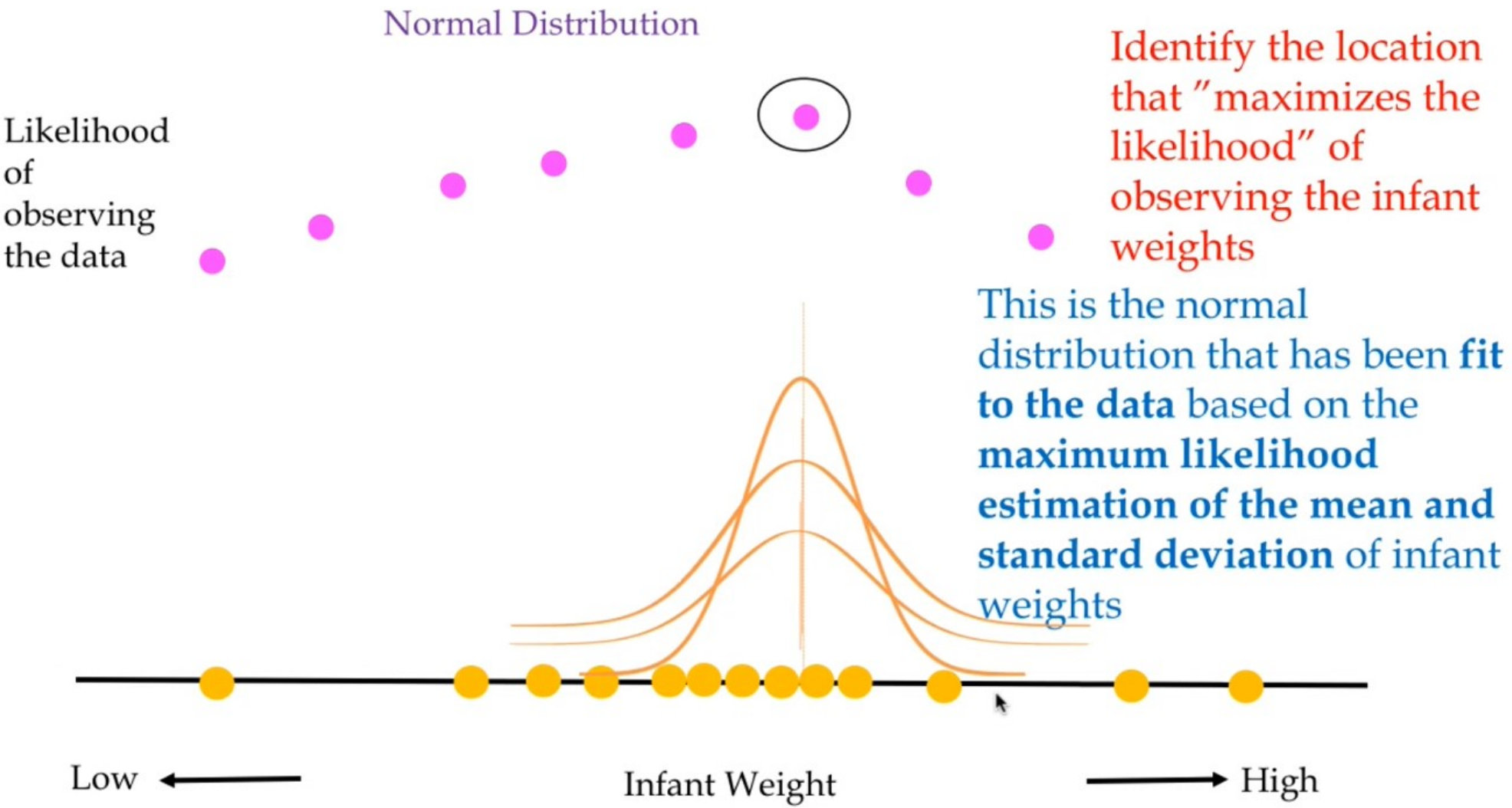
The probability of observing the weights is relatively low again




Normal Distribution
Likelihood of observing the data
Identify the location that "maximizes the likelihood" of observing the infant weights
Low
Infant Weight
High


Normal Distribution
Likelihood of observing the data
Identify the location that "maximizes the likelihood" of observing the infant weights
This is the **maximum likelihood estimate of mean** of the infant weights that we measured.
Low
Infant Weight
High


Normal Distribution
Likelihood of observing the data
Identify the location that "maximizes the likelihood" of observing the infant weights
This is the **maximum likelihood estimate of standard deviation** of the infant weights that we measured.
Low
Infant Weight
High


Normal Distribution
Likelihood of observing the data
Identify the location that "maximizes the likelihood" of observing the infant weights
This is the normal distribution that has been **fit to the data** based on the **maximum likelihood estimation of the mean and standard deviation** of infant weights
Low
Infant Weight
High

Maximum Likelihood Method

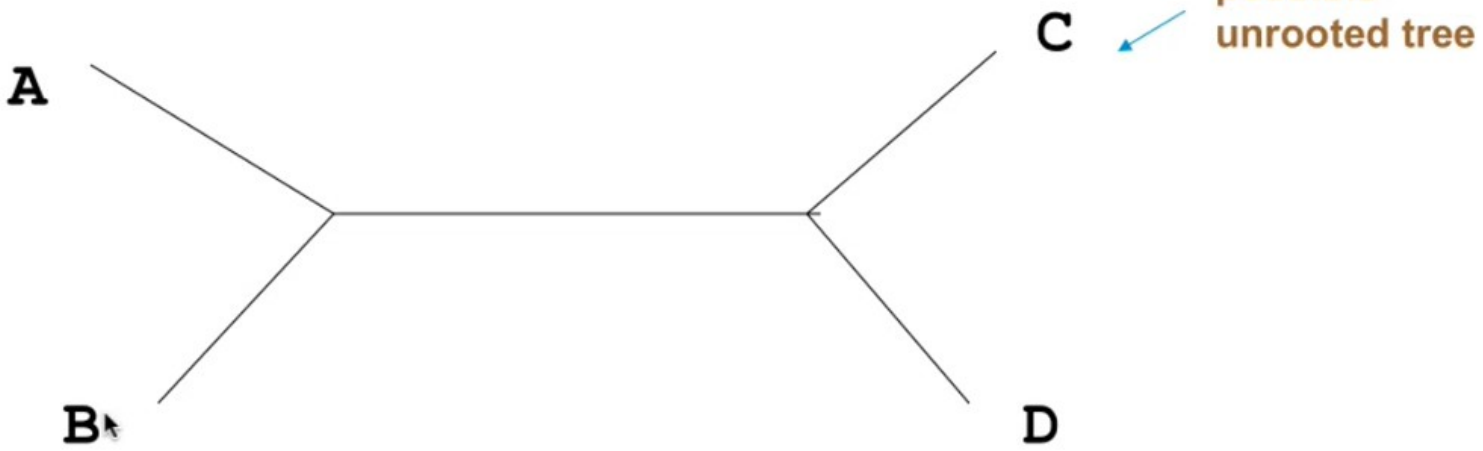
- Uses probability calculations to find a tree that best accounts for the variation in a set of sequences
- All possible trees are considered
- No. of sequence changes/mutations that might have occurred is considered for each tree
- Method can be used to explore relationships between the most diverse sequences
- Computationally intense

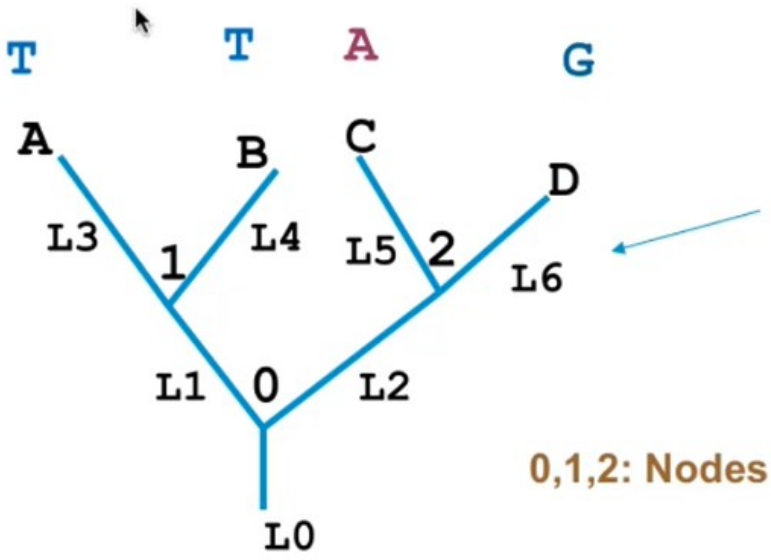
Seq. A: ACGCGTTGGG

Seq. B: ACGCGTTGGG

Seq. C: ACGCAATGAA

Seq. D: ACACAGGGAA





One of the possible rooted tree

0,1,2: Nodes

Four possible bases in three nodes (0,1,2)

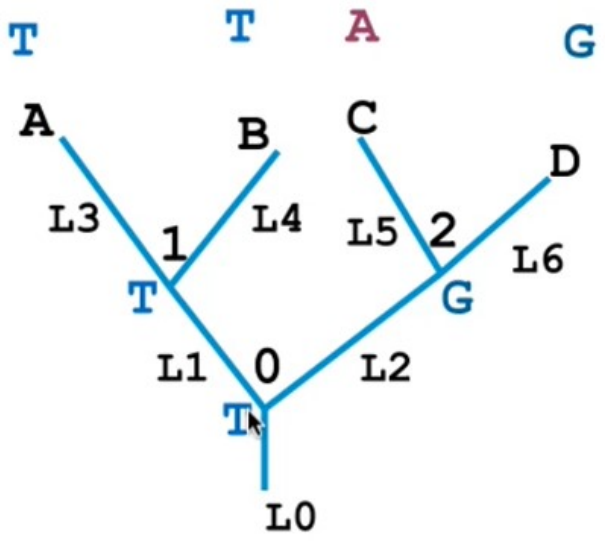
4 x 4 x 4 = 64 possible combinations

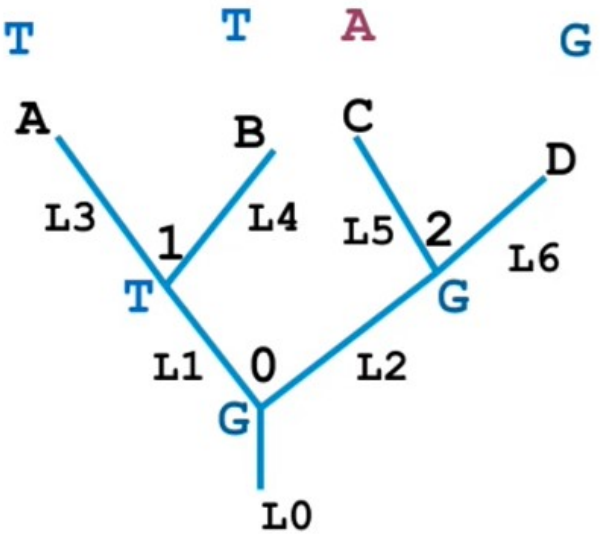
L1-L5: Likelihood values - probability of base change per site

Likelihood of the tree (L) =

L0 * L1 * L2 * L3 * L4 * L5 * L6

Transversion: Purine ↔ Pyrimidine
Transition: Pyrimidine ↔ Pyrimidine
Purine ↔ Purine





L(Tree) = L(Tree1) + L(Tree2) + … + L(Tree64)

Calculation is repeated for all other columns

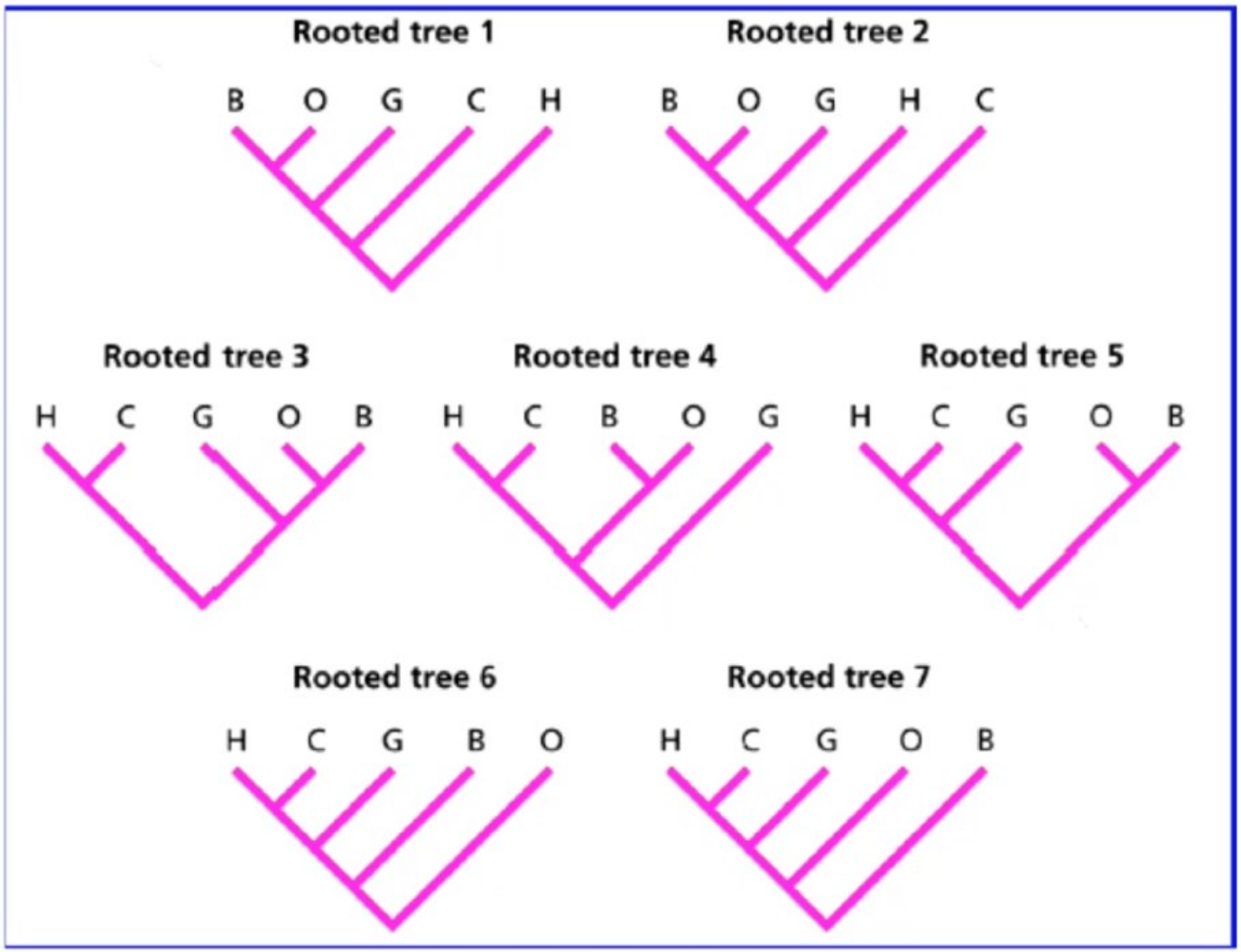
Maximum Likelihood Method

Find the likelihood of the trees for all columns

Each of the three possible trees is evaluated

One with the highest likelihood score is identified

Assessing Tree Topologies



Comparing Tree Topologies

- To resolve ambiguities in tree topology
- Useful when the trees produced by several methods have been applied to a single data set or by the same method applied to different data sets.
- All trees are treated equally in the analysis
- The analysis can identify support across a range of techniques or data for a given topology
- To estimate support for tree topology for a given data set constructed by a particular tree construction method bootstrap analysis is employed

Boot strapping

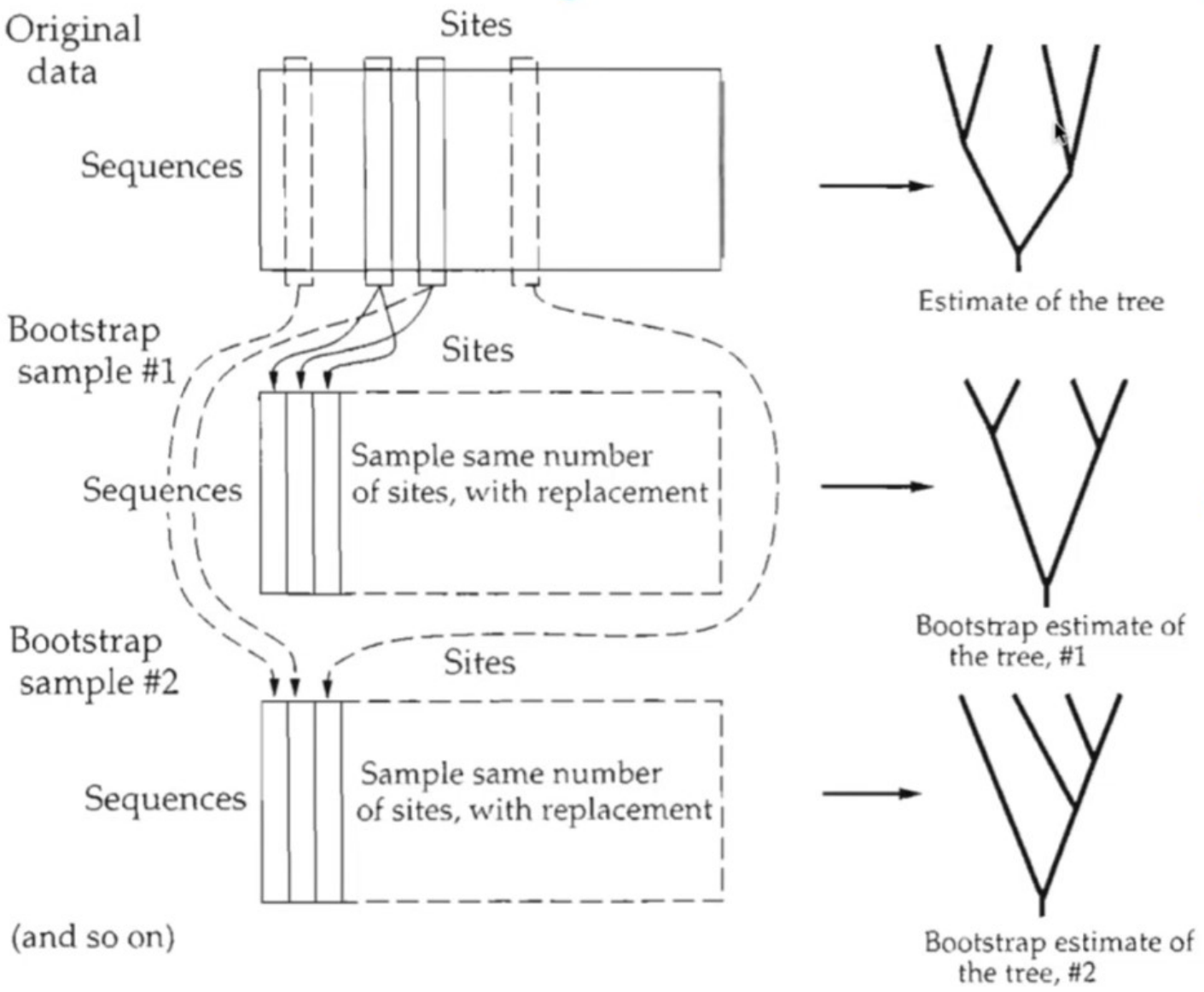
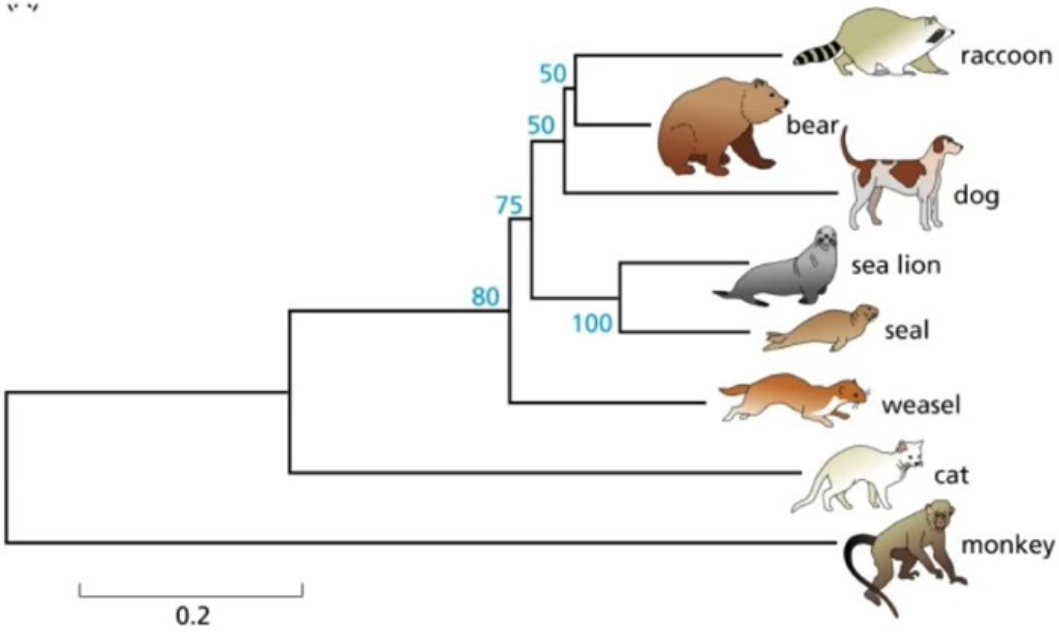
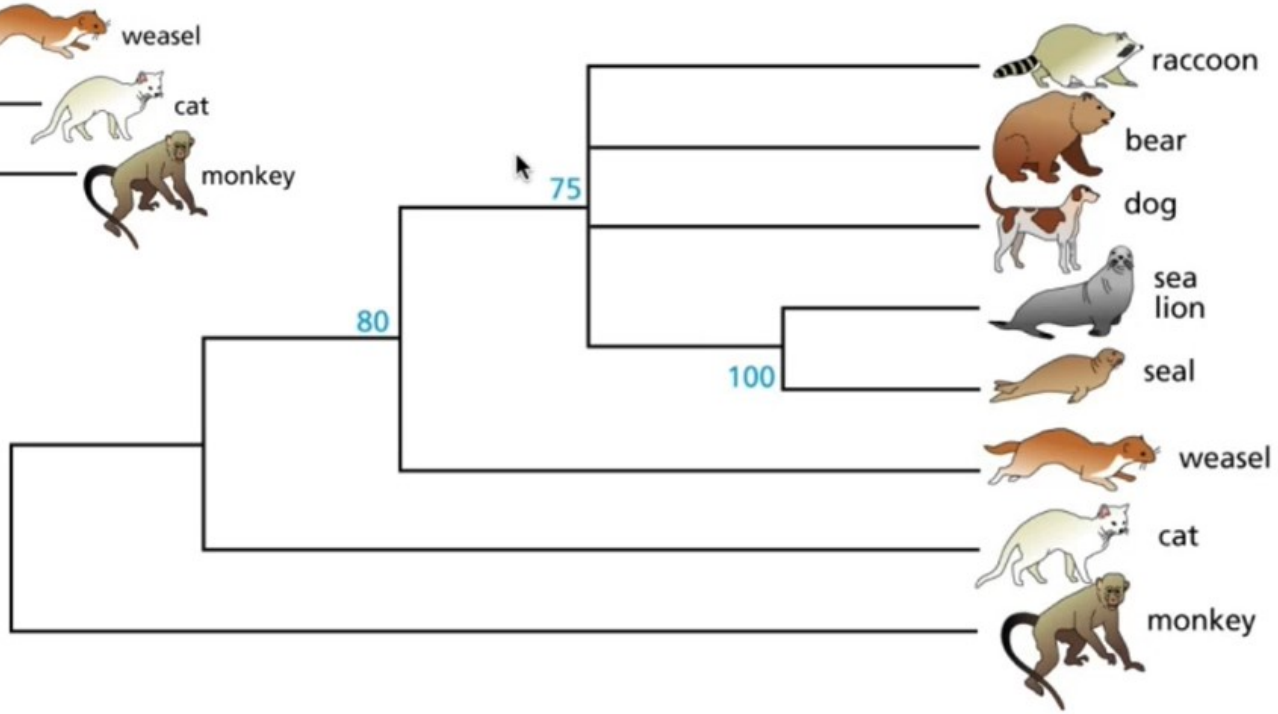


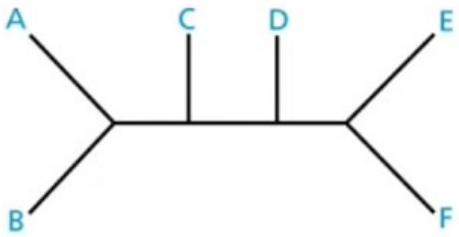
Image Credit: Felsenstein, 2004

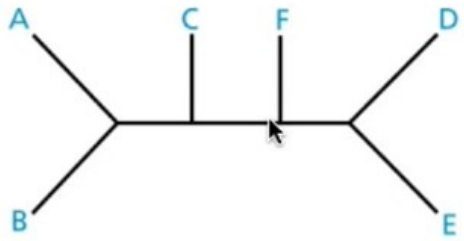
Condensed tree - internal branches not supported by bootstrap values removed

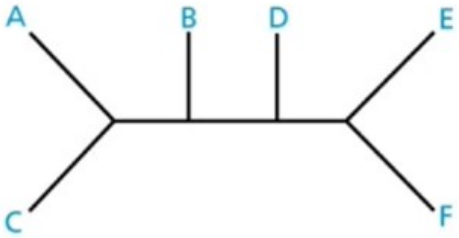


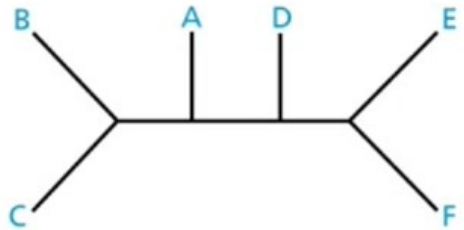


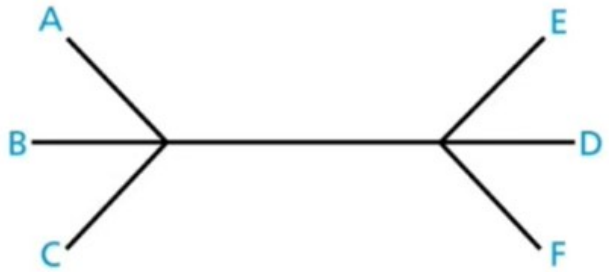
Consensus trees – show features that are consistent between trees



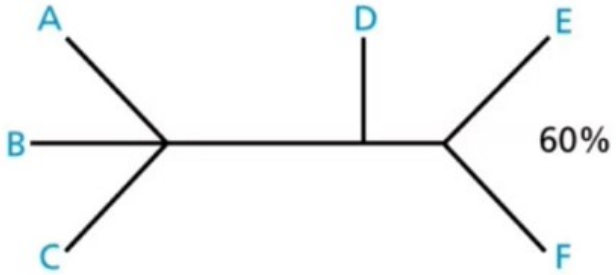


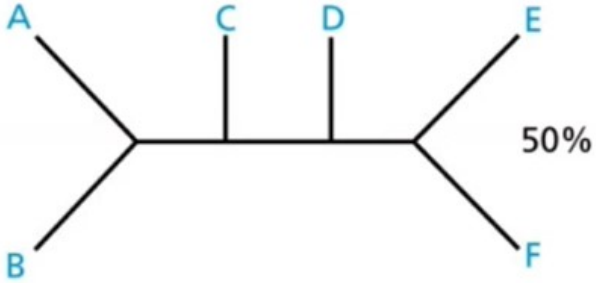






Strict Consensus





Majority Rule Consensus

Genome Analysis

Five approaches to genomics

Approach I: cataloguing genomic information

Genome size; number of chromosomes; GC content; isochores; number of genes; repetitive DNA; unique features of each genome

Approach II: cataloguing comparative genomic information

Orthologs and paralogs; COGs; lateral gene transfer

Approach III: function; biological principles; evolution

How genome size is regulated; polyploidization; birth and death of genes; neutral theory of evolution; positive and negative selection; speciation; epigenetics

Approach IV: Human disease relevance

Approach V: Bioinformatics aspects

Algorithms, databases, websites

Prominent web resources

Ensembl genomes

http://www.ensembl.org
http://ensemblgenomes.org

NCBI Genome

https://www.ncbi.nlm.nih.gov/genome

Dept. of Energy Joint Genome Institute (DOE JGI)

http://jgi.doe.gov

Genomes On Line Database (GOLD)

https://gold.jgi.doe.gov

UCSC

https://genome.ucsc.edu
http://microbes.ucsc.edu

Chronology of genome sequencing projects

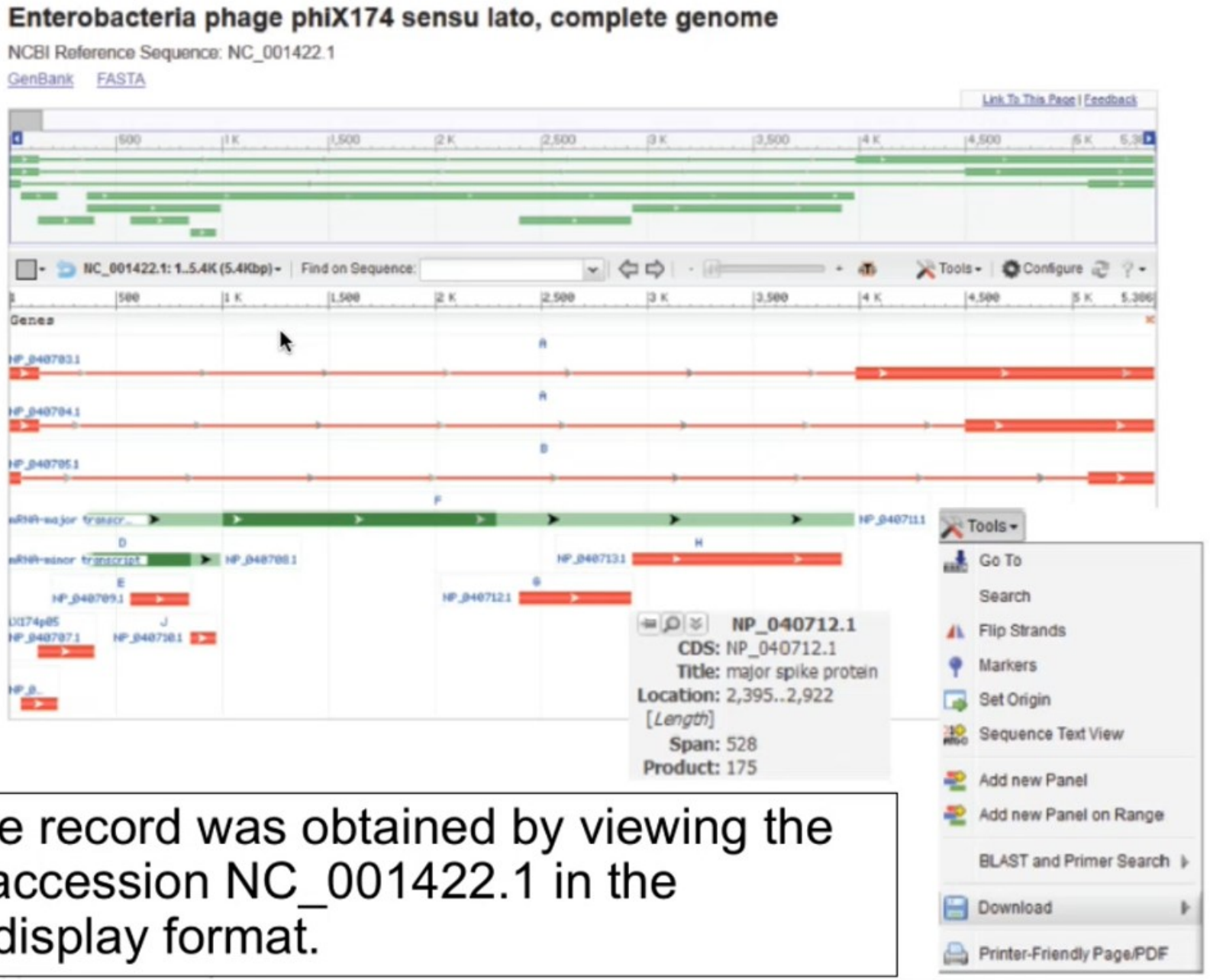
1976: first viral genome

Fiers et al. sequence bacteriophage MS2 (3,569 base pairs, accession NC_001417). Bacteriophage are viruses that infect bacteria.

1977:

Sanger et al. sequence bacteriophage ϕX174.
This virus is 5,386 base pairs (encoding 11 genes).
See accession J02482; NC_001422.

NCBI data for bacteriophage φX174



Nucleotide record was obtained by viewing the entry for accession NC_001422.1 in the graphics display format.

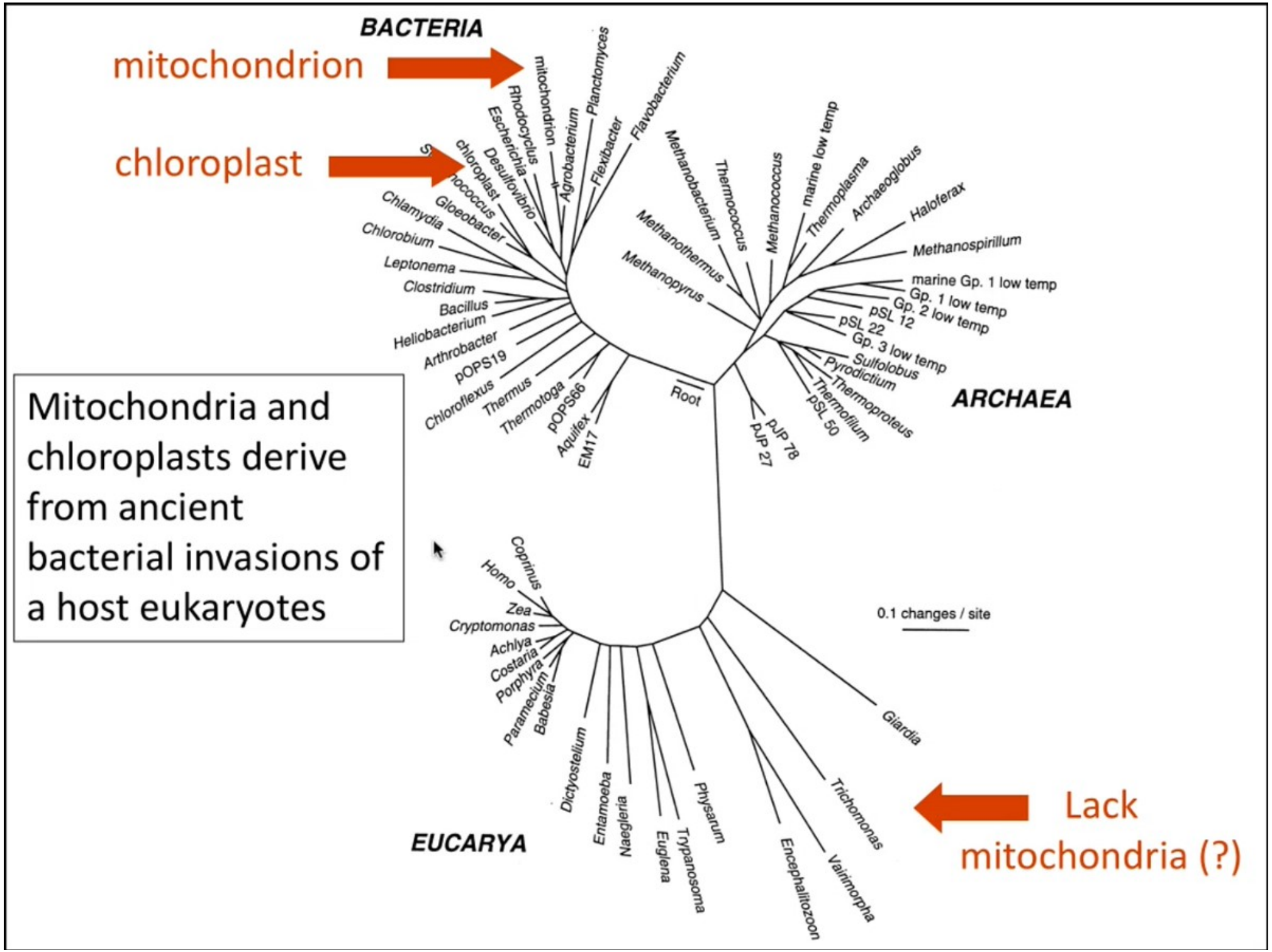
NCBI data for bacteriophage φX174

NCBI Genome record includes a summary of the accession number, length, number of proteins (11), sequence neighbors (*n* = 77), and host species.

Chronology of genome sequencing projects

1981
Human mitochondrial genome: first eukaryotic organellar genome. 16,500 base pairs (encodes 13 proteins, 2 rRNA, 22 tRNA). Today (2021), over 12589 mitochondrial genomes sequenced.

1986
Chloroplast genome
156,000 base pairs (most are 120 kb to 200 kb)
>6000 chloroplast sequences (2021)


BACTERIA
mitochondrion
chloroplast
mitochondrion
Rhodocyclus
Escherichia
Desulfovibrio
chloroplast
Agrobacterium
Planctomyces
Flexibacter
Flavobacterium
Chlamydia
Chlorobium
Leptonema
Clostridium
Bacillus
Heliobacterium
Arthrobacter
pOPS19
Chloroflexus
Thermus
Thermotoga
pOPS66
Aquifex
EM17
Root
Methanobacterium
Thermococcus
Methanococcus
marine low temp
Thermoplasma
Archaeoglobus
Haloferax
Methanospirillum
Methanothermus
Methanopyrus
marine Gp. 1 low temp
Gp. 1 low temp
Gp. 2 low temp
pSL 12
pSL 22
Gp. 3 low temp
Sulfolobus
Pyrodictium
Thermoproteus
Thermofilum
pSL 50
pJP 78
pJP 27
ARCHAEA
Mitochondria and chloroplasts derive from ancient bacterial invasions of a host eukaryotes
Coprinus
Homo
Zea
Cryptomonas
Achlya
Costaria
Porphyra
Paramecium
Babesia
Dictyostelium
Entamoeba
Naegleria
Euglena
Trypanosoma
Physarum
Encephalitozoon
Vairimorpha
Trichomonas
Giardia
0.1 changes / site
EUCARYA
Lack mitochondria (?)

MitoMap: resource for organelle genomes

MITOMAP

A human mitochondrial genome database

A compendium of polymorphisms and mutations of the human mitochondrial DNA

INDEXED IN
ISI Web of Knowledge℠
Current Web Contents

Scout Report SELECTION

BEAGLE WEBPICK

Search MITOMAP for information on: [ ] Perform search Clear

Gene, disease, enzyme names may be abbreviated, truncated, etc. Examples of search words: ND1, ND4, NARP, LHON, 11778, 3243, Leu, Lys, etc.

Mitomap Quick Reference

The Human Mitochondrial Sequence (updated)

Amino Acid Translation Tables

Mitochondrial References (A-Z) (>1 MB)
or view only authors A-L or authors M-Z

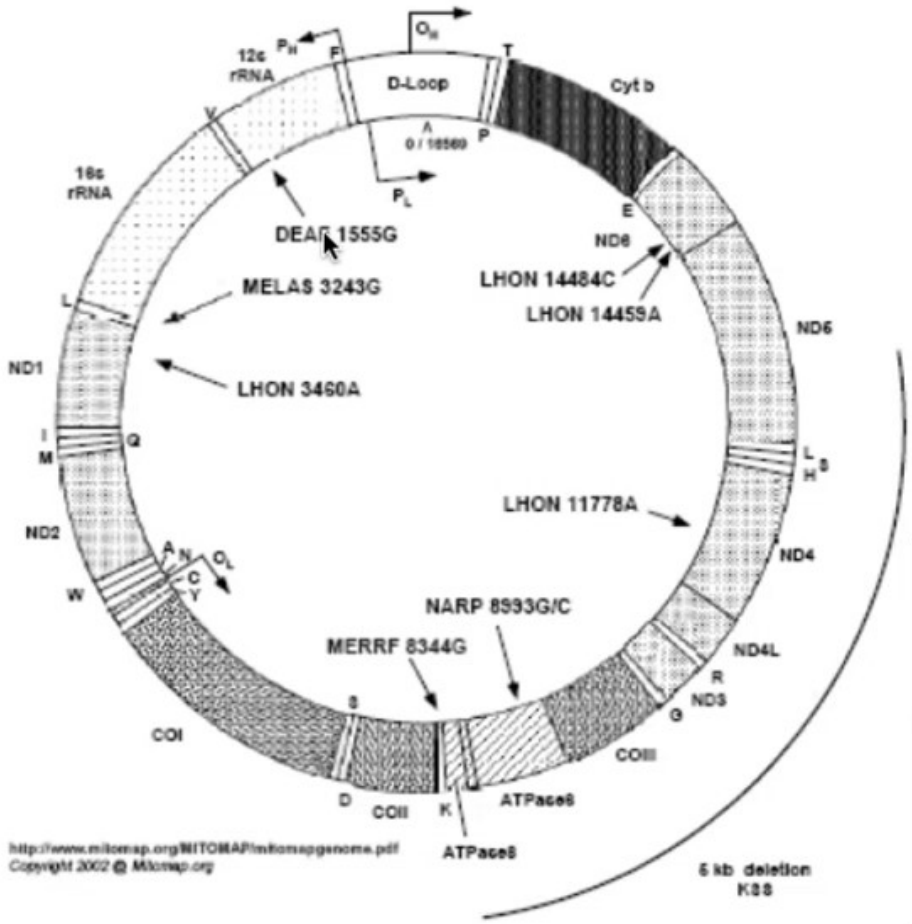
Mitochondrial DNA Function Locations (Gene Loci)

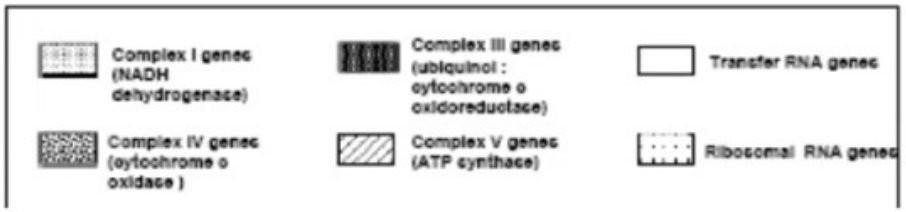
Illustrations

-Mitochondrial DNA Map
-Eleven pathological mutations in $tRNA^{Leu(UUR)}$
-Mitochondrial energetics
-Diabetes metabolism & the mitochondria
-World migrations
-mtDNA Tree

http://www.mitomap.org/

It is possible to map mutations in human mitochondrial DNA that are responsible for disease





Morbid map of the mitochondrial genome showing diseases associated with particular mutations.

Chronology of genome sequencing projects

1992: first eukaryotic chromosome.
Chromosome III of the budding yeast *S. cerevisiae* was sequenced.

1995: first genome of a free-living organism, the bacterium *Haemophilus influenzae.*
Accession: NC_000907.1
Size:1,830,138 bp (1.8 Mb or megabase pairs)

Chronology of genome sequencing projects

1997:
More bacteria and archaea
Escherichia coli
4.6 megabases, 4200 proteins (38% of unknown function)

1998: first multicellular organism
Nematode *Caenorhabditis elegans*
97 Mb; 19,000 genes.

1999: first human chromosome
Chromosome 22 (49 Mb, 673 genes)

Chronology of genome sequencing projects

2000:
Fruitfly *Drosophila melanogaster* (13,000 genes)

Plant *Arabidopsis thaliana*

2001: draft sequence of the human genome
(public consortium and Celera Genomics).

Chronology of genome sequencing projects

2002: More completed genomes
2003: HapMap (catalog variation in human genome)
2004: chicken, rat; finished human chromosomes
2005: chimpanzee, dog
2006: sea urchin, honeybee
2007: rhesus macaque, first individual human
2008: platypus, first cancer genome
2009: bovine, methylome map
2010: 1000 Genomes pilot, Neanderthal
2011: genomics vision
2012: Denisovan, bonobo
2013: comb jelly, 700,000 year old horse
2014: primates, plants, ancient hominids
2015: diversity in Africa
2021: Complete Human Genome

Overview of genome analysis

- Selection of genomes for sequencing
- Sequence one individual genome, or several?
- How big are genomes?
- Genome sequencing centers
- Sequencing genomes: strategies
- When has a genome been fully sequenced?
- Repository for genome sequence data
- Genome annotation

Applications of genome sequencing

Purpose	Template	Example
De novo sequencing	Genome sequencing	Sequencing >1000 influenza genomes
	Ancient DNA	Extinct Neandertal genome
	Metagenomics	Human gut
Resequencing	Whole genomes	Individual humans
	Genomic regions	Assessment of genomic rearrangements or disease-associated regions
	Somatic mutations	Sequencing mutations in cancer
Transcriptome	Full-length transcripts	Defining regulated messenger RNA transcripts
	Noncoding RNAs	Identifying and quantifying microRNAs in samples
Epigenetics	Methylation changes	Measuring methylation changes in cancer

Genome sequencing has many applications and purposes.

Large-scale human genome sequencing projects

Recent large-scale projects include:

- Human-centered projects
- Human Microbiome Project
- For nonhuman species, large numbers of strains, inbred lines, or related organisms

Large-scale human genome sequencing projects

Project	URL	Goal
The 1000 Genomes Project	http://www.1000genomes.org/	Find human genetic variants having frequencies >1%
International Cancer Genome Consortium (ICGC)	http://www.icgc.org/	Catalog mutations in tumors from 50 cancer types
UK10K	http://www.sanger.ac.uk/about/press/2010/100624-uk10k.html	Sequence the genomes of 10,000 UK individuals
100,000 Genomes Project	http://www.genomicsengland.co.uk/	Sequence 100,000 individuals in the UK
Autism Genome 10K Project	http://autismgenome10k.org/	Sequence 10,000 autism-related genomes
Personal Genome Project	http://www.personalgenomes.org/	Effort to sequence 100,000 human genomes

Currently separate projects involving sequencing of hundreds of thousands of exomes/genomes are in progress.

Large-scale human genome sequencing projects

Project	URL	Goal
The 1000 Genomes Project	http://www.1000genomes.org/	Find human genetic variants having frequencies >1%
International Cancer Genome Consortium (ICGC)	http://www.icgc.org/	Catalog mutations in tumors from 50 cancer types
UK10K	http://www.sanger.ac.uk/about/press/2010/100624-uk10k.html	Sequence the genomes of 10,000 UK individuals
100,000 Genomes Project	http://www.genomicsengland.co.uk/	Sequence 100,000 individuals in the UK
Autism Genome 10K Project	http://autismgenome10k.org/	Sequence 10,000 autism-related genomes
Personal Genome Project	http://www.personalgenomes.org/	Effort to sequence 100,000 human genomes

Currently separate projects involving sequencing of hundreds of thousands of exomes/genomes are in progress.

Criteria for selecting genomes for sequencing

Criteria include:

- genome size (some plants are >>>human genome)
- cost
- relevance to human disease (or other disease)
- relevance to basic biological questions
- relevance to agriculture

Diversity of genome sizes

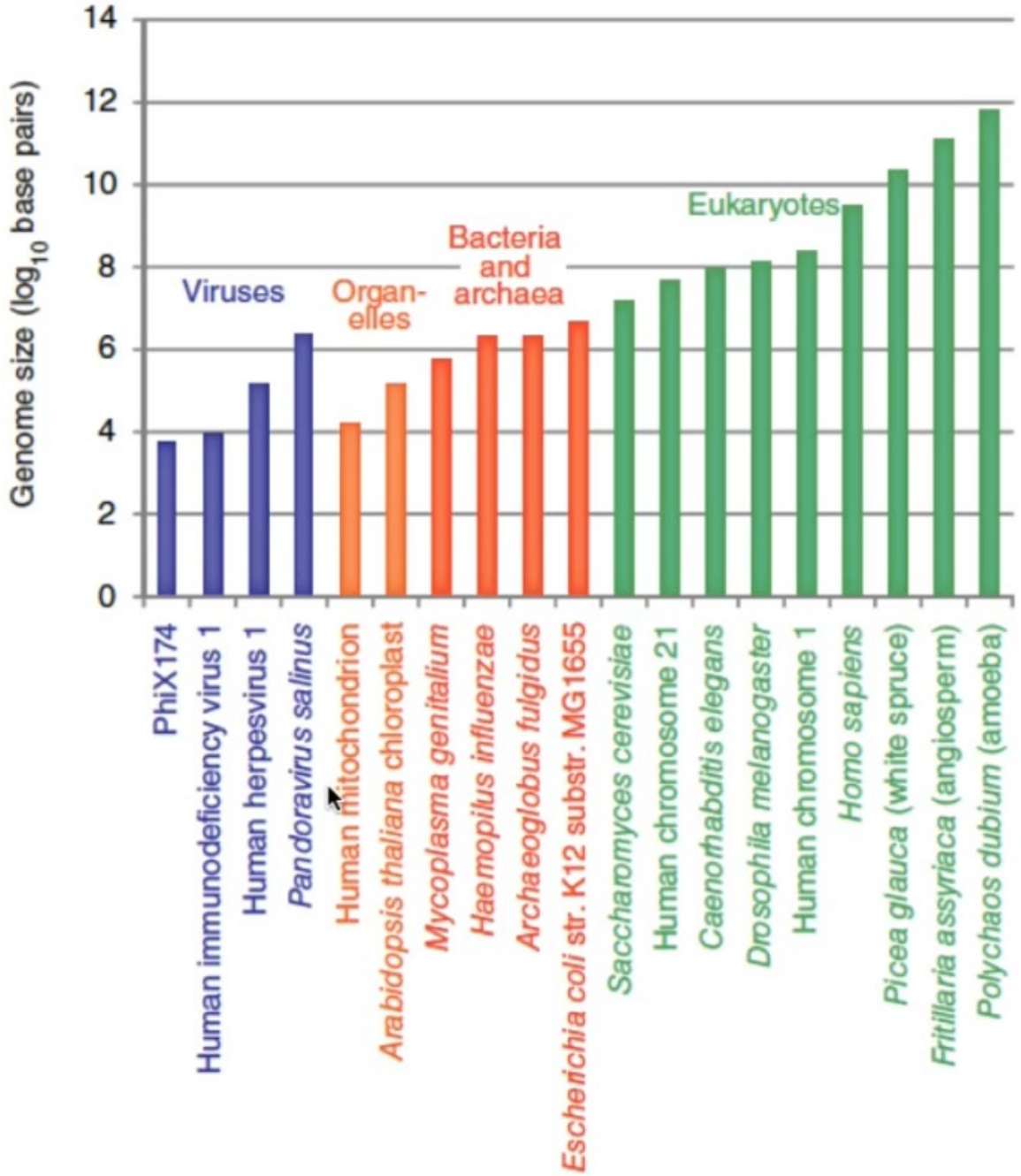
How big are genomes?

Viral genomes: 1 kb to 350 kb (Mimivirus: 1181 kb)

Bacterial genomes: 0.5 Mb to 13 Mb

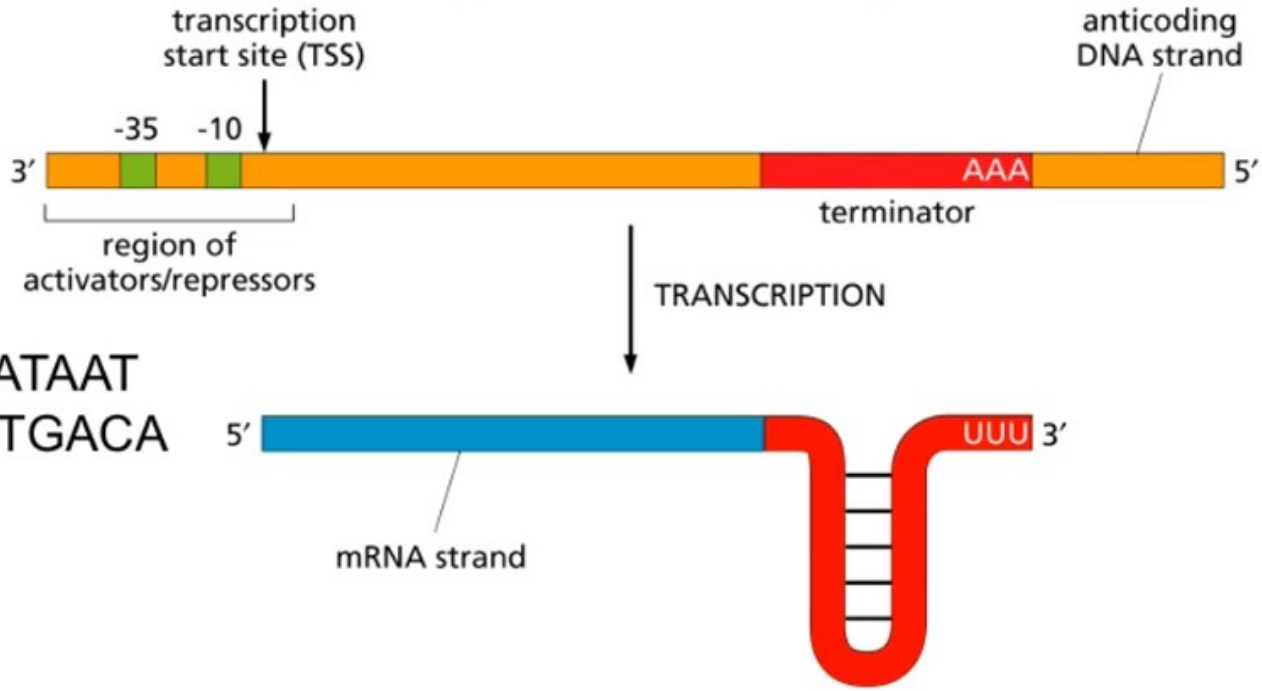
Eukaryotic genomes: 8 Mb to 686 Gb (human: ~3 Gb)

Genomes sizes from viruses to eukaryotes



Genome Annotation

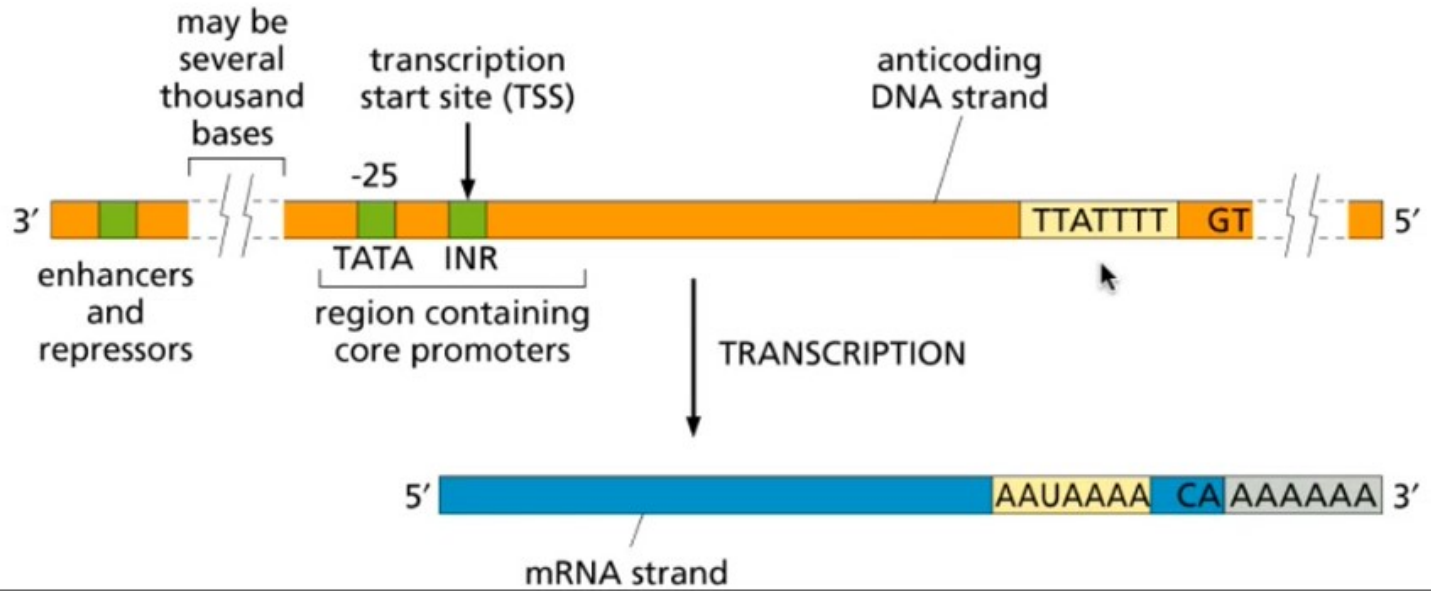
Prokaryotic Transcription

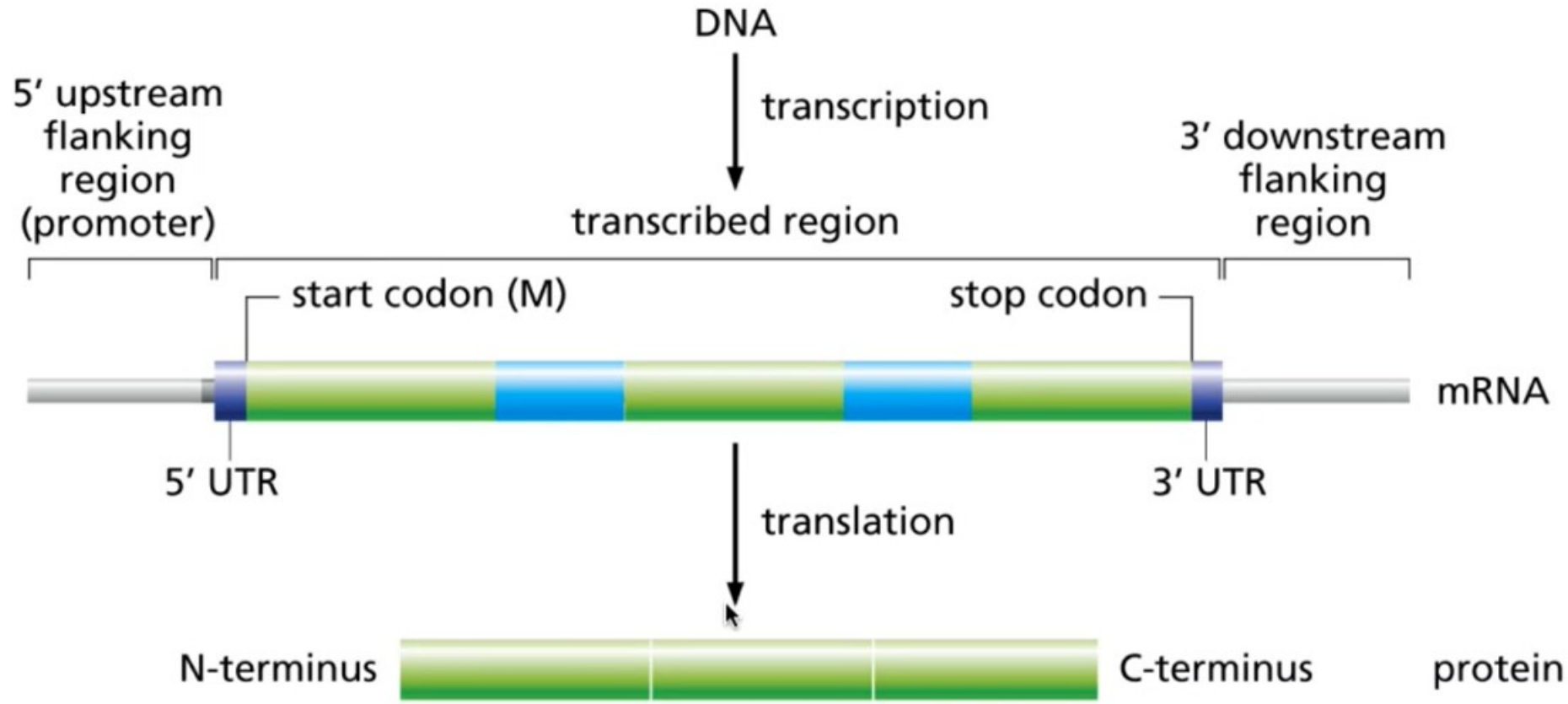


-10 Motif – TATAAT

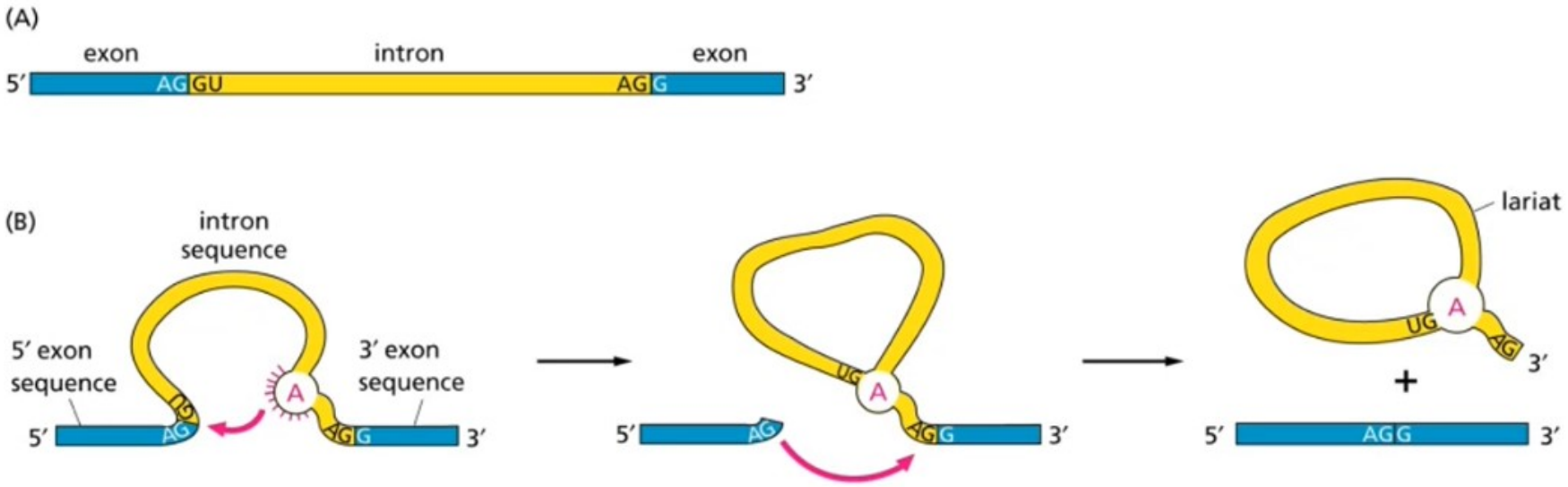
-35 Motif - TTGACA

Eukaryotic Transcription

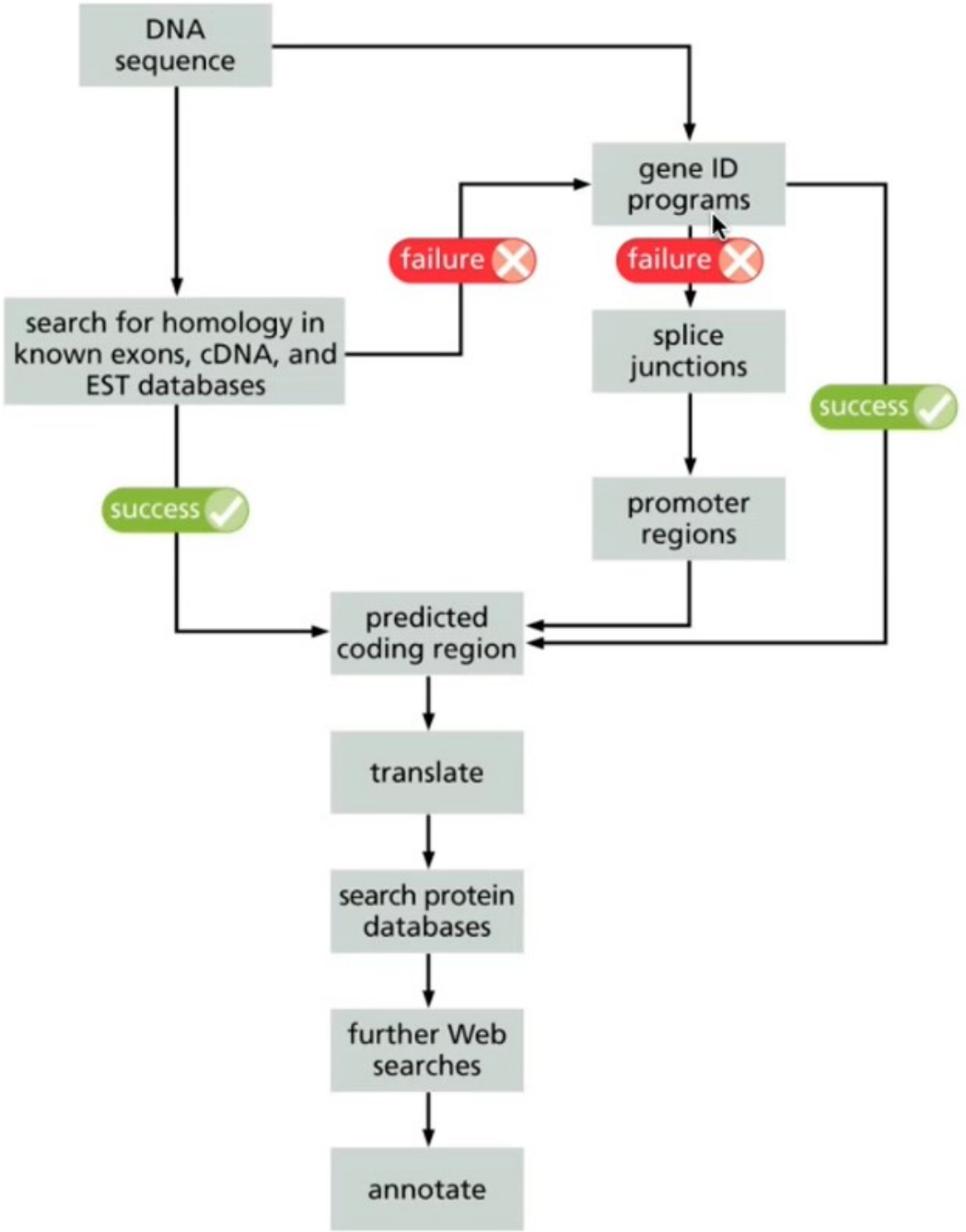



DNA
transcription
5′ upstream flanking region (promoter)
transcribed region
3′ downstream flanking region
start codon (M)
stop codon
mRNA
5′ UTR
3′ UTR
translation
N-terminus
C-terminus
protein

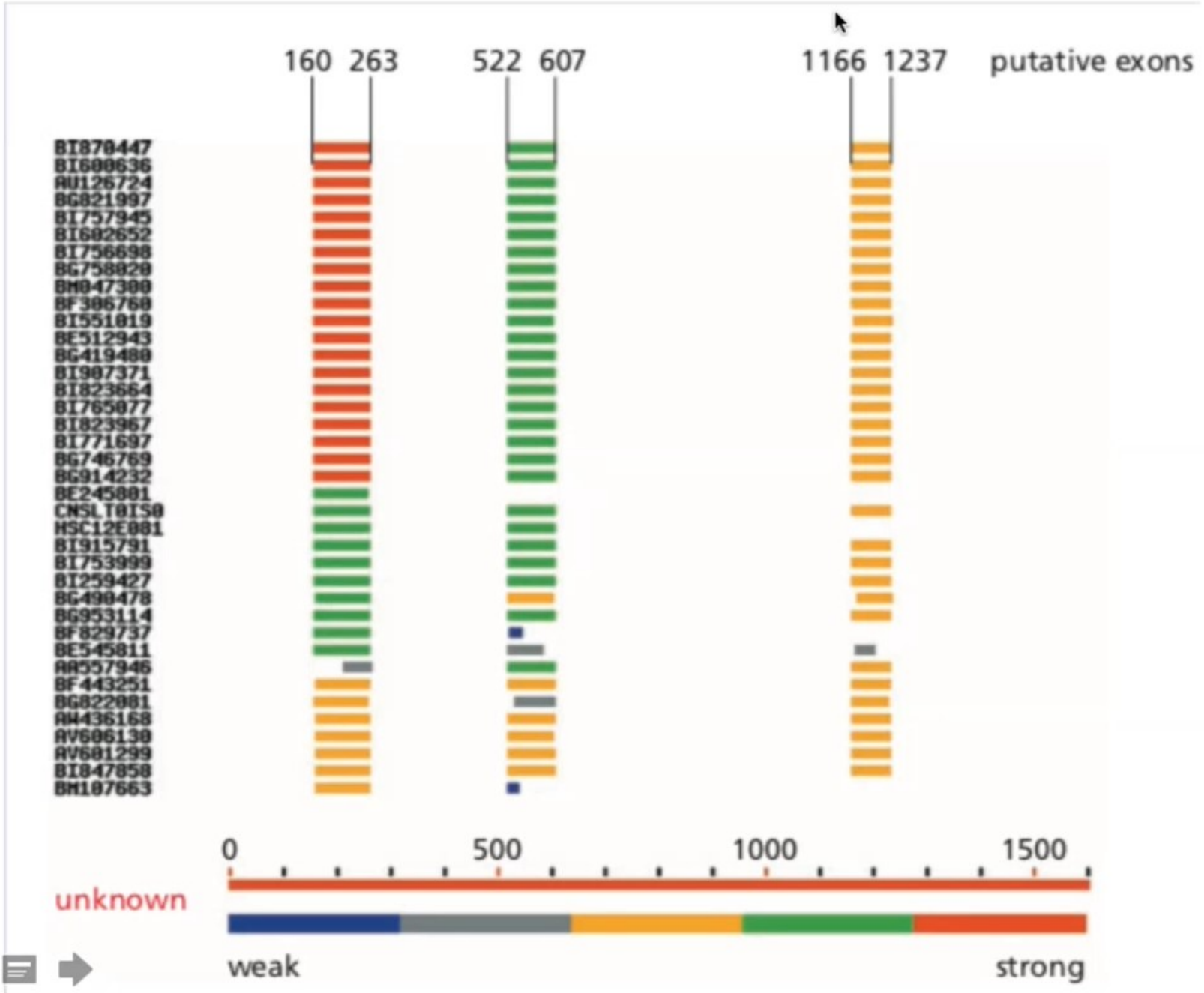
Modification



Steps Involved in the identification and annotation of gene sequences



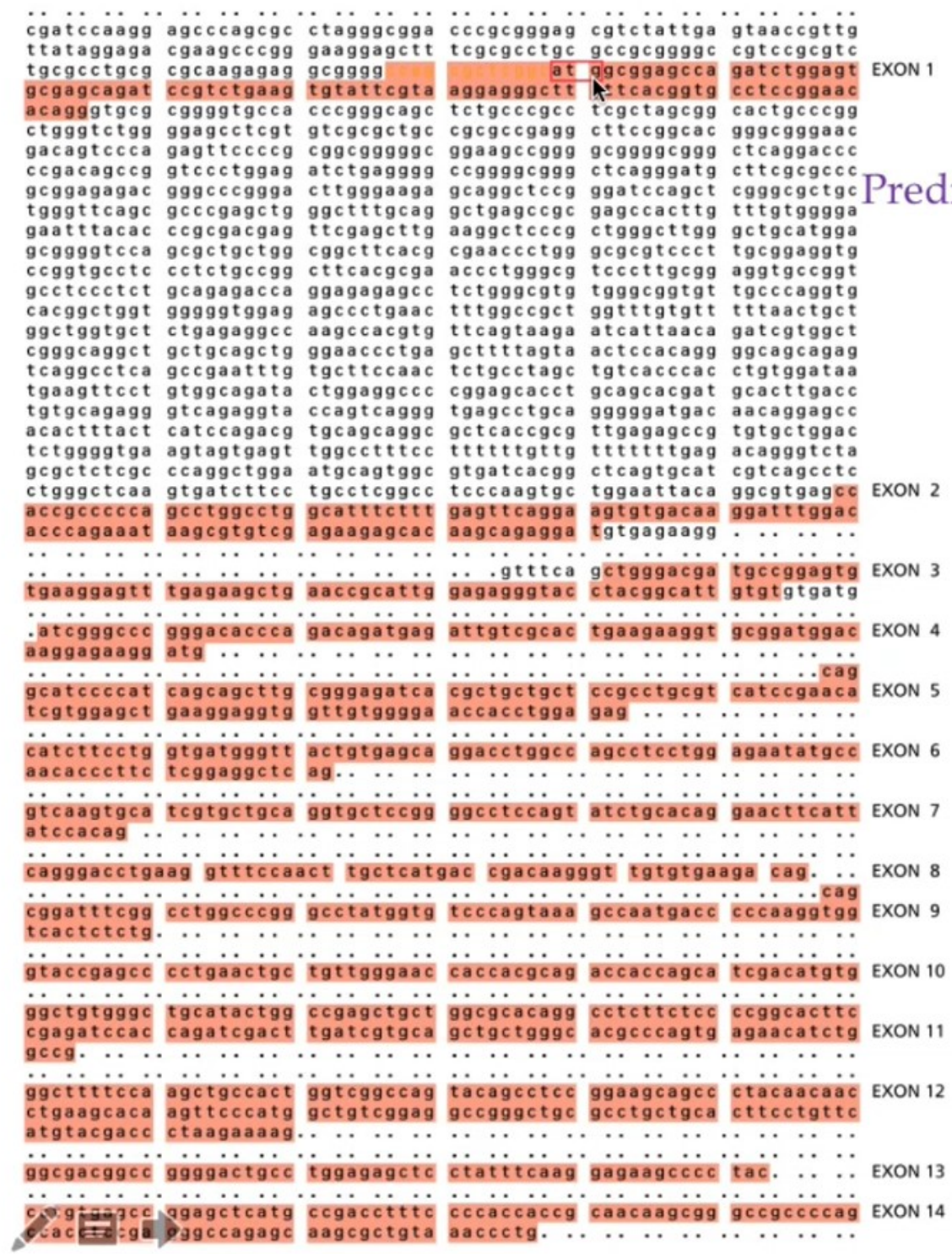
Similarity search against EST database



Typical Series of Steps in Eukaryotic Gene Prediction

1. Submit DNA sequences to exon prediction programs
2. Take average or consensus exon prediction
3. Translate predicted exon into protein sequences in all frames and directions
4. Take translation with least or no stop codons
5. Search protein database with the translated segment
6. If hit found use protein homolog to delineate exon(s)
7. Repeat for all other exons
8. Annotate and splice exon to obtain putative protein sequence

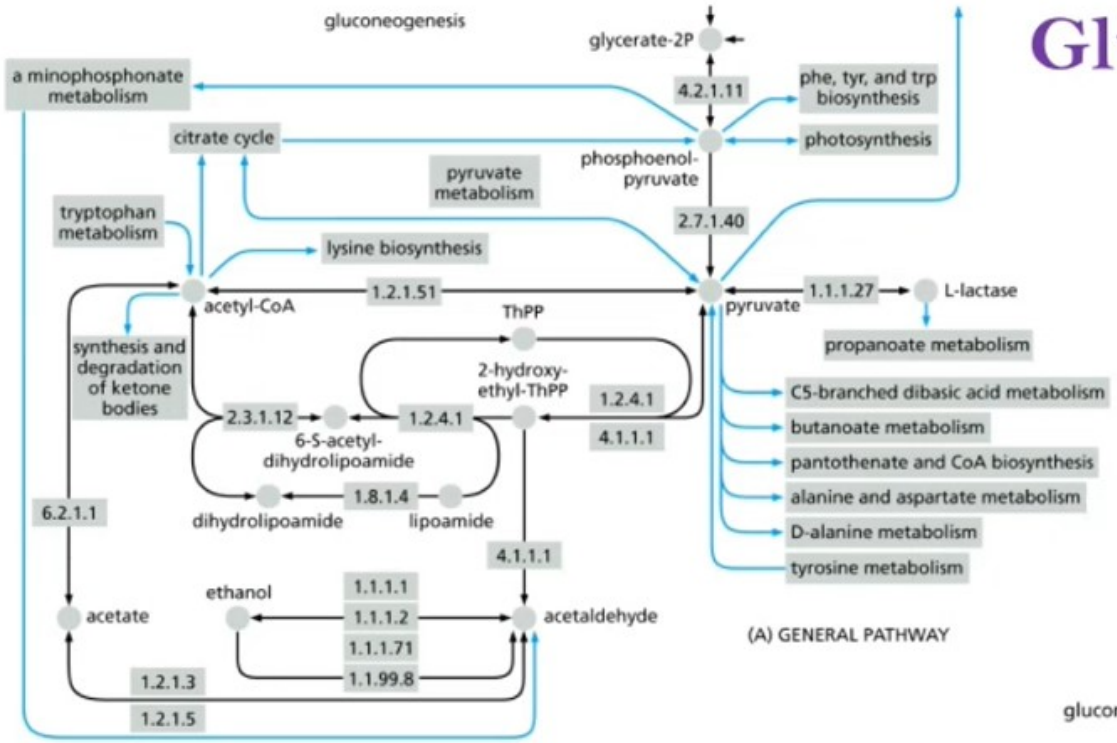
Gene Prediction



Prediction of start site is a nontrivial task

```
tttaaaaggg aggagcgggc tggaggggaa agagggagaa catggtcatt actgaatcca      60
cacattgcac aaatagaaaa aggaacaggc aggggaatag tcaattatgt atttgcctcc     120
tgctgtgtaa atcagcactt cagtaagata aggtgaggac agagcagcta cctgtgggga     180
catttaacct tttatctgta gctatctgct tagggacata gagaaaggca gnttcttgca     240
tgactcagct ttttgcttaa ttttttcctt ttggcatatg aattgagctc ccacgngntt     300
ttggttggtt ttgggcataa gtggagagtt caattggggc cagggccccg agttattttc     360
ttttcacaaa tatacccttt agagttcaga ggaaaggctg ggattagagc cttctttgag     420
cactattcat ggattacatg agggagtngc tgtgaaaaga gggccccgga ctaagccctg     480
ggcacctgaa caagtgagag gtcaggaagg ggagaaaaat ccagcaaagg aaccccagag     540
gacccggcca gtagctaaaa agaaaagtgc ggaagagtgc agggaaatgg taaagaaaag     600
aaagctttcc acaagacagg accgatcagc tgagccaact gctgctgagt acagtgatga     660
gagctgaagt ctctcgcccg aaaggcaata aggcgggagg gggaaacctg agggcagcgg     720
gttctctctt gggagcggga acacggaacc atgggcagcg aggcaaatgc agcctgggga     780
gaagctctga gtaacggacg tgcgaggatg atgcttgtgc aagaaggaaa tgaccgctgt     840
cccggctccg cgggagacga accccggctt cccgccctca gggactagct ctccagggaa     900
ctgggacggt cagtgccggt cgaggcagct cctcgctgaa ggaaggagca ggggacaggg     960
aacggaatgg ggagcgctag cccccaacta cgtccatctg gccatgtttg aagagccana    1020
aaatggagga gggaacccct agagcgtgcc agacggagac tgctctccgt ngcagtcggg    1080
gcgcttccgg cagggcgccn actcccagcc gagcgccctc cgcctgctcc tccaggattc    1140
ctcttcgccc tttctggggc cgccccgggg cgctctcagn aggatggcca acaccttccc    1200
tccatcccta caccccgccg ccccctgccc gtggccgcgc tcggctcccg cactgctcac    1260
tccacccccc acatcccagc ccgctgccag agccggggag agggcggggg ccgcgtgggc    1320
gagaccgtga acagcggctg tcacgtgggc cgcccaggcc aataggggtg aggctttggg    1380
tccagctcag tcctcccccg gcgcctccga ctggcagtgg gactcagcgg gcgtggaggt    1440
cgcggctgag cgagcgagcc ctgggcgagt gaattgtggc tgtgggttga cggtggagac    1500
acccccccga gggaggcgga gggaagggag gcgaggcctg cacctgcatg cttcccgcct    1560
cccactcccc agcgcccccg gaccgtgcag ttctctgcag gaccaggccA TGGAGCTCGA    1620
AGTCCGGCGG GTCCGACAGG CGTTCCTGTC CGGCCGGTCG CGACCTCTGC GGTTTCGGCT    1680
GCAGCAGCTG GAGGCCCTGC GGAGGATGGT GCAGGAGCGC GAGAAGGATA TCCTGACGGC    1740
CATCGCCGCC GACCTGTGCA AGgtancacg cgtgcggcgg ggtgtgggga aactggcccc    1800
cgccgngcac ttgtggactg gagtcttcgg ctgggttttg tttttgcttt tacatttngg    1860
attactccac cactgggagt atgatctcca gcgatacaga taaagccaaa gttcccgcag    1920
actttccagg tcctctagca ctcagaaggg catatgttac ctagcttctg tggttccttt    1980
tctgtatatt agagaattag caagccctta ccagggcgtg aagggtgcaa aaggagtctg    2040
aatggcaaac agctagtctg ataatgccag ttgttgtcac tacaggtgta cctggtnnnt    2100
gttctgacat tnagggccaa gtgtatcata cttacnctgn aagnttaact gtgattctct    2160
tataacagAG TGAATTCAAT GTGTACAGTC AGGAAGTCAT TACTGTCCTT GGGGAAATTG    2220
ATTTTATGCT TGAGAATCTT CCTGAATGGG TTACTGCTAA ACCAGTTAAG AAGAACGTGC    2280
TCACCATGCT GGATGAGGCC TATATTCAGC CACAGCCTCT GGGAGTGGTG CTGATAATCG    2340
GAGCTTGGAA TTACCCCTTC GTTCTCACCA TTCAGCCACT GATAGGAGCC ATCGCTGCAG    2400
gtctggttgc caccttatgt ctatatacct ttttagggag gcttattttc tcatattaat    2460
tggnattaag gatagtggct aattaaatac atttacttgg tgatttgcct ttgtttacac    2520
caccagtgta ctggaattca tacatccata cata
```

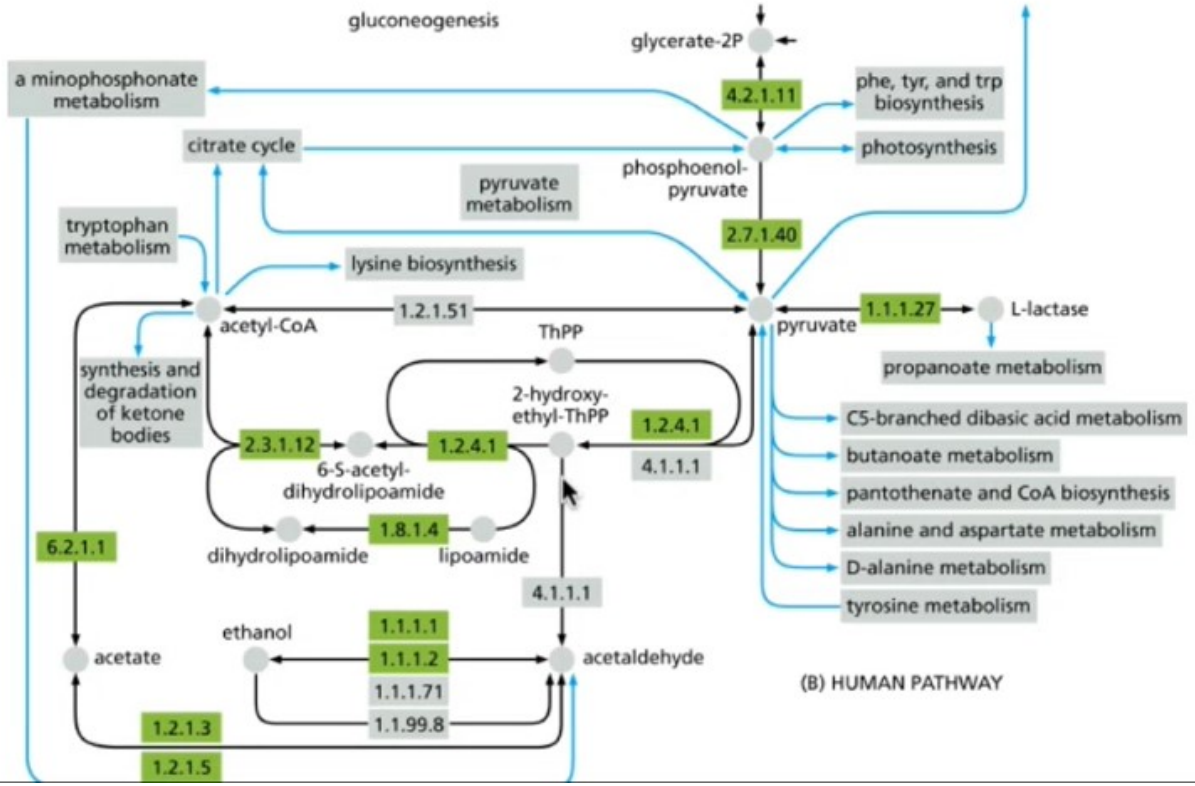
Gluconeogenesis Pathway



(A) GENERAL PATHWAY

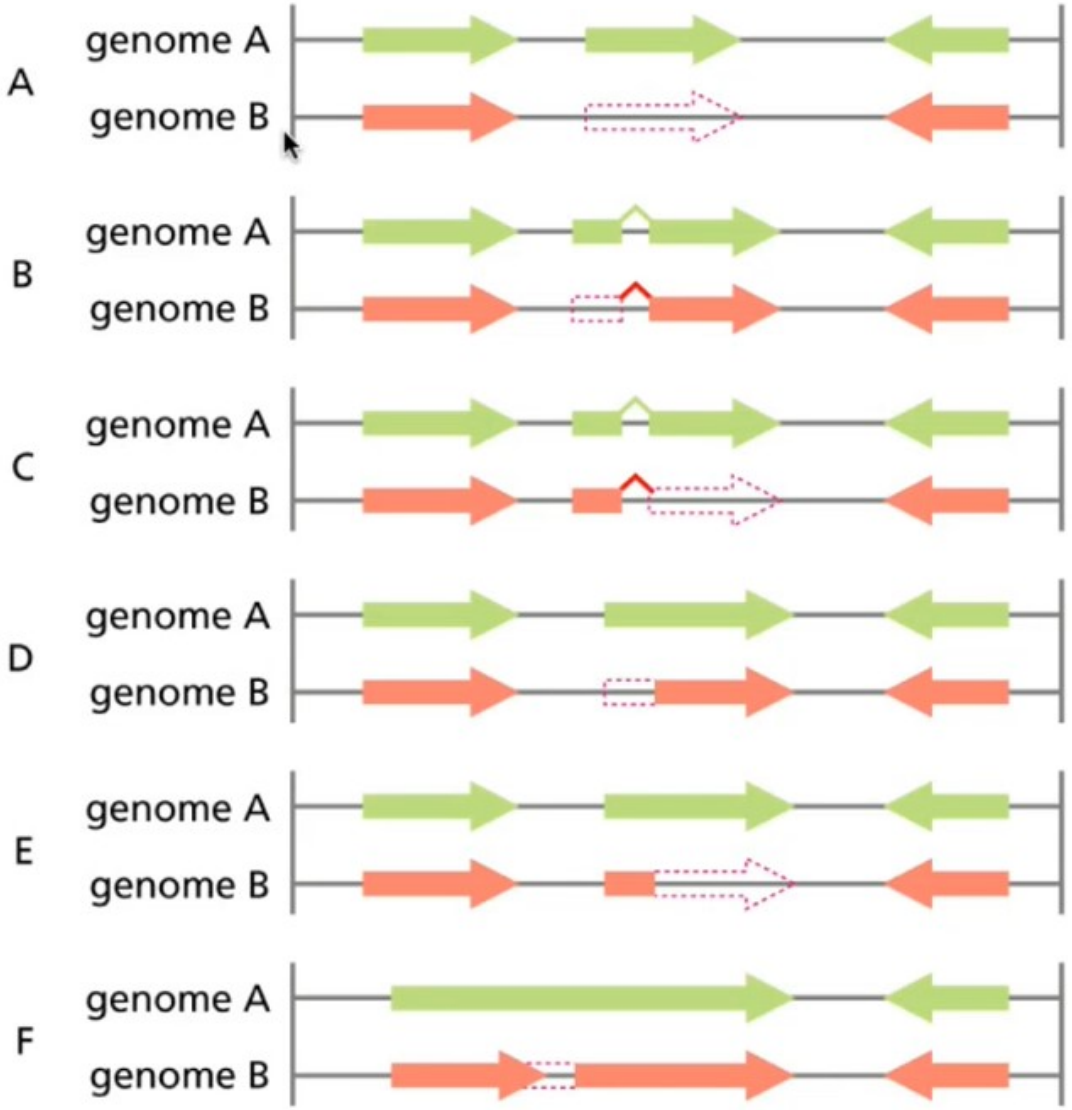
General Pathway

Human Pathway



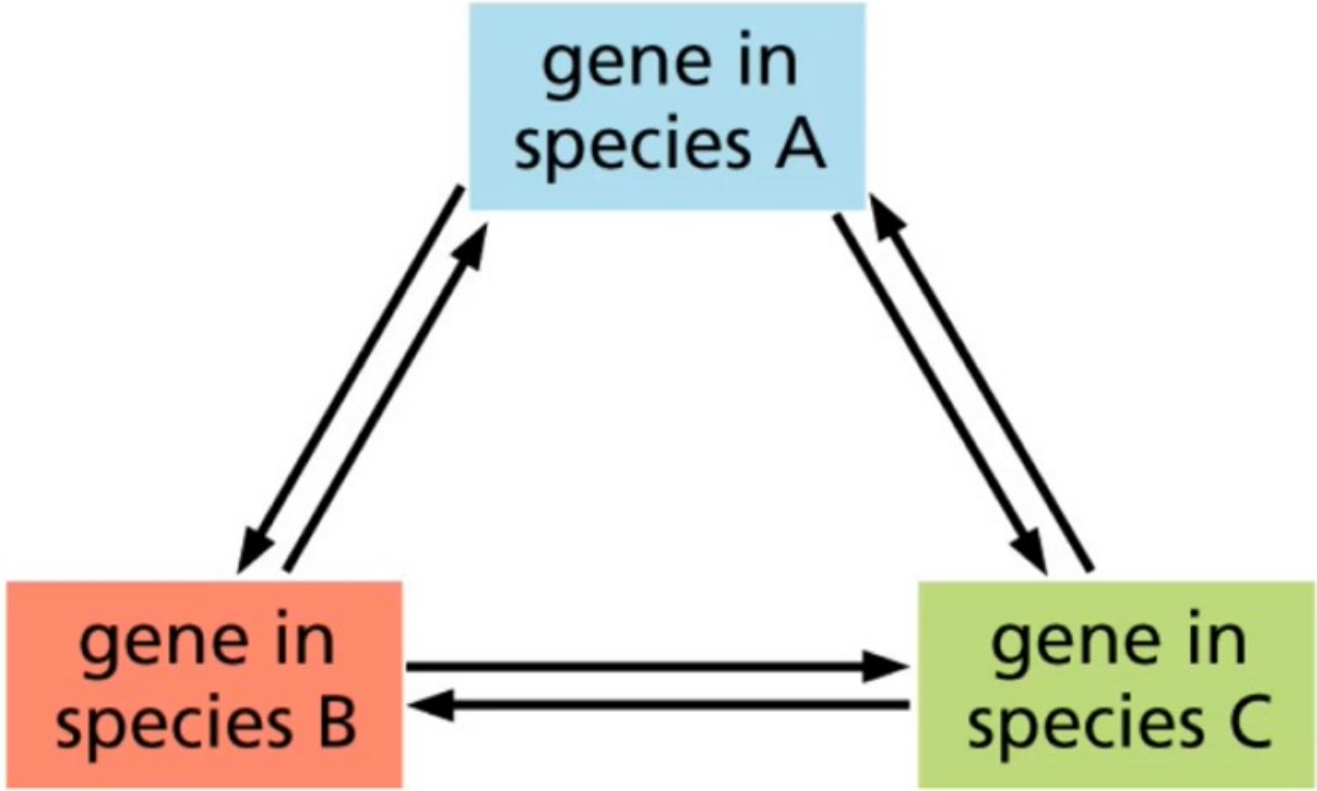
(B) HUMAN PATHWAY

Analysis of related genomic region helps to detect errors in annotation

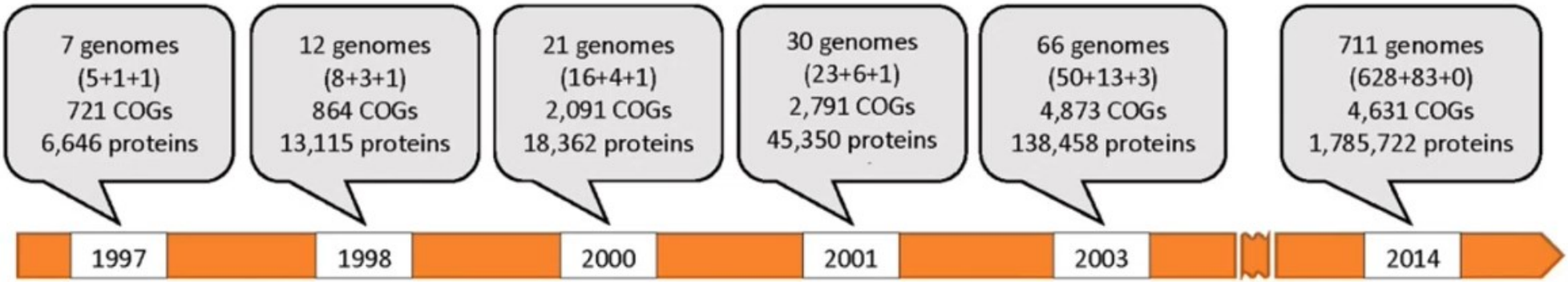


Comparative Genomics

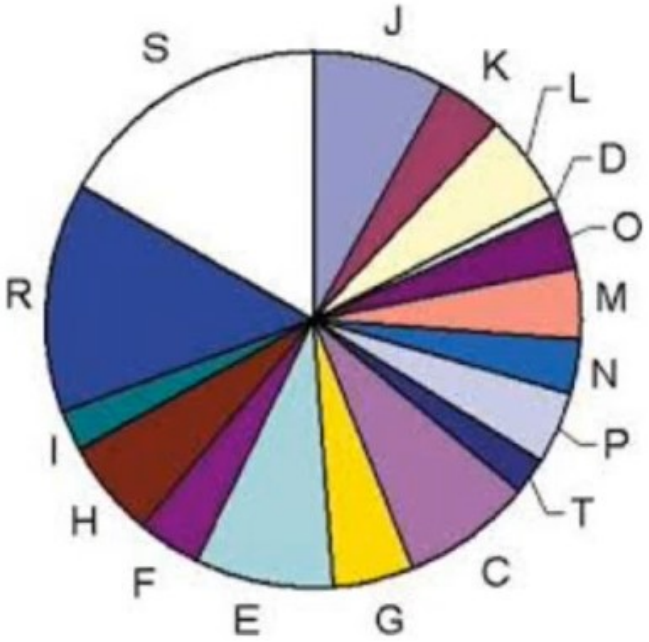
Cluster of orthologous Groups (COG)



Orthology is established by bidirectional Best-scoring BLAST Hit (BeT).



2020 = 1300 genomes + 4877 COGs

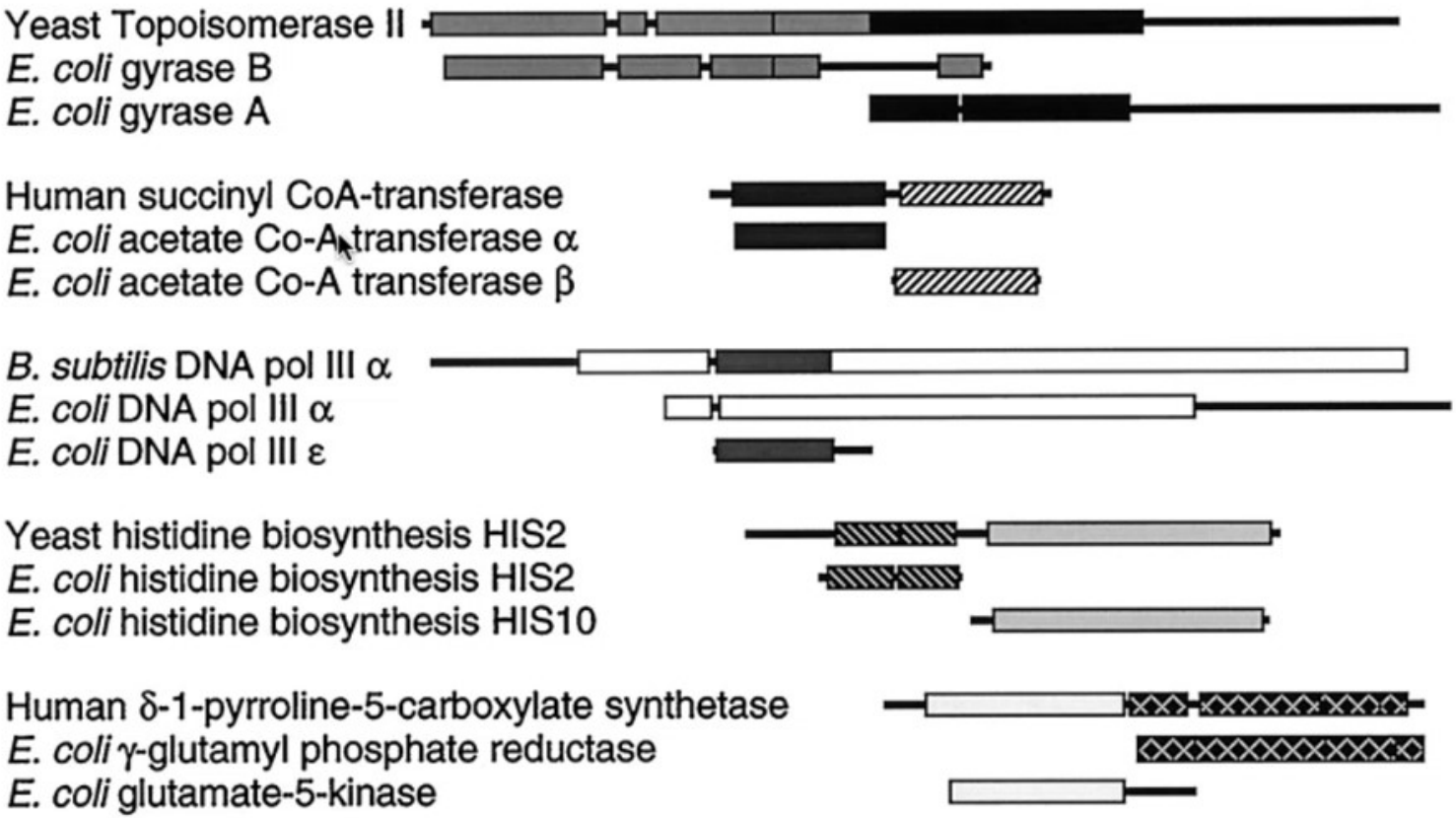


One-letter abbreviations for the functional categories:

J, translation, including ribosome structure and biogenesis;
L, replication, recombination and repair;
K, transcription;
O, molecular chaperones and related functions;
M, cell wall structure and biogenesis and outer membrane;
N, secretion, motility and chemotaxis;
T, signal transduction;
P, inorganic ion transport and metabolism;
C, energy production and conversion;
G, carbohydrate metabolism and transport;
E, amino acid metabolism and transport;
F, nucleotide metabolism and transport;
H, coenzyme metabolism;
I, lipid metabolism;
D, cell division and chromosome partitioning;
R, general functional prediction only;
S, no functional prediction

Genomic Context Based Methods

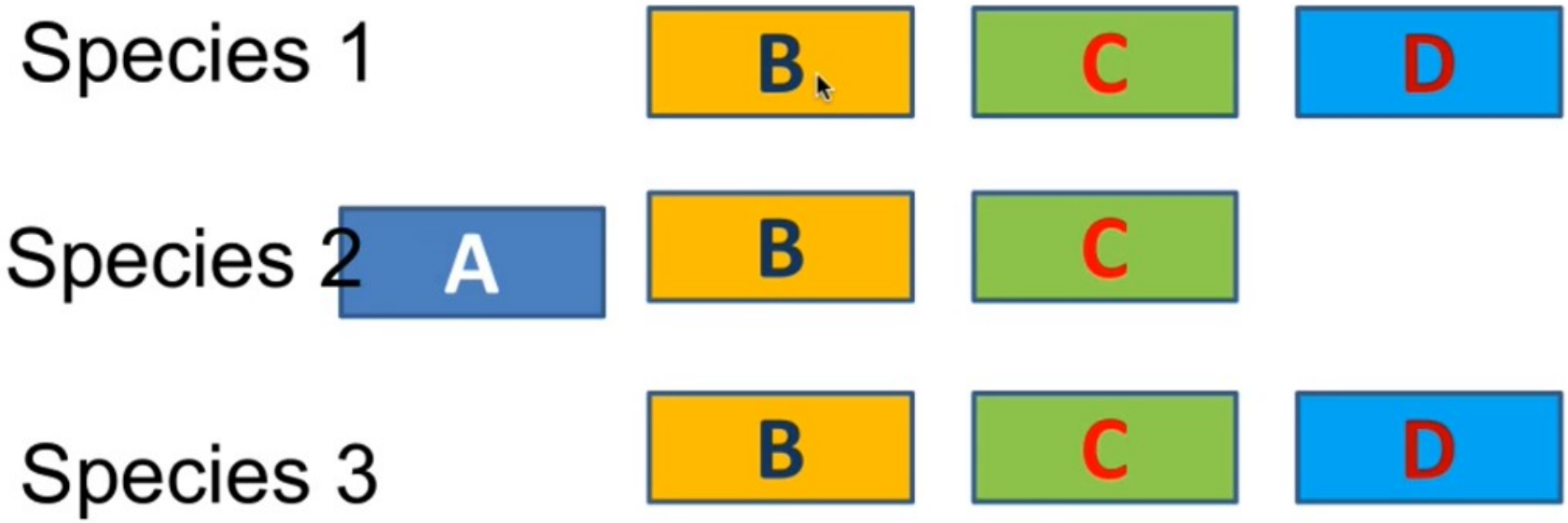
Gene Fusion/Fission



Inference: The observation that two independent genes in one organism are fused into a single gene in another organism suggests a direct physical interaction or a functional link between the genes.

Marcotte et al, 1999

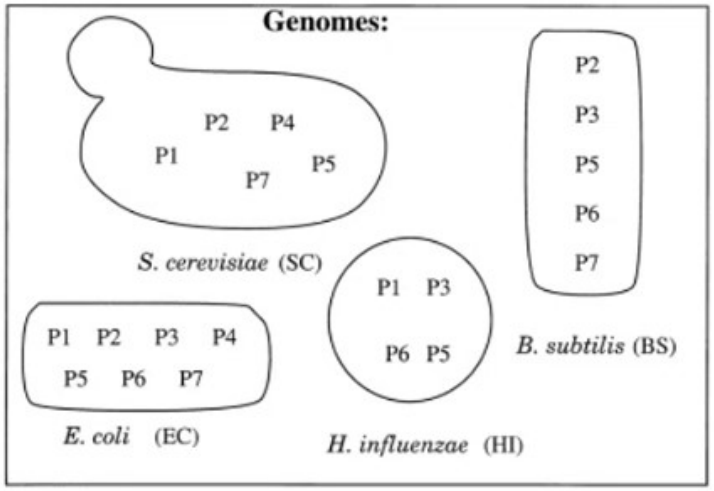
Co-occurrence of Genes



Inference: The observation that Gene B and Gene C have conserved their gene order in more than one organism suggests a selective pressure to regulate them together. Hence one can infer that the two genes must be functionally linked.

Overbeek et al, 1999

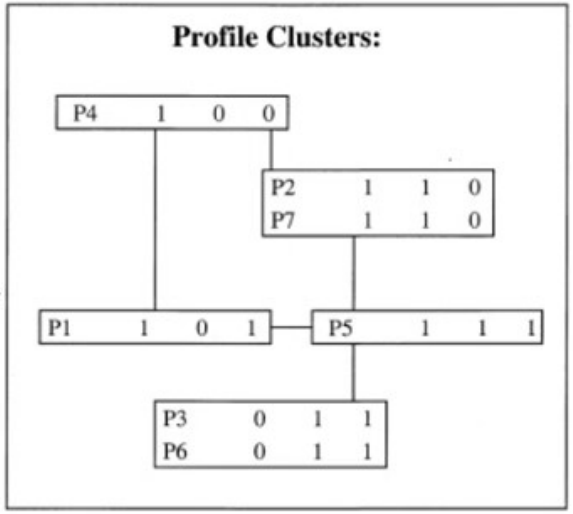
Phylogenetic Profiles



Inference: Since Protein 2 and Protein 7cluster together, suggesting coevolution, it may be inferred that they are functionally linked.

Phylogenetic Profile:

	EC	SC	BS	HI
P1		1	0	1
P2		1	1	0
P3		0	1	1
P4		1	0	0
P5		1	1	1
P6		0	1	1
P7		1	1	0



Conclusion: P2 and P7 are functionally linked.
P3 and P6 are functionally linked

Pellegrini *et al ., 1999*