SCIE2100 | BINF6000

Bioinformatics

Genome Analysis I

Atefeh Taherian Fard, PhD

Australian Institute for Bioengineering and Nanotechnology

a.taherianfard@uq.edu.au

Outline

Lecture 1:

- Overview genome sequencing and sequencing technologies
- Genome re-sequencing
- De-novo genome assembly

Lecture 2:

- Gene features in prokaryotes
- Gene features in eukaryotes
- Computational approaches for gene prediction
- Functional genome annotation

Why Do We Sequence Genomes?

Slides adapted from Lutz Krause

Why Do We Sequence Genomes?

Genome resequencing:

- Characterise genotype-phenotype associations
- Understand genetics of complex diseases
- Genome-based diagnosis; non-invasive prenatal testing
- Personalised medicine, genome-based prediction of optimal treatment (e.g. targetable mutations in cancer) and side-effects. Future: optimal diet, metabolic deficiencies, disease risk, …

De-novo sequencing

- Understand molecular biology of organisms, identify genes, gene functions, encoded pathways, metabolic capabilities, gene regulation and genome evolution

A Brief History of DNA Sequencing

1953 Watson and Crick publish structure of DNA double helix

1971 First DNA sequence determined (all 12 bp!)

1977 Sanger et al establish “Sanger” sequencing and sequence first ever genome (virus 5 Kb genome); state of the art until early 2000s

1990 The Human Genome Project (HGP) begins – large scale project to sequence human genome

1995 First genome of free living organism (bacteria *H. influenza*) by Craig Venter and Hamilton Smith

1997 First complete eukaryotic genome (yeast, 12 Mb)

1998 Sequencing of HGP begins; Fist animal genome (roundworm *C. elegans* 100 Mb); ~22 bacterial genomes

A Brief History of DNA Sequencing Continued

1999 First human chromosome sequenced (chr22)

2001 Draft human genome by HGP; Fruit fly (*Drosophila*)

2003 Completion of Human Genome Project

2006 Sequencing shake up! Massively parallel (next-generation) sequencing by 454 Life Sciences and Illumina

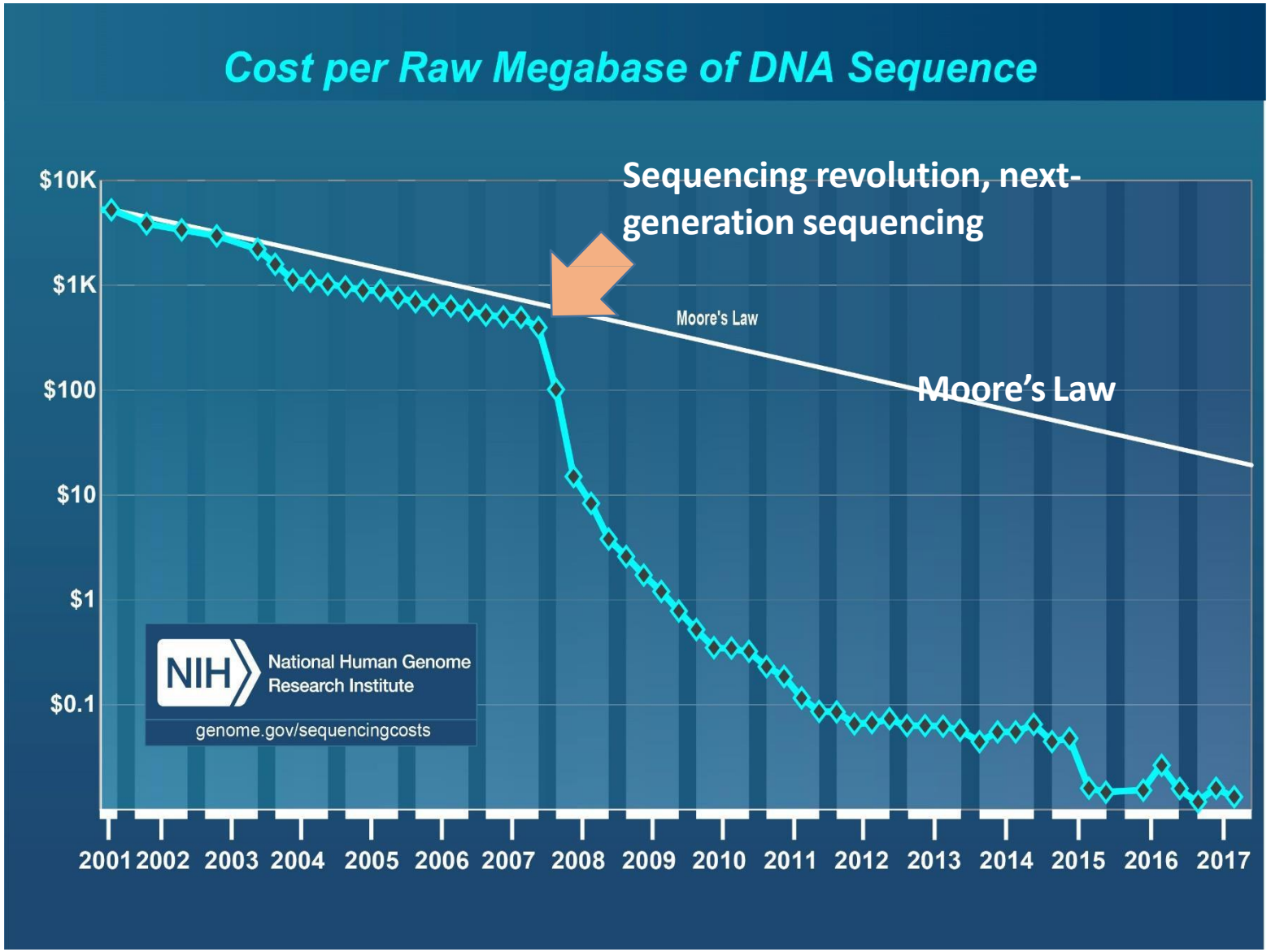
2008 First personal genome of James Watson

2009 Era of bioinformatics analysis and personalized medicine

2015 >200,000 human genomes sequenced

Now: High-throughput genome sequencing has initiated a new area in biomedicine and will (soon) transform clinical practice

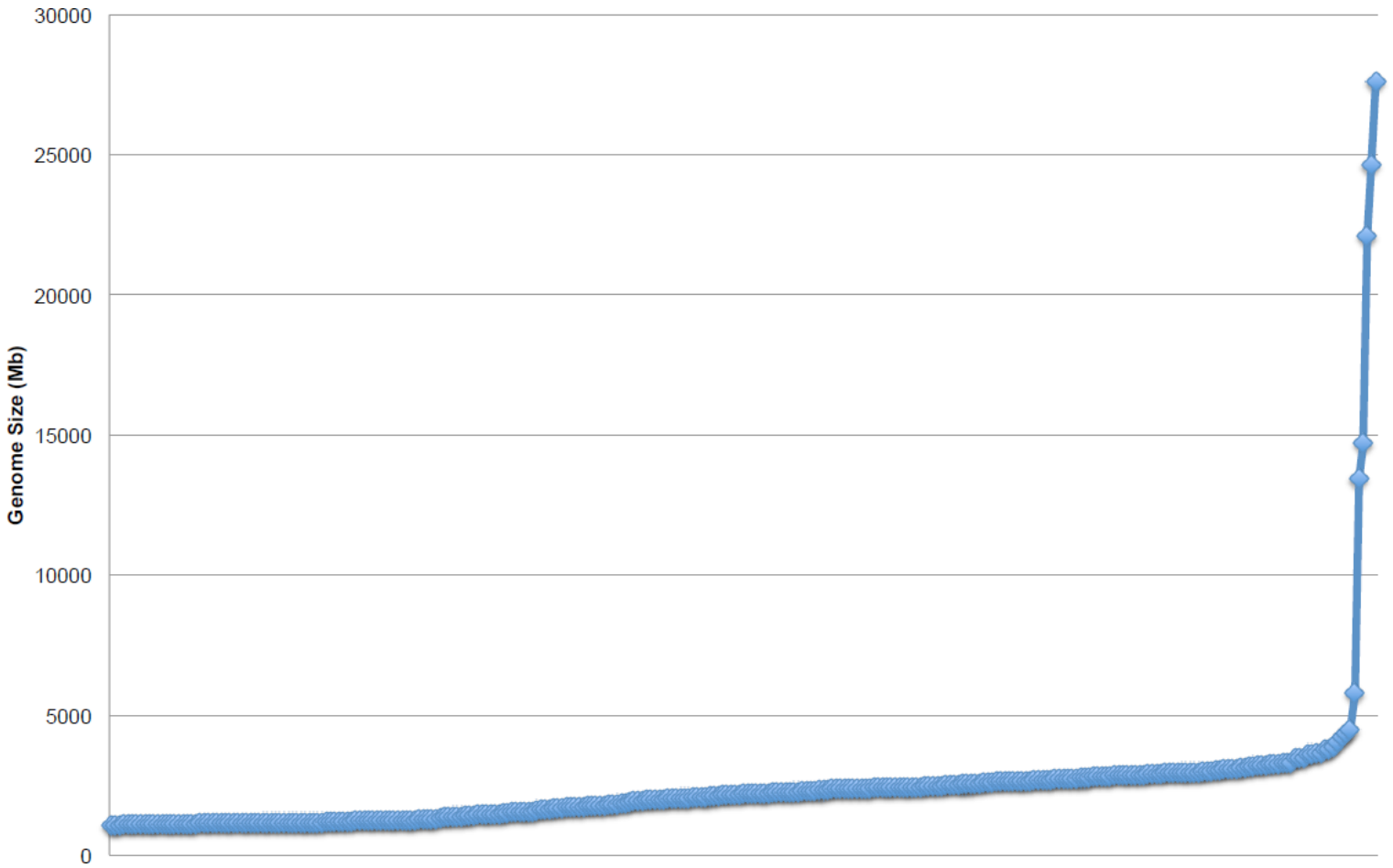
Cost of DNA Sequencing



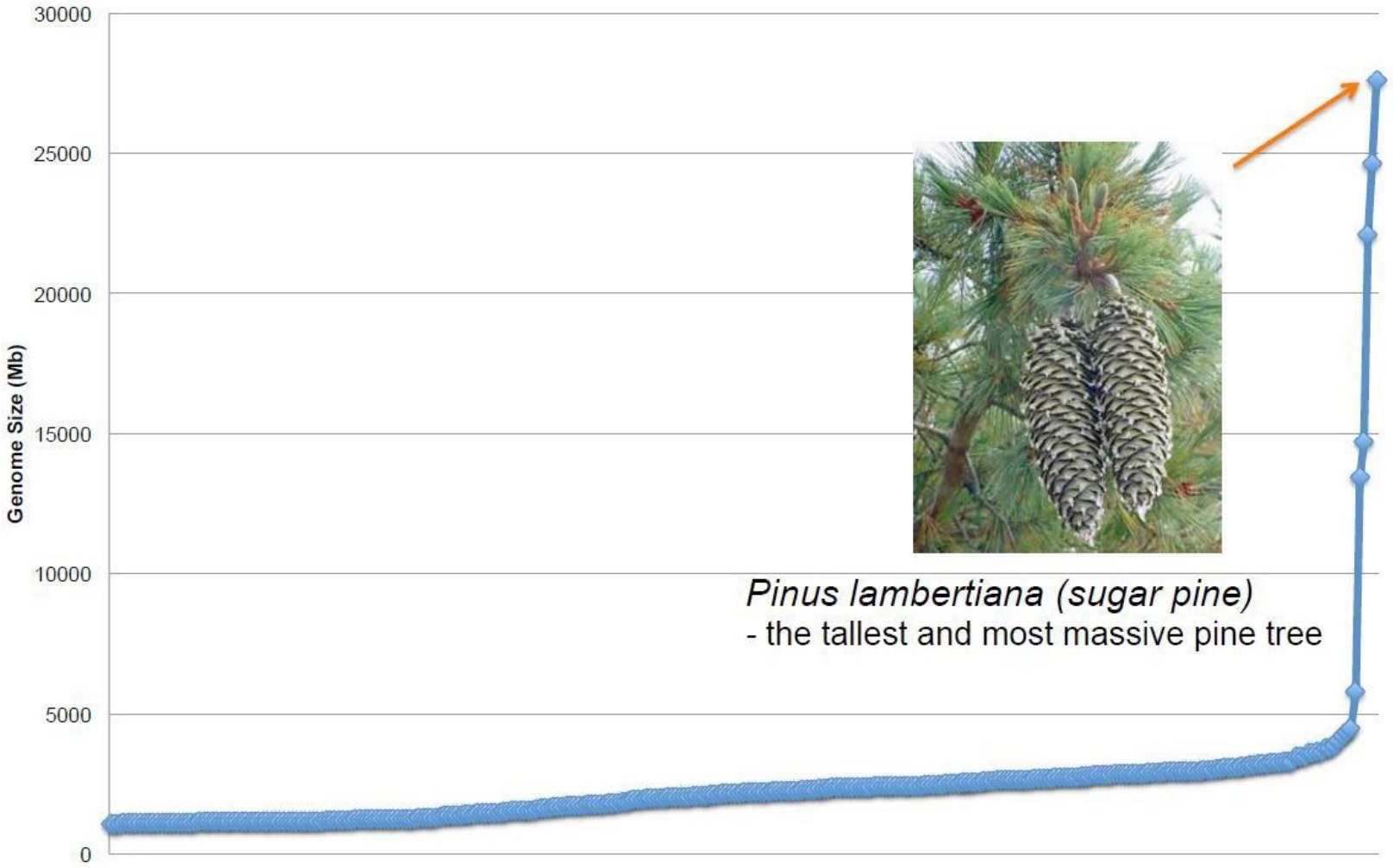
The problem:

- First human genome took 15 years and $2.7 billion
- Current costs: ~$1,000, soon ~$100?
- Moore's law: describes a long-term trend in the computer hardware industry
- 'Compute power' doubles every two years
- DNA sequencing outpaces Moore's law posing major challenges to bioinformatics

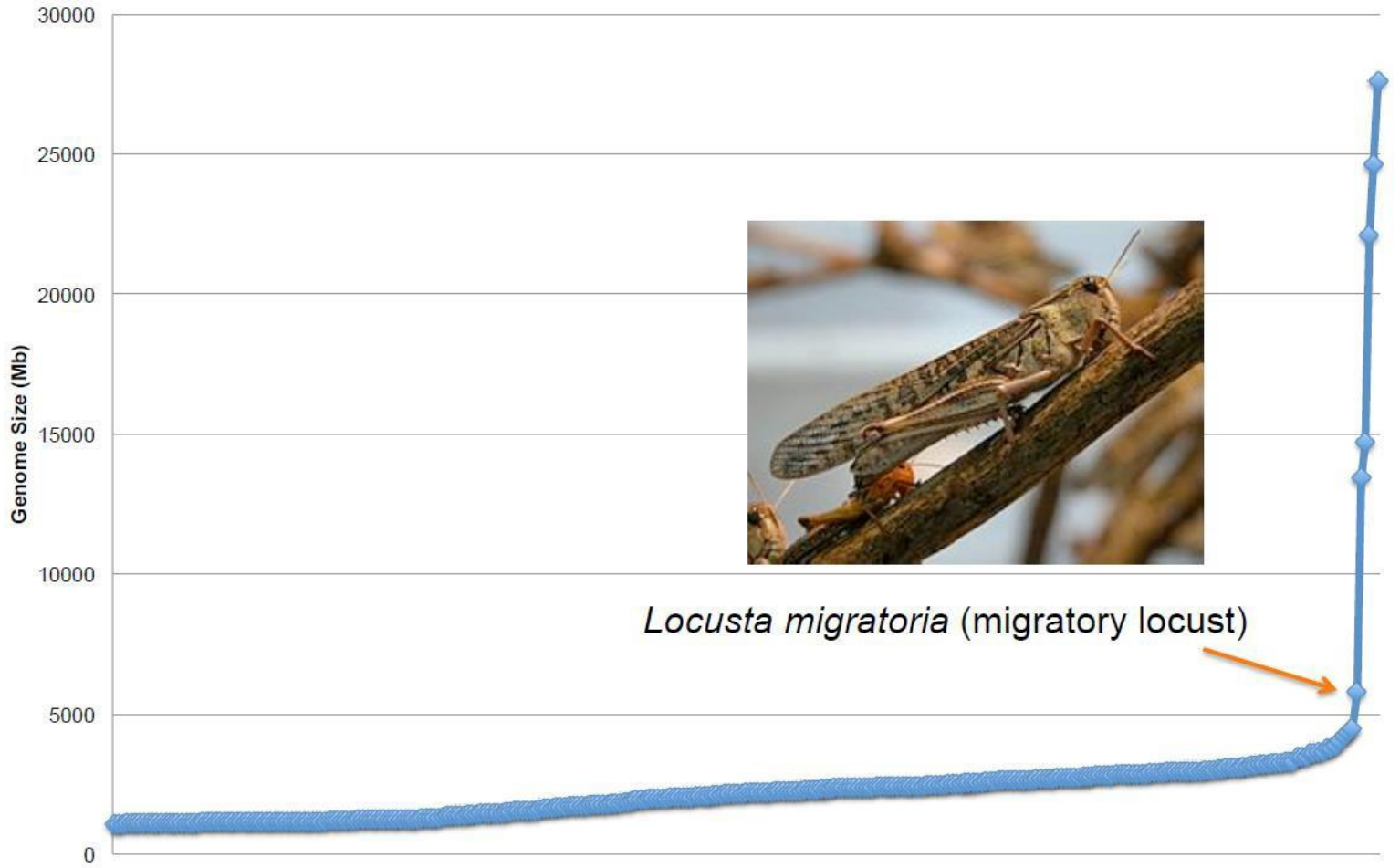
Top 200 Largest Sequenced Genomes



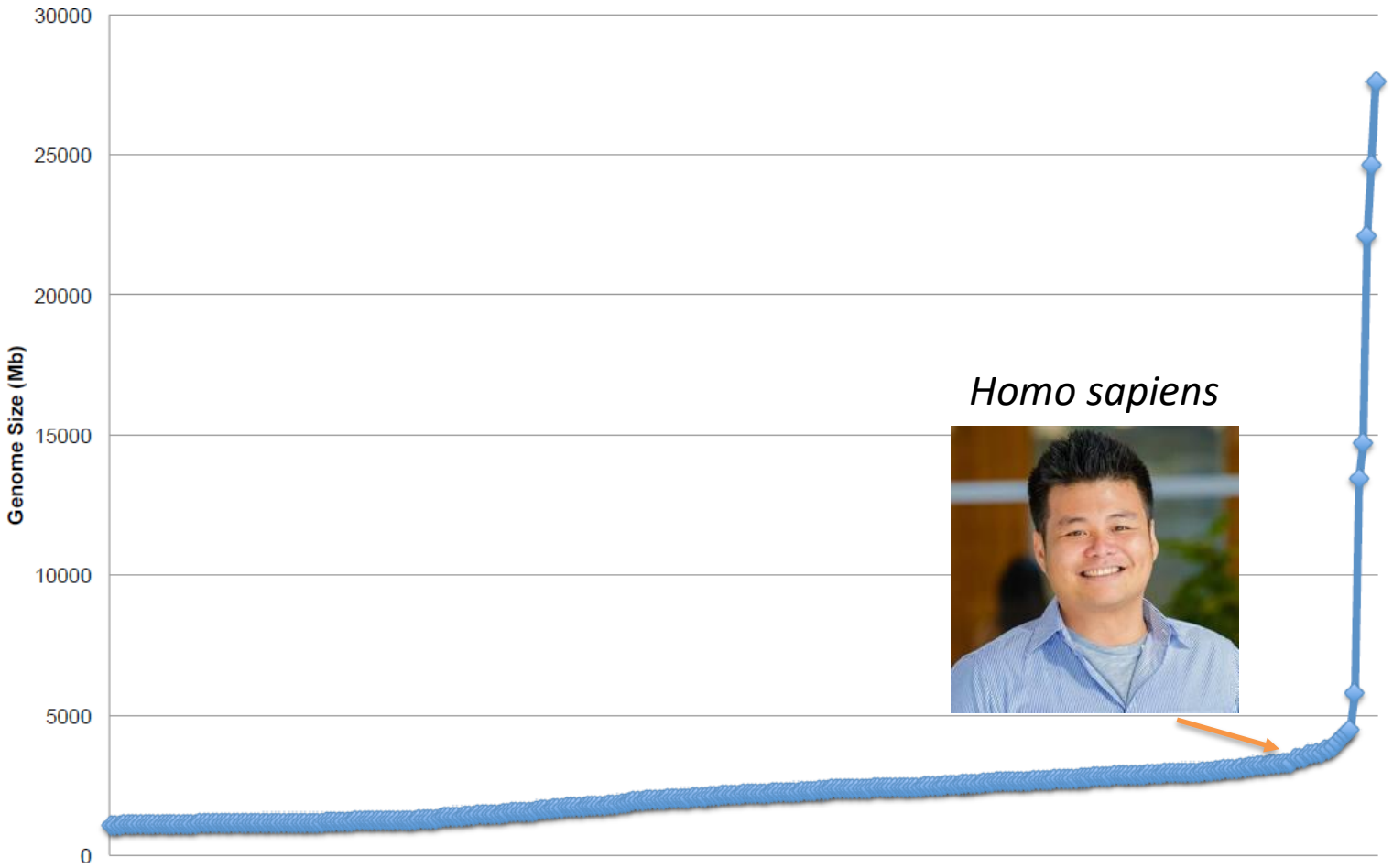
Top 200 Largest Sequenced Genomes



Top 200 Largest Sequenced Genomes



Top 200 Largest Sequenced Genomes



Overview Genome Sizes

Virus, Plasmid, Phage

- 1 kbp to 100 kbp … HIV 9181 bp

Bacteria, Archaea

- 1 Mbp to 14 Mbp … *E. coli* 4.6 Mbp

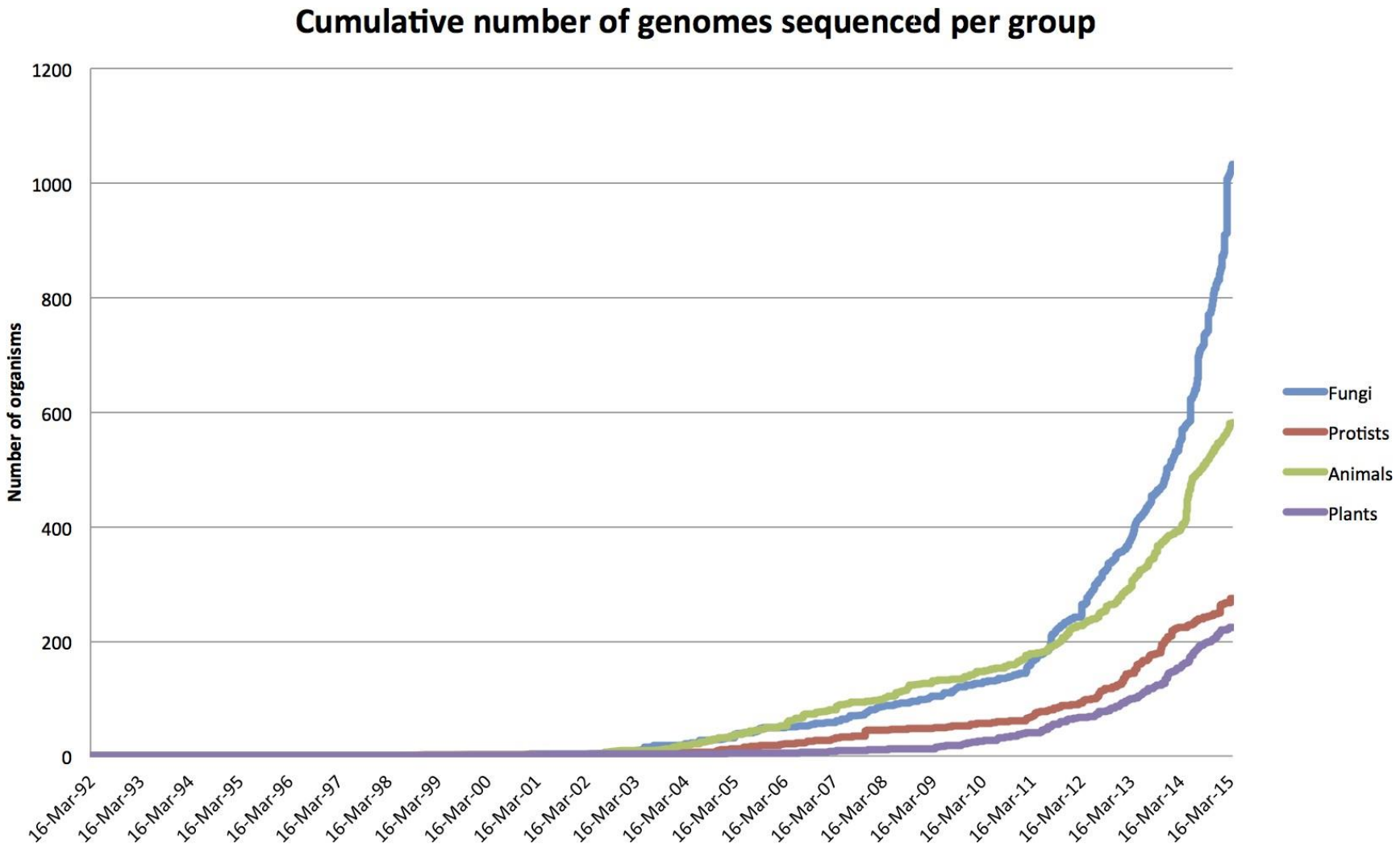
Simple Eukaryotes

- 10 Mbp to 100 Mbp … Malaria 23 Mbp

Animals, Plants

- 100 Mbp to >100 Gbp!
- Human 3.2 G

What Eukaryotic Genomes are Being Sequenced Now?



How do we sequence a genome?

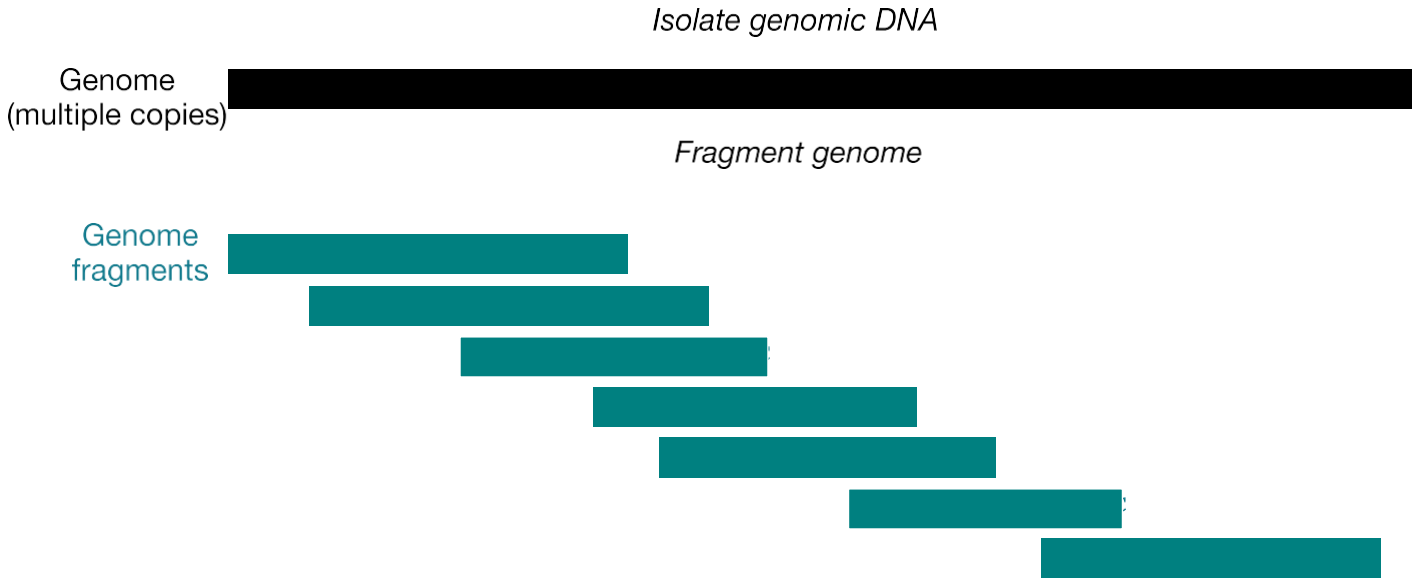
Whole Genome Shotgun (“WGS”)

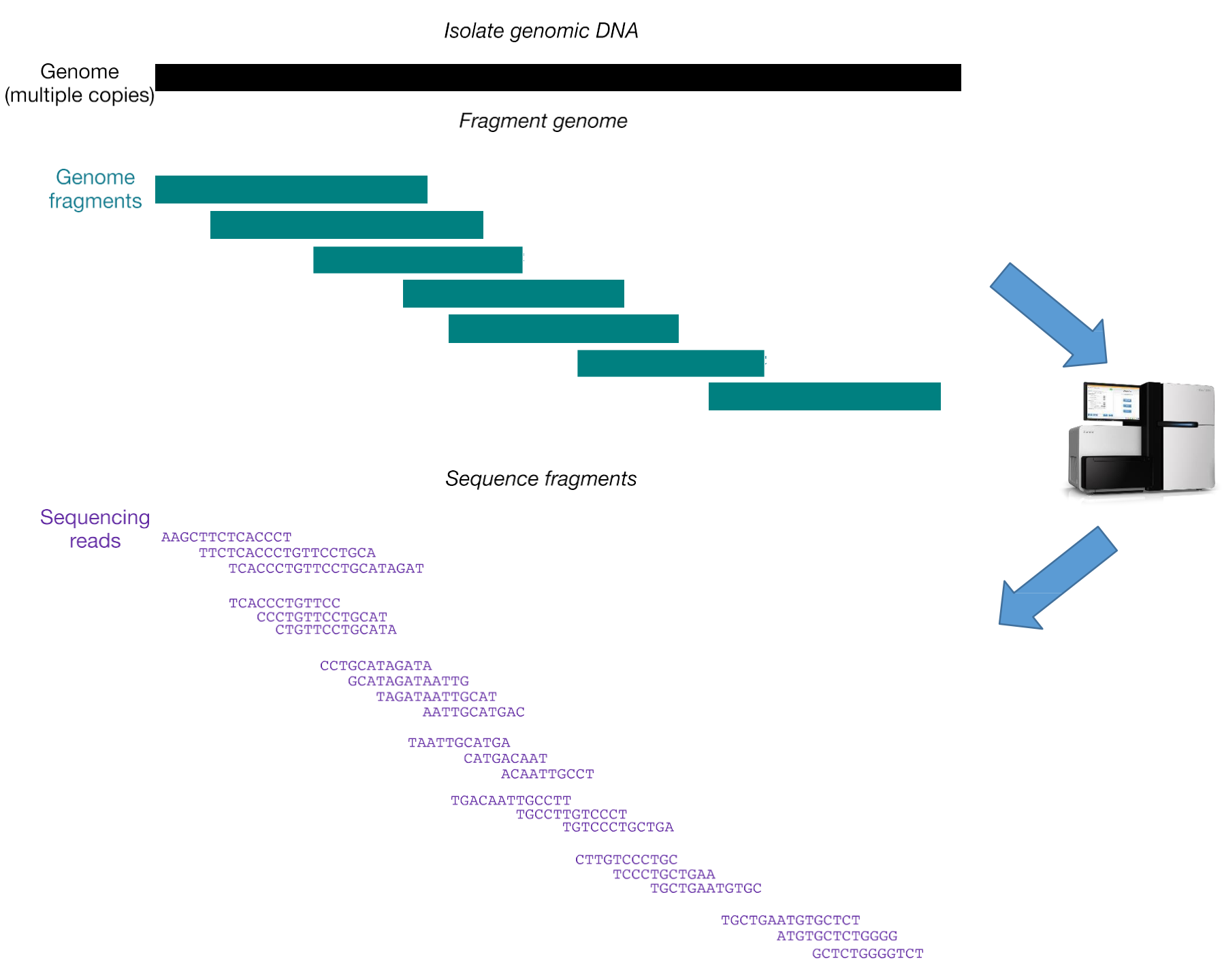
- Shear DNA to appropriate size
- Do some library preparation
- Put in sequencing machine
- Get some big text files with your reads
- Panic when NOTEPAD.EXE won’t load them

Adapted from Torsten Seemann

Isolate genomic DNA

Genome
(multiple copies)

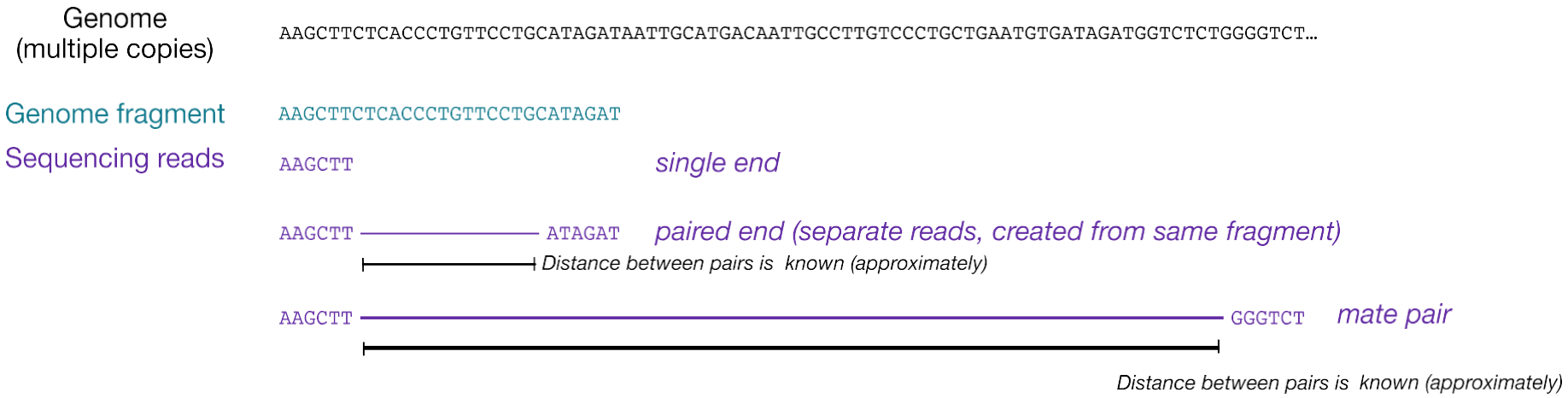

Isolate genomic DNA
Genome
(multiple copies)
Fragment genome
Genome
fragments



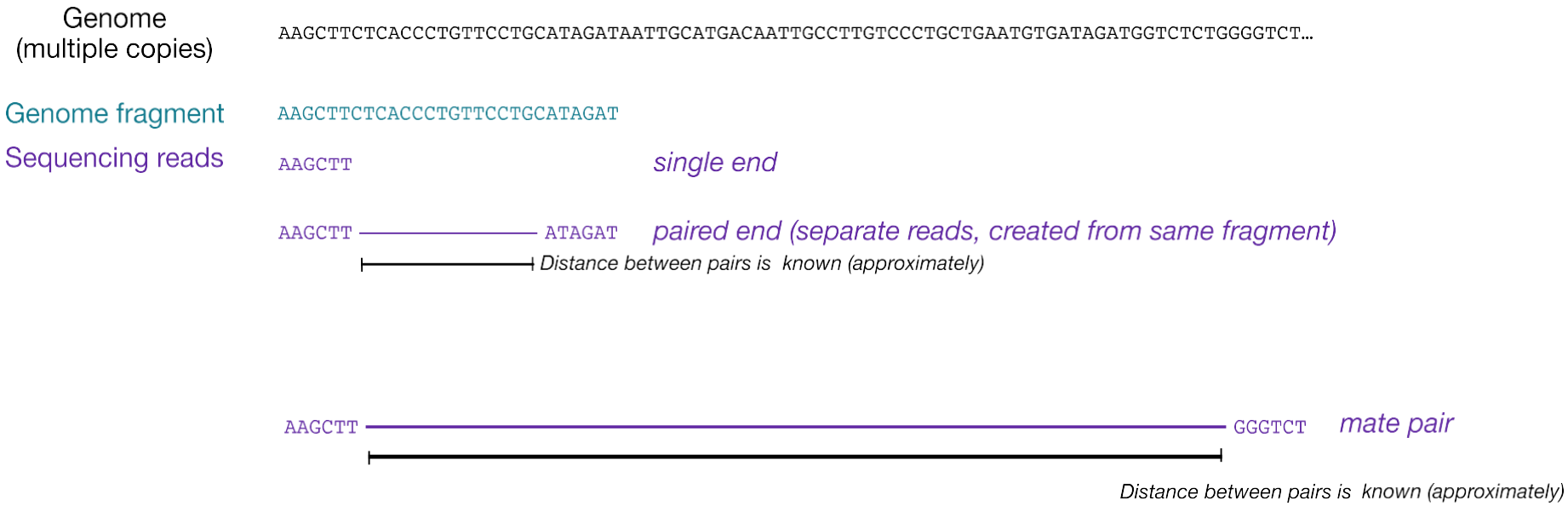
This approach is called ‘Shot Gun’ sequencing


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
ATAGAT
ATAGAT

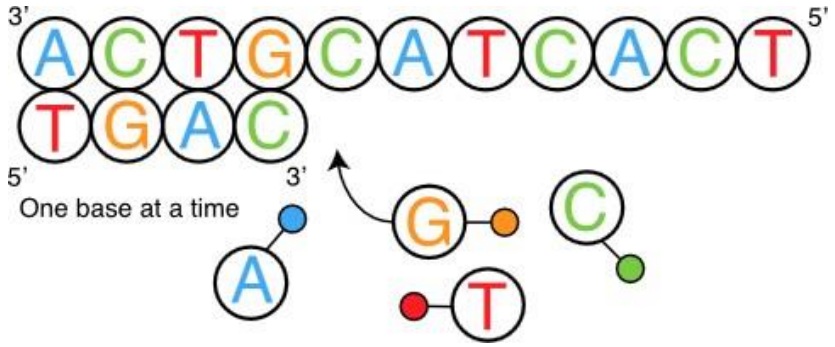

Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
ATAGAT
ATAGAT

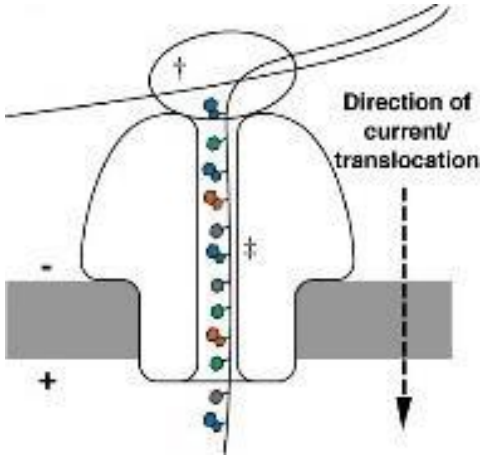

Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)

Sequencing Technologies

Sequencing by synthesis (e.g. Illumina):



Nanopore (e.g. MinION):



Heather et al, Genomics 2016

DNA Sequencing Technologies

Method	Read length	Accuracy	Cost per 1 million bases	Advantages
PacBio	~15 Kb	87%	$0.05–$0.08	Long sequence reads
Ion Torrent	100 bp	99.60%	$1	Less expensive equipment. Fast.
Sequencing by synthesis (Illumina)	150 bp	99.90%	$0.05 to $0.15	Large scale sequencing
Nanopore (MinION)	Varies, up to 500 kb	~92–97%	Varies	Longest reads. Portable, palm sized
Chain termination (Sanger)	1200 bp	99.90%	$2,400	State of the art until early 2000s

Run time: 20 min – 11 days
Human genome: 3.2 billion bases
E. coli genome: 4.6 million bases
Adapted from http://en.wikipedia.org/wiki/DNA_sequencing

Genome Re-Sequencing

- Reference genome sequence available, *representative* of organisms
- Sequence genomes of individuals to characterise their individual genotype

Applications:

- Understand genome evolution (e.g. bacteria)
- Characterize genotype-phenotype associations (e.g. malaria drug response)
- Understand genetics of complex diseases
- Genome-based diagnosis; non-invasive prenatal testing
- Personalized medicine, genome-based prediction of optimal treatment (e.g. targetable mutations in cancer) and side-effects. Future: optimal diet, metabolic deficiencies, disease risk, …

Genome Re-Sequencing

Reference genome sequence AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGCTCTGGGGTCTCTGGGGTCT…

Target genome

DNA fragments

Sequencing reads

TGTCCCTGCTGA
CTTGTCCCTGC
TGCCTTGTCCCT
CCTGCATAGATA
TCACCCTGTTCCTGCATAGAT
GCTCTGGGGTCT
TAGATAATTGCAT
CATGACAAT
AATTGCATGAC
CCCTGTTCCTGCAT
TGCTGAATGTGCTCT
TAATTGCATGA
TGCTGAATGTGC
TTCTCACCCTGTTCCTGCA
ACAATTGCCT
GCATAGATAATTG
CTGTTCCTGCATA
TGACAATTGCCTT
TCACCCTGTTCC
AAGCTTCTCACCCT
TCCCTGCTGAA
ATGTGCTCTGGGG

AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGCTCTGGGGTCTCTGGGGTCT…

Sequencing reads

TGTCCCTGCTGA

CTTGTCCCTGC

TGCCTTGTCCCT

CCTGCATAGATA

TCACCCTGTTCCTGCATAGAT

GCTCTGGGGTCT

TAGATAATTGCAT

CATGACAAT

AATTGCATGAC

CCCTGTTCCTGCAT

TGCTGAATGTGCTCT

TAATTGCATGA

TGCTGAATGTGC

TTCTCACCCTGTTCCTGCA

ACAATTGCCT

CTGTTCCTGCATA

GCATAGATAATTG

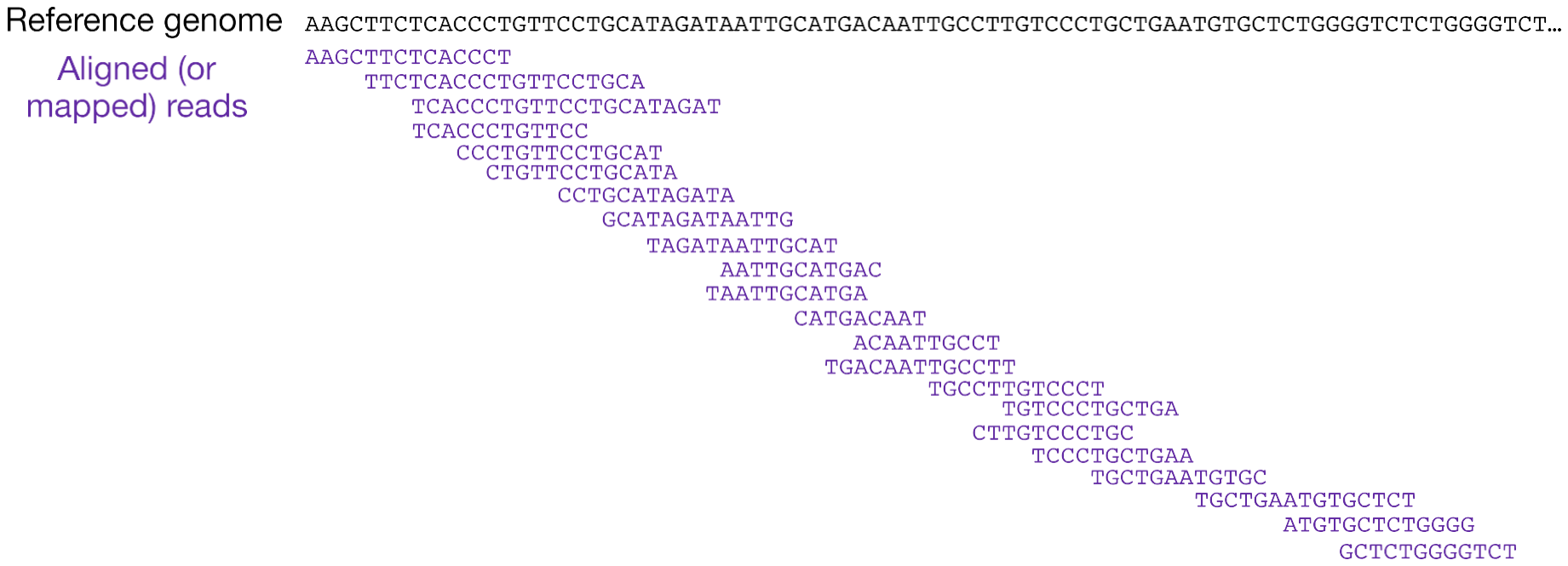
TGACAATTGCCTT

TCACCCTGTTCC

AAGCTTCTCACCCT

TCCCTGCTGAA

ATGTGCTCTGGGG


Reference genome
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGCTCTGGGGTCTCTGGGGTCT…
Aligned (or mapped) reads
AAGCTTCTCACCCT
TTCTCACCCTGTTCCTGCA
TCACCCTGTTCCTGCATAGAT
TCACCCTGTTCC
CCCTGTTCCTGCAT
CTGTTCCTGCATA
CCTGCATAGATA
GCATAGATAATTG
TAGATAATTGCAT
AATTGCATGAC
TAATTGCATGA
CATGACAAT
ACAATTGCCT
TGACAATTGCCTT
TGCCTTGTCCCT
TGTCCCTGCTGA
CTTGTCCCTGC
TCCCTGCTGAA
TGCTGAATGTGC
TGCTGAATGTGCTCT
ATGTGCTCTGGGG
GCTCTGGGGTCT

Genome Re-Sequencing

Reference genome sequence AAGCT T CTCACCTACTGCTCGCTAGACTCGATAGCTCAGATCGCTA……

Sequencing reads

AAGCT C CTCACCT
AGCT C CTCACCTAC
CT C CTCACCTACTGCTC
T C CTCACCTACTGCTC
TCACCTACTGCTCGCTAGACTCG
ACTGCTCGCTAGACTCGATAGC
ACTCGATAGCTCAGATCGCTA

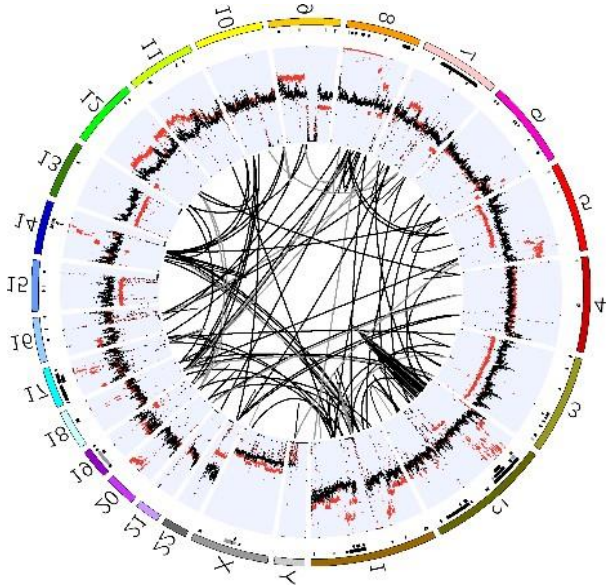
Single nucleotide substitution (SNP) in target genome

Example:
Identify point mutations in cancer genomes by comparing tumor genomes with genome sequence of normal tissue from same patient

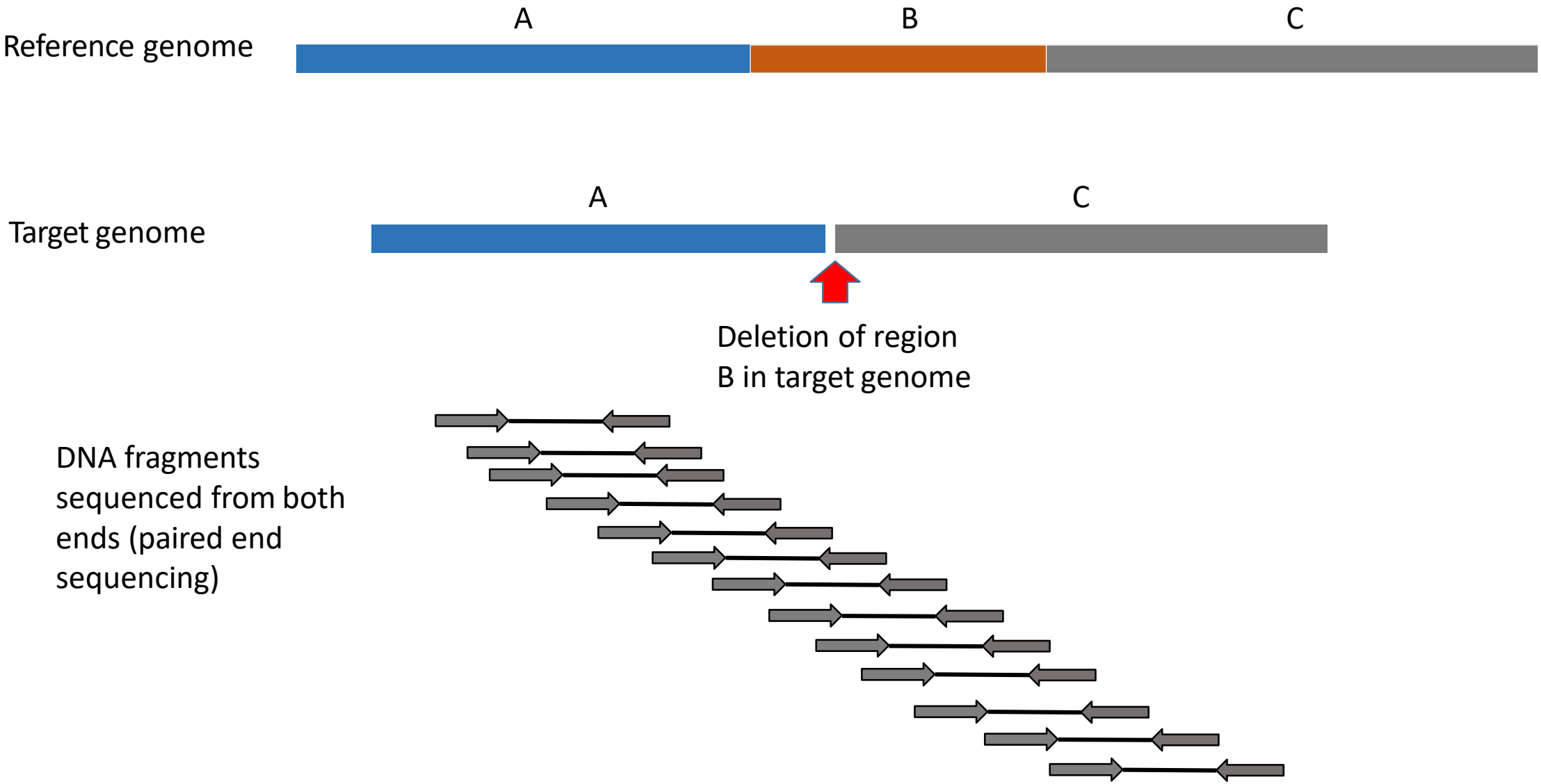
How Can We Identify Structural Variations?

- Structural variations are structural changes in the genome sequencing, including insertions, deletions or translocations
- Can be inferred from paired-end sequencing reads

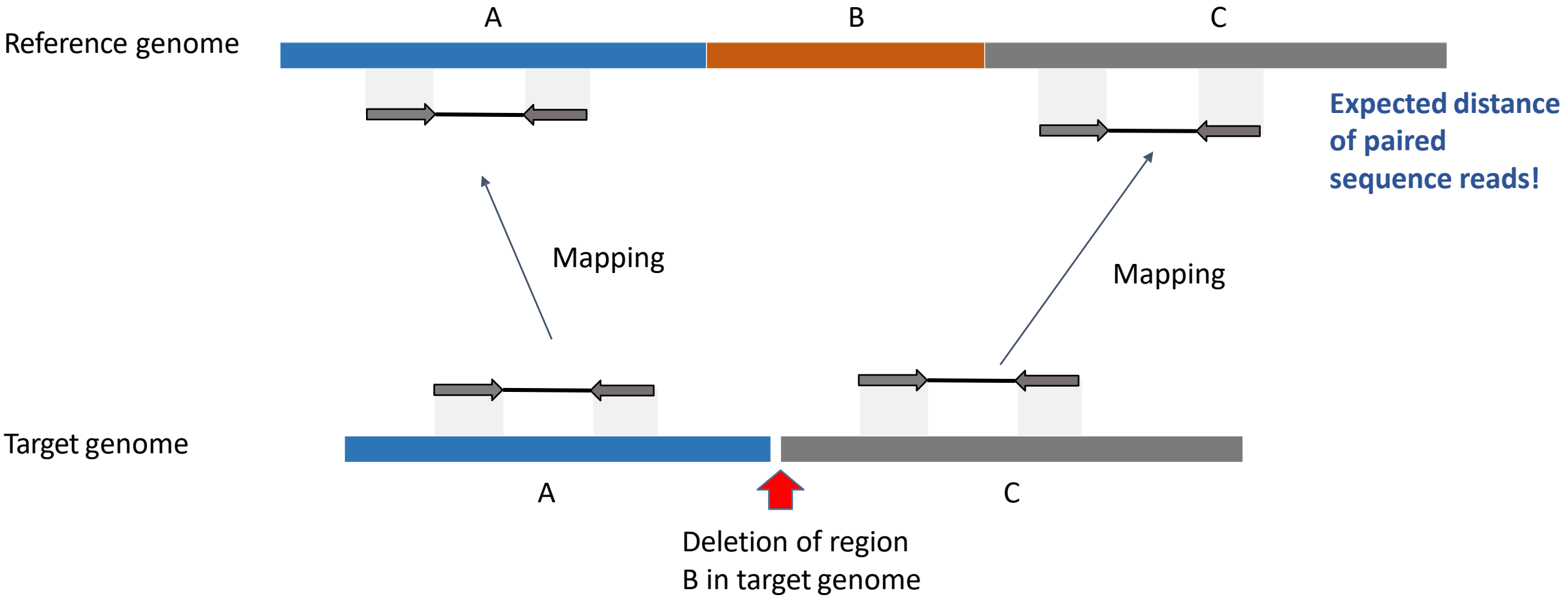
Example:
Lung cancer
Each line represents translocation in tumor genome compared to normal tissue from same patient



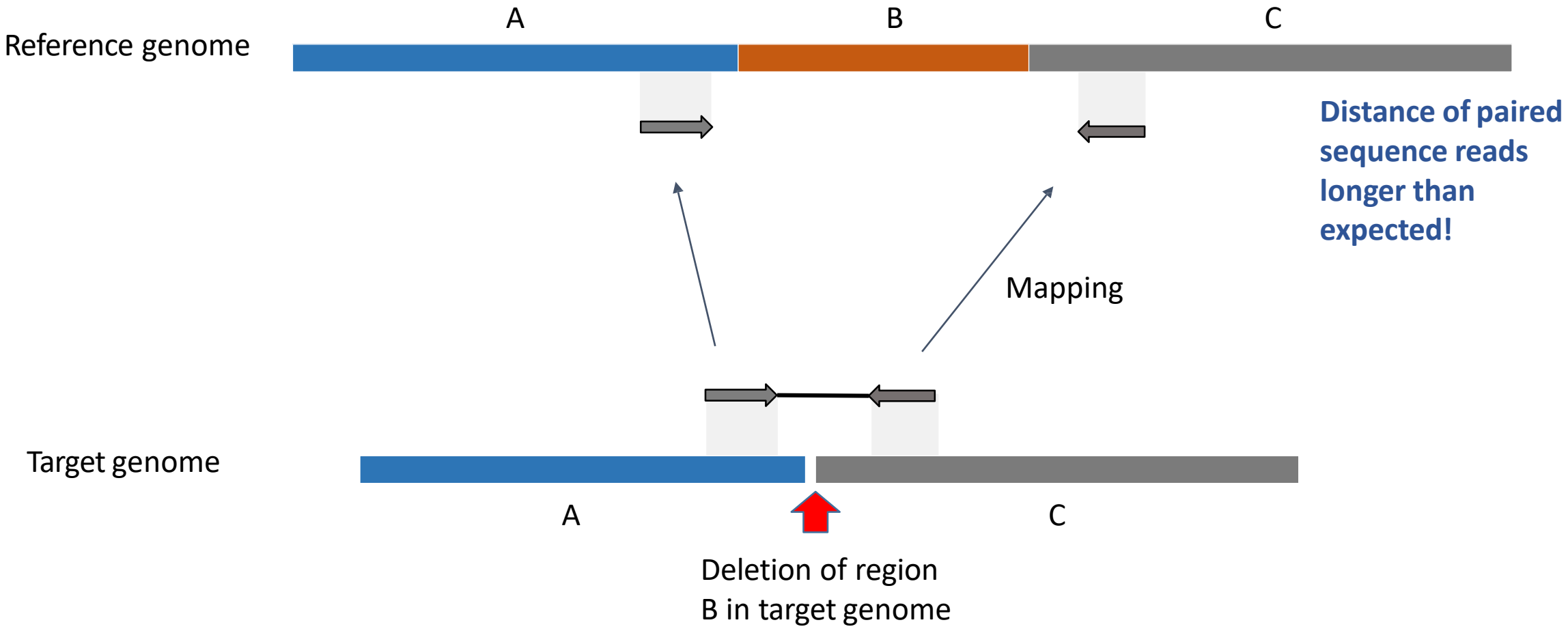
Structural Variations



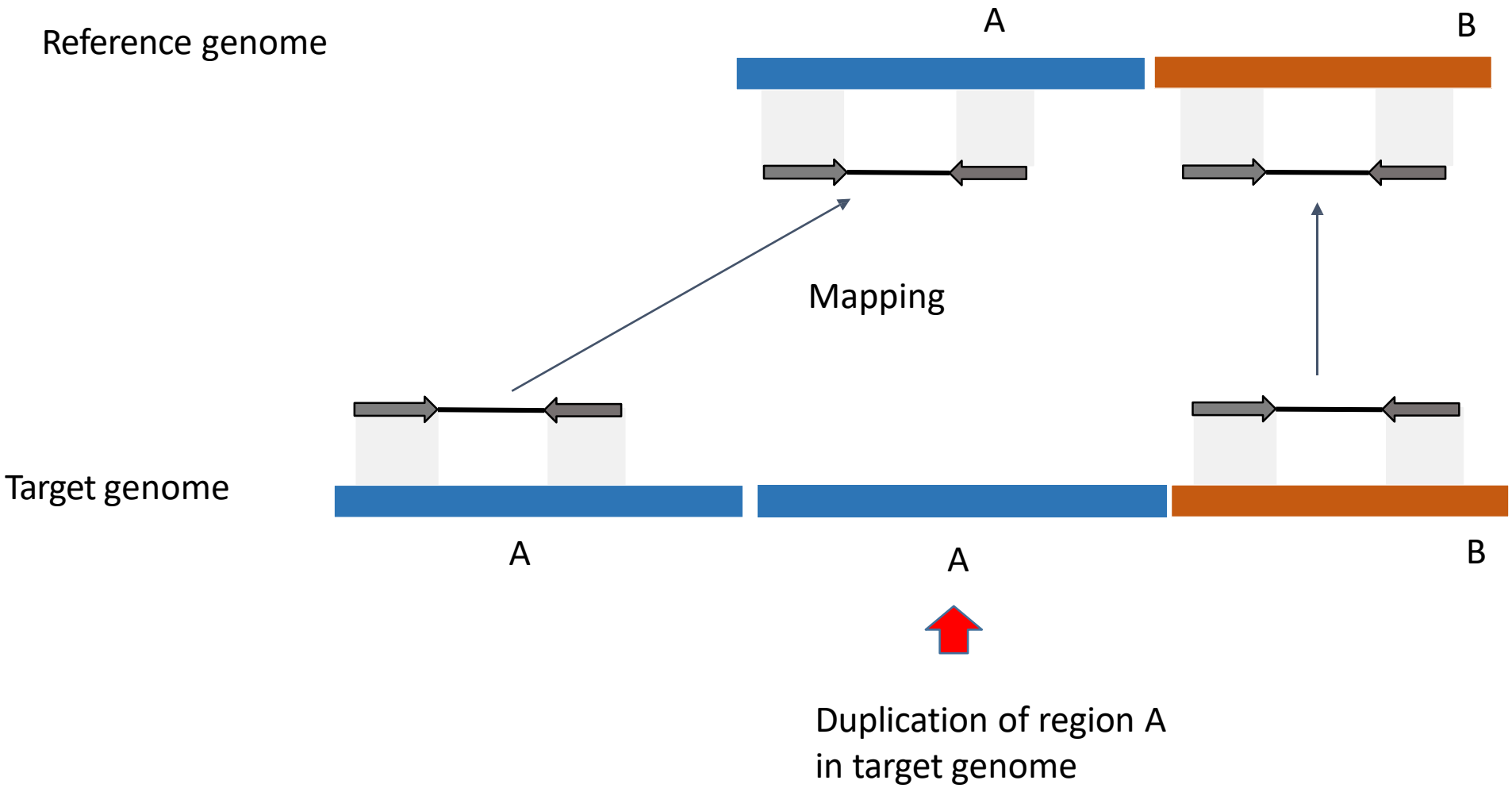
Structural Variations



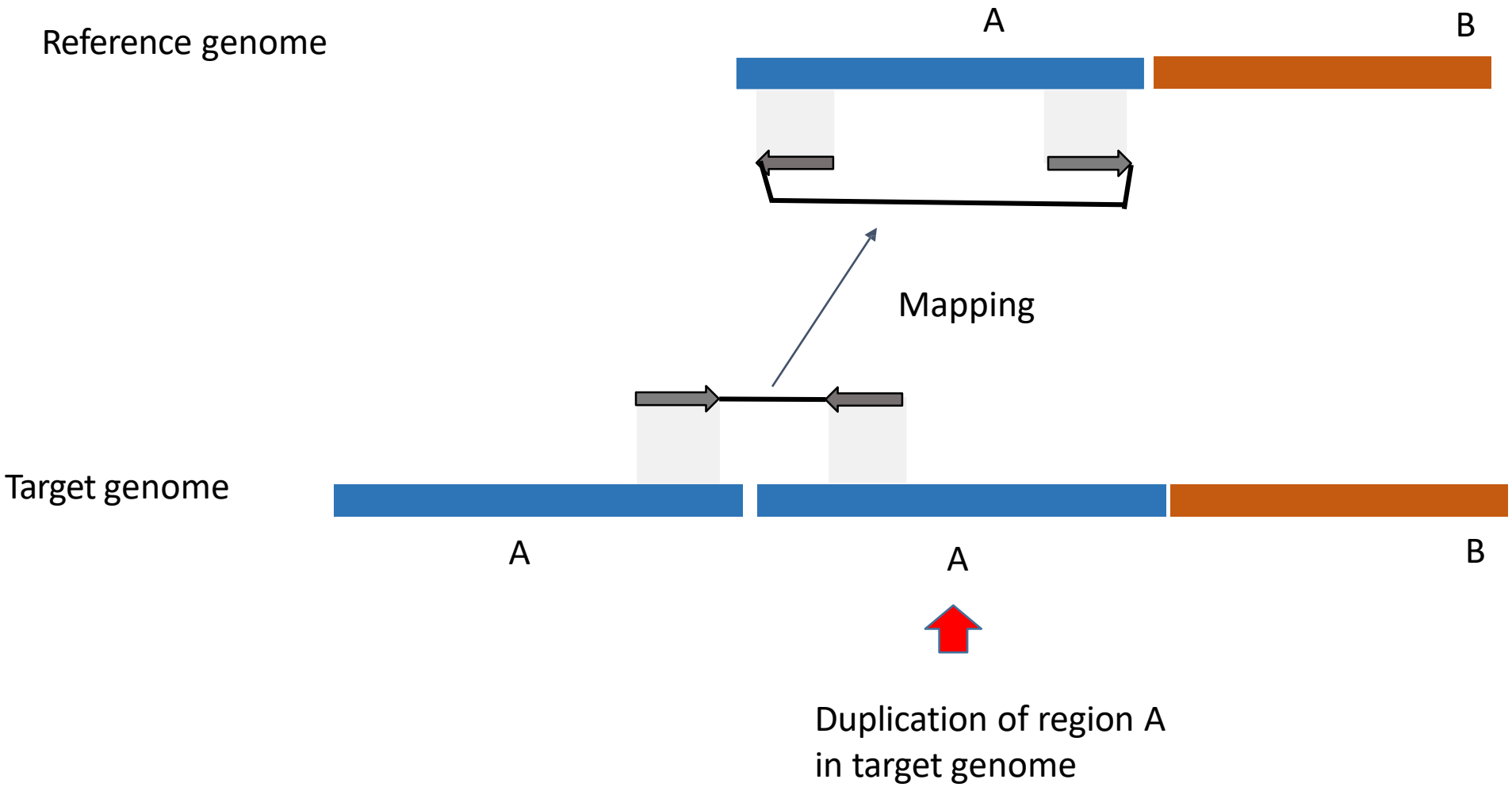
Structural Variations



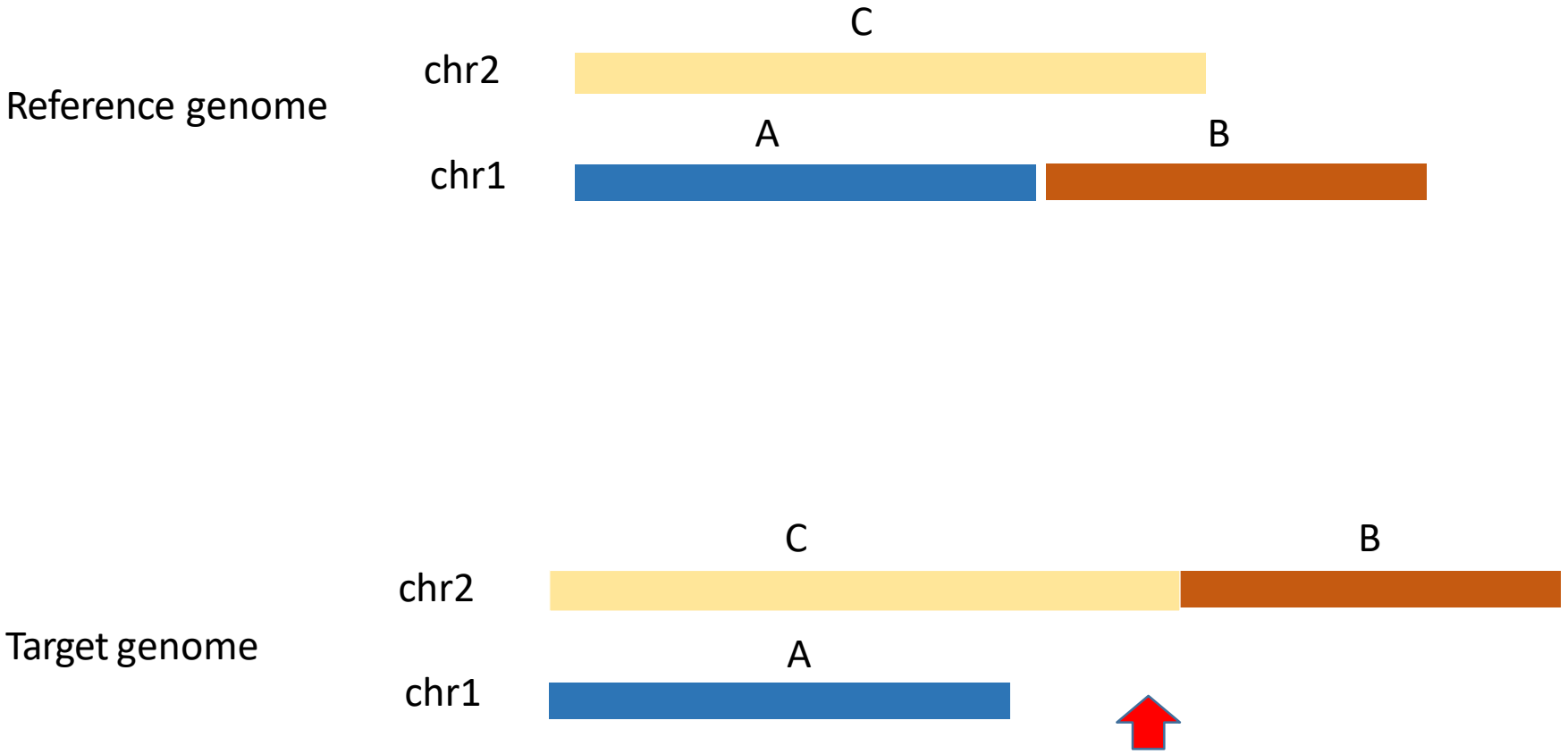
Duplications



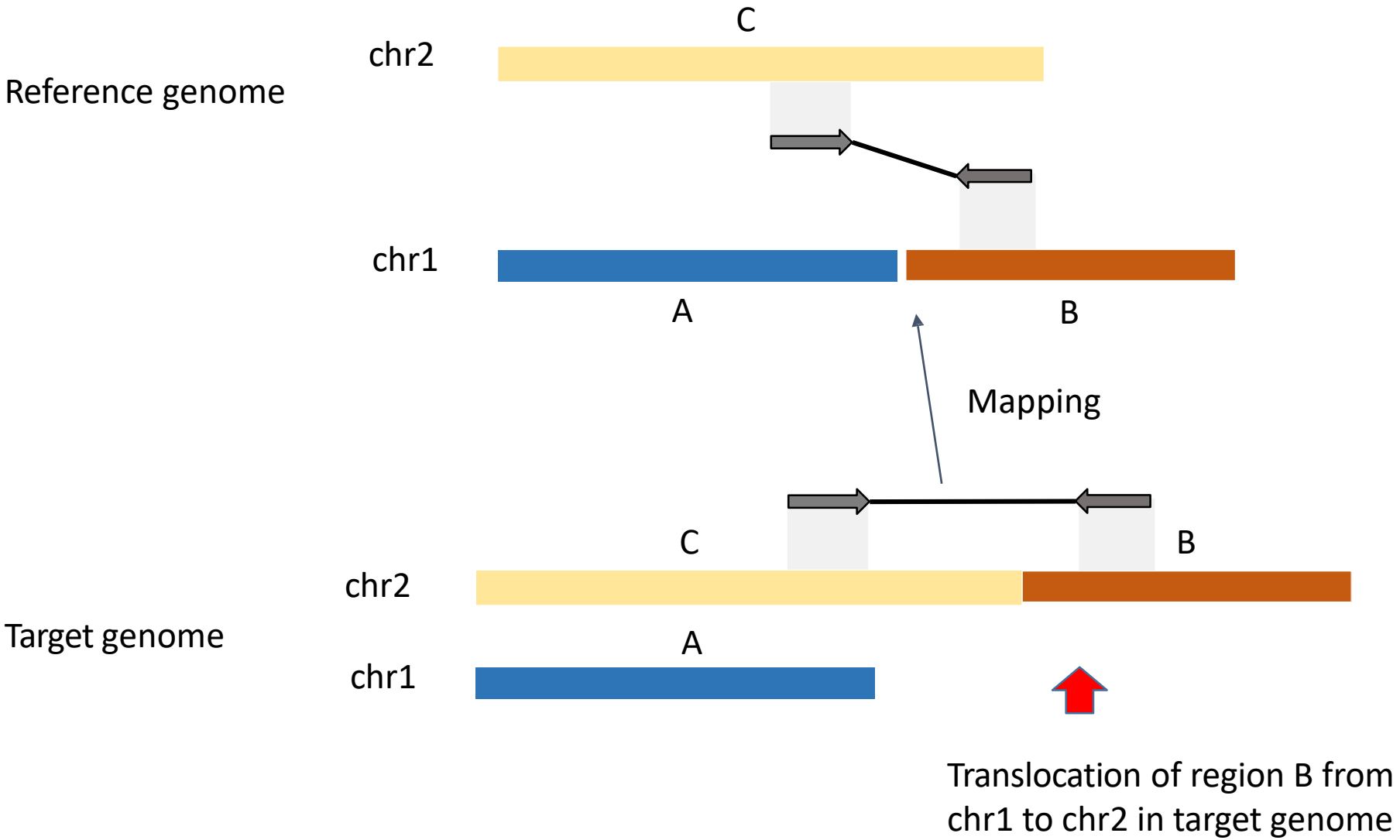
Duplications



Translocations



Translocations



Genome Re-Sequencing

- Aligning to a reference genome is significantly faster and easier than generating a de novo assembled genome
- If you work with eukaryotes, you will probably spend most of your effort on aligning and comparing to a reference
- Genome assembly is more common in organisms with small genomes (single-celled organisms)

De Novo Genome Assembly

- No prior information about the genome (no reference genome)
- Only sequencing reads supplied
- Necessary for novel genomes (e.g. parasites)
- Reconstruct the genome sequences of an organism from its read sequences alone

Genome Assembly

DNA copies of the genome

Sequence Reads

Assembled genome

The Call–Chronicle–Examiner

EARTHQUAKE AND FIRE: SAN FRANCISCO IN RUINS

NO HOPE LEFT FOR SAFETY OF ANY BUILDINGS

BLOW BUILDINGS UP TO CHECK FLAMES

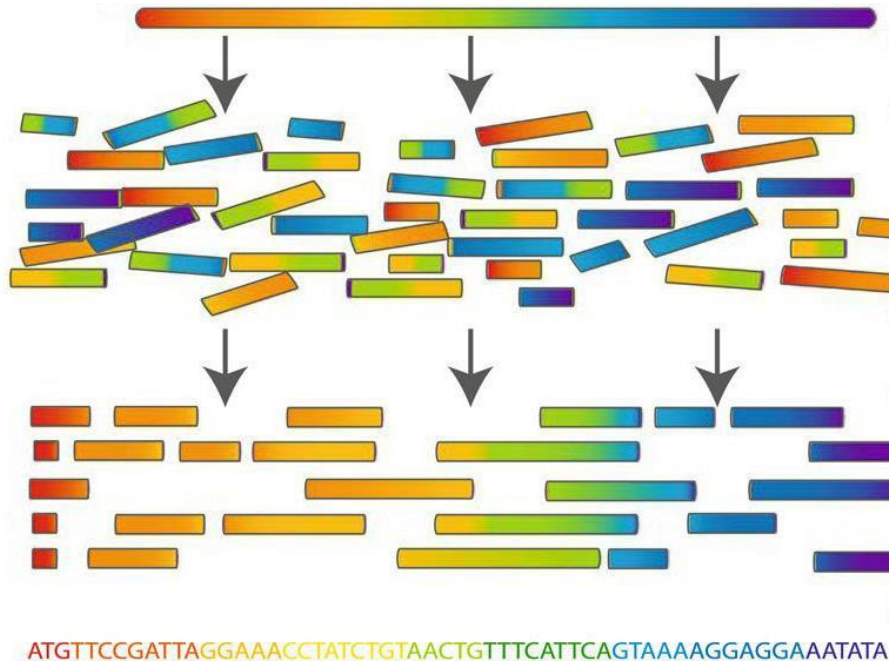
WHOLE CITY IS ABLAZE

CHURCH OF SAINT IGNATIUS IS DESTROYED

MAYOR CONFERS WITH MILITARY AND CITIZENS

Adapted from Torsten Seemann

Genome Assembly



Commins et al, Biol Proced Online (2009)

Example:

True sequence (7bp):

- AGTCTAT

Reads (3 x 4bp):

- AGTC, GTCT, CTAT

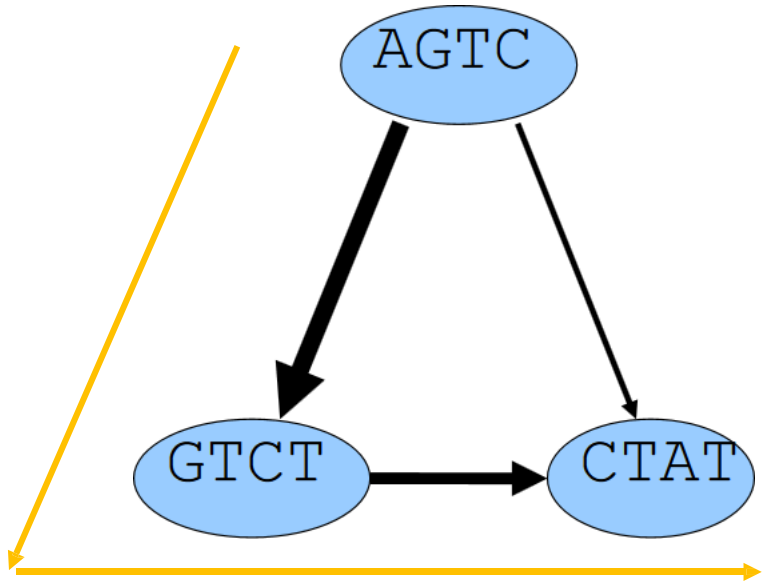
Overlaps:

AGTC -	AGTC - - -	GTCT - -
- GTC T	- - - CTAT	- - CTAT
(good)	(poor)	(ok)

Adapted from Torsten Seemann

Overlap Graph

- Nodes represent sequencing reads
- Edge width represent overlap score
- Consensus is generated by aligning reads along consensus graph (orange)
- aGTCTCTat



Adapted from Torsten Seemann

Assembly Graph

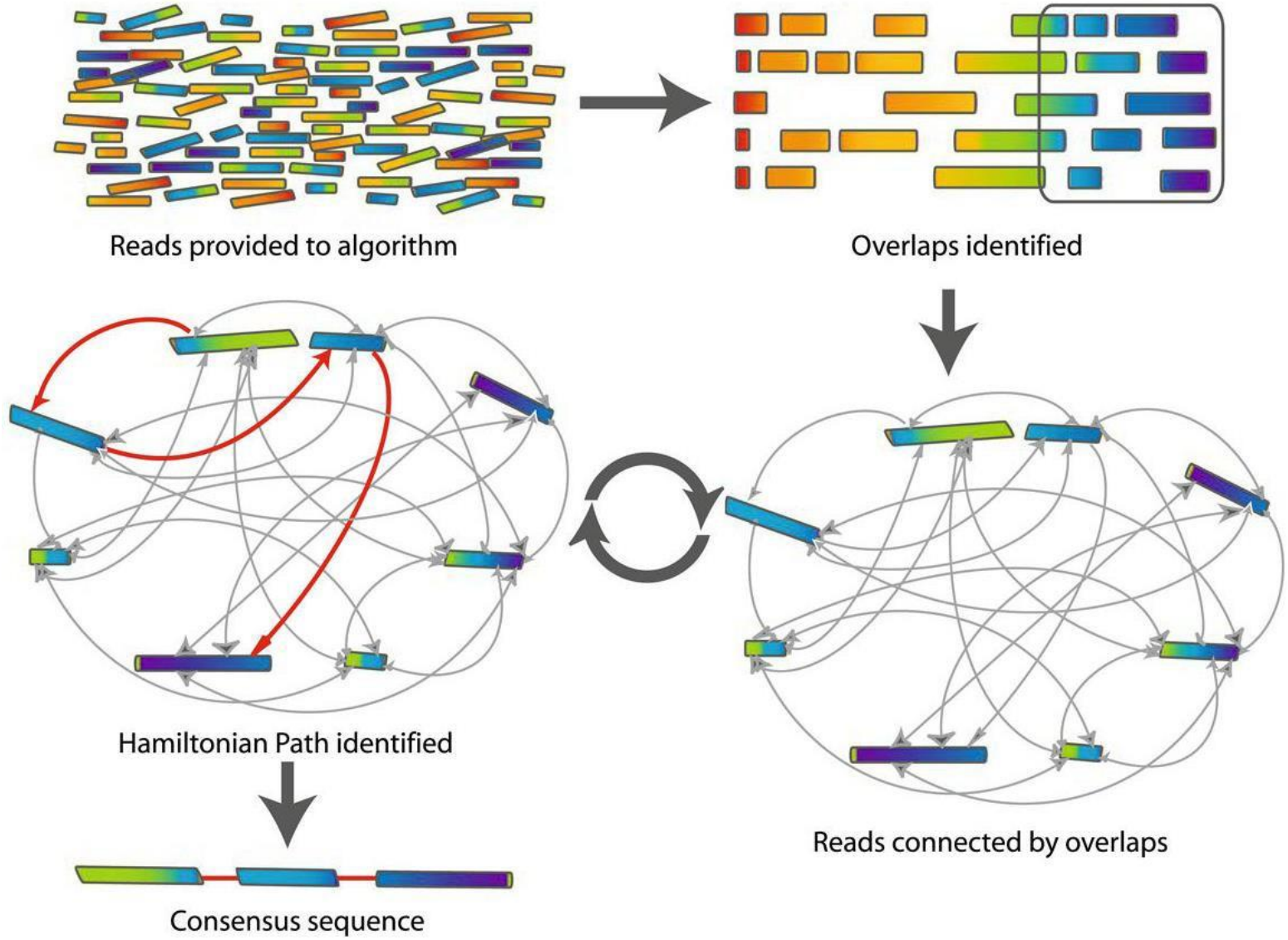
Steps:

1)Overlap: All against all pair-wise comparison

2)Build assembly graph: Nodes=reads, edges=overlaps

3)Hamilton path: Path that visits each node exactly once

4) Consensus: Align reads along assembly path



Commins et al, Biol Proced Online (2009)

Example: De Novo Genome Sequencing of Blood Flook *Schistosoma spp.*

- Infections by *Schistosoma spp*. significant health problem in Africa and Southeast Asia

Genome assembly challenging:

- Large genome size: 451 Mb
- High number of repetitive regions (>30% of genome repetitive)

Sequencing of *Schistosoma spp*

- 100 fold coverage on Illumina HiSeq and low-coverage PacBio

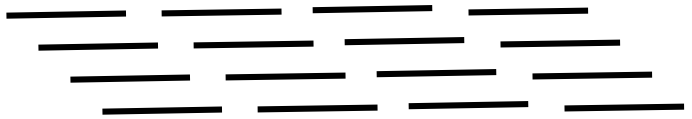
Every base of genome covered by ~100 sequence reads

	Average read length	Number of reads	Bases
Illumina	90bp	623M	56.1Gbp
PacBio	**3,205bp**	714K	2.3Gbp

Library	Insert size (bp)	Reads	Sequenced bases (Gb)
Small insert/paired-end	200	158M	14.3
Small insert/paired-end	500	174M	15.7
Small insert/paired-end	800	91M	8.2
Large insert/paired-end	2K	130M	11.8
Large insert/paired-end	5K	68M	6.1

Genome Assembly

DNA Extraction and Sequencing



Short reads

Assembly (ABYSS)

Contigs (20,949)

Scaffolding (SSPACE)

Scaffolds (4,780)

Mate pairs

Gene Identification

Functional Annotation

Next lecture

Scaffold statistics

Number	4,780
Largest scaffold (Mb)	1.1
Mean scaffold size (kb)	78.2

SCIE2100 | BINF6000

Bioinformatics

Genome Analysis II

Atefeh Taherian Fard, PhD

Australian Institute for Bioengineering and Nanotechnology

a.taherianfard@uq.edu.au

Outline

Lecture 1:

- Overview genome sequencing and sequencing technologies
- Genome re-sequencing
- De-novo genome assembly

Lecture 2:

- Gene features in prokaryotes
- Gene features in eukaryotes
- Computational approaches for gene prediction
- Functional genome annotation

Gene finding Approaches

- Physical, genetic or other ***experimental approaches***
  - *e.g.* Genetic knockouts

- ***Computational approaches***

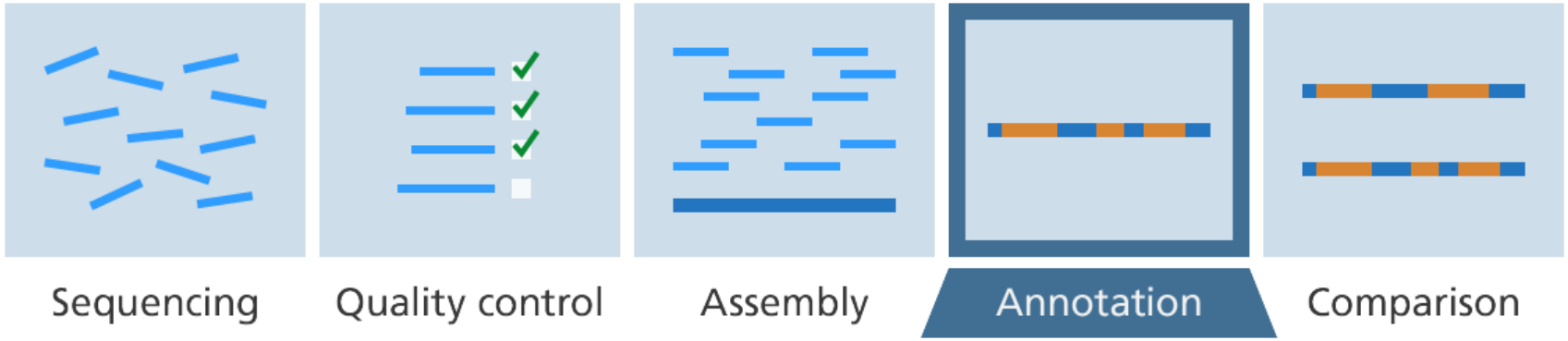
1) Identity search

2) Similarity search Homology based

3) *Ab initio* approaches

Slides adapted from Lutz Krause

www.yourgenome.org



Using computational methods, find all genes (or other elements) in a long, unannotated string of nucleotides.

```
ACCGGTCAATAGCCGCAGACTACGGCATTTCAGAGGGACAGGCACTATAGCAACTAGCAACCCCCGTATAATACAAGGAGGCT
CAAGCTCCACTCTGACTCTCAACTTATTACGCTGTCACTCGATACGGCAGGGGCATTTAGACTTACGGCATATACCCGCCCGA
TCCAGCTTACGATACTACTGCTACTGGATACCCTGATAGCCAATCATTACGACTACTACTACGGCATTTCAGACCCGACAGGC
ACTAGAGCAACTAGCAACACCCGTATAATACAAGGAGGCTCAAGCTCCAGCTCTCACTGCAGCTATGTGGTGCACACATGTGC
ATCGTATGACTCAGTCGATGCTATCACGTACATCGTGTGGGTGCACACCACCCATGCCCTGATAGCCCCTGATTTTAGCCCCA
GCATTATTTTTCCGACGAGATCACGTACCCCCTACGGCATTTCAGAGGGACAGGGGACGCGCCCAATTACGACTACTACTACG
GCATTTCAGACCCGACAGGCACTAGAGCAACTAGCAACACCCGTATAATACAAGGAGGCTCAAGCTCCAGCCTTTTCAACAGA
CCGGGCGTTACGGTAAAAAAAAATCCGGCCGTACGGACTACTGGATACCGCAGACTACGGCATTTCAGAGGGACAGGCACTAT
AGCAACTAGCAACCCCCGTATAATACAAGGAGGCTCAAGCTCCACTCTGACTCTCAACTTATGACAGGGGACGATGACTCAGT
CGATTTCGCTATCACGTAAAACATCGTGTGGGGTGCACACCACCGCATGCCCTTTCAGGATAGCCCCTGATTTTAAGCCCCAG
CATTATTTGGTTCCGACGAGATCACGTACCCCACTACGGCATTTCAGAGGGACACTCAGTCGATGCTATCACGTACATCGTGT
GGGTGCTTACACCACGCCATGCCCTGATAGCCCCTGGGGATTTTAGCCCCAGCATTAATTTCTTCCGACGAGCCCTCAGACCC
GACAGGGGCACTAGAGCAACTATATAAGCAACACCCGTATACCCATACCAAGGAGGCTCAAGCTCCAGCCTTCTTCAACAGGA
CCGGGCGGGATTCCACATCATTCATGGGCAGCATCCCAGCAAACCCACGGCATAAGGACCACCCCTCGGCTAAGCAATCGCAT
AATACGGCGCTGCGCTCTACGTCTAGAGCTCACCATCTTACGAGGCTCTACCTCTATG...[3 BILLION or so MORE]
```

Aim: To identify transcriptional unit

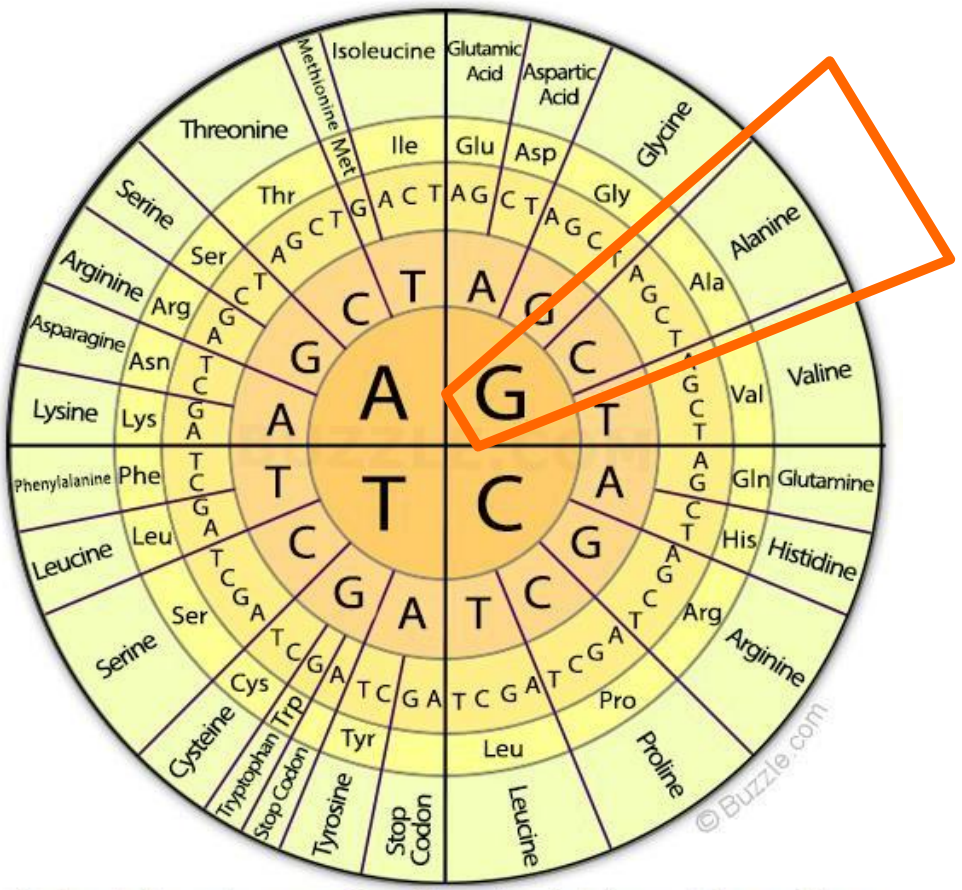
What do we know? Know only approximately what they look like

How? Find their locations and boundaries as accurately as possible, overlook as few as possible, and report as few non-genes as possible.

Adapted from Mark Ragan

Codon bias

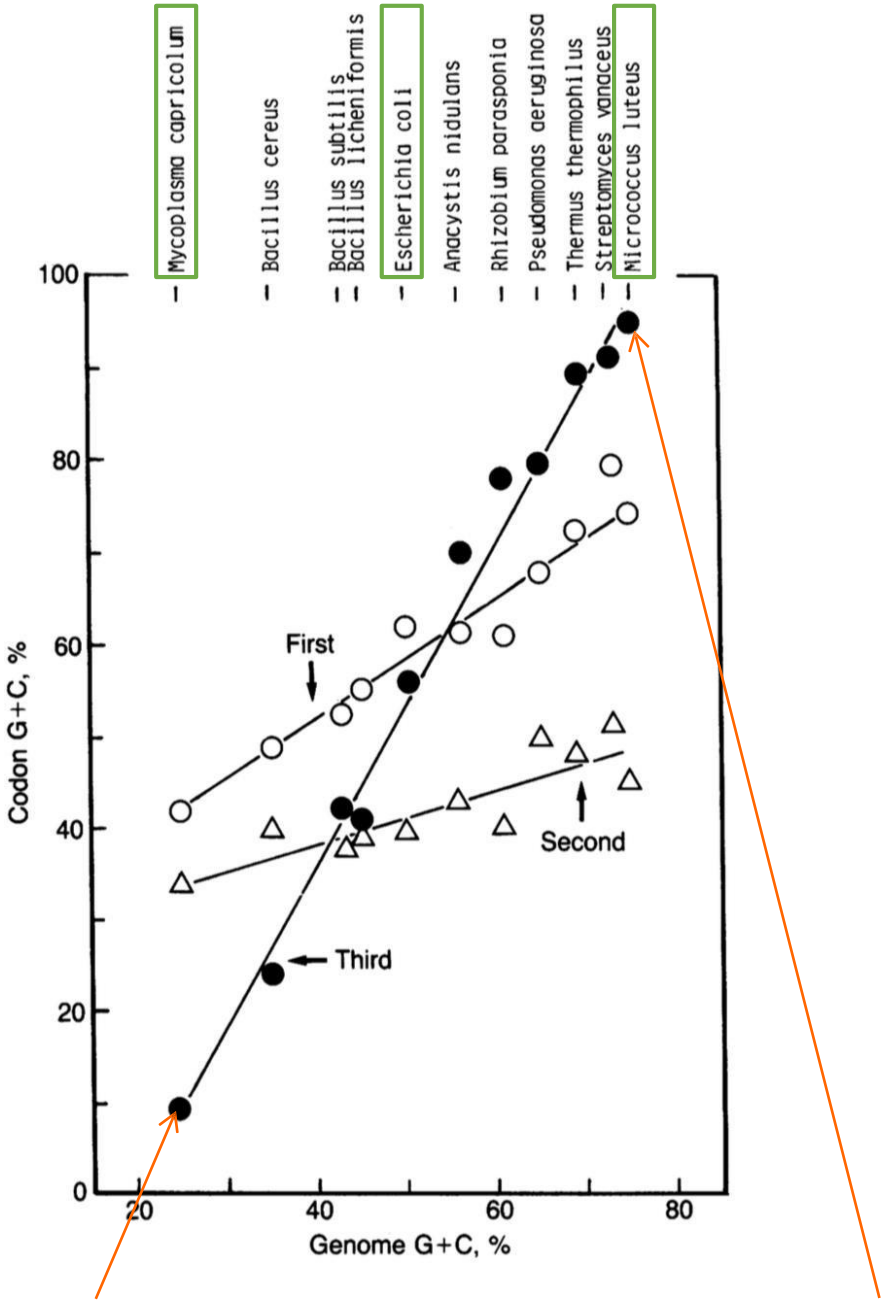
- 61 codons encoding 20 amino acids, genetic code is redundant
- Codon bias: Each organism seems to prefer a different set of codons over others

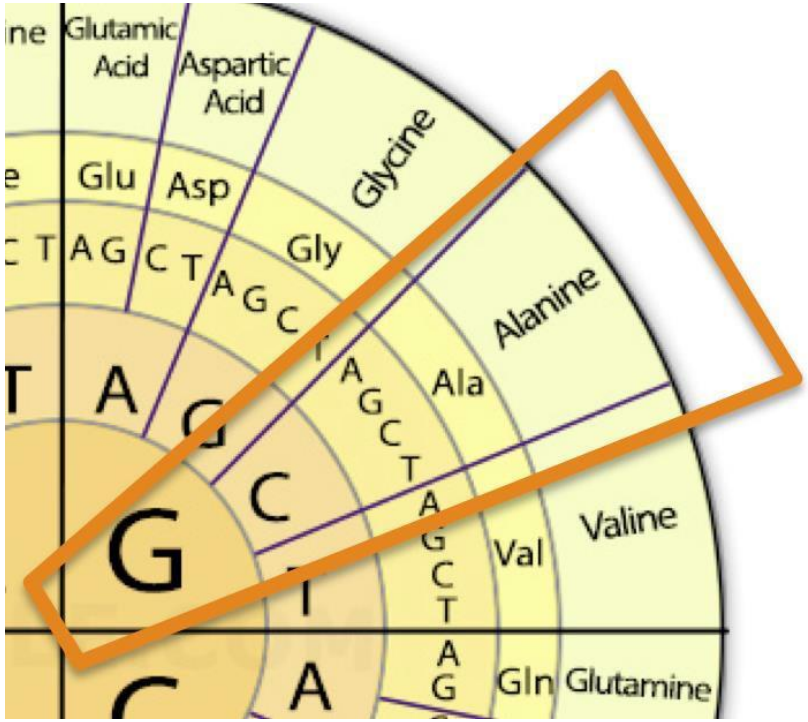


To decode the codon, move from the center circle towards the periphery.

- This redundancy is mainly at the 3rd codon position
- Example 1: The codons for alanine can have ANY base in the 3rd position
- Examples 2: Leu is encoded by 6 codons. But human in nuclear genes it's most frequently coded by **CTG**, and only rarely by **TTA** or **CTA**

http://www.buzzle.com/articles/codon-chart.html





Very high correlation between **genomic** GC content and **3rd codon position GC** usage

=> Codons in GC-rich genomes will have higher frequency of GC at 3rd position

GCA, GCT usage more common

GCC, GCG usage more common

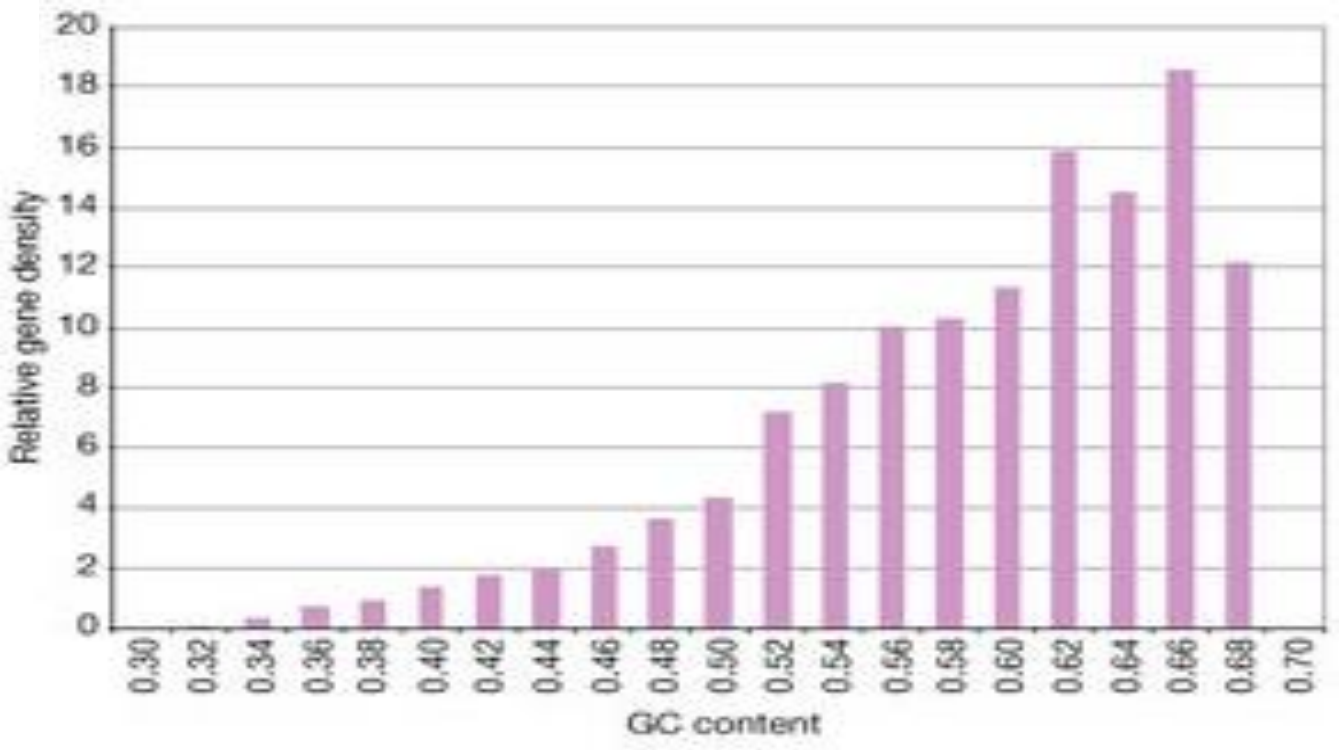
Muto and Osawa, 1987

CpG islands, (G+C)-rich regions and genes

- For slightly more than half of human genes, transcriptional start site occurs in a CpG-rich region
- *This is particularly true for "housekeeping" genes:*

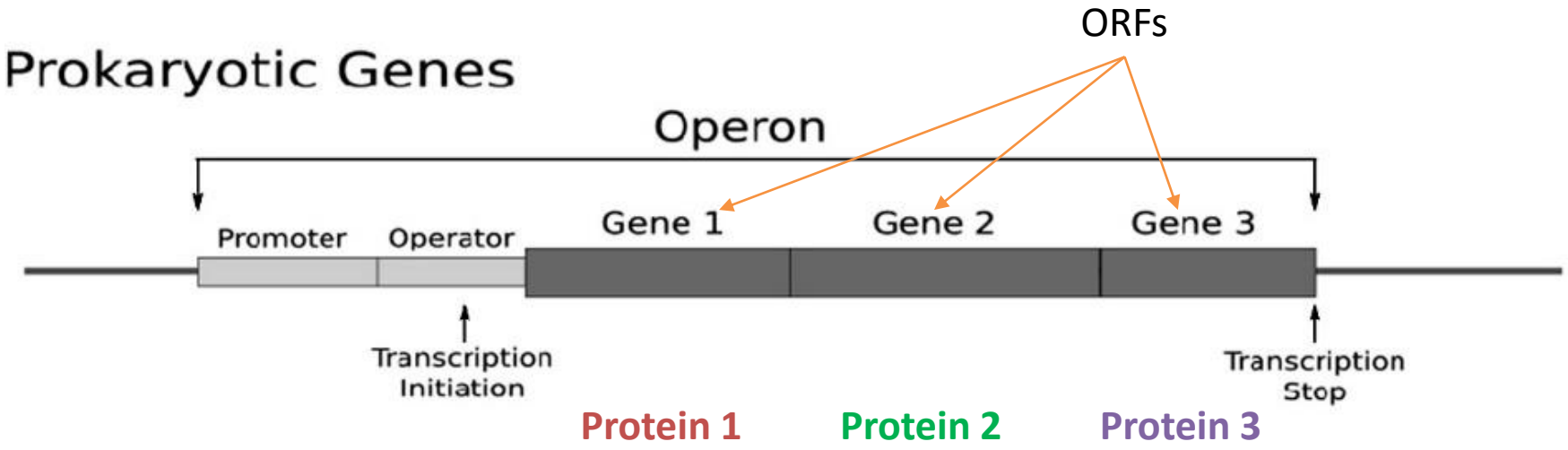
the first coding exon usually occurs in a CpG island

- Mammalian genomes show elevated gene density in (G+C)-rich regions



Brown, Genomes 2

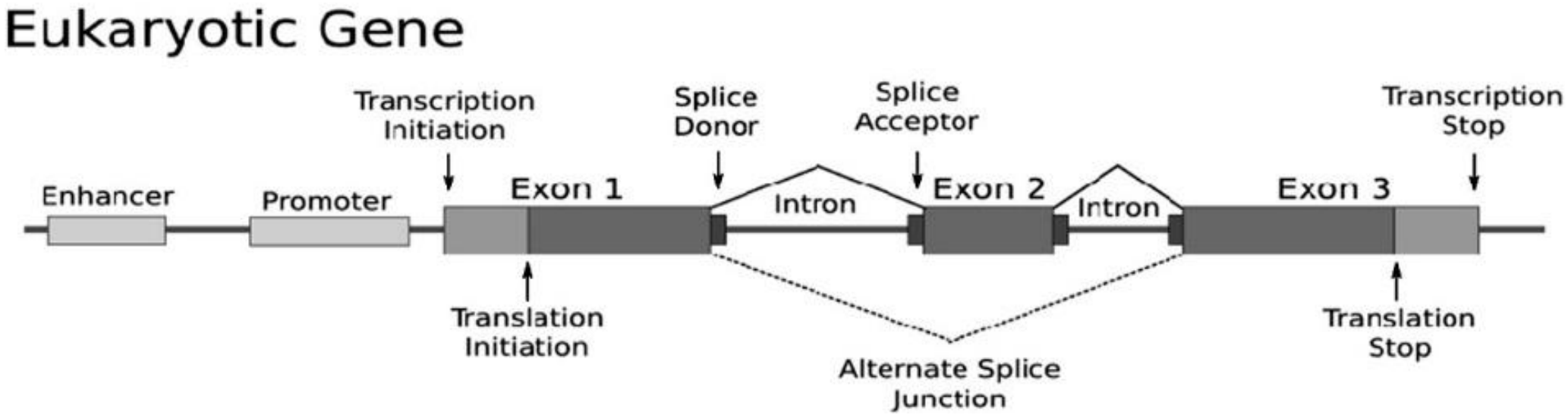
Gene features in Prokaryotes



Protein coding genes in prokaryotes:

- Are open reading frames (ORFs)
- Begin with start codon (usually ATG for Methionine)
- End with stop codon
- No in-frame stop-codon

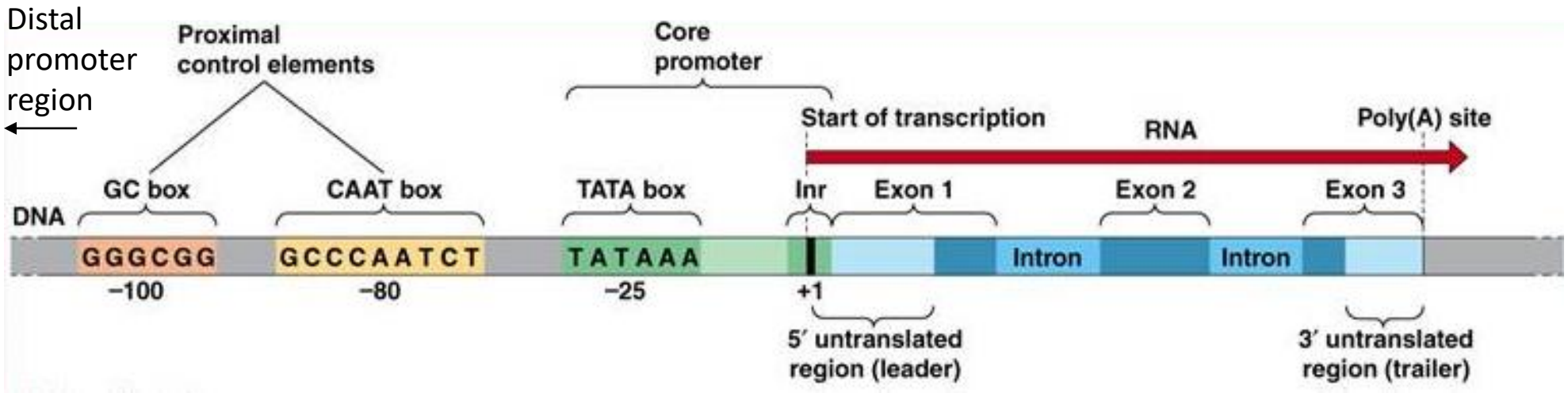
Gene features in Eukaryotes



Protein coding genes in Eukaryotes:

- Exons and introns
- Regulatory sequences (promoter and enhancer)
- Splice donor and accepter sides
- 5’ Cap and 3’ PolyA tail

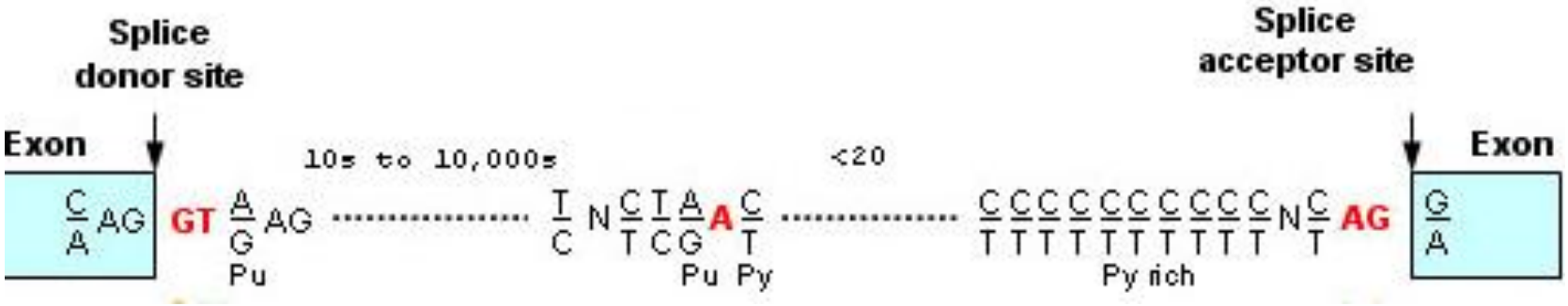
Regulatory promoter regions in Eukaryotic genes



- Regulatory sequences have distinct features, recognised by DNA-binding proteins to regulate gene expression
- Sequence motifs, may be conserved
- In each species, different sets of genes have different motifs

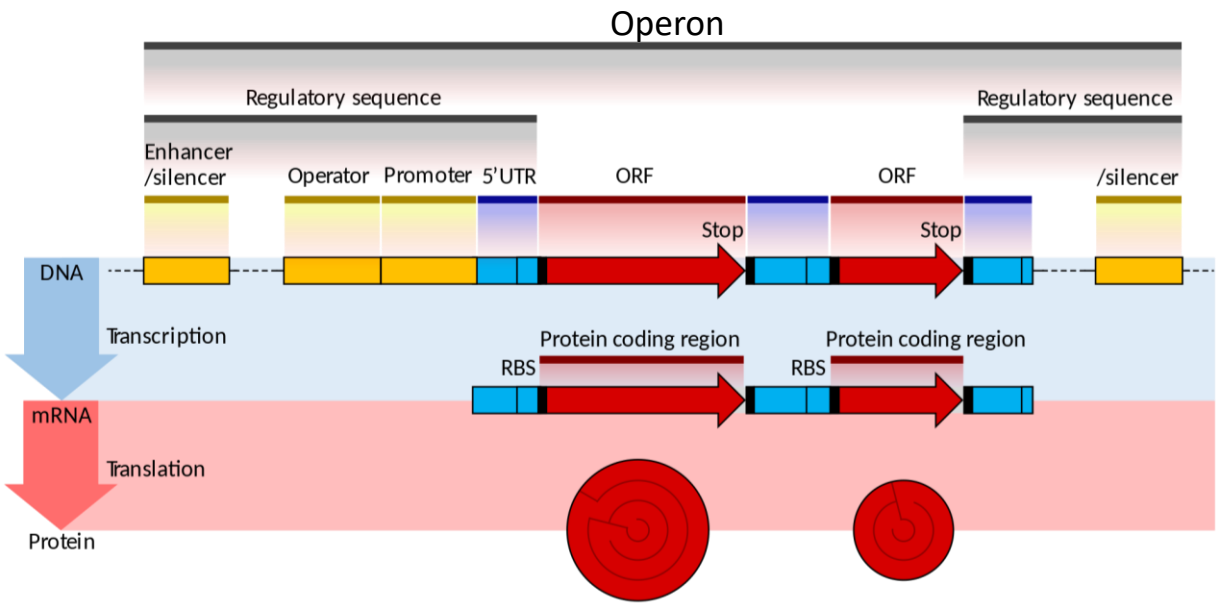
http://www.mun.ca/biology/

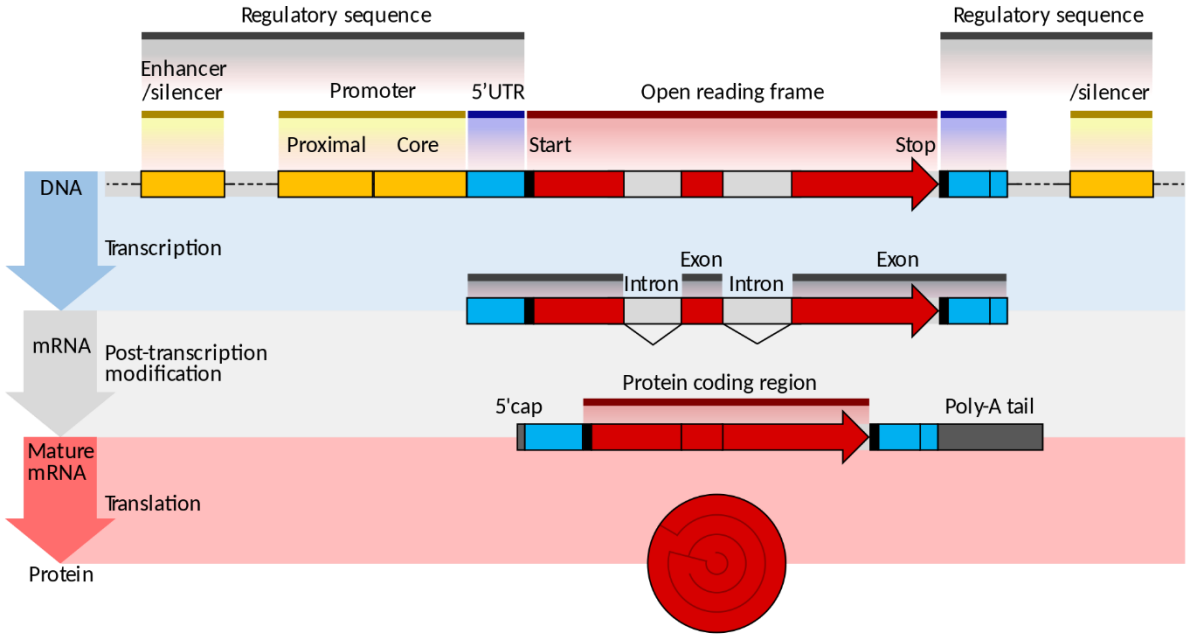
Intron-exon splice sites



- The GT-AG rule
- Numbers indicate the number of nucleotide bases between each conserved region
- Pu: Purines (A and G)
- Py: Pyrimidines (C and T)
- N: Any nucleotide

Summary: Gene features in Prokaryotics vs Eukaryotes





- Codon bias and GC rich regions
- Transcriptional start and stop sites
- ORFs: Start and stop codons
- 5’ UTR: Ribosomal Binding Sequence site
- 3’UTR

- Codon bias and GC rich regions
- Promoter regions
- Intro and Exon splice site
- ORFs: Start and Stop codons
- 5’ UTR: 5’ Cap (G cap site)
- 3’ UTR: PolyAs

Shafee & Lowe (2017)

Gene finding Approaches

- ***Computational approaches***

1) Identity search

- Looks for exact match the query and target sequence!

2) Similarity search- Homology based

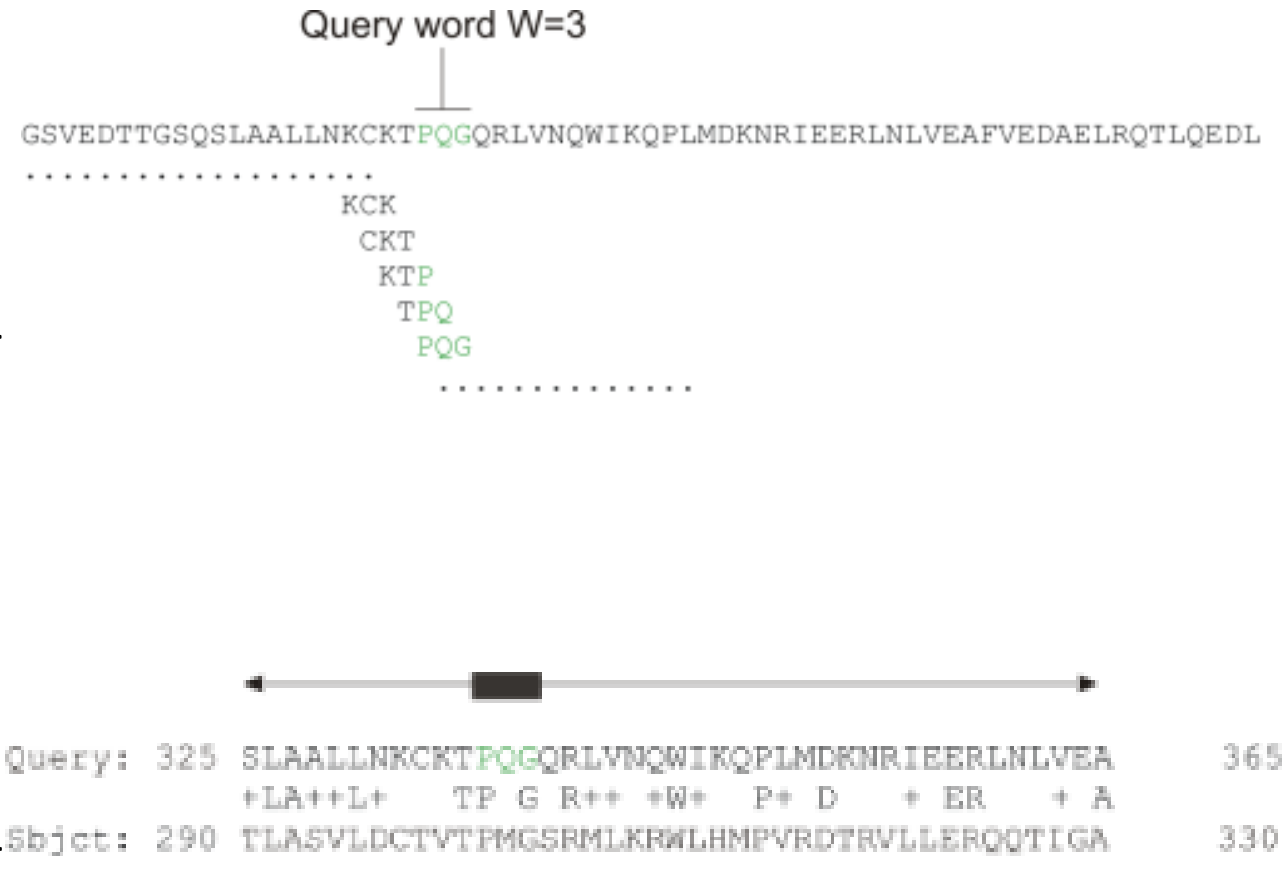
- Translate genome in all 6 reading frames, map translated sequence to known proteins (*e.g.* using BLAST)
- Only identifies known genes
- Species specific
- Unknown genes and evolved genes with low sequence identity are missed
- Compare genome to close relative organism and identify conserved regions

3) *Ab initio* approaches

- Evaluate ***specific sequence properties of genes*** to distinguish between non-coding and coding regions
- Identifies also new genes, but may produce many false predictions

Similarity search (e.g. BLAST)

- BLAST identifies homologous sequences using a heuristic method which initially finds short matches between two sequences.
- When finding a match between a query sequence and a hit sequence, the starting point is the *words* that the two sequences have in common.
- After initial finding of words (seeding), the BLAST algorithm will extend the (only 3 residues long) alignment in both directions.
- Each time the alignment is extended, an alignment score is increases/decreased.
- When the alignment score drops below a predefined threshold, the extension of the alignment stops.
- If the obtained alignment receives a score above a certain threshold, it will be included in the final BLAST result.

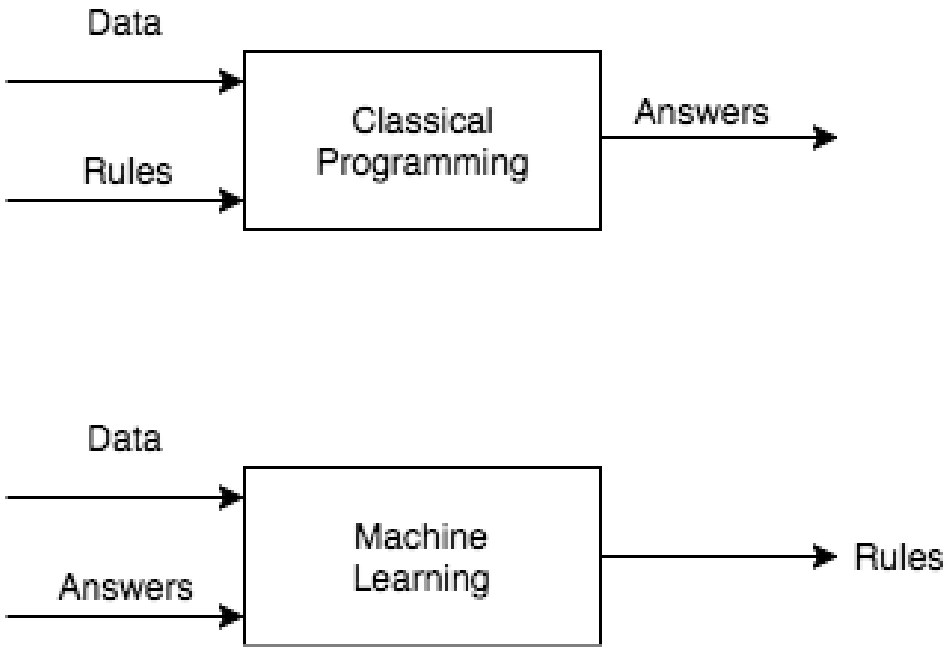


http://resources.qiagenbioinformatics.com/

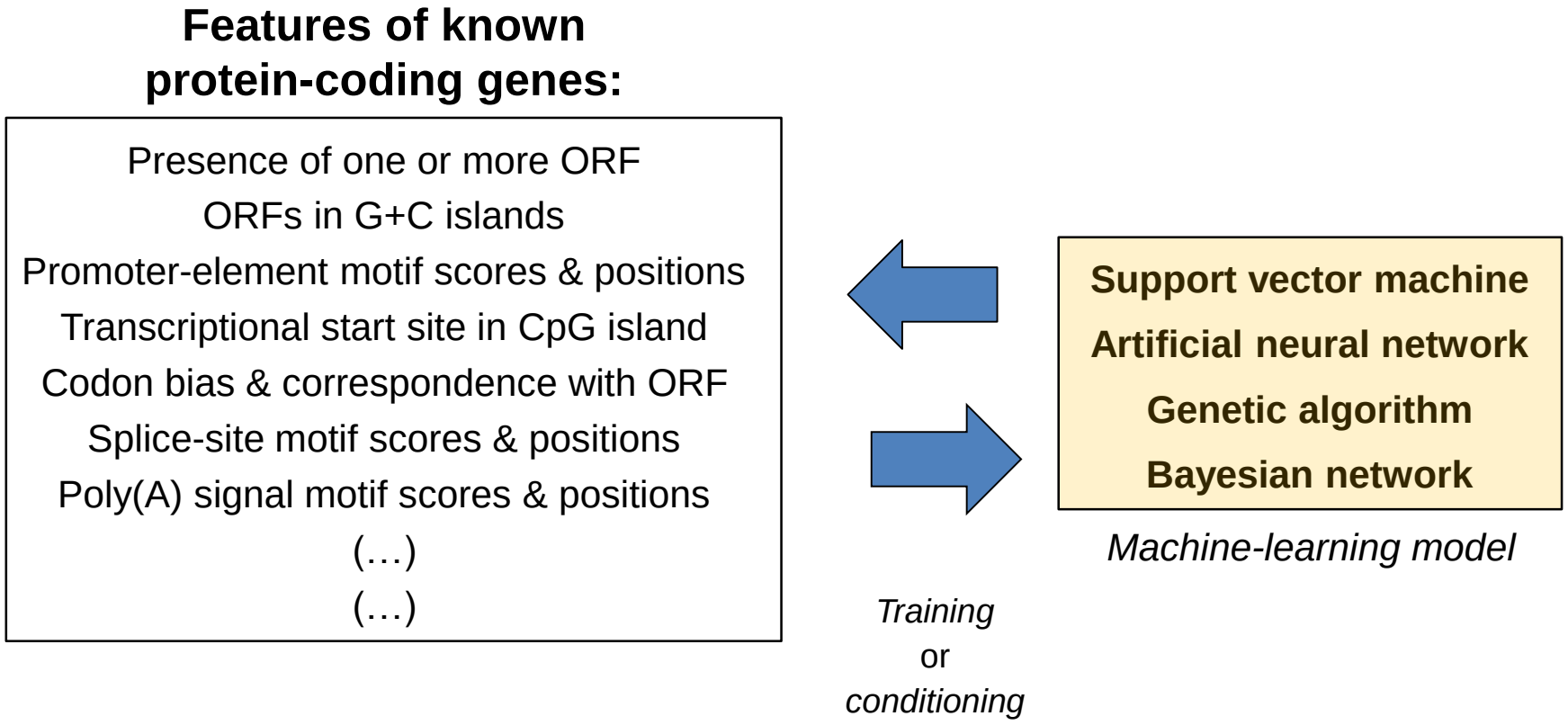
Machine learning approach to *ab initio* gene finding

- A branch of artificial intelligence
- Enabling computers make successful predictions using past experiences
- Based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention

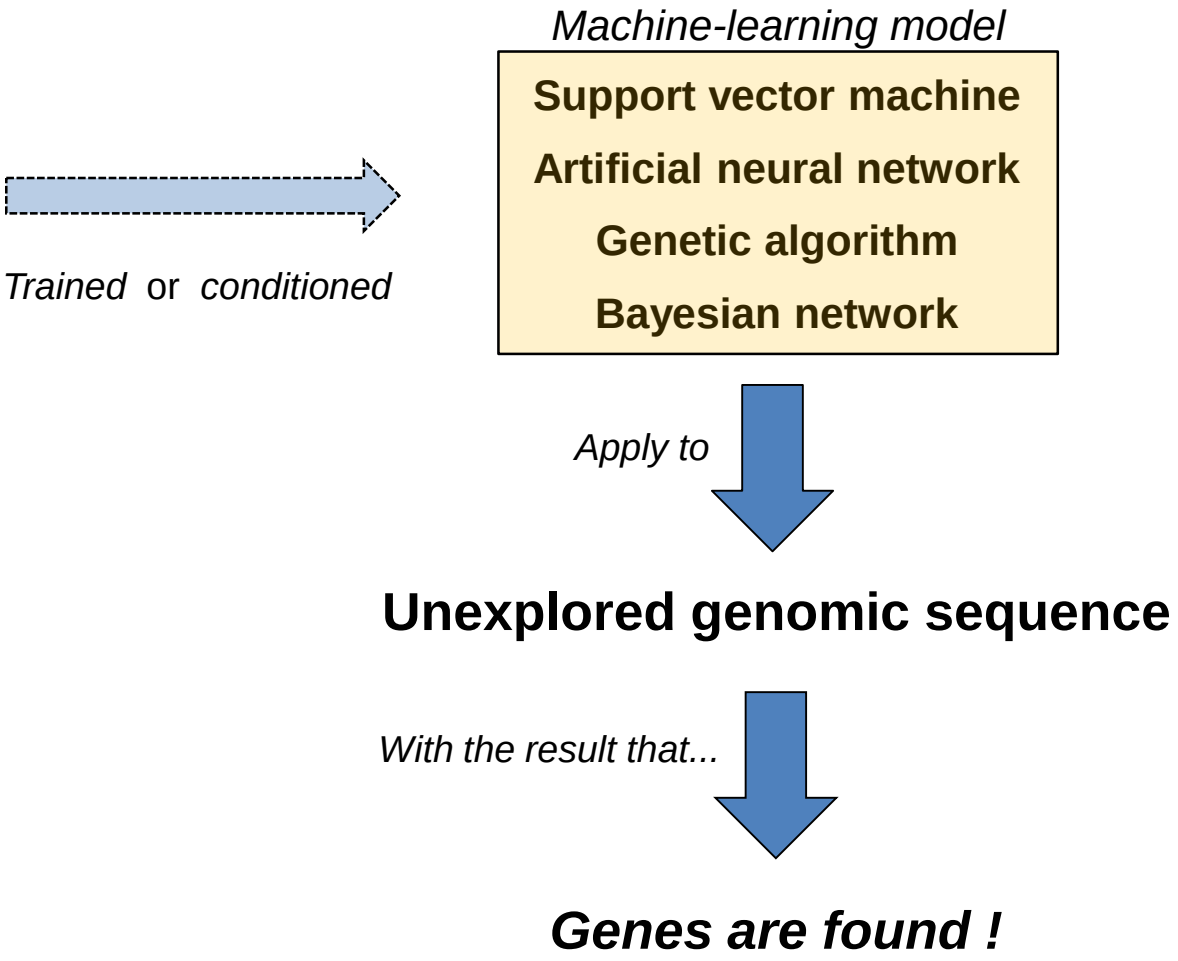
Machine Learning vs Traditional programming



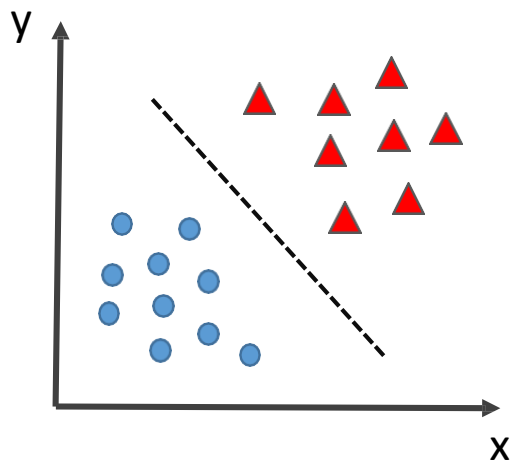
Machine learning approach to ab initio gene finding



Machine learning approach to ab initio gene finding



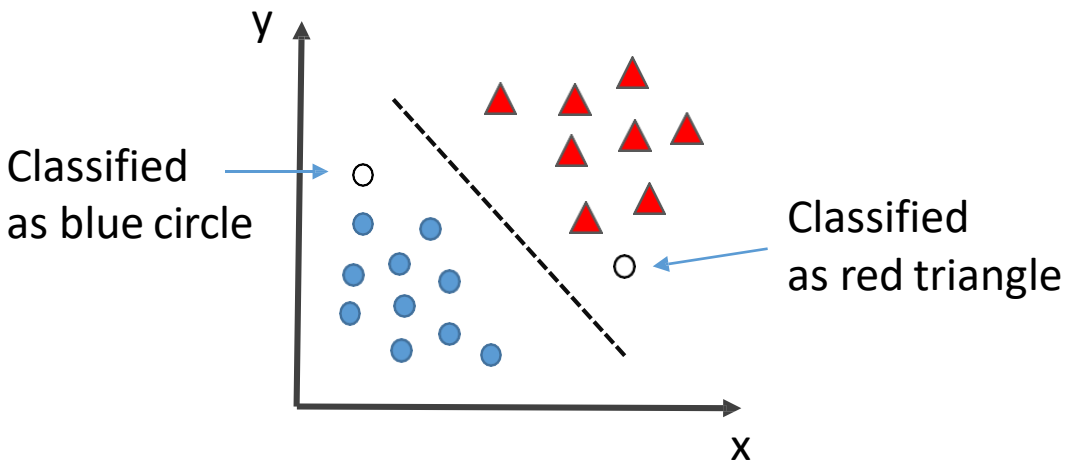
Support Vector Machines (SVMs)



Training:

- Given a training set of items from two classes a SVM learns a hyperplane that separates the items from both classes

Support Vector Machines (SVMs)



Training:

- Given a training set of items from two classes a SVM learns a hyperplane that separates the items from both classes

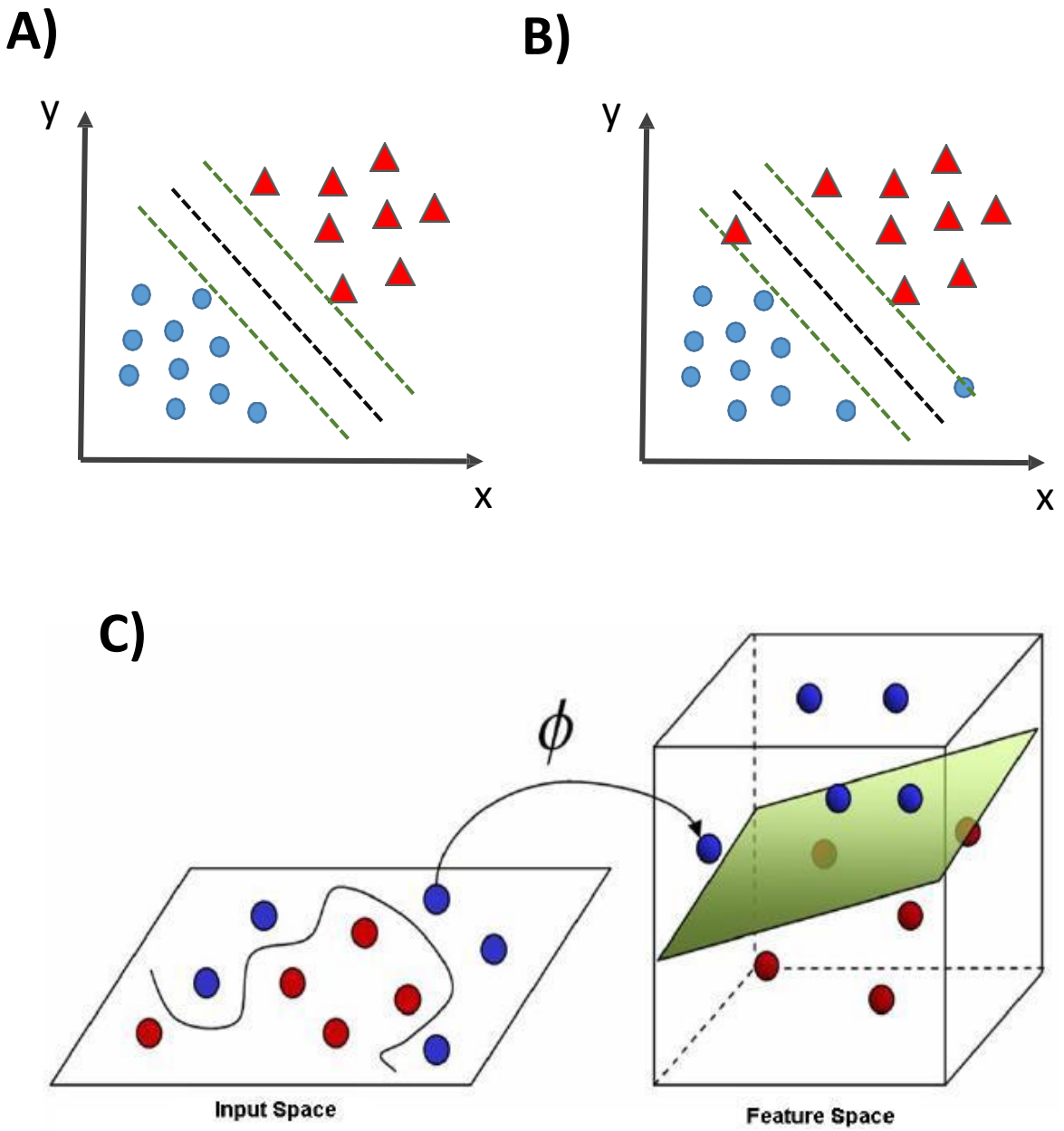
Classification:

- New items with unknown class affiliation classified depending on the side of the hyperplane

a) SVMs learn hyperplanes with maximal margins to improve generalization ability

b) Over fitting avoided by allowing misclassification of outliers of training set (called softmargin hyperplane)

c) Non-linear classifier achieved by mapping data to higher dimensional feature space via non- linear mapping function (kernel). By learning hyperplane in feature space, non-linear classifier achieved in input space



SVMs for Gene Identification

- SVMs can be used to classify ORFs into coding and non-coding (*e.g. Krause et al. NAR 2008*)

Training:

- Generate training set of ORFs known to be coding (positive training set) and ORFs known to be non-coding (negative training set)
- For each ORF, count codon frequencies
- Represent each ORF as 61 dimensional vector of codon frequencies
- Train SVM to distinguish between to sets of ORFs

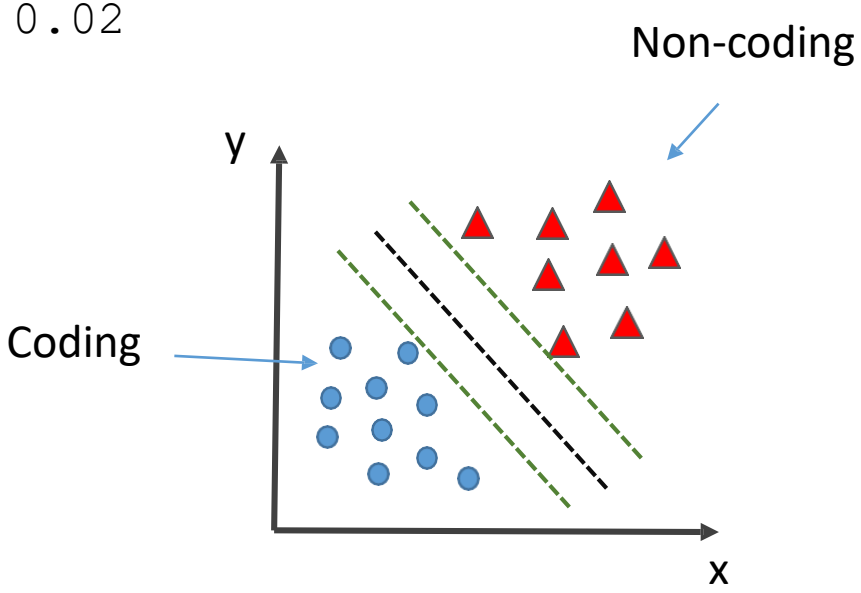
ORF: ATGGCTATCGACGAAAACAAACAGAA…

Codons:

ATG
GCT
ATC
GAC
etc.

Our **ORF** has the frequencies:

```
ATG = 0.1
GCT = 0.02
ATC = 0.001
GAC = 0.02
```



SVMs for Gene Identification

Classification (gene prediction):

- Extract all ORFs of genome
- Represent each ORF as 61 dimensional vector of codon frequencies
- Apply trained SVM to classify all ORFs into coding and non-coding depending on side of hyperplane
- scikit-learn library in Python
  - sklearn.svm

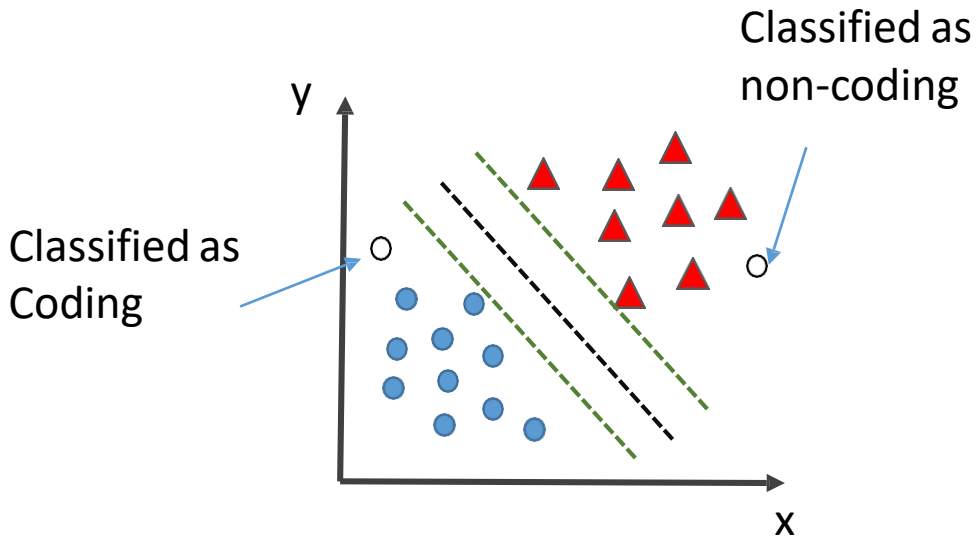
ORF: ATGGCTATCGACGAAAACAAACAGAA…

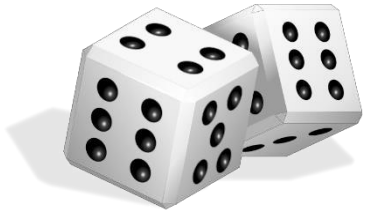
Codons:

ATG
GCT
ATC
GAC
etc.

Our **ORF** has the frequencies:

```
ATG = 0.1
GCT = 0.02
ATC = 0.001
GAC = 0.02
```





Hidden Markov Models

The dishonest casino:

Known information:

- Casino has 2 die, fair dice, loaded dice
- Casino player switches back & forth between dies
- Once either of the dice is used, it will continue to be used for a while

Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**

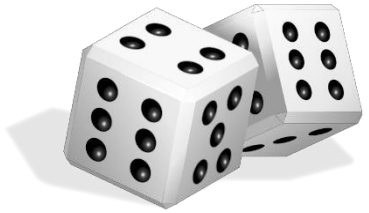
Observations:

- Sequence of roles:

  3 5 3 1 3 6 3 6 4 4 1 6 2 …

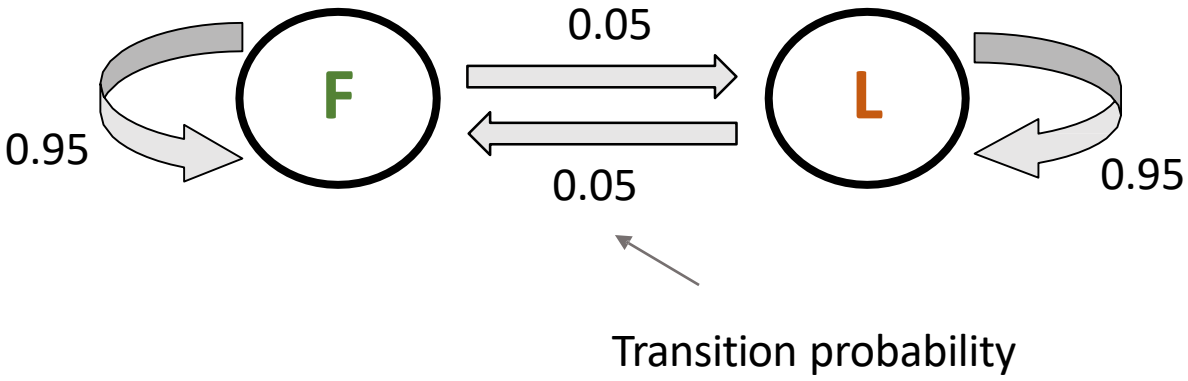
Question:

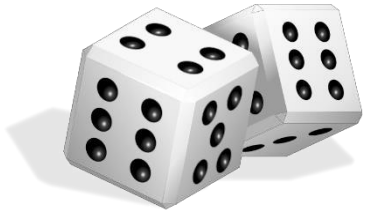
- Which dice used for each role?



Dishonest Casino Example

Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	1/2



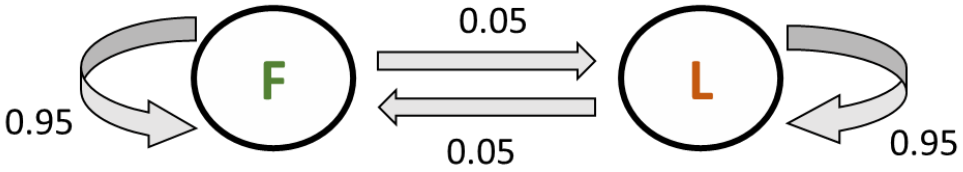


Dishonest Casino Example

Observation:

Sequence of roles:

Obs: 3 1 **6** 2 5 2 3 1 3 **6** 3 **6 6** 4 **6 6** 2 **6** …



Hidden information:

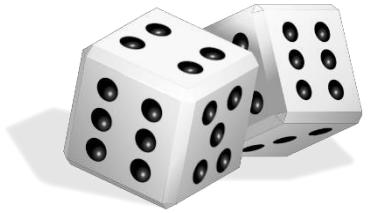
Sequence of states, e.g.

S1: F F F F F F F F F L L L L L L L L L….

S2: F F F F F F F F F F F F F F F F F F….

S3: L L L F F F F F F L L L L L L L L L…..

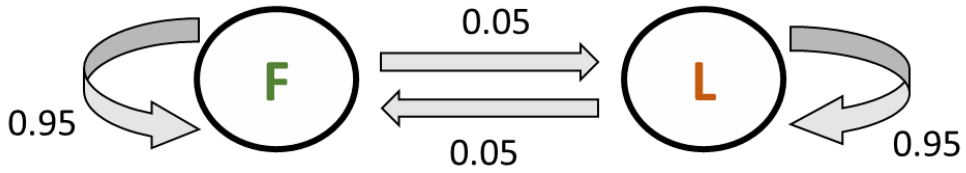
Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**



Dishonest Casino Example

Obs: 3 1 **6** 2 5 2 3 1 3 **6** 3 **6** **6** 4 **6** **6** 2 **6**

S1: F F F F F F F F F L L L L L L L L L



Transition to L state

P(Obs|S1) = 1/6 * 0.95 * 1/6 * 0.95 … * 0.05 * 1/2 * 0.95 * 1/10 * 0.95 * 1/2 … = 3.4e-14

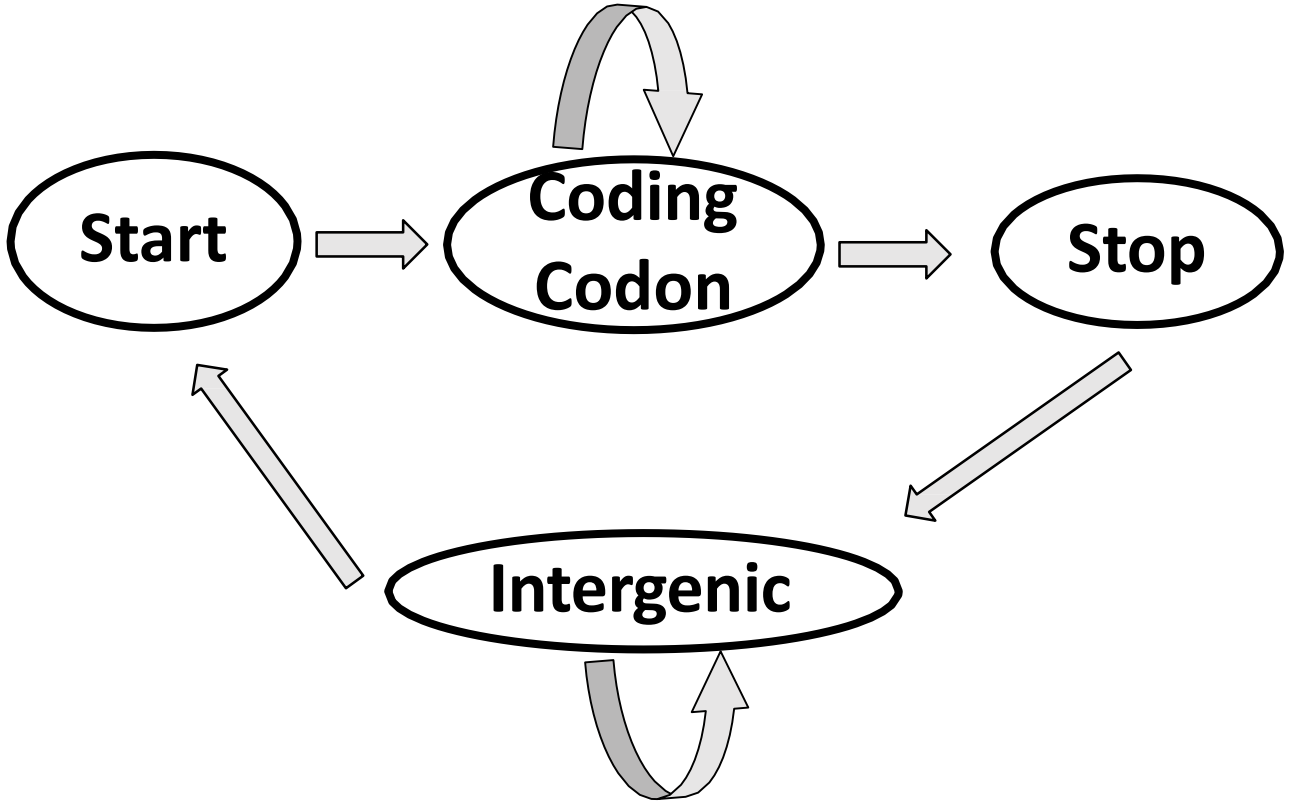
P(Obs|S2) = 4.1e-15

…..

Aim: Identify most likely path through model, which is S1 in this case, 9 roles fair dice, 9 roles loaded dice

Number	Probability Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**

Simple HMM for Gene Identification in Prokaryotes



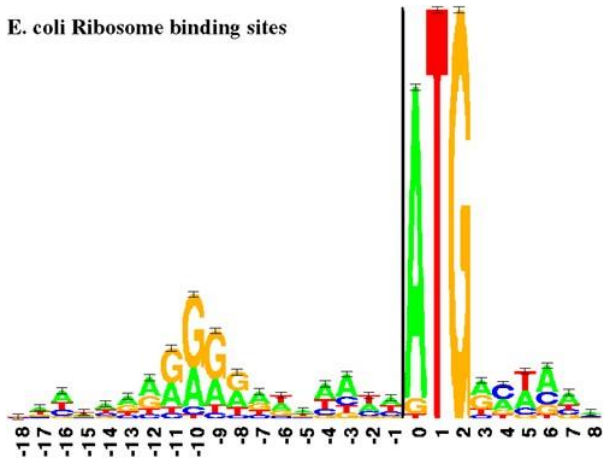
Training: Train model to learn codon frequencies of coding and non-coding sequences

Classification: Given observed DNA sequence, find most likely path through model to divide sequence into coding and non-coding regions

…CCTATC **ATG** GCT ATC GAC GAA AAC AAA … **TAA** CCTTATACTAG…

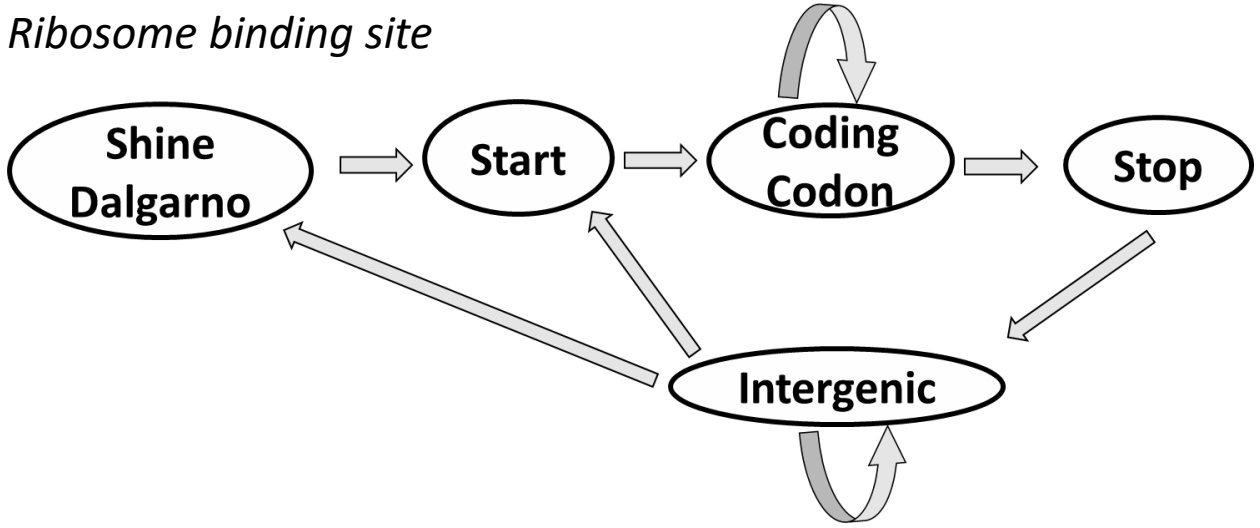
More Complex HMM for Gene Identification in Prokaryotes

Include signal for ribosomal binding site



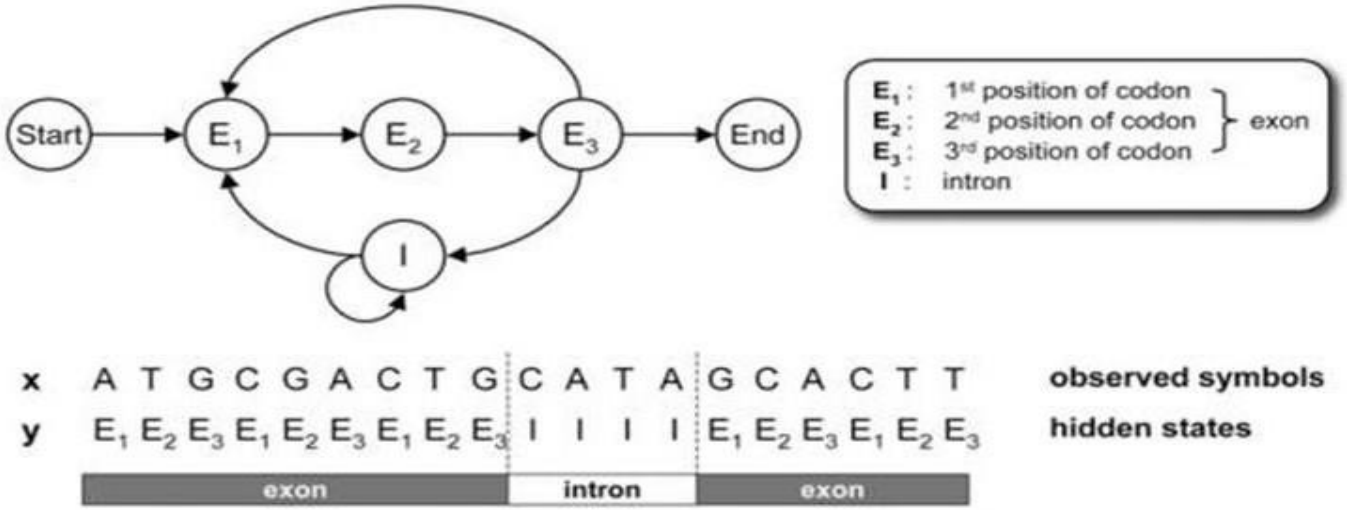
A/G-rich region about 10 bases upstream of the start codon
Helps recruit ribosome to mRNA

Ribosome binding site

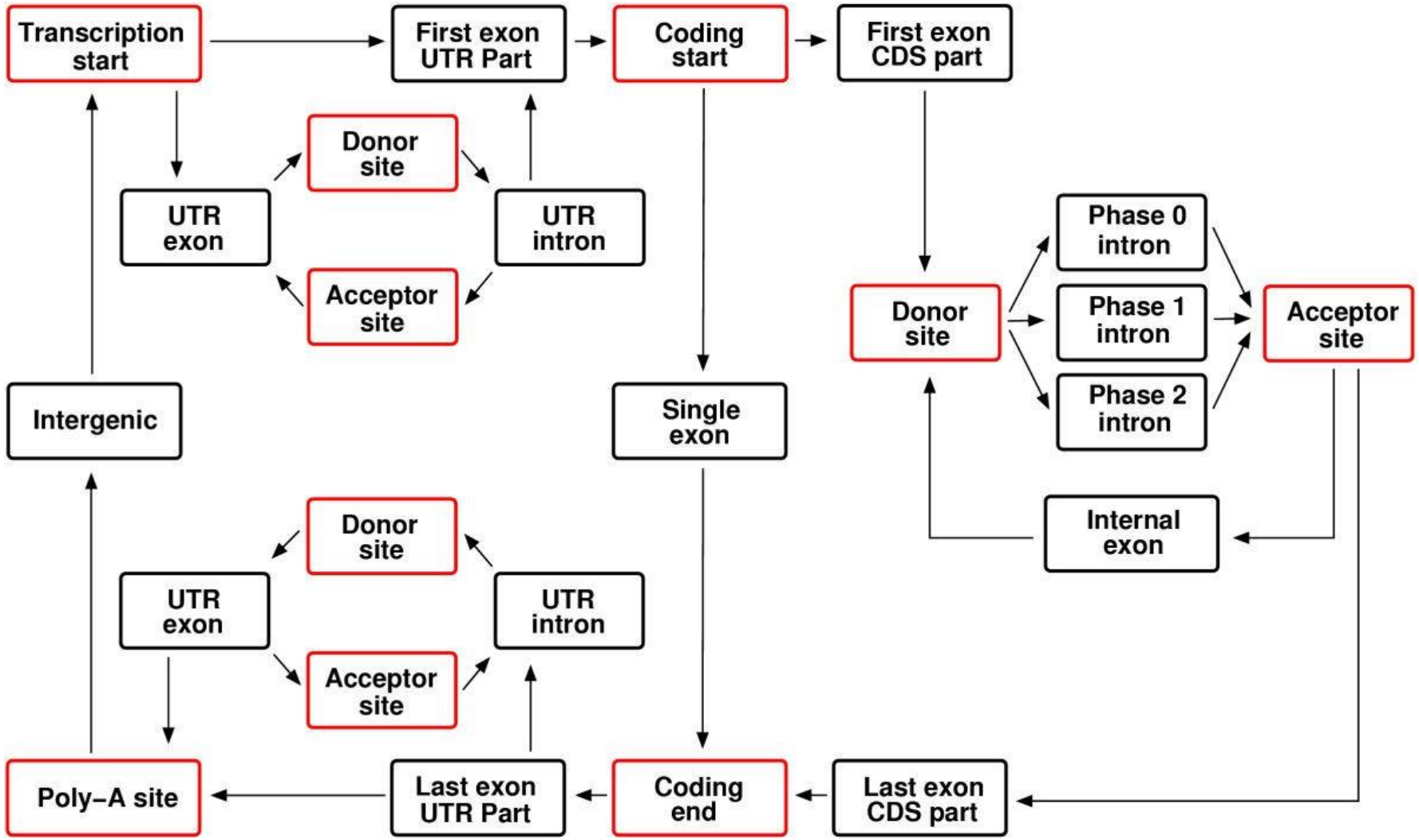


Gene Prediction in Eukaryotic Genomes

- **A Simple HMM for Modeling Eukaryotic Genes**



- hmmlearn() Python package. https://github.com/hmmlearn/hmmlearn

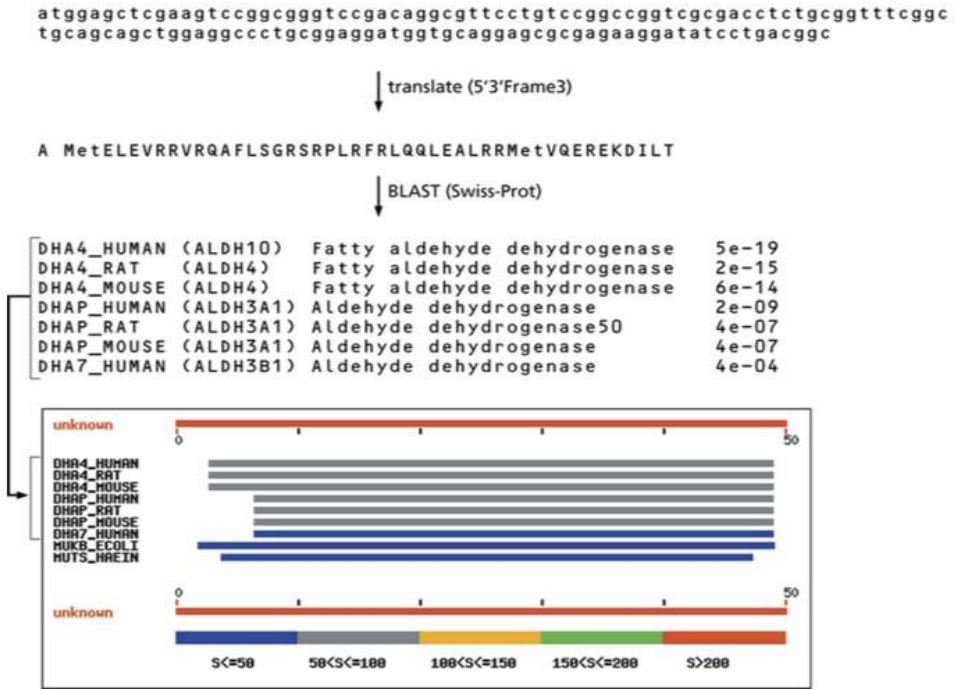

Transcription start
First exon UTR Part
Coding start
First exon CDS part
Donor site
UTR exon
UTR intron
Acceptor site
Phase 0 intron
Donor site
Phase 1 intron
Acceptor site
Phase 2 intron
Intergenic
Single exon
Internal exon
Donor site
UTR exon
UTR intron
Acceptor site
Poly–A site
Last exon UTR Part
Coding end
Last exon CDS part

Overview Characteristics that Can be Incorporated into HMM

- Coding regions begin with start codon and end with stop codon
- Coding regions do not contain stop codons
- Motifs of splice sites
- Sequence characteristics of UTRs
- Sequence characteristics of promoter regions
- Length distribution of coding regions, exons and introns
- polyA site

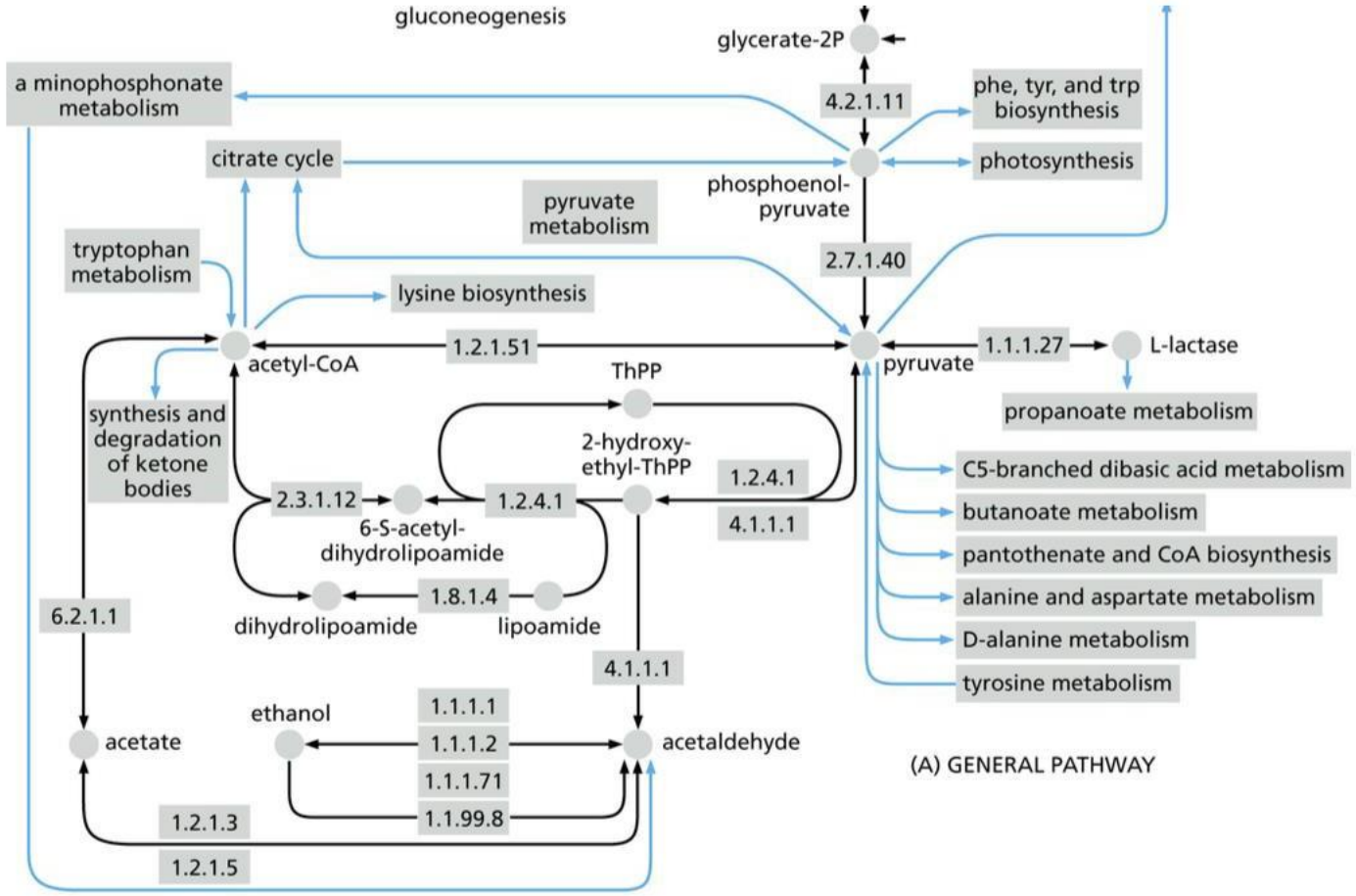
Genome annotation

- Once you have identified bona fide genes, what do you do with them?
- Use homology searchers (e.g. BLAST) to identify putative functions for genes, as well as identify other genetic elements such as functional RNAs and repeats
- Gene ontology provides a standard vocabulary to describe gene functions
- Annotations can be done for a whole genome/set of genomes using online services (e.g. IMG-M, KAAS)



Pathway analysis

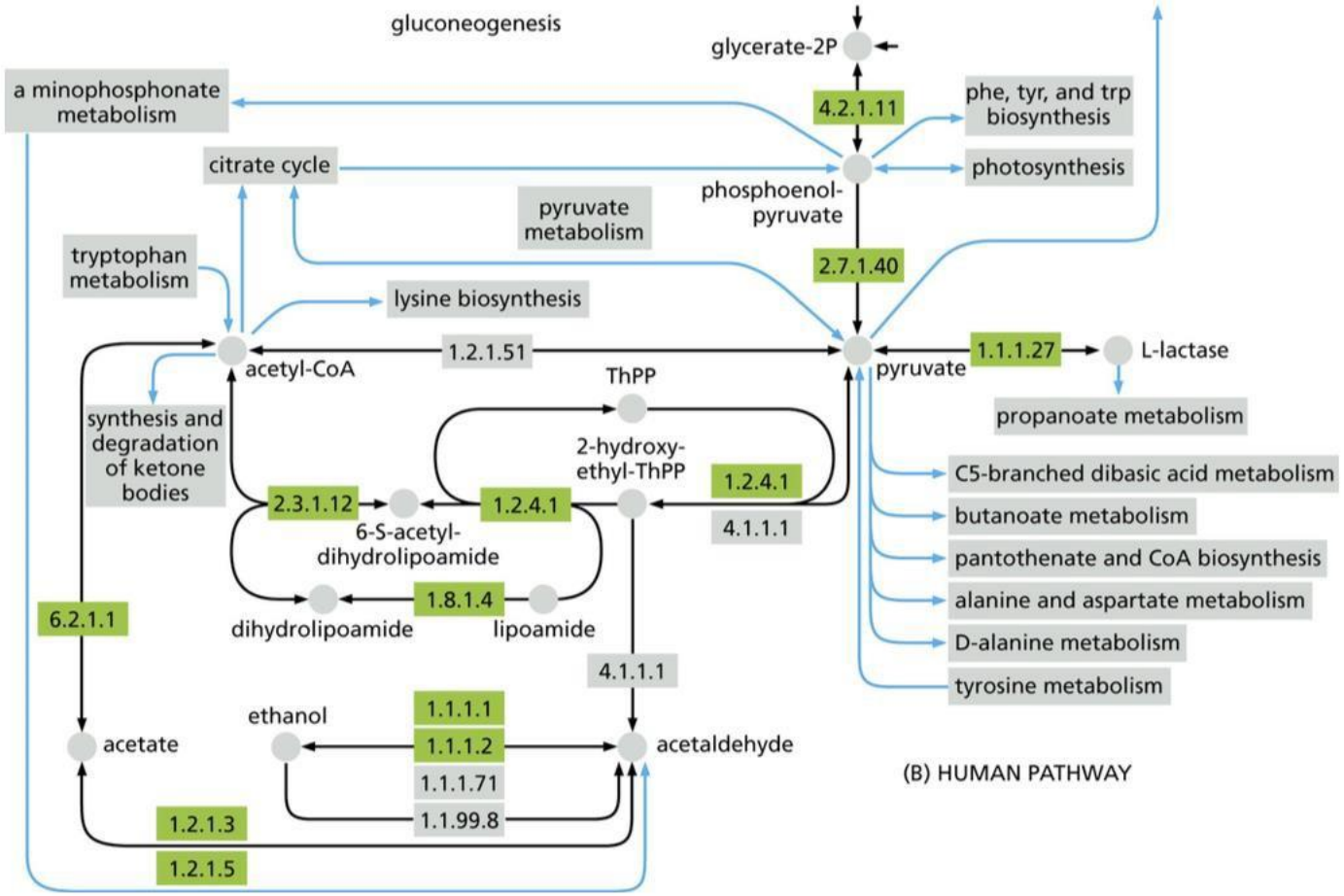
Once genes have been annotated, they are then fit into metabolic pathways to try and decipher the metabolism of the organism



Each gray box with a number represents an enzyme/protein in the pathway

Pathway analysis

We can color in the boxes based on which proteins we find in the organism



Also, if you have most of the proteins in the pathway, but one or two are missing, it suggests functions for some of your “unknown”/hypothetical proteins

Pathway Databases

- KEGG: Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/)
- Reactome (https://reactome.org/)
- GO: Gene Ontology (http://geneontology.org/)

Take-Home points

- Eukaryotic gene-finding is more difficult because of the presence of introns and exons, and also because exons are short and some exons may have no coding sequence
- Exon finding can be aided by using transcriptional signals and splice sites
- Determining correct starting positions and length of exons is essential as all predicted exons must eventually be linked together to form a total coding sequence
- Once you have found genes, they can be annotated and used to infer metabolism
- Practical things you might want to know:
  - How to identify ORFs
  - Differences between prokaryotic and eukaryotic genes, and
  - Bioinformatics strategies for eukaryotic gene finding
  - How to read pathway maps

Genome Analysis Lectorial

Atefeh Taherian Fard

Australian Institute for Bioengineering and Nanotechnology

The University of Queensland

BINF6000 I SCIE2100 I Bioinformatics I – Introduction

Summary from the Two Genome Analysis lectures

- **Lecture 1:**
- Overview genome sequencing and sequencing technologies
- Genome re-sequencing
- De-novo genome assembly

- **Lecture 2:**
- Gene features in prokaryotes
- Gene features in eukaryotes
- Computational approaches for gene prediction
- Functional genome annotation

Outline for today's Lectorial

- Overview of De novo and genome re-sequencing
- Identifying structural variations
- Gene feature in prokaryote and eukaryotes
- Computational approaches for gene finding
  - Support Vector Machine (SVM)
  - Hidden Markov model
- Past exam questions

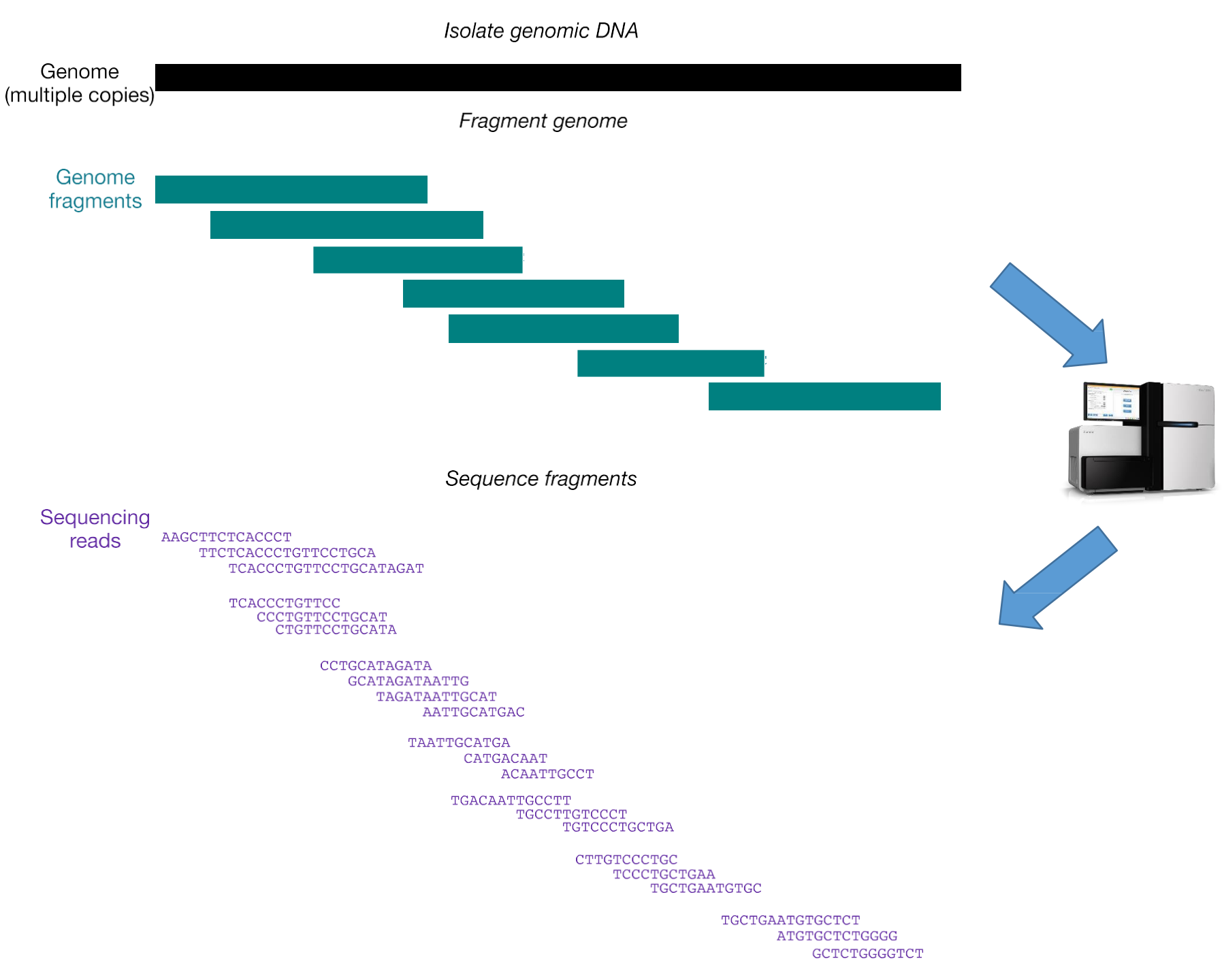
Why Do We Sequence Genomes?

Genome resequencing:

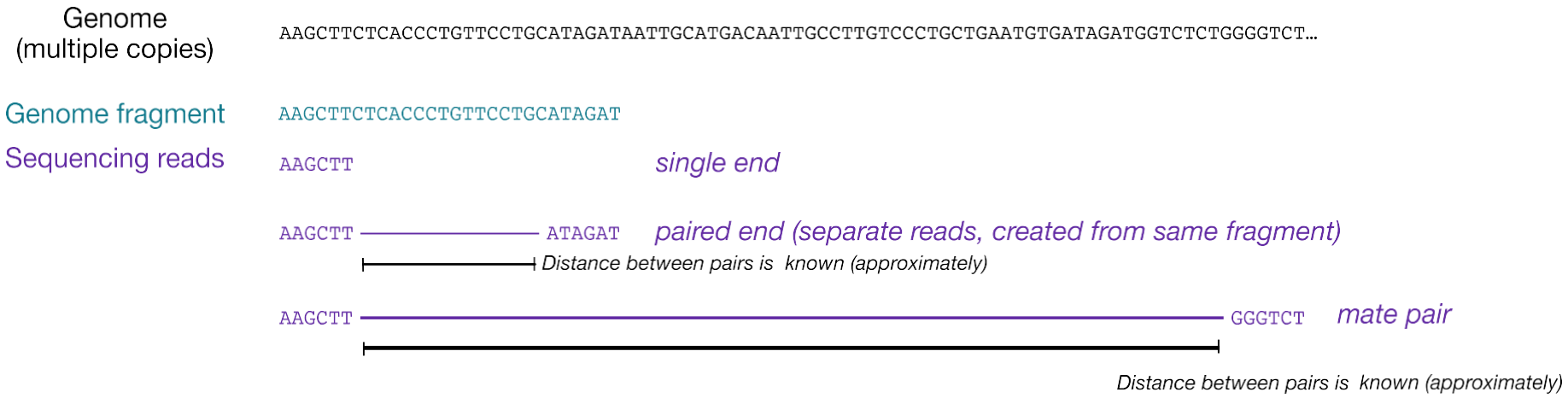
- Characterise genotype-phenotype associations
- Understand genetics of complex diseases
- Genome-based diagnosis; non-invasive prenatal testing
- Personalised medicine, genome-based prediction of optimal treatment (e.g. targetable mutations in cancer) and side-effects. Future: optimal diet, metabolic deficiencies, disease risk, …

De-novo sequencing

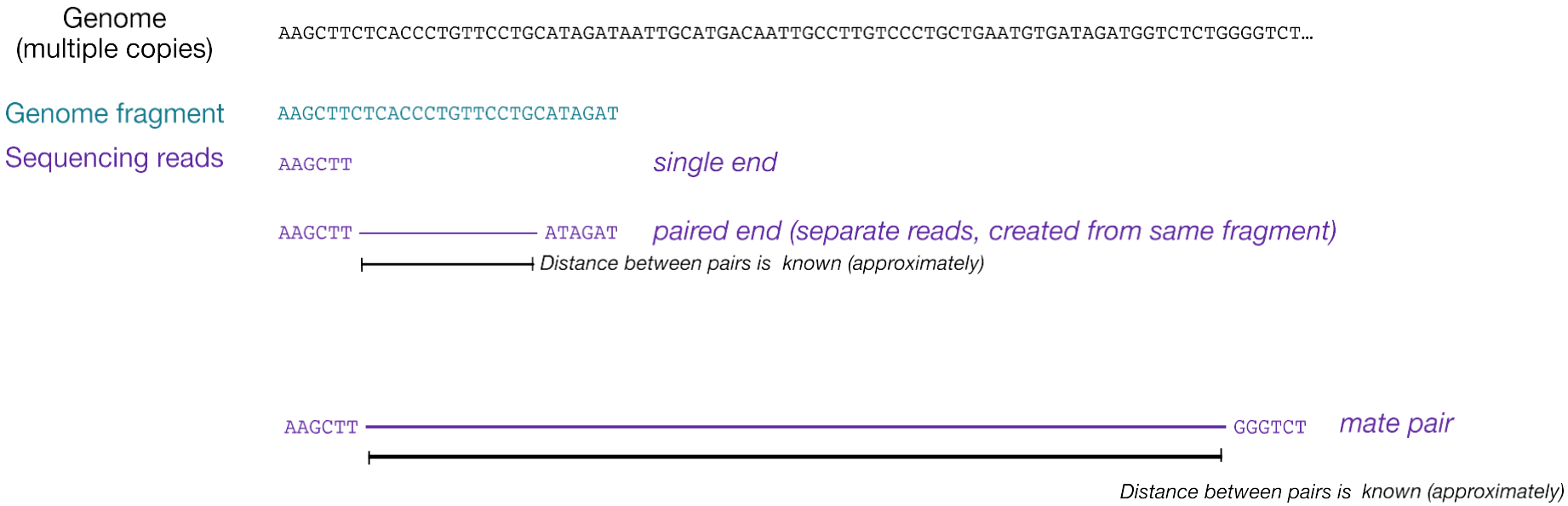
- Understand molecular biology of organisms, identify genes, gene functions, encoded pathways, metabolic capabilities, gene regulation and genome evolution



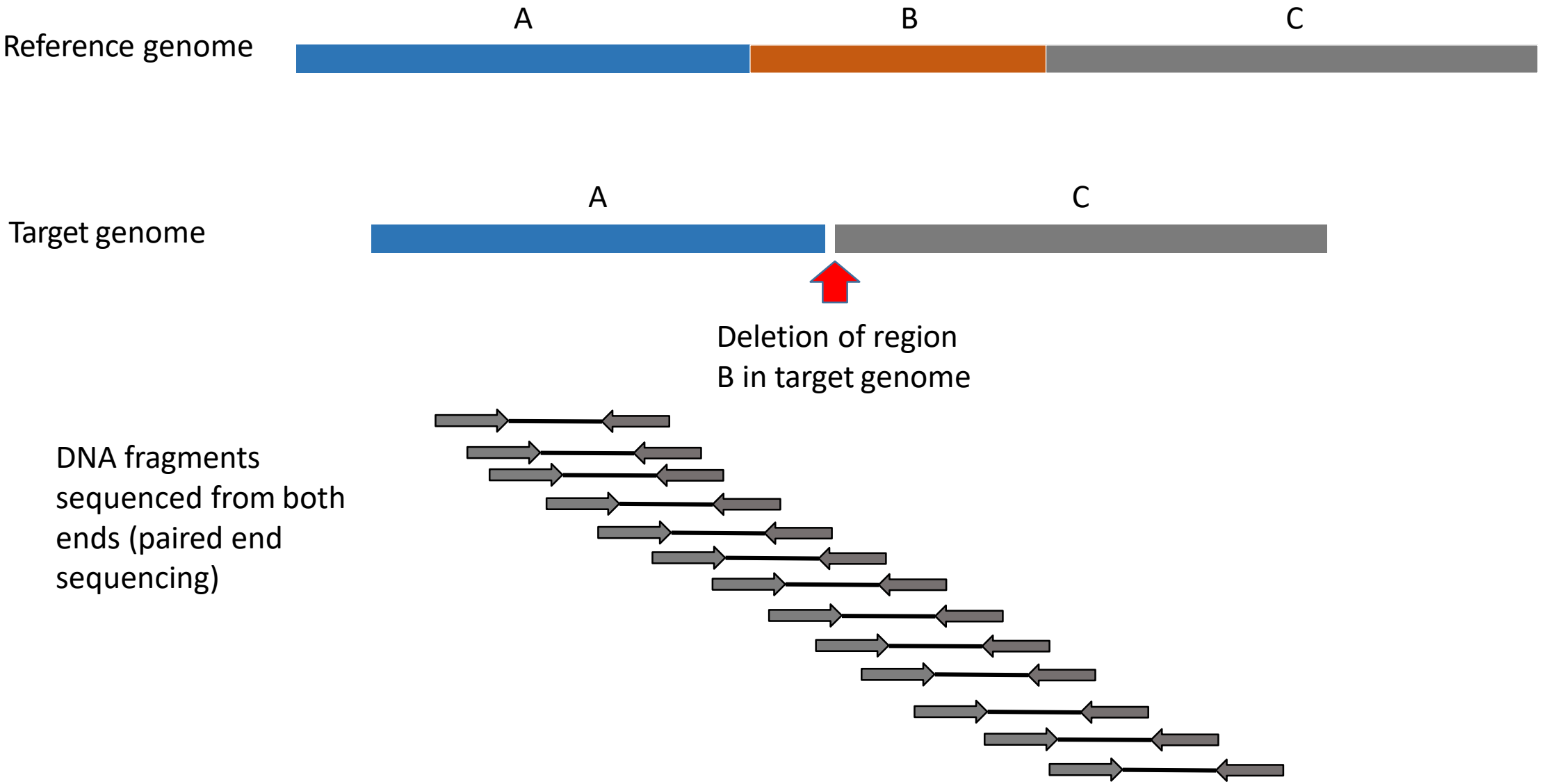
This approach is called ‘Shot Gun’ sequencing


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)

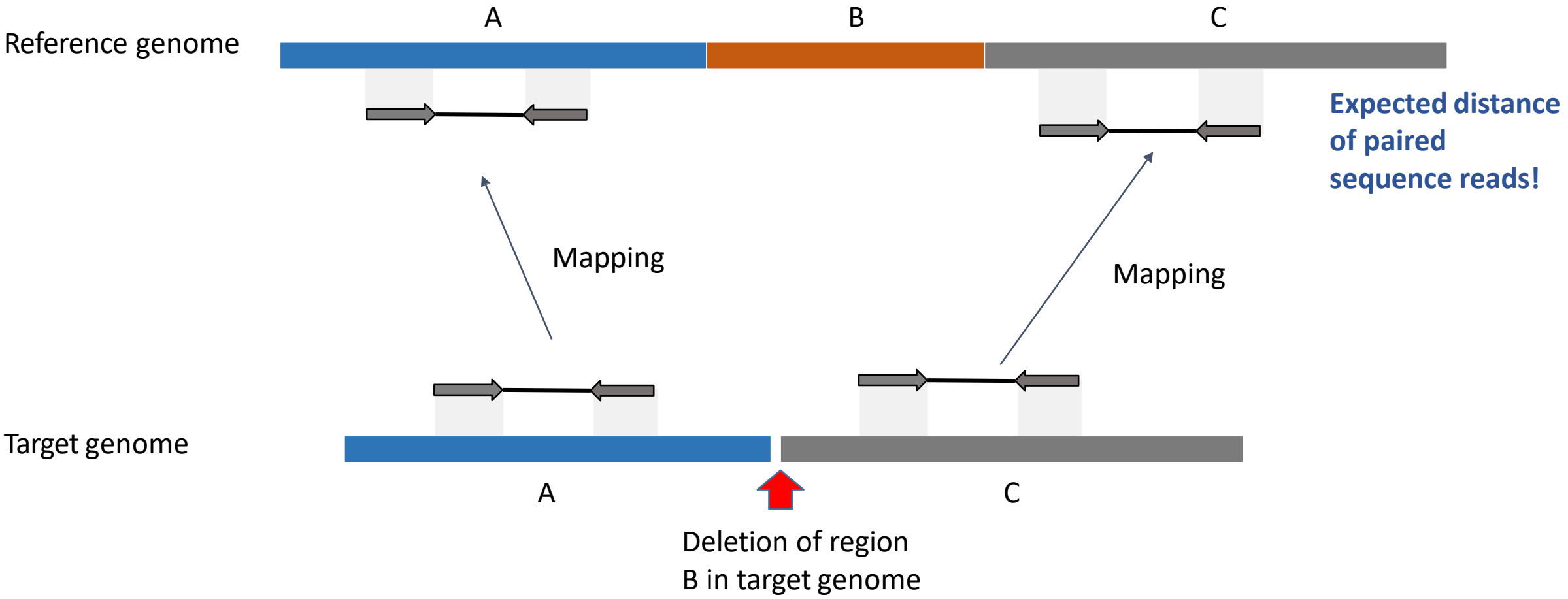

Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
ATAGAT
ATAGAT


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)

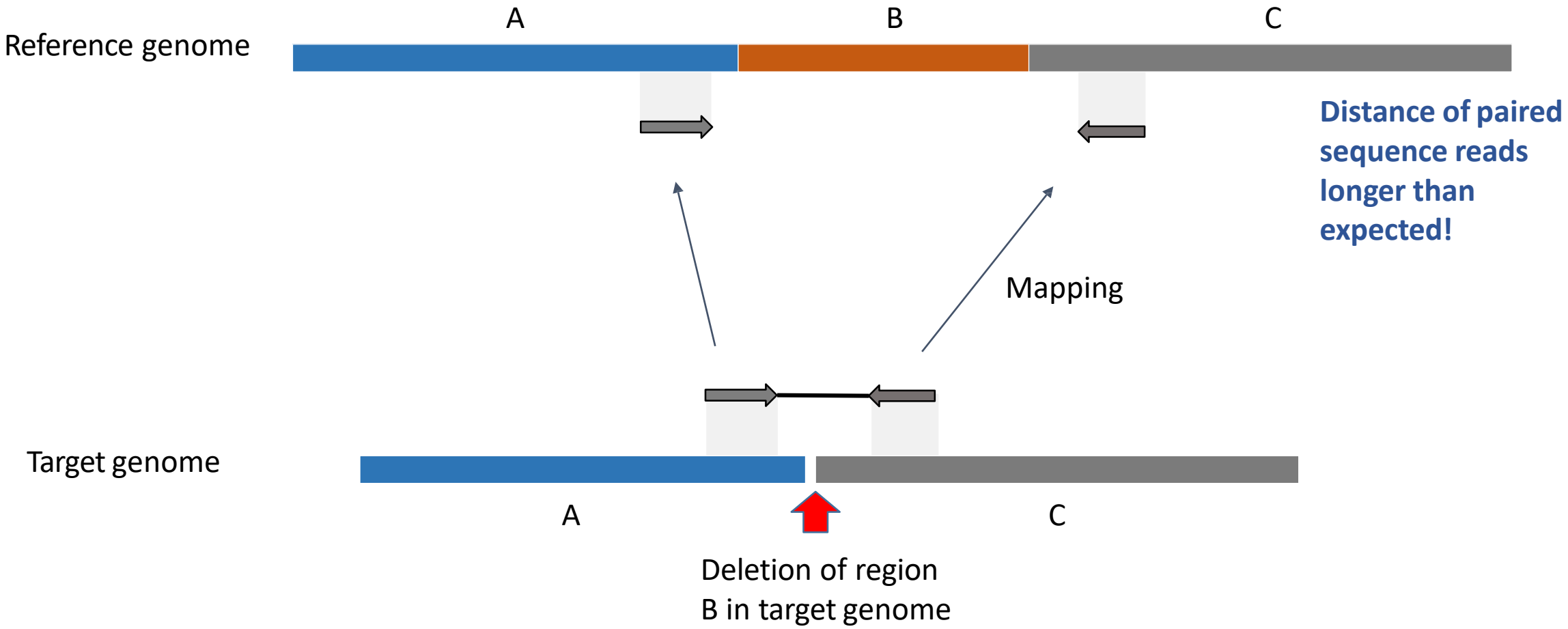
Structural Variations



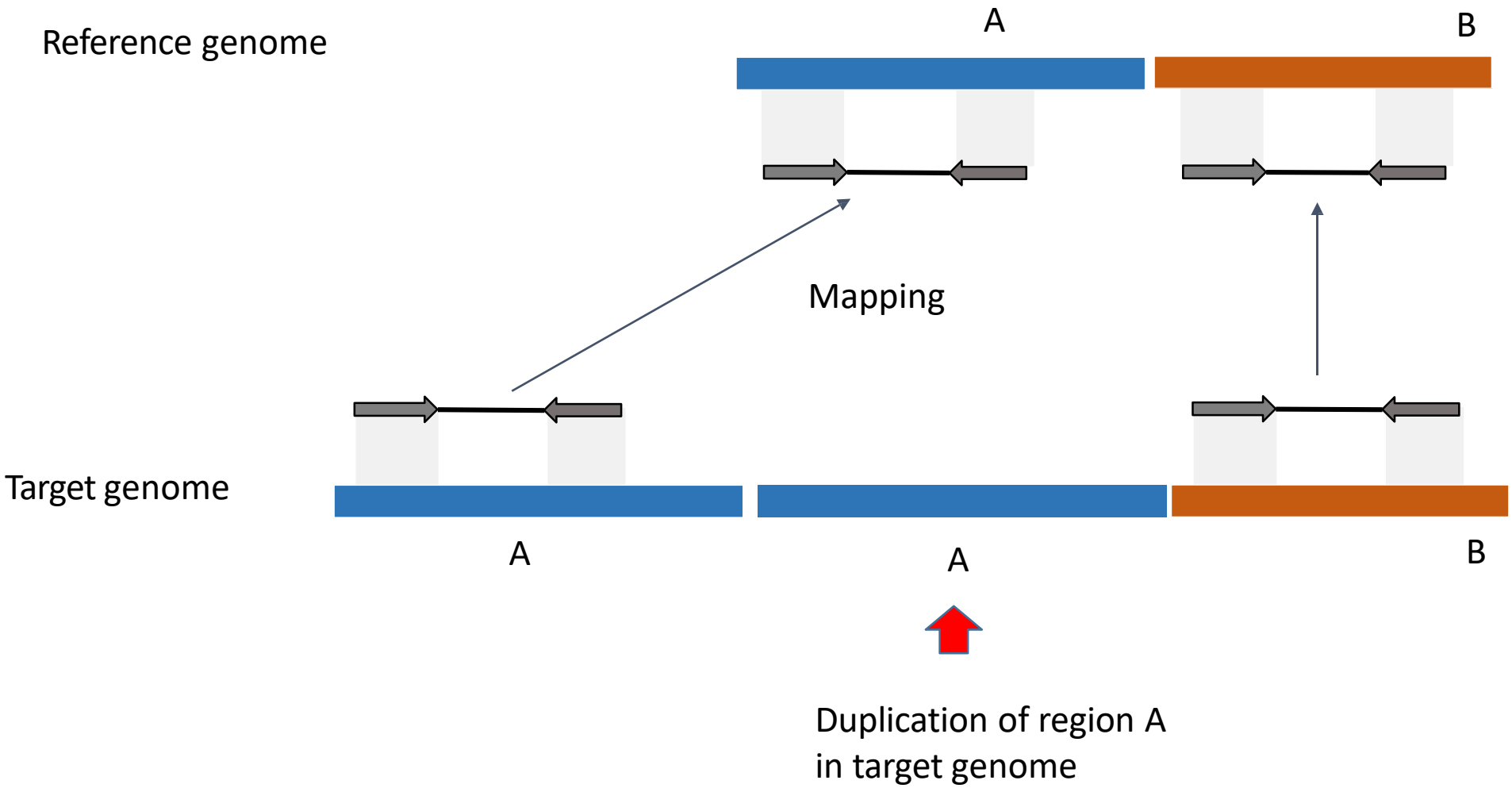
Structural Variations



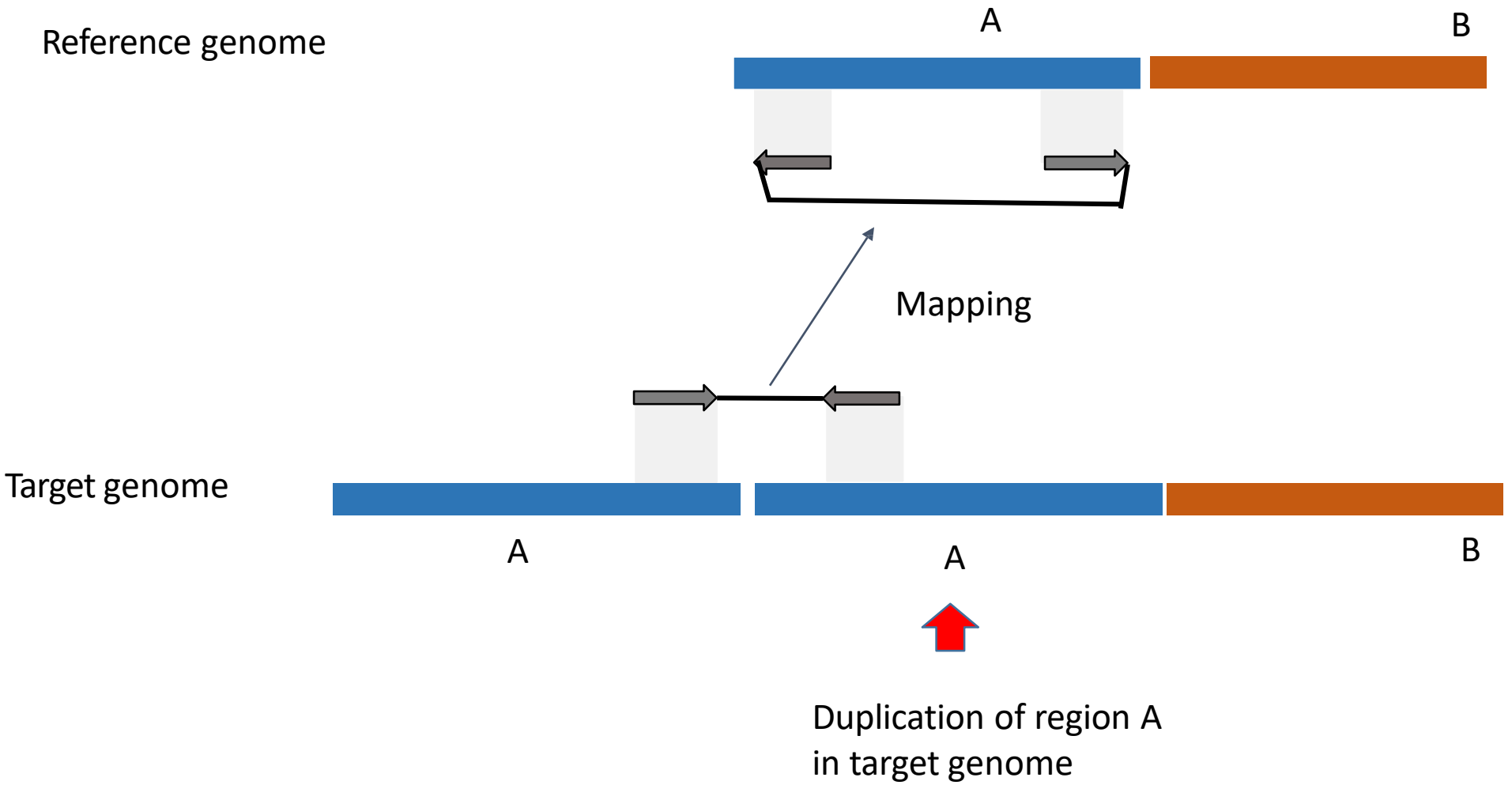
Deletion


Reference genome
A
B
C
Distance of paired sequence reads longer than expected!
Mapping
Target genome
A
C
Deletion of region
B in target genome

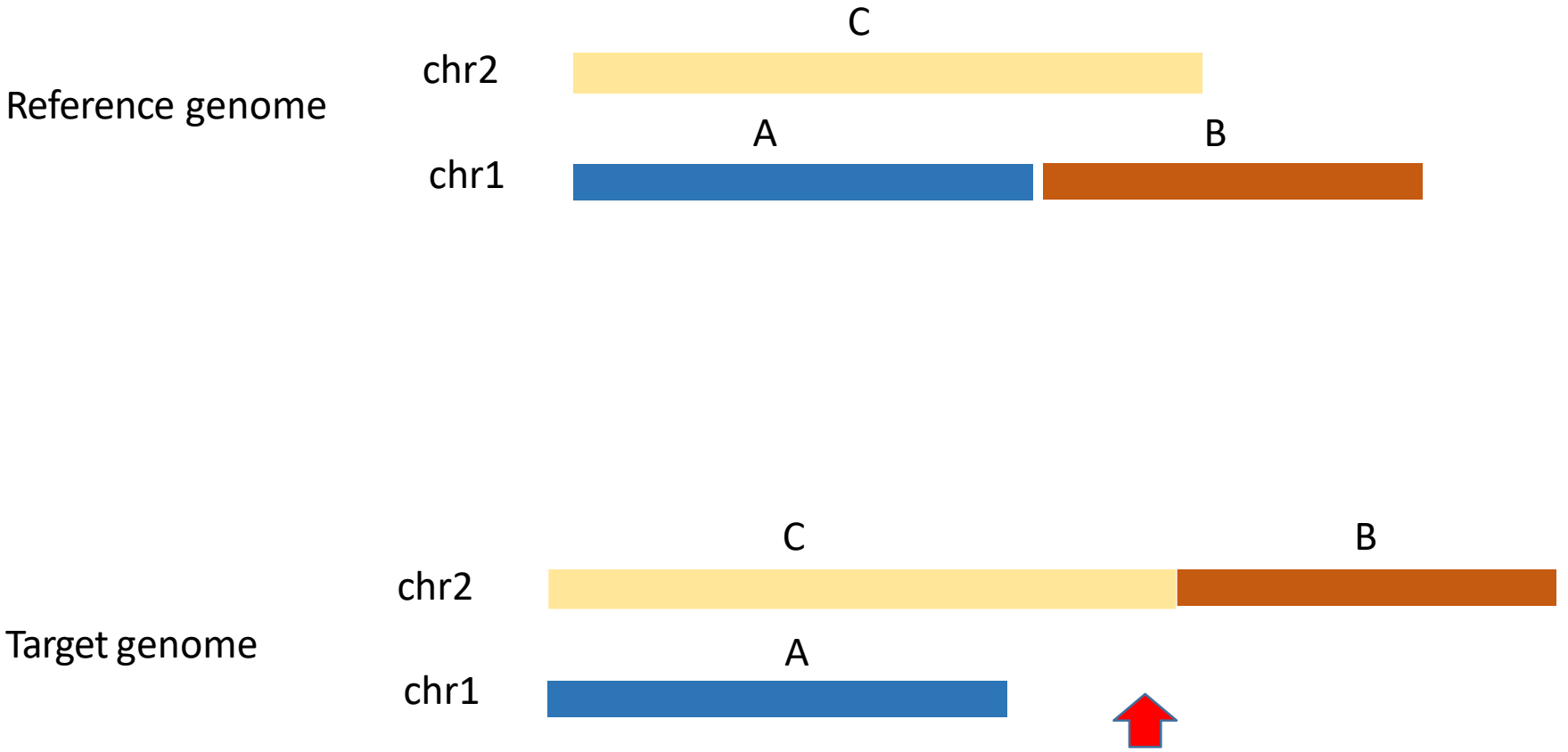
Duplications



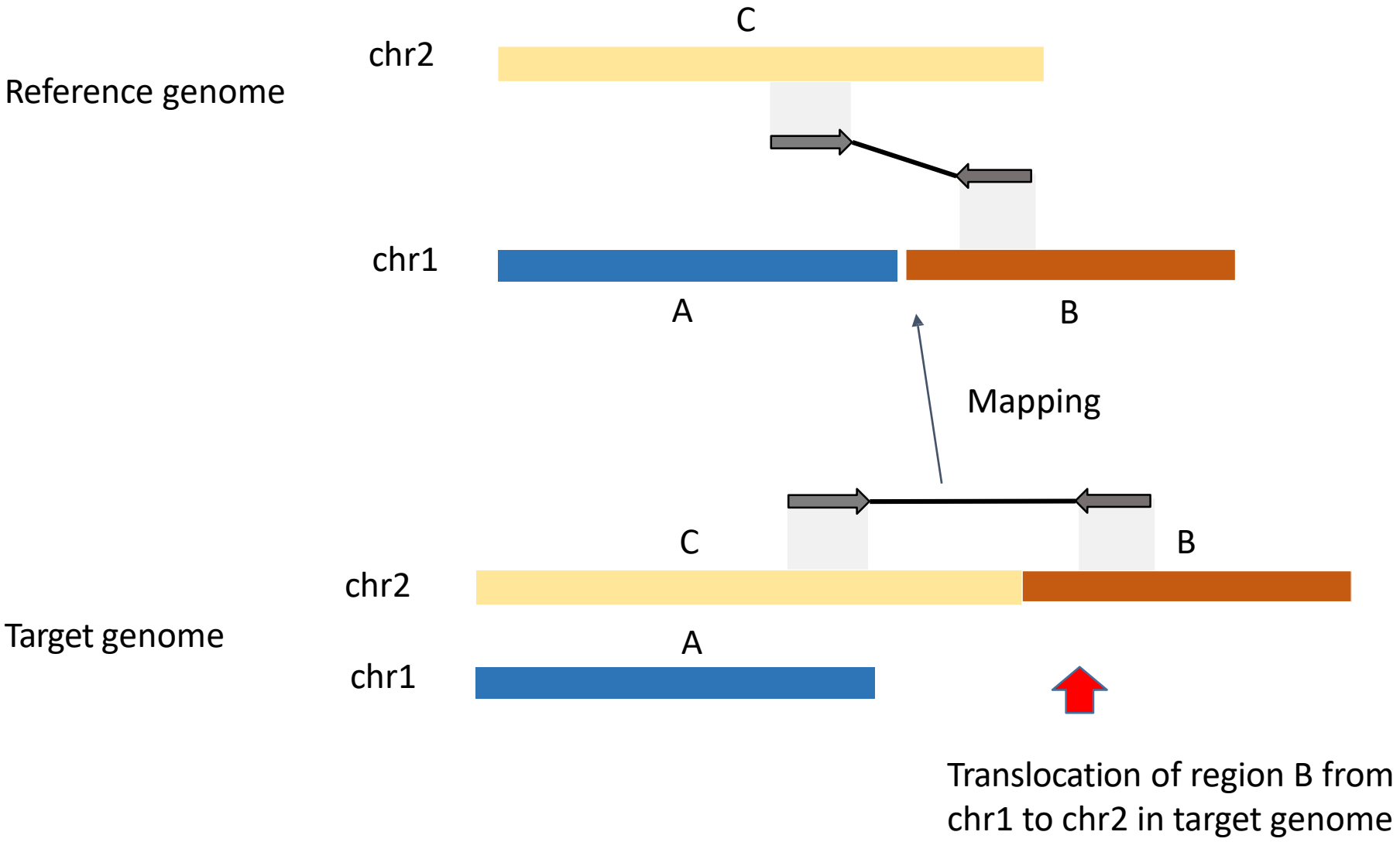
Duplications



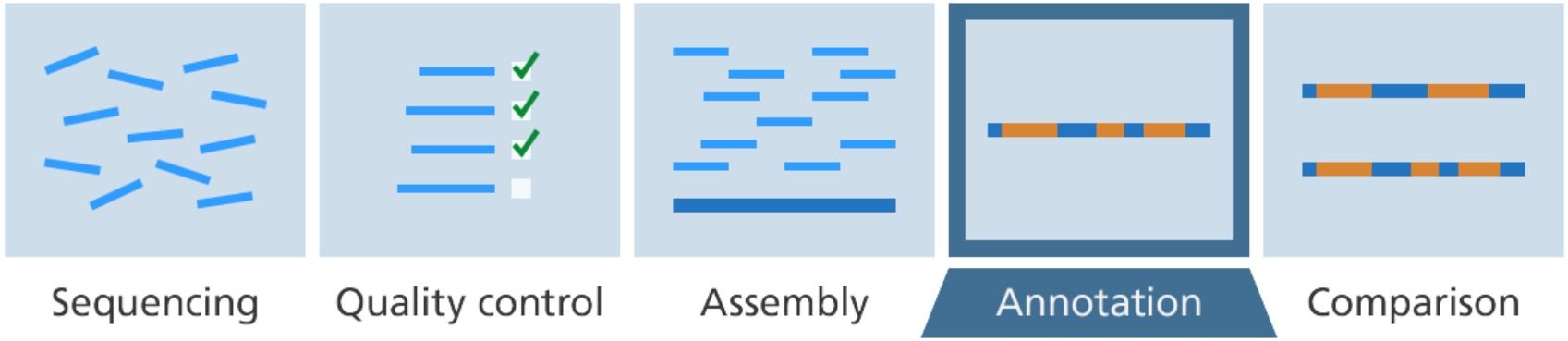
Translocations



Translocations



www.yourgenome.org



Using computational methods, find all genes (or other elements) in a long, unannotated string of nucleotides.

```
ACCGGTCAATAGCCGCAGACTACGGCATTTCAGAGGGACAGGCACTATAGCAACTAGCAACCCCCGTATAATACAAGGAGGCT
CAAGCTCCACTCTGACTCTCAACTTATTACGCTGTCACTCGATACGGCAGGGGCATTTAGACTTACGGCATATACCCGCCCGA
TCCAGCTTACGATACTACTGCTACTGGATACCCTGATAGCCAATCATTACGACTACTACTACGGCATTTCAGACCCGACAGGC
ACTAGAGCAACTAGCAACACCCGTATAATACAAGGAGGCTCAAGCTCCAGCTCTCACTGCAGCTATGTGGTGCACACATGTGC
ATCGTATGACTCAGTCGATGCTATCACGTACATCGTGTGGGTGCACACCACCCATGCCCTGATAGCCCCTGATTTTAGCCCCA
GCATTATTTTTCCGACGAGATCACGTACCCCCTACGGCATTTCAGAGGGACAGGGGACGCGCCCAATTACGACTACTACTACG
GCATTTCAGACCCGACAGGCACTAGAGCAACTAGCAACACCCGTATAATACAAGGAGGCTCAAGCTCCAGCCTTTTCAACAGA
CCGGGCGTTACGGTAAAAAAAAATCCGGCCGTACGGACTACTGGATACCGCAGACTACGGCATTTCAGAGGGACAGGCACTAT
AGCAACTAGCAACCCCCGTATAATACAAGGAGGCTCAAGCTCCACTCTGACTCTCAACTTATGACAGGGGACGATGACTCAGT
CGATTTCGCTATCACGTAAAACATCGTGTGGGGTGCACACCACCGCATGCCCTTTCAGGATAGCCCCTGATTTTAAGCCCCAG
CATTATTTGGTTCCGACGAGATCACGTACCCCACTACGGCATTTCAGAGGGACACTCAGTCGATGCTATCACGTACATCGTGT
GGGTGCTTACACCACGCCATGCCCTGATAGCCCCTGGGGATTTTAGCCCCAGCATTAATTTCTTCCGACGAGCCCTCAGACCC
GACAGGGGCACTAGAGCAACTATATAAGCAACACCCGTATACCCATACCAAGGAGGCTCAAGCTCCAGCCTTCTTCAACAGGA
CCGGGCGGGATTCCACATCATTCATGGGCAGCATCCCAGCAAACCCACGGCATAAGGACCACCCCTCGGCTAAGCAATCGCAT
AATACGGCGCTGCGCTCTACGTCTAGAGCTCACCATCTTACGAGGCTCTACCTCTATG...[3 BILLION or so MORE]
```

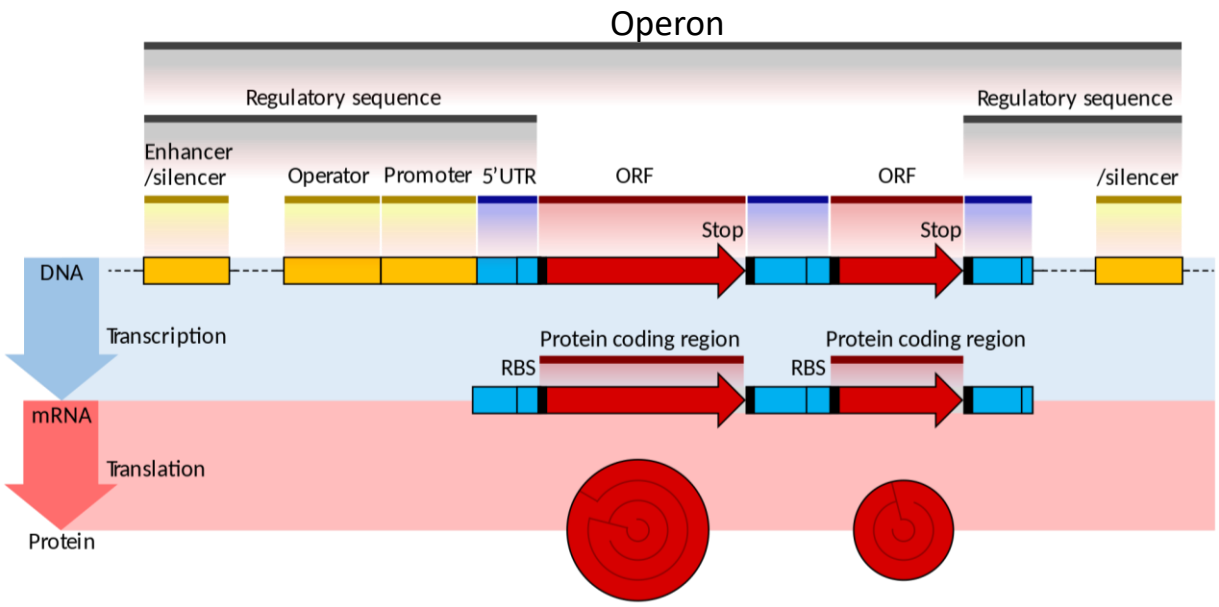
Aim: To identify transcriptional unit

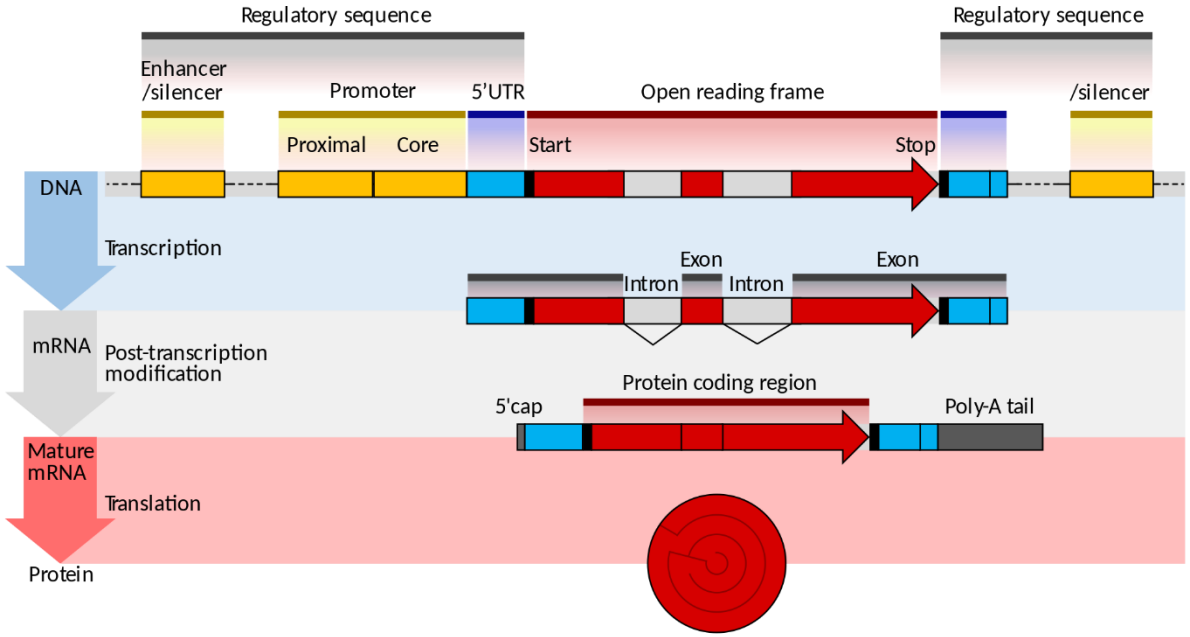
What do we know? Know only approximately what they look like

How? Find their locations and boundaries as accurately as possible, overlook as few as possible, and report as few non-genes as possible.

Adapted from Mark Ragan

Summary: Gene features in Prokaryotics vs Eukaryotes





- Codon bias and GC rich regions
- Transcriptional start and stop sites
- ORFs: Start and stop codons
- 5’ UTR: Ribosomal Binding Sequence site
- 3’UTR

- Codon bias and GC rich regions
- Promoter regions
- Intro and Exon splice site
- ORFs: Start and Stop codons
- 5’ UTR: 5’ Cap (G cap site)
- 3’ UTR: PolyAs

Shafee & Lowe (2017)

Gene finding Approaches

- Physical, genetic or other ***experimental approaches***
  - *e.g.* Genetic knockouts

- *Computational approaches*

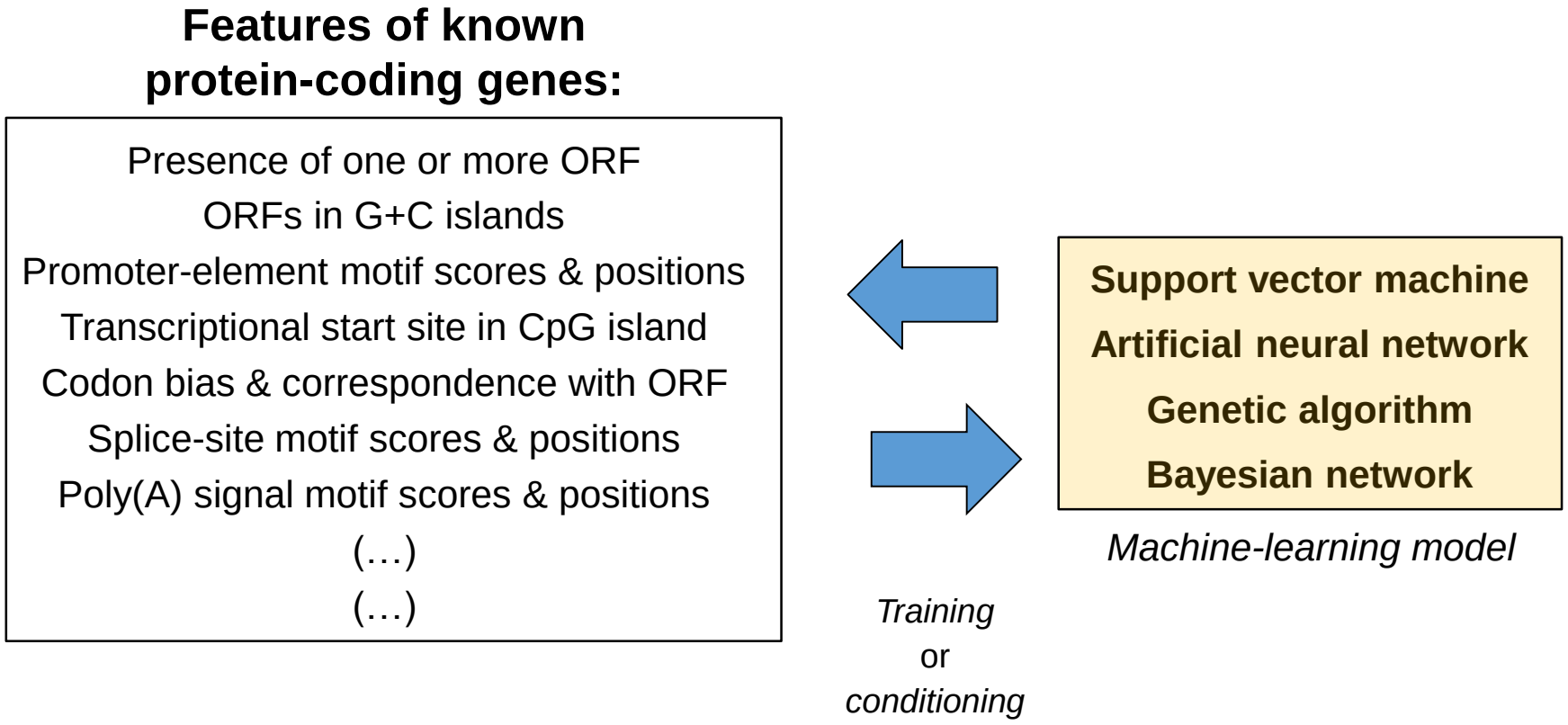
1) Identity search

2) Similarity search Homology based

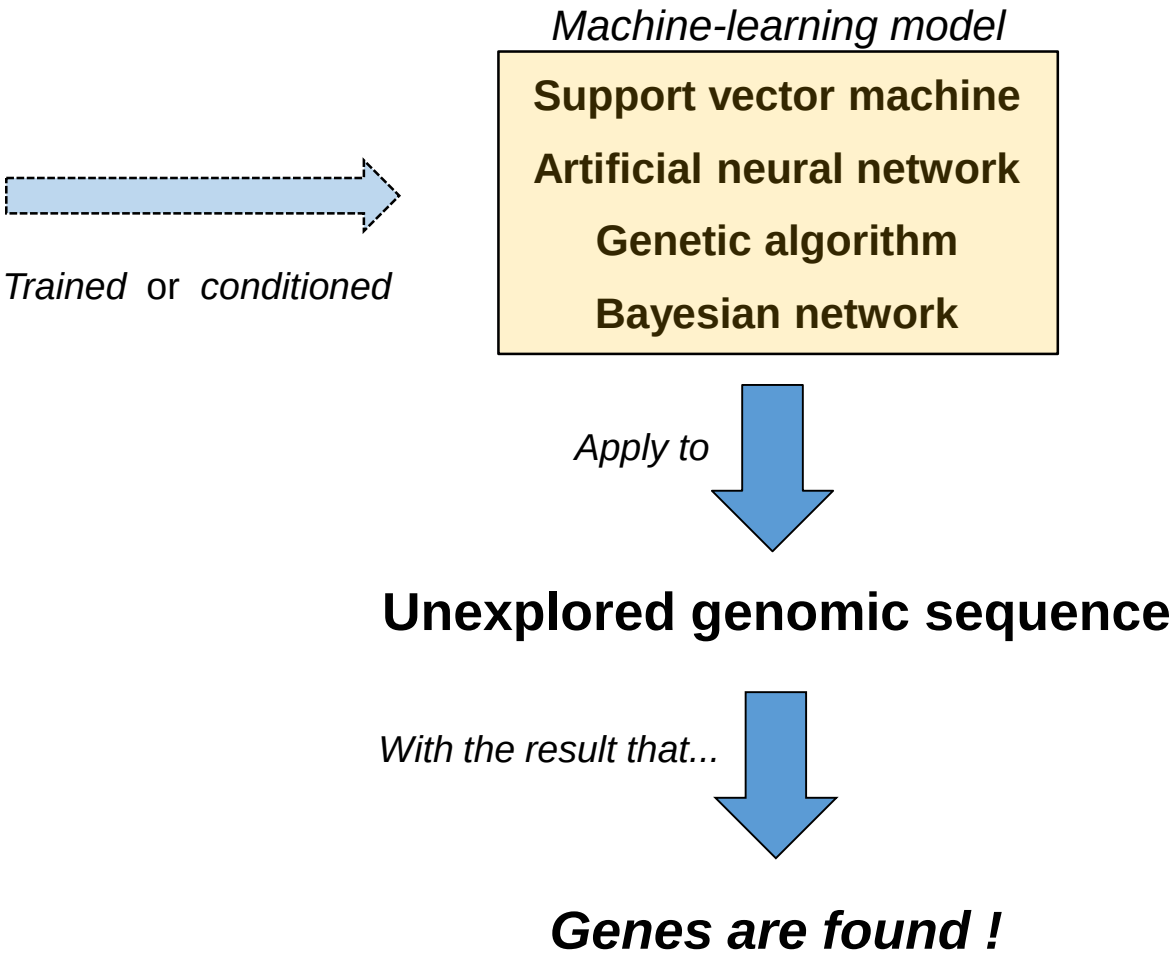
3) *Ab initio* approaches

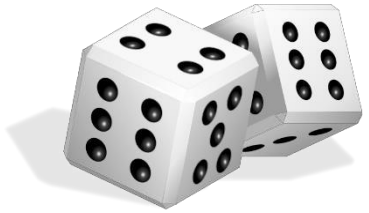
Slides adapted from Lutz Krause

Machine learning approach to ab initio gene finding



Machine learning approach to ab initio gene finding





Hidden Markov Models

The dishonest casino:

Known information:

- Casino has 2 die, fair dice, loaded dice
- Casino player switches back & forth between dies
- Once either of the dice is used, it will continue to be used for a while

Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**

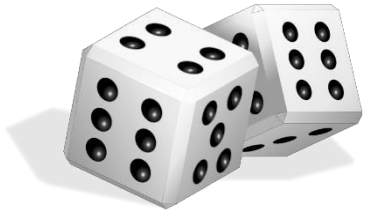
Observations:

- Sequence of roles:

  3 5 3 1 3 6 3 6 4 4 1 6 2 …

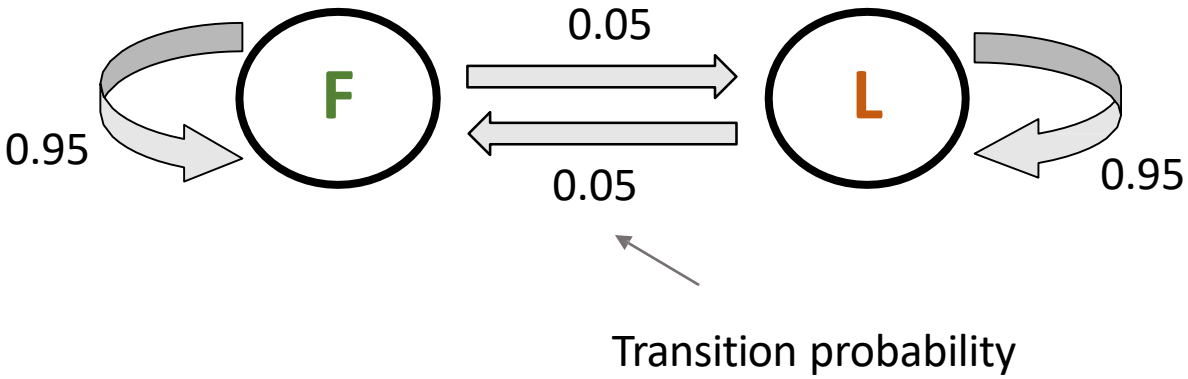
Question:

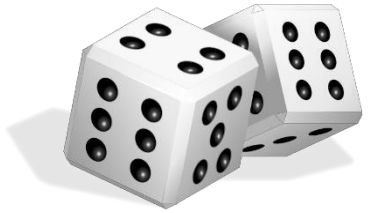
- Which dice used for each role?



Dishonest Casino Example

Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	1/2



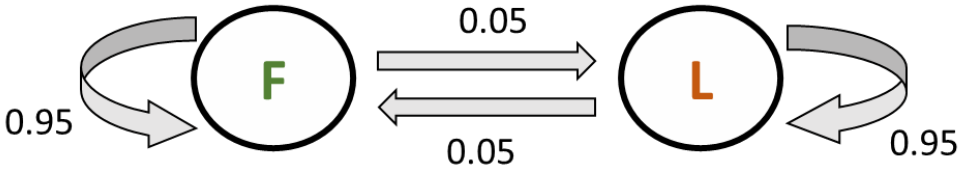


Dishonest Casino Example

Observation:

Sequence of roles:

Obs: 3 1 **6** 2 5 2 3 1 3 **6** 3 **6** **6** 4 **6** **6** 2 **6** …



Hidden information:

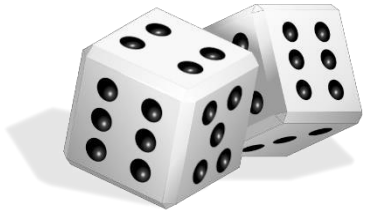
Sequence of states, e.g.

S1: F F F F F F F F F L L L L L L L L L….

S2: F F F F F F F F F F F F F F F F F F….

S3: L L L F F F F F F L L L L L L L L L…..

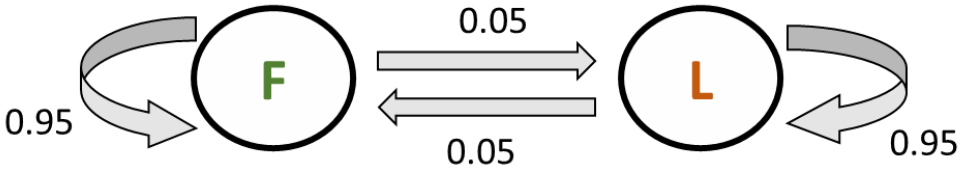
Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**



Dishonest Casino Example

Obs: 3 1 **6** 2 5 2 3 1 3 **6** 3 **6** **6** 4 **6** **6** 2 **6**

S1: F F F F F F F F F L L L L L L L L L



Transition to L state

P(Obs|S1) = 1/6 * 0.95 * 1/6 * 0.95 … * 0.05 * 1/2 * 0.95 * 1/10 * 0.95 * 1/2 … = 3.4e-14

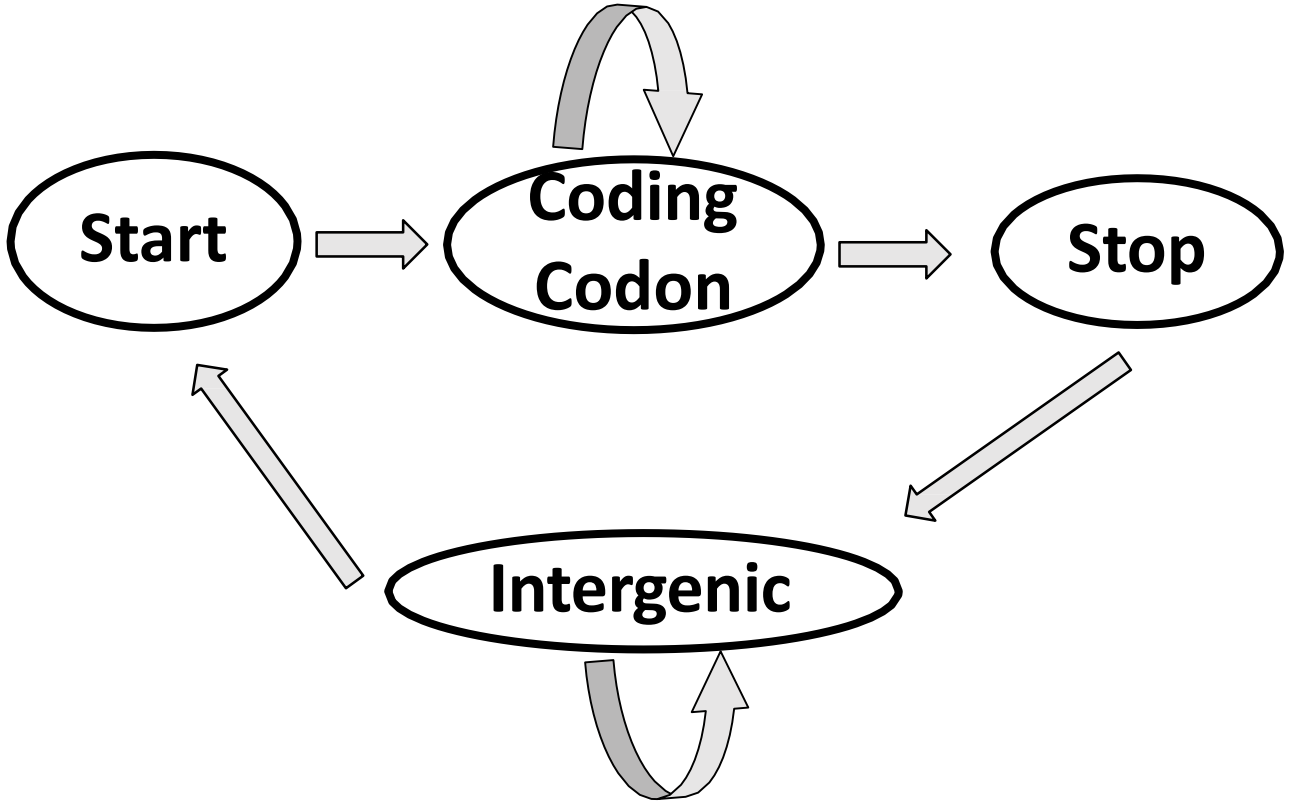
P(Obs|S2) = 4.1e-15

…..

Aim: Identify most likely path through model, which is S1 in this case, 9 roles fair dice, 9 roles loaded dice

Number	Probability Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**

Simple HMM for Gene Identification in Prokaryotes



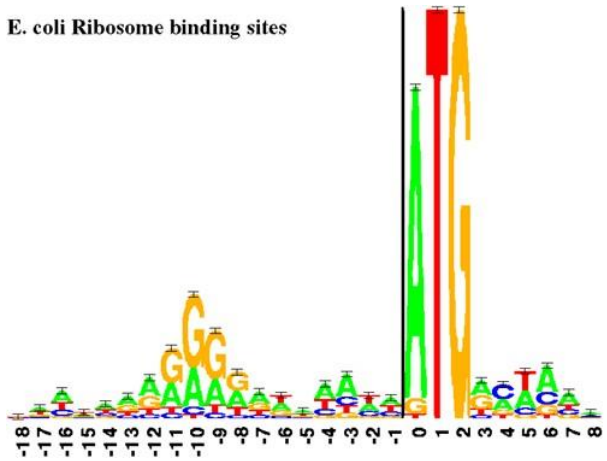
Training: Train model to learn codon frequencies of coding and non-coding sequences

Classification: Given observed DNA sequence, find most likely path through model to divide sequence into coding and non-coding regions

…CCTATC **ATG** GCT ATC GAC GAA AAC AAA … **TAA** CCTTATACTAG…

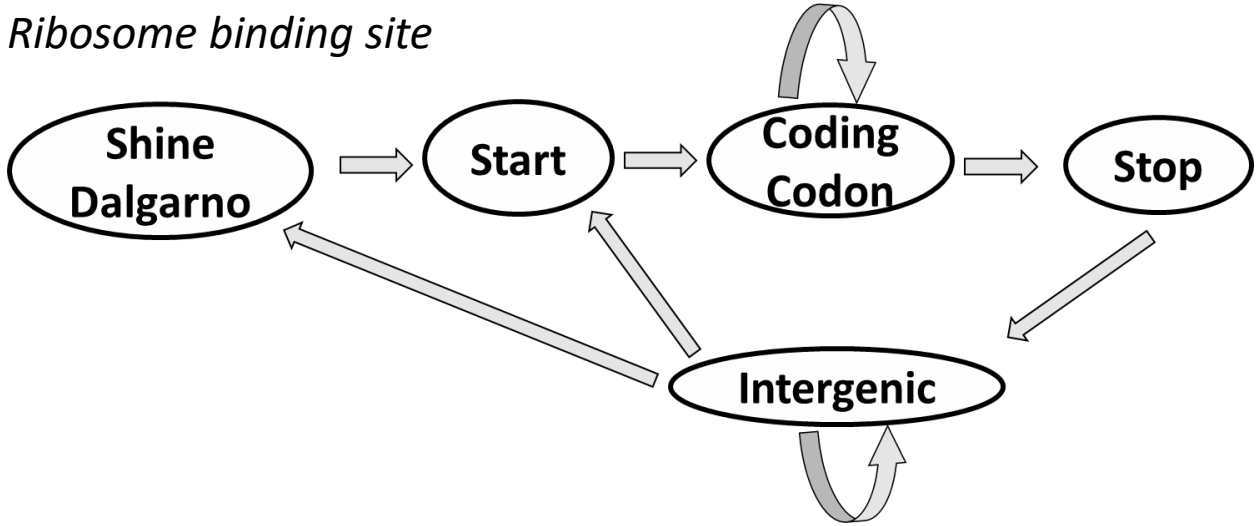
More Complex HMM for Gene Identification in Prokaryotes

Include signal for ribosomal binding site



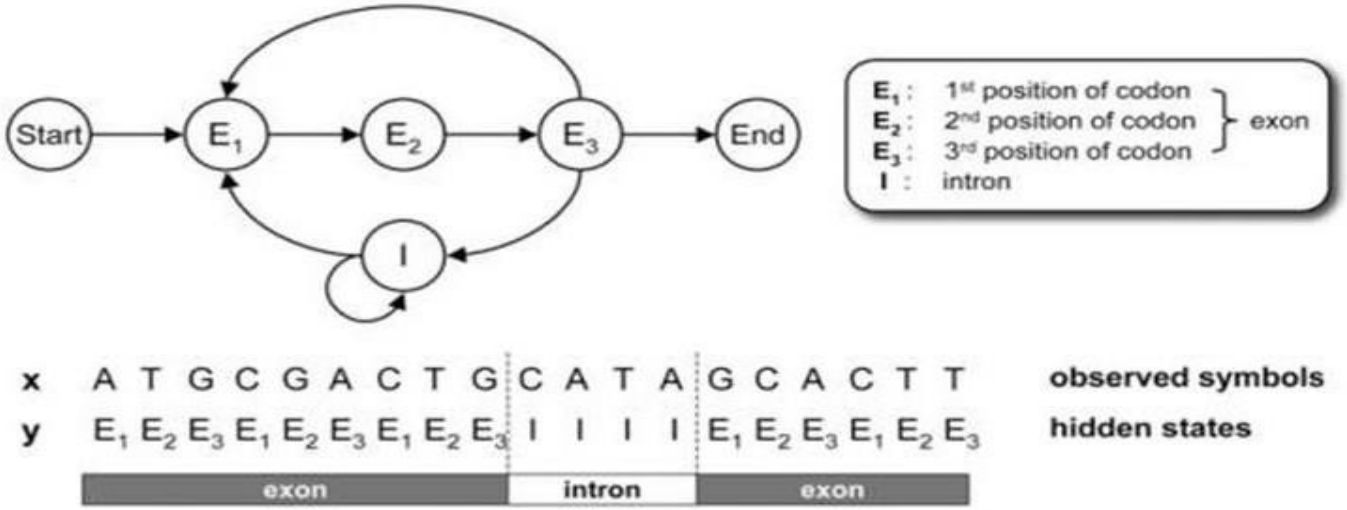
A/G-rich region about 10 bases upstream of the start codon
Helps recruit ribosome to mRNA

Ribosome binding site

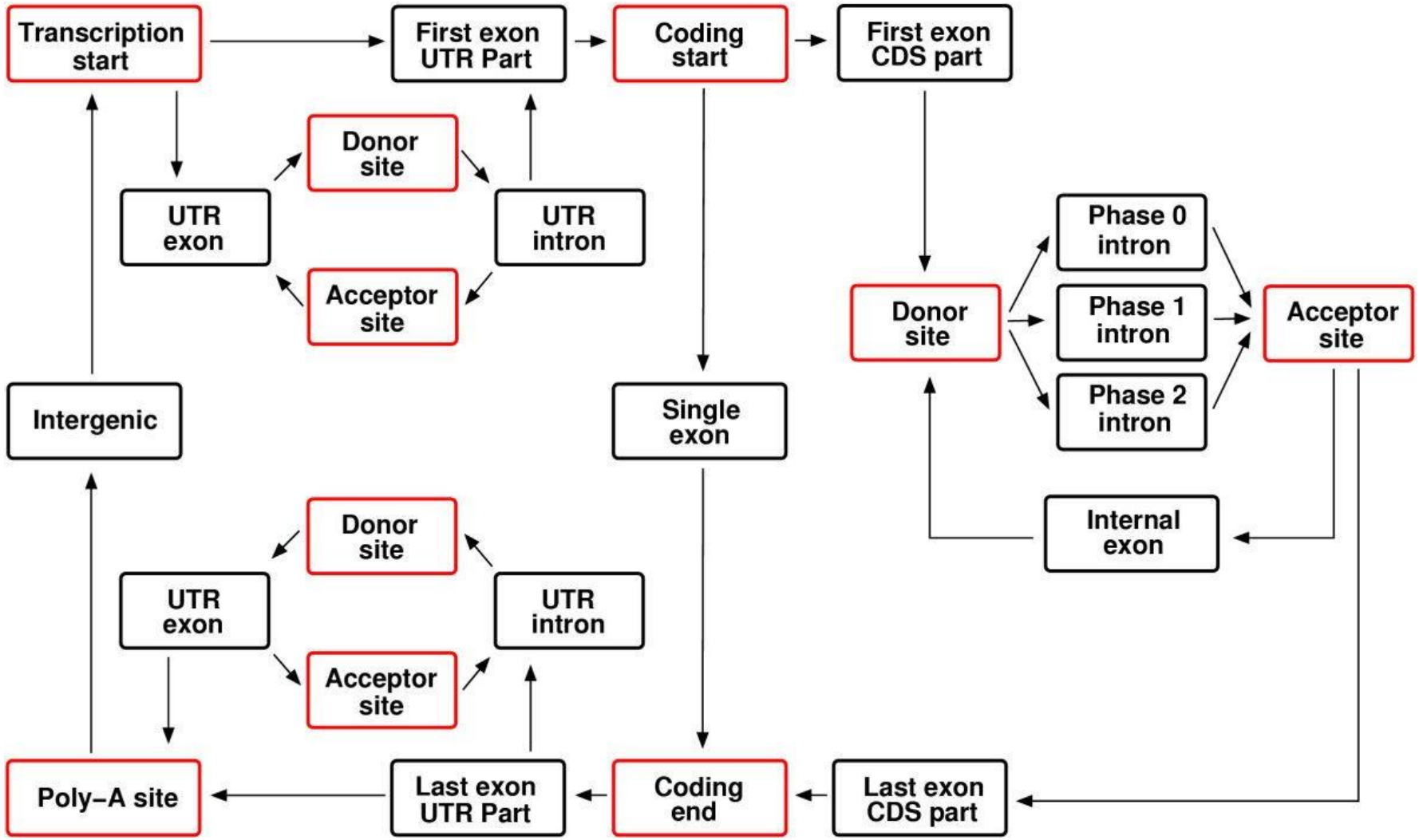


Gene Prediction in Eukaryotic Genomes

- **A Simple HMM for Modeling Eukaryotic Genes**



- hmmlearn() Python package. https://github.com/hmmlearn/hmmlearn


Transcription start
First exon UTR Part
Coding start
First exon CDS part
Donor site
UTR exon
UTR intron
Acceptor site
Phase 0 intron
Donor site
Phase 1 intron
Acceptor site
Phase 2 intron
Intergenic
Single exon
Internal exon
Donor site
UTR exon
UTR intron
Acceptor site
Poly–A site
Last exon UTR Part
Coding end
Last exon CDS part