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

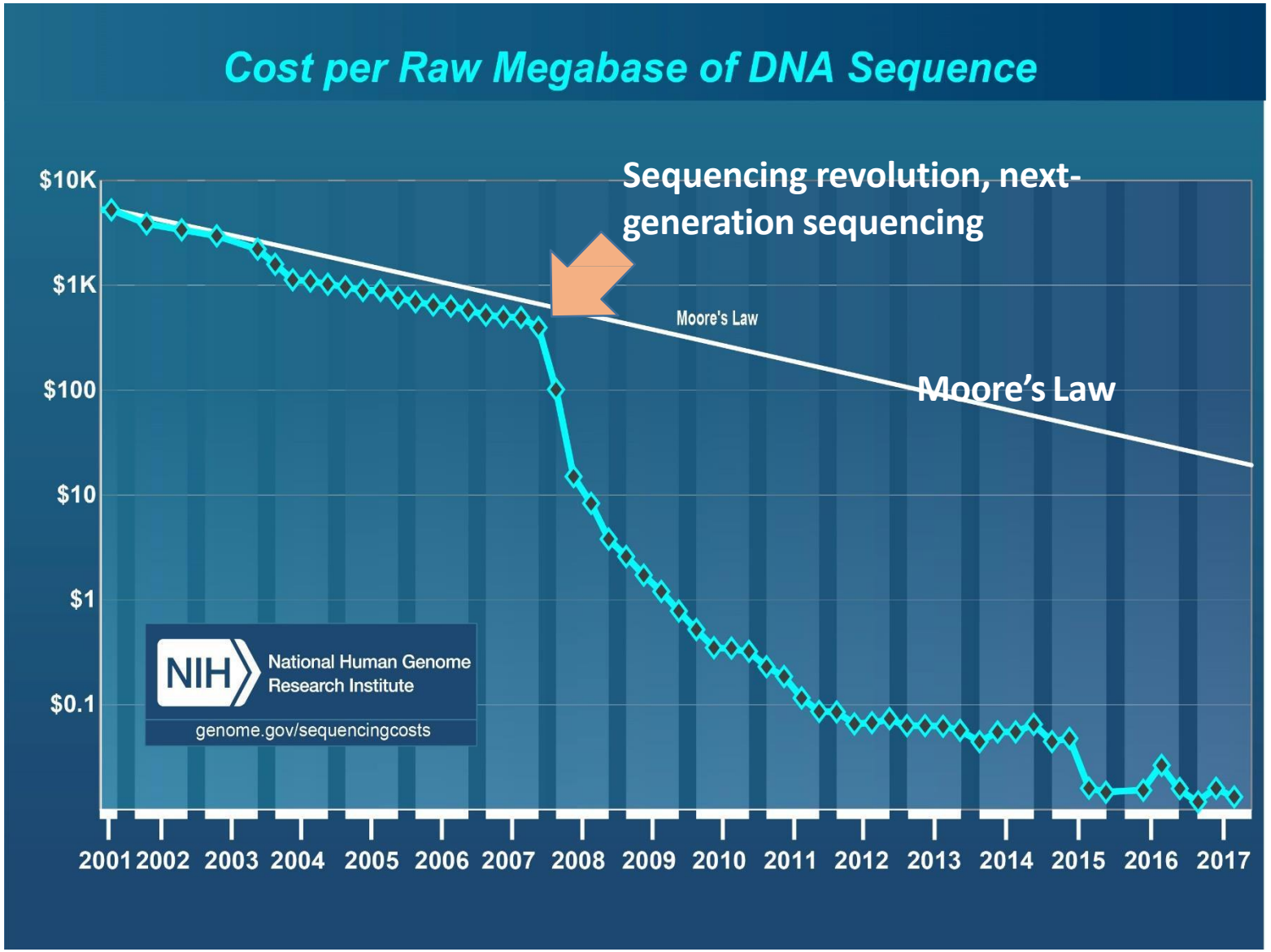
2008 First personal genome of James Watson

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

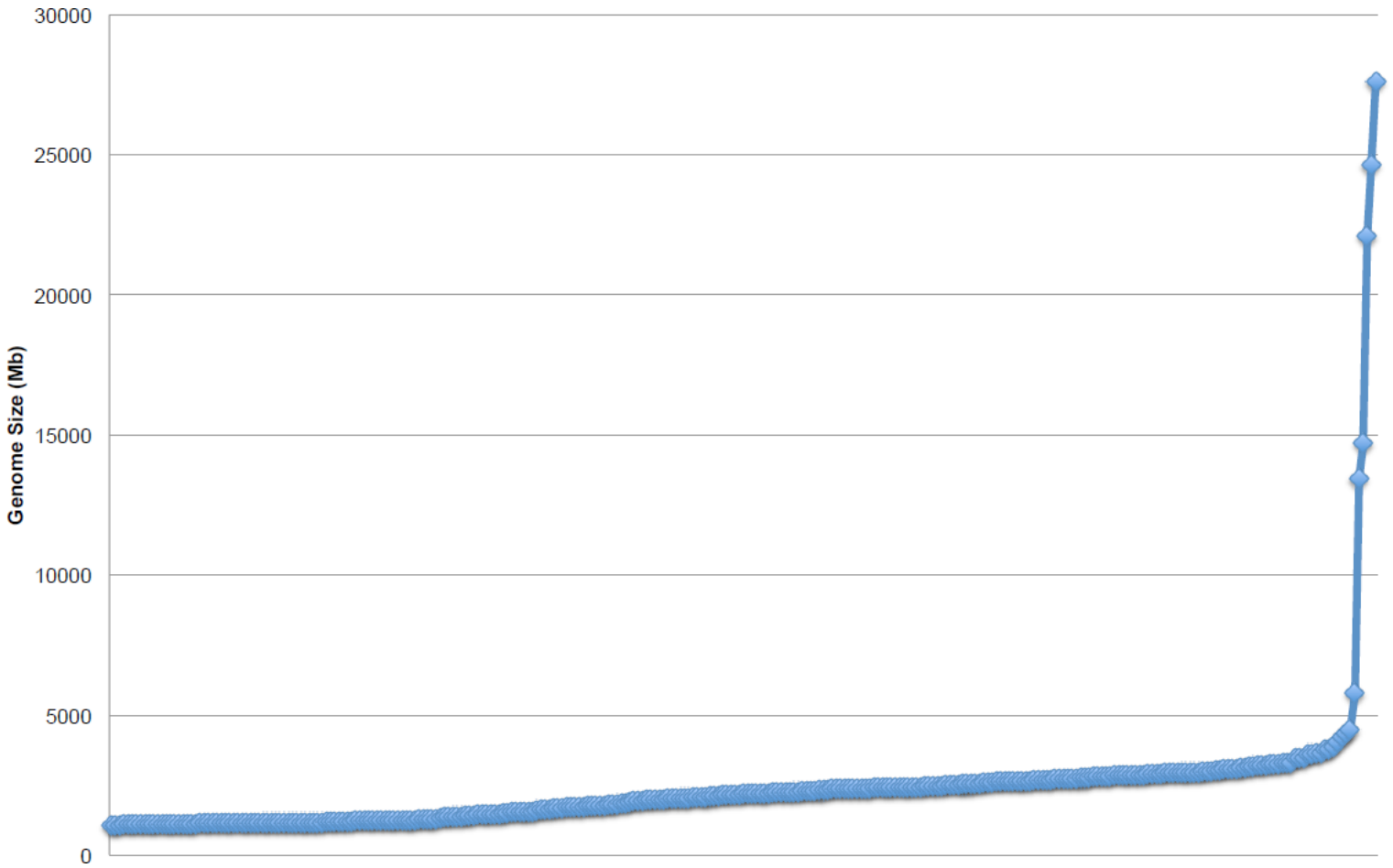
Cost of DNA Sequencing



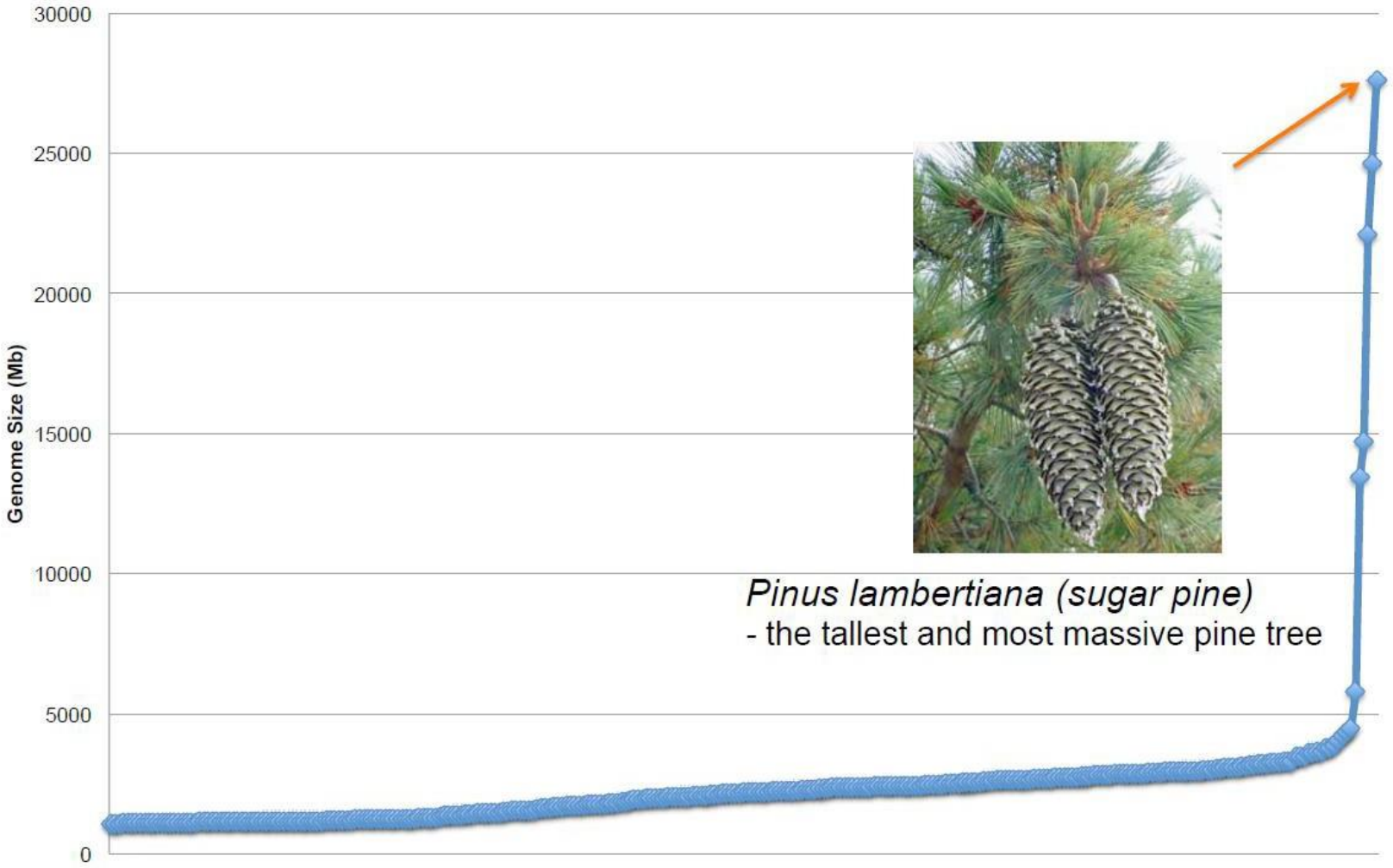
The problem:

- 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

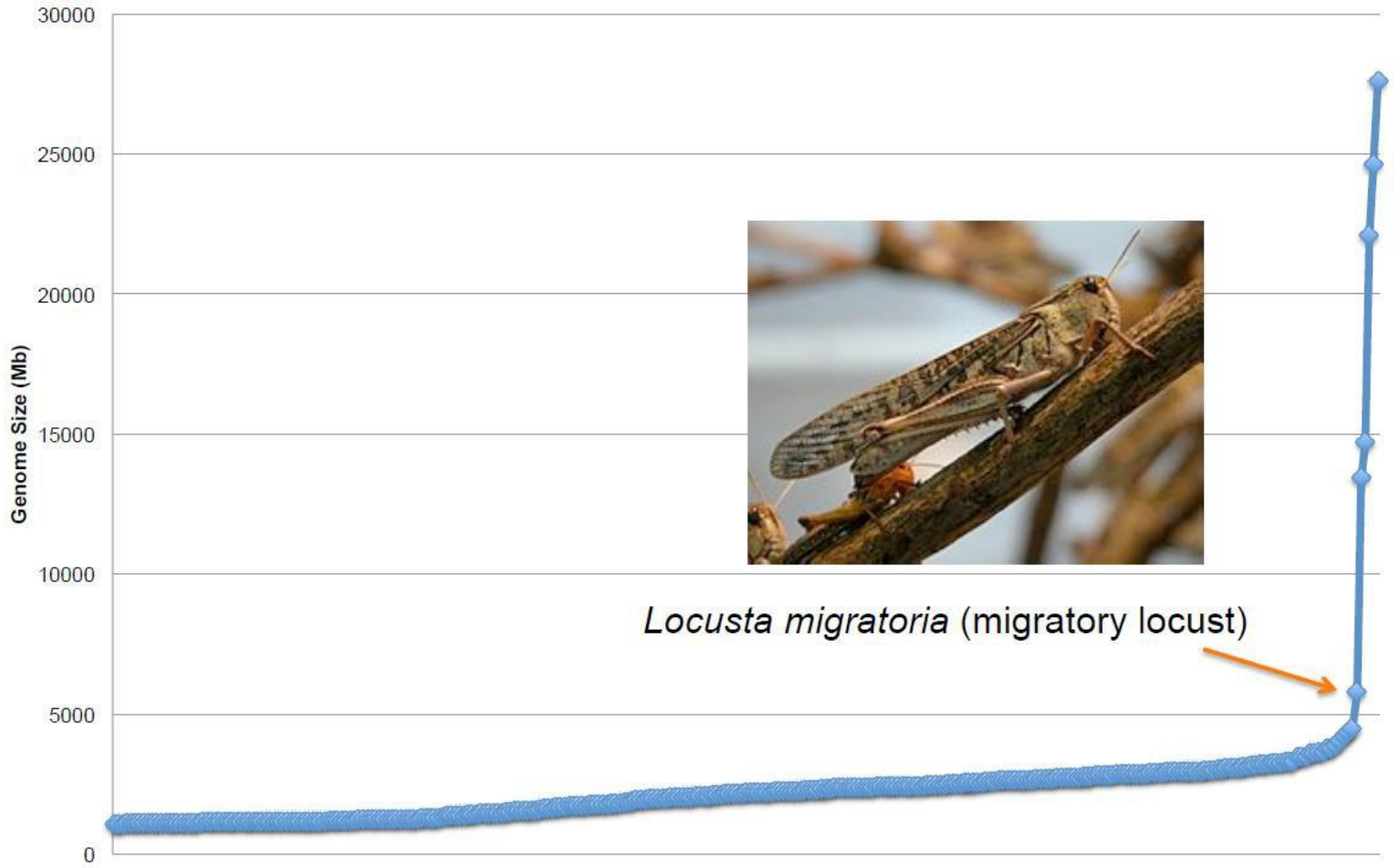
Top 200 Largest Sequenced Genomes



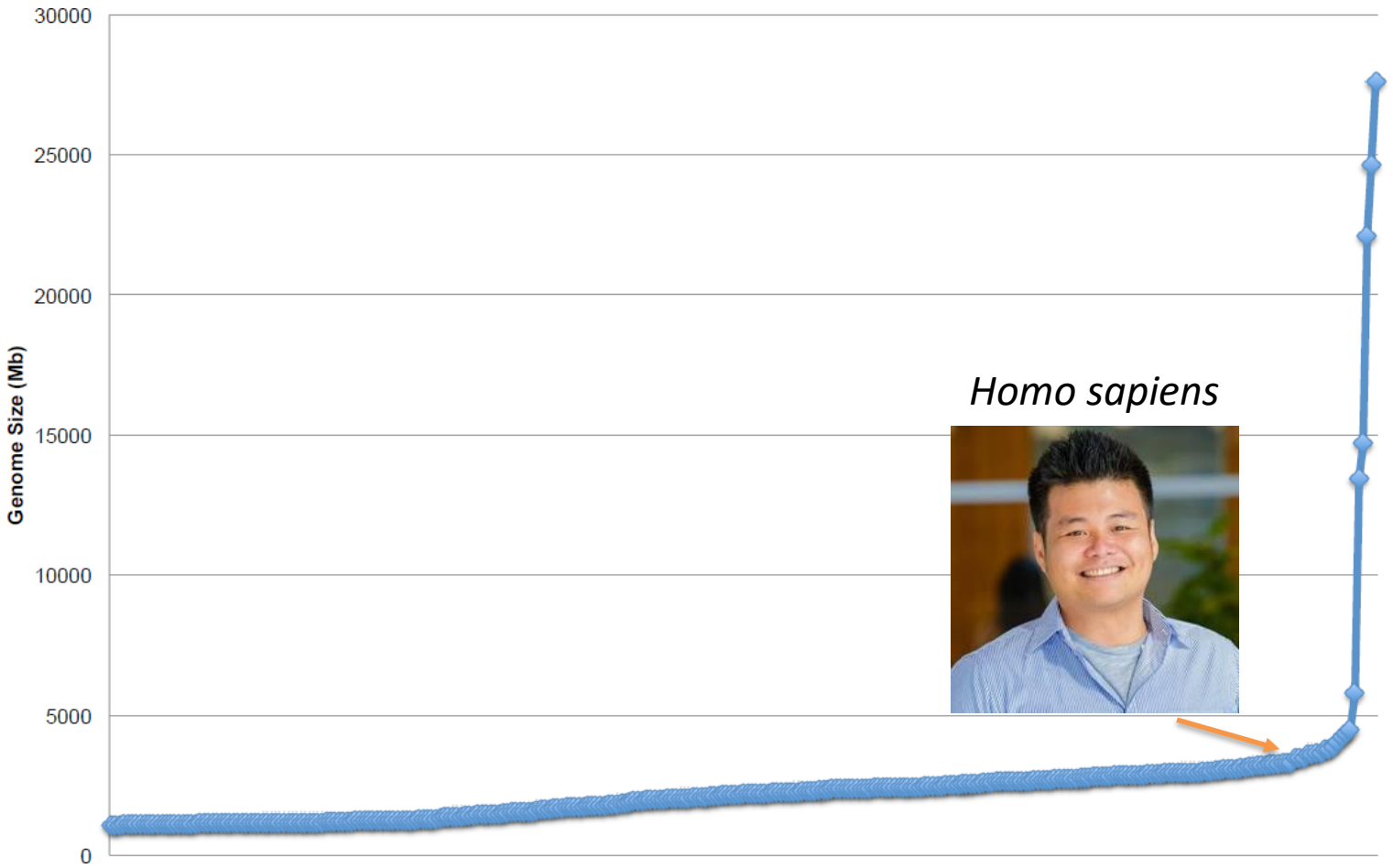
Top 200 Largest Sequenced Genomes



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

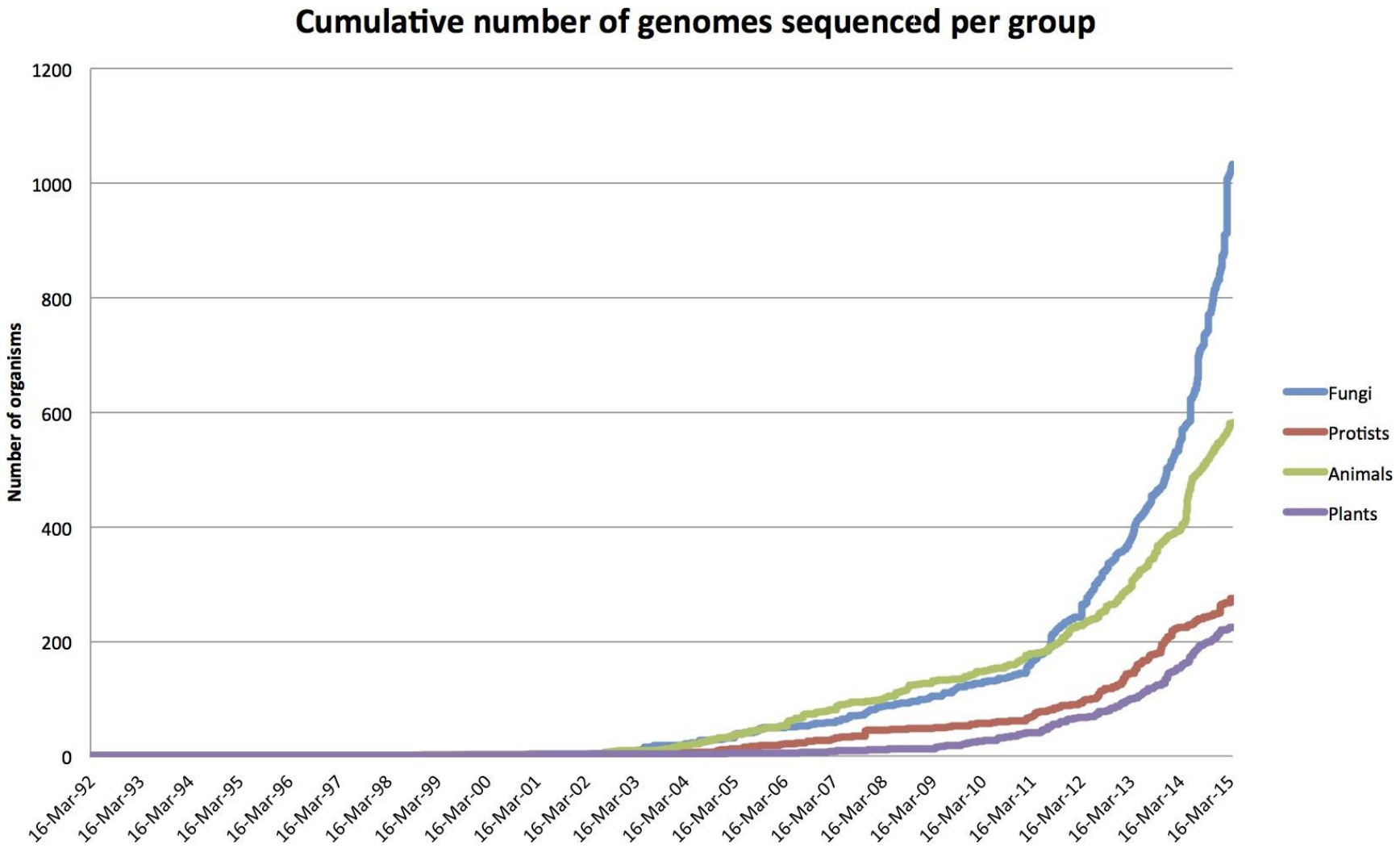
Simple Eukaryotes

- 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?

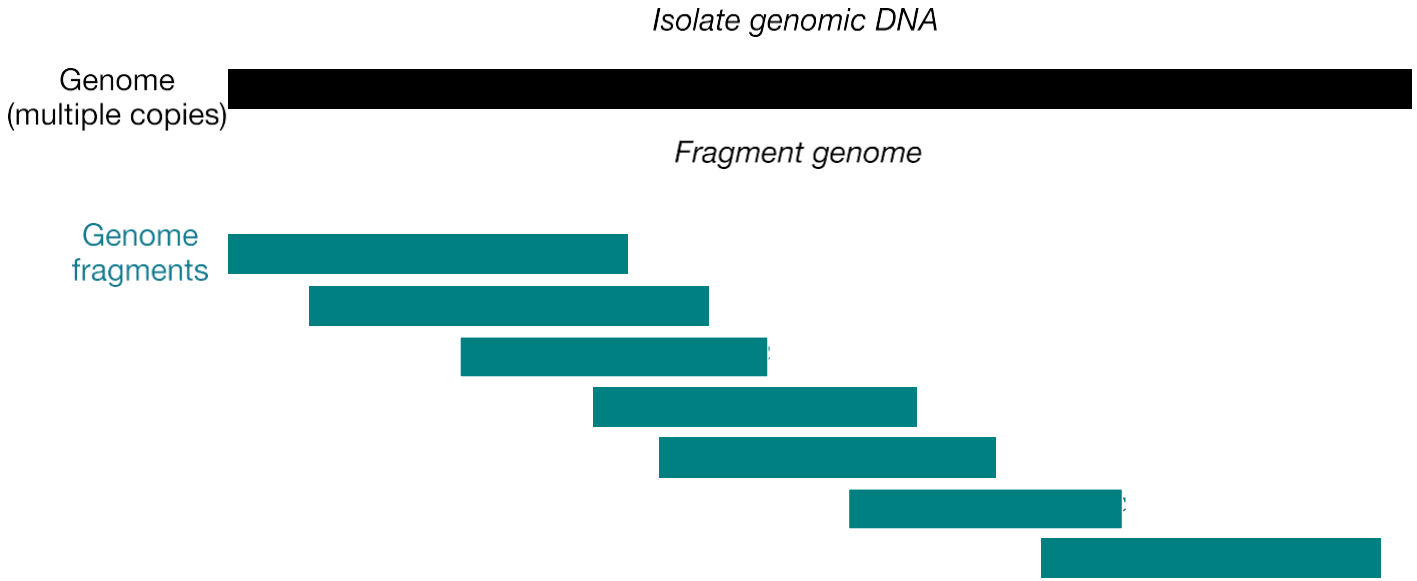
Whole Genome Shotgun (“WGS”)

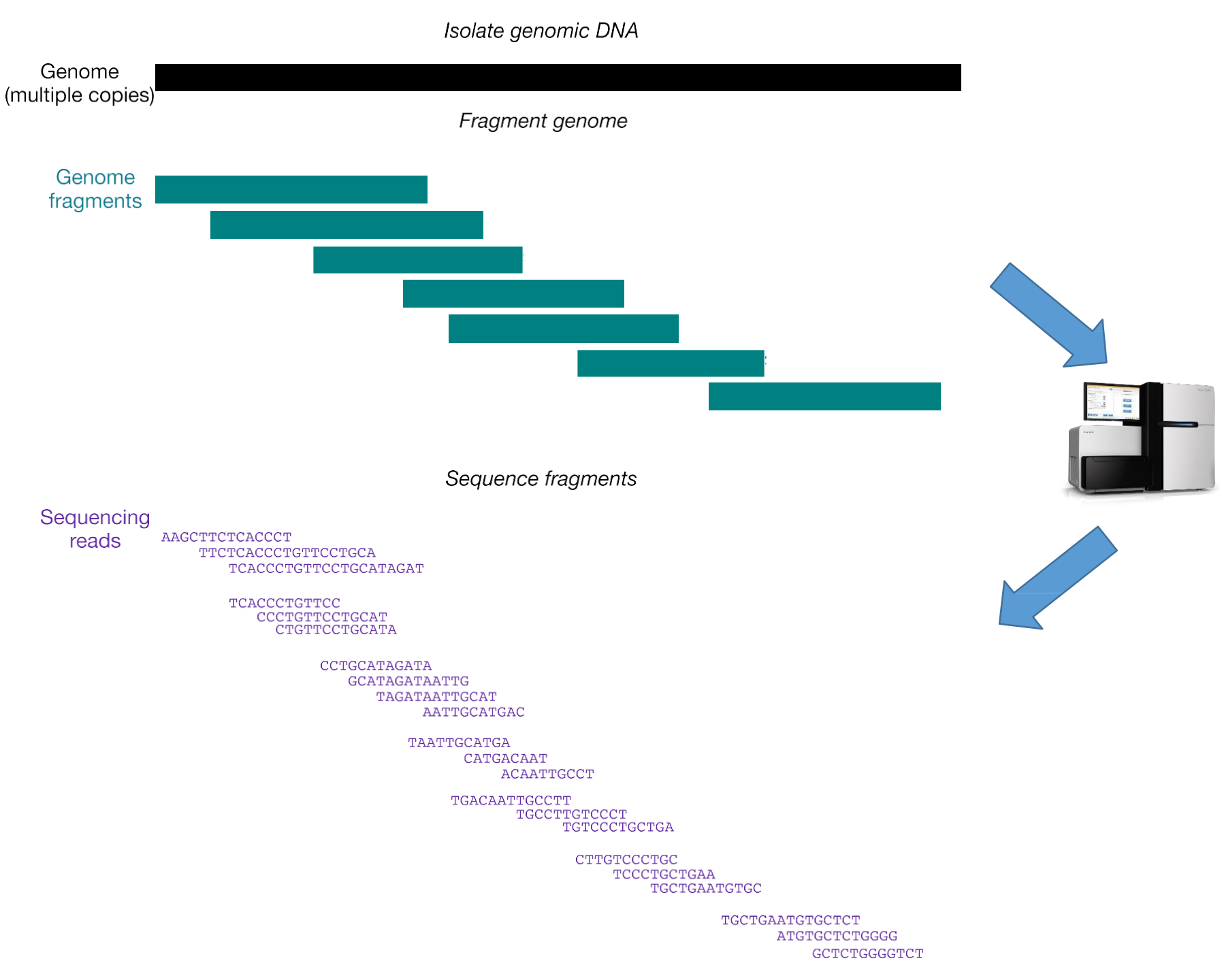
- Shear DNA to appropriate size
- Do some library preparation
- Put in sequencing machine
- Get some big text files with your reads
- Panic when NOTEPAD.EXE won’t load them

Adapted from Torsten Seemann

Isolate genomic DNA

Genome
(multiple copies)

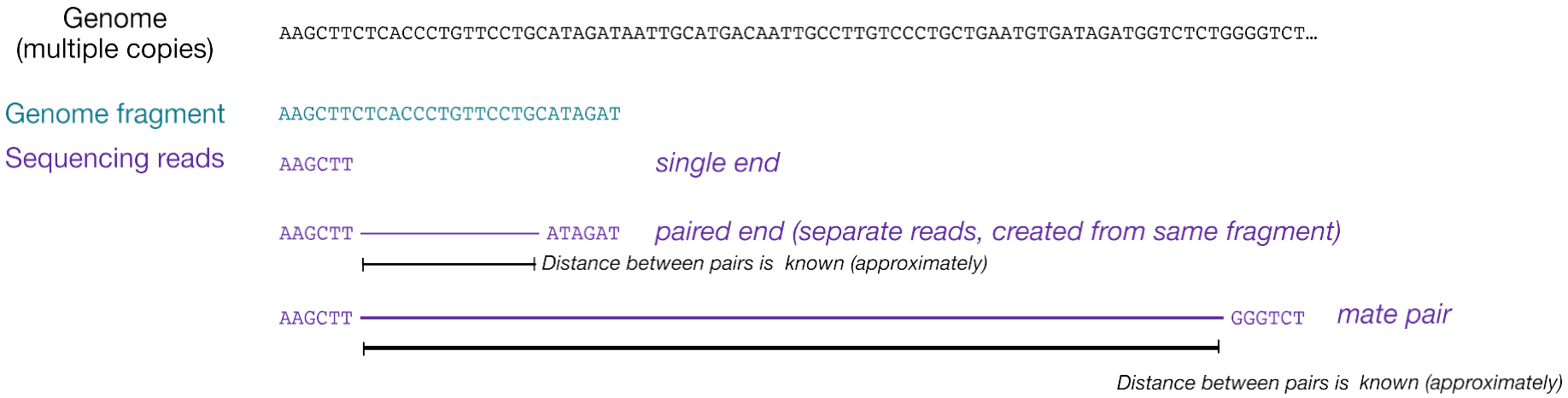

Isolate genomic DNA
Genome
(multiple copies)
Fragment genome
Genome
fragments



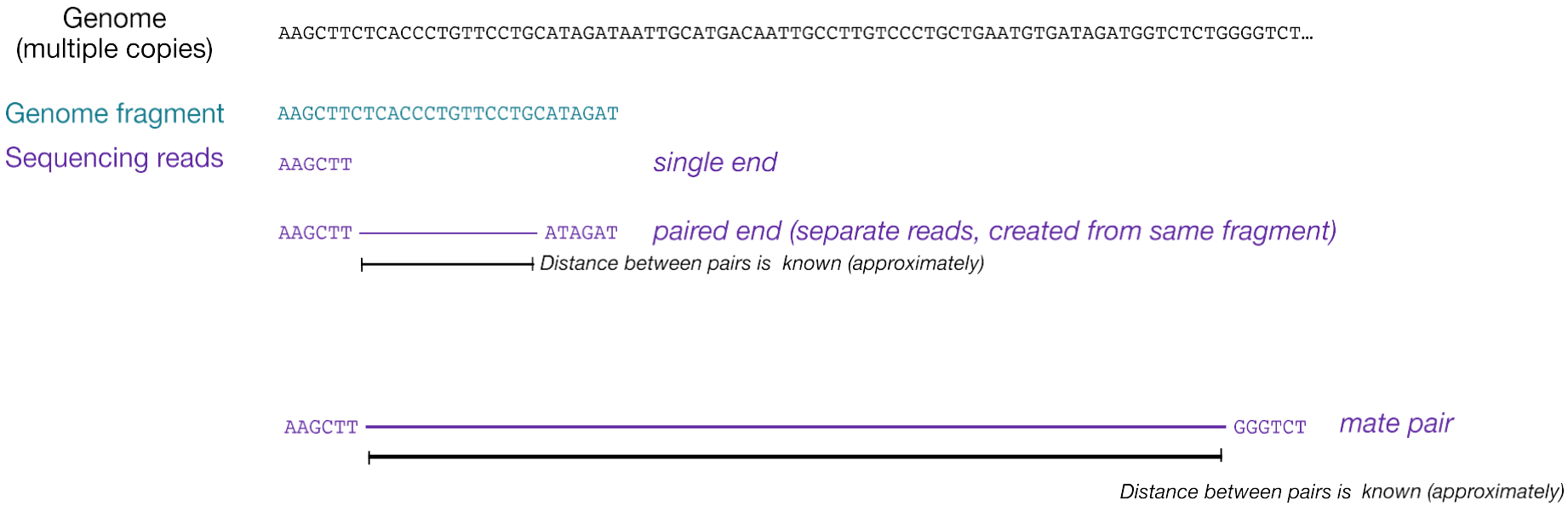
This approach is called ‘Shot Gun’ sequencing


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

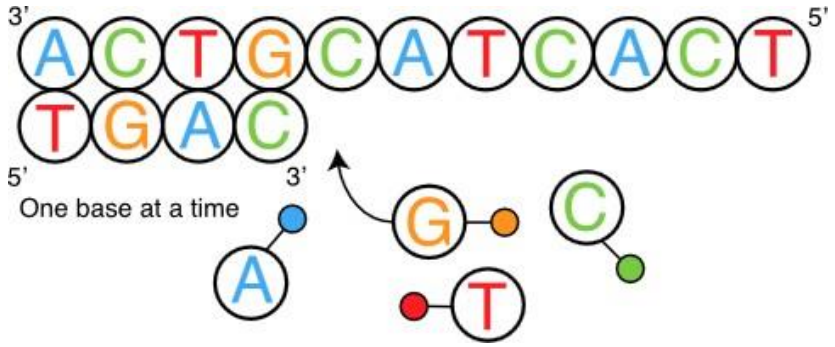

Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)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


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

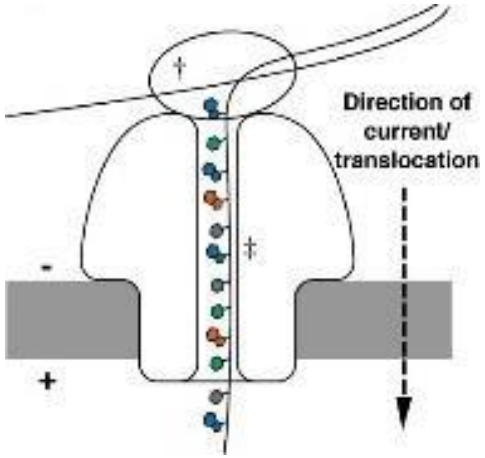

Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)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

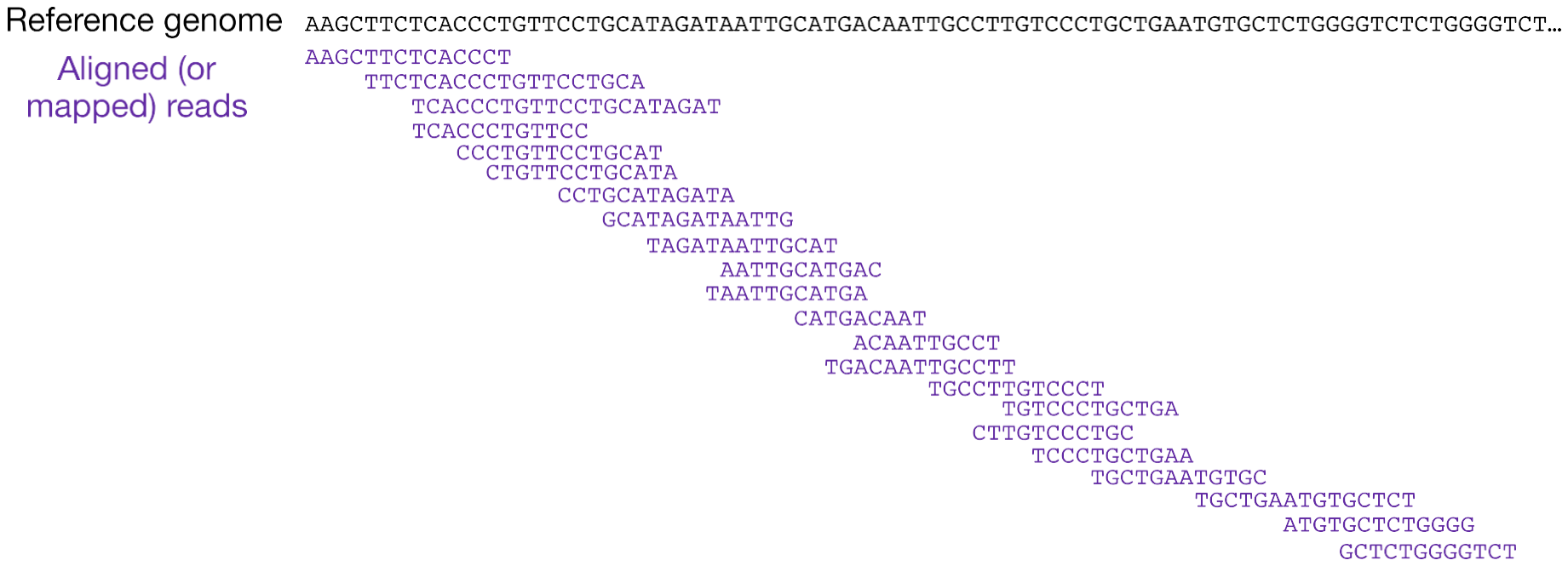
TGACAATTGCCTT

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

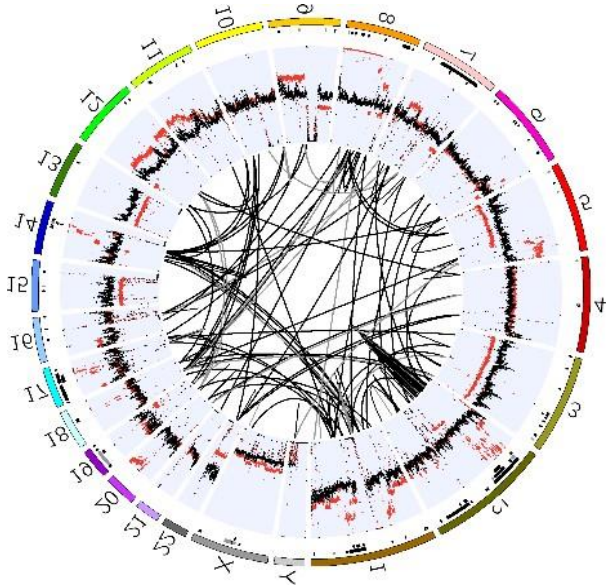
Single nucleotide substitution (SNP) in target genome

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

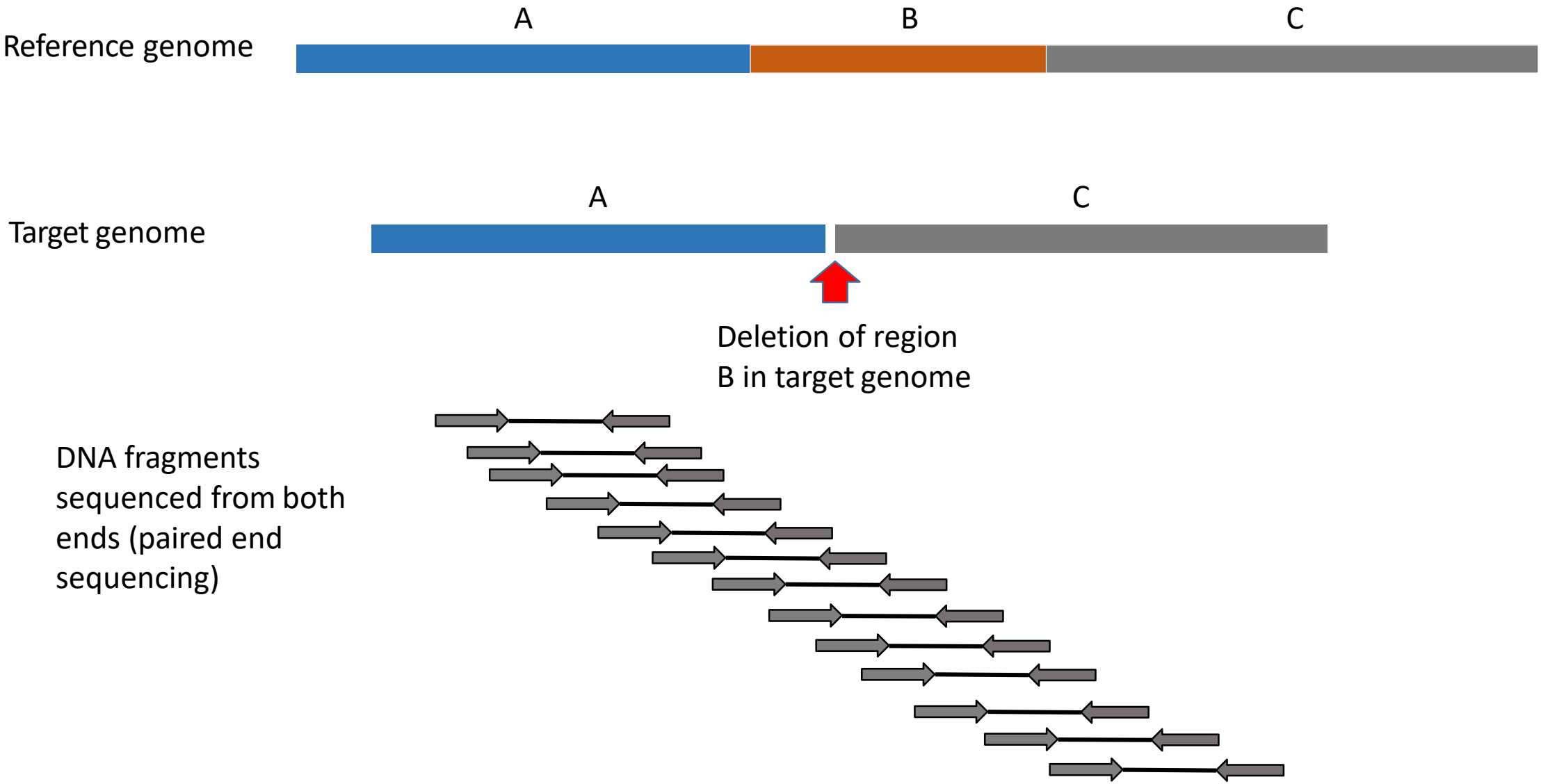
How Can We Identify Structural Variations?

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

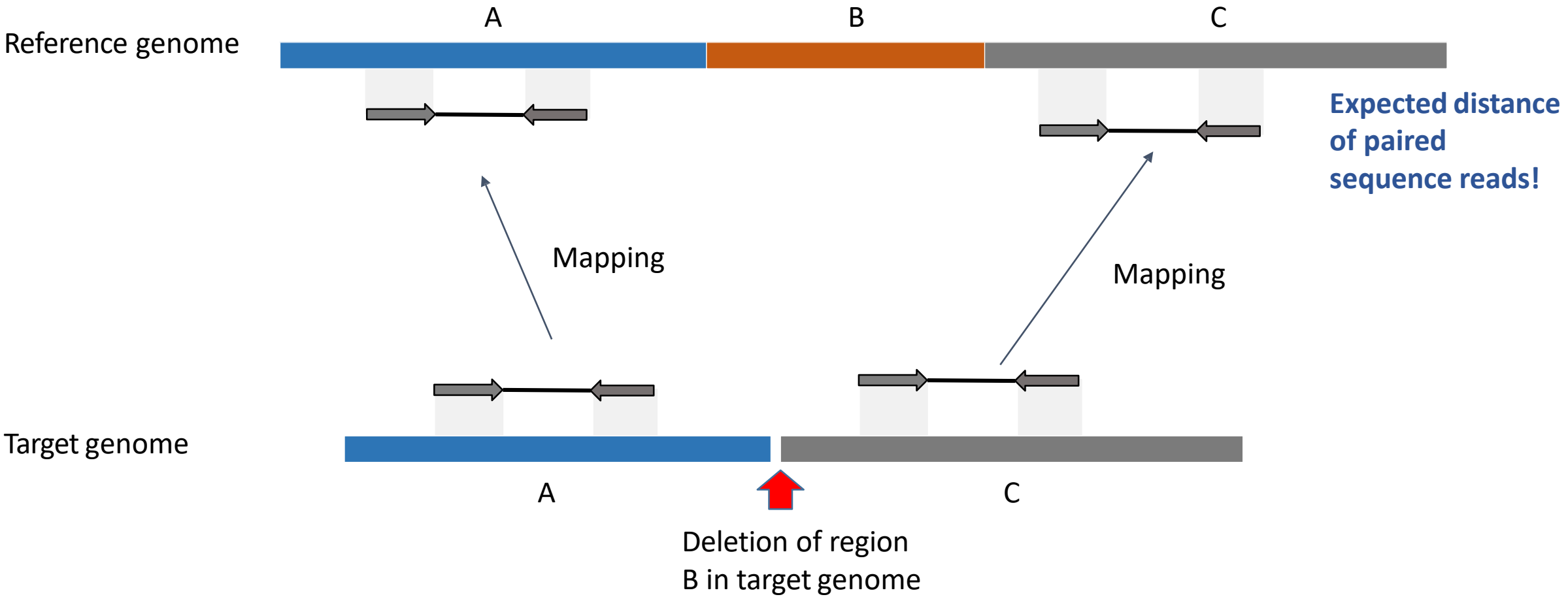
Example:
Lung cancer
Each line represents translocation in tumor genome compared to normal tissue from same patient



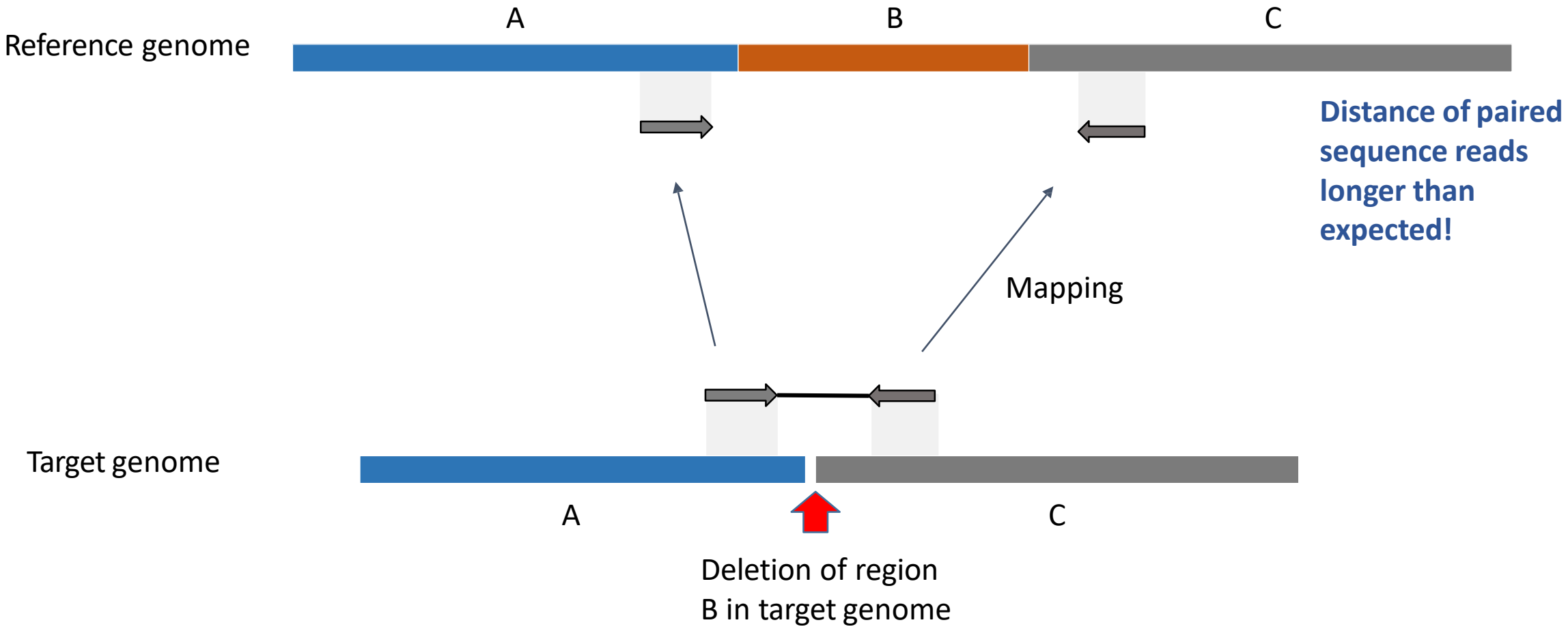
Structural Variations



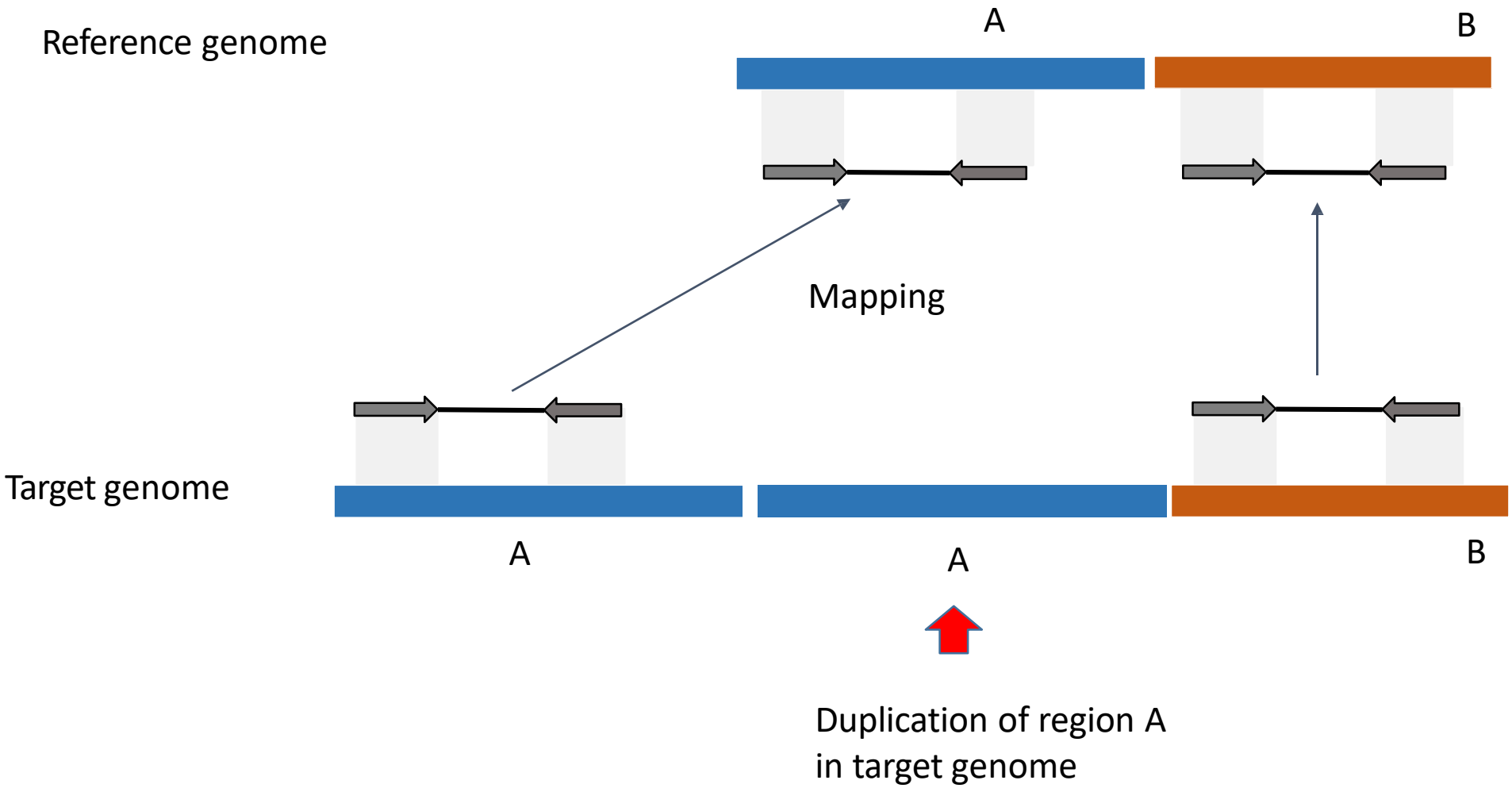
Structural Variations



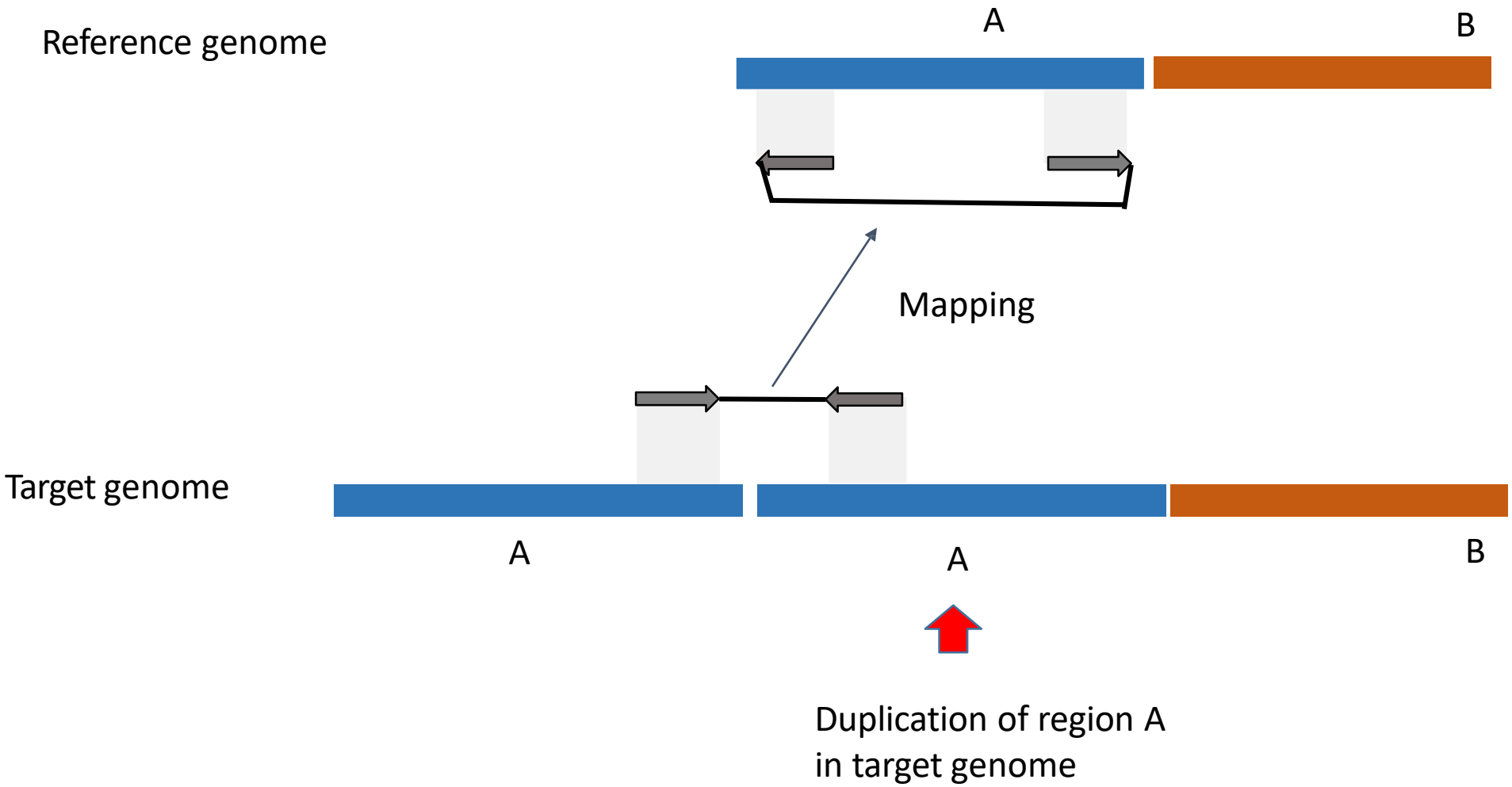
Structural Variations



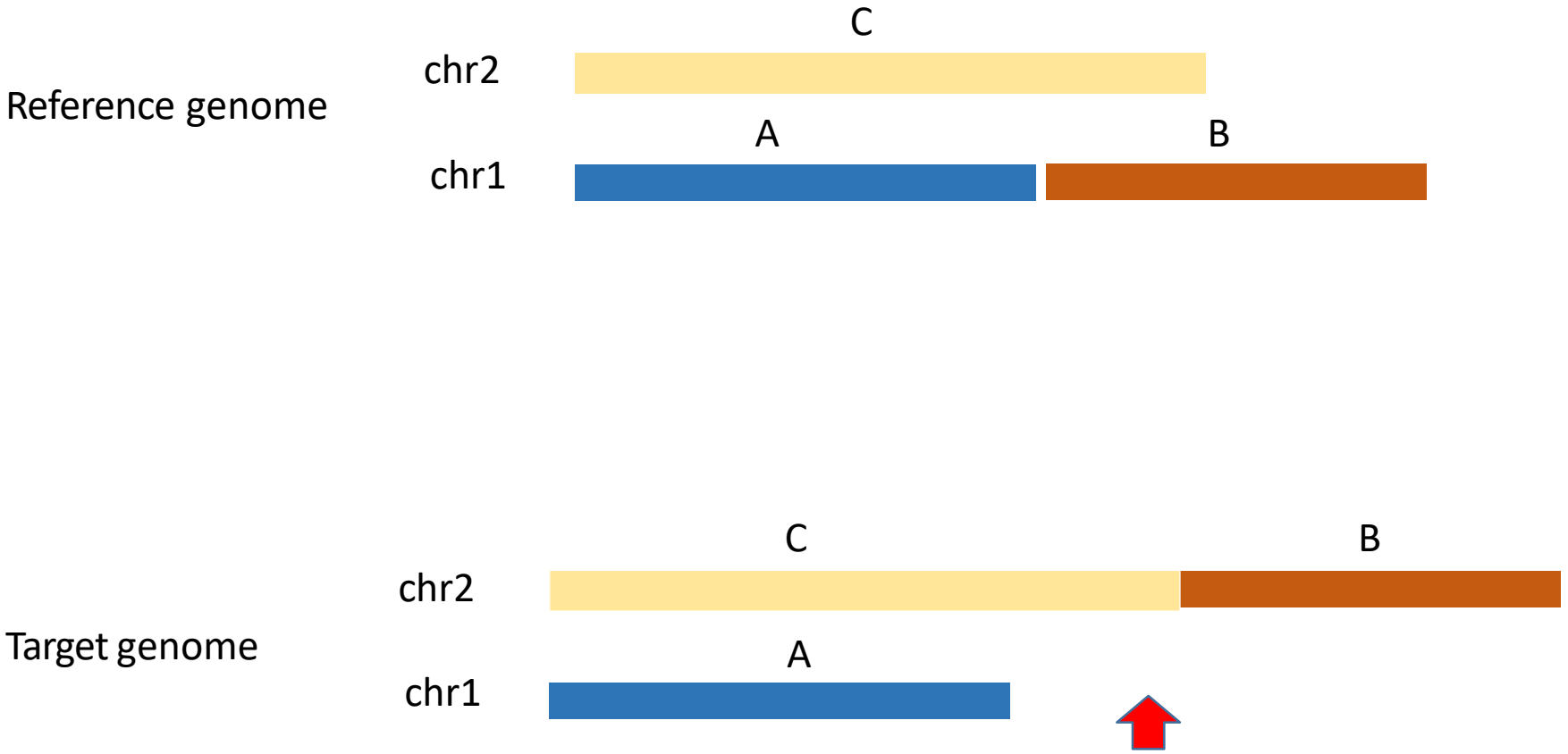
Duplications



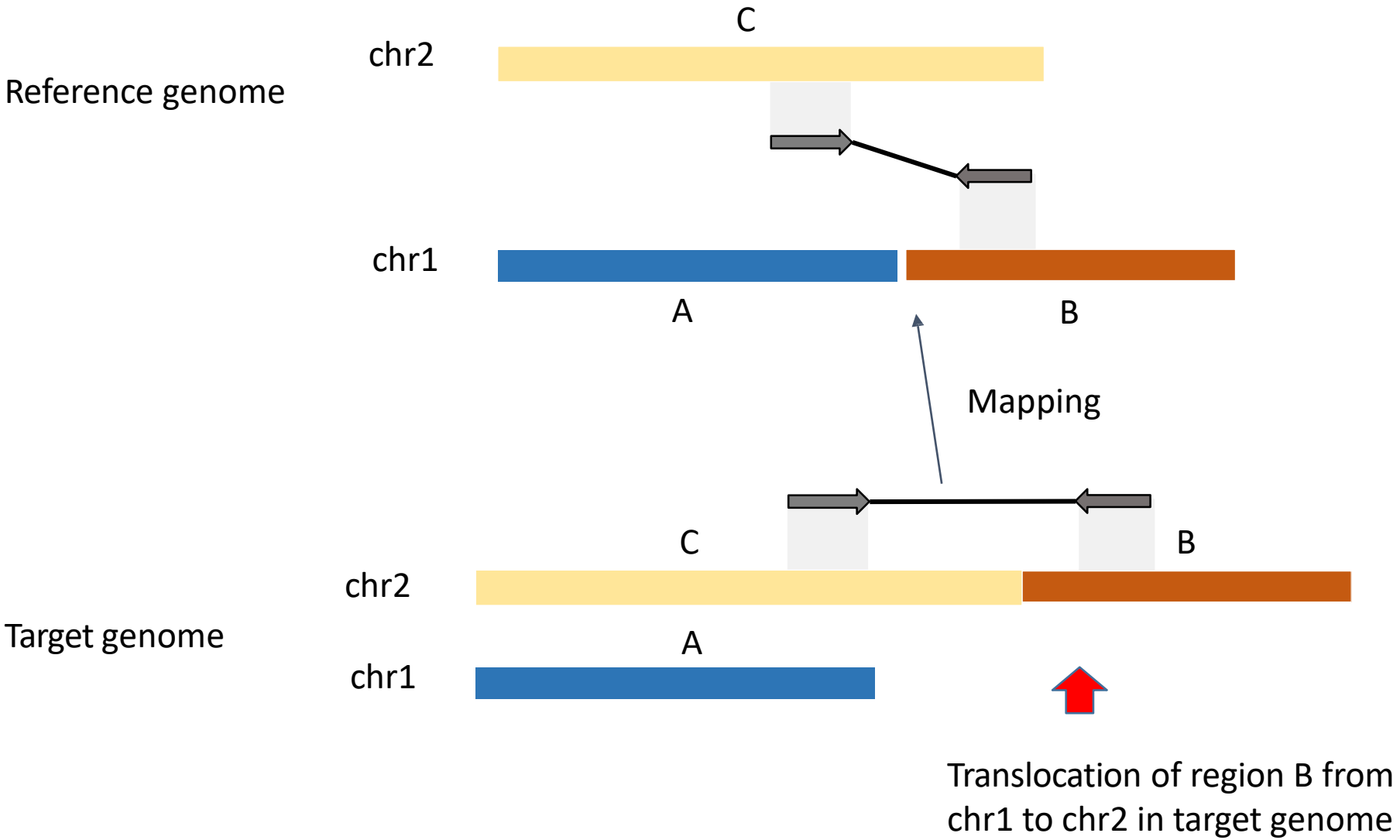
Duplications



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

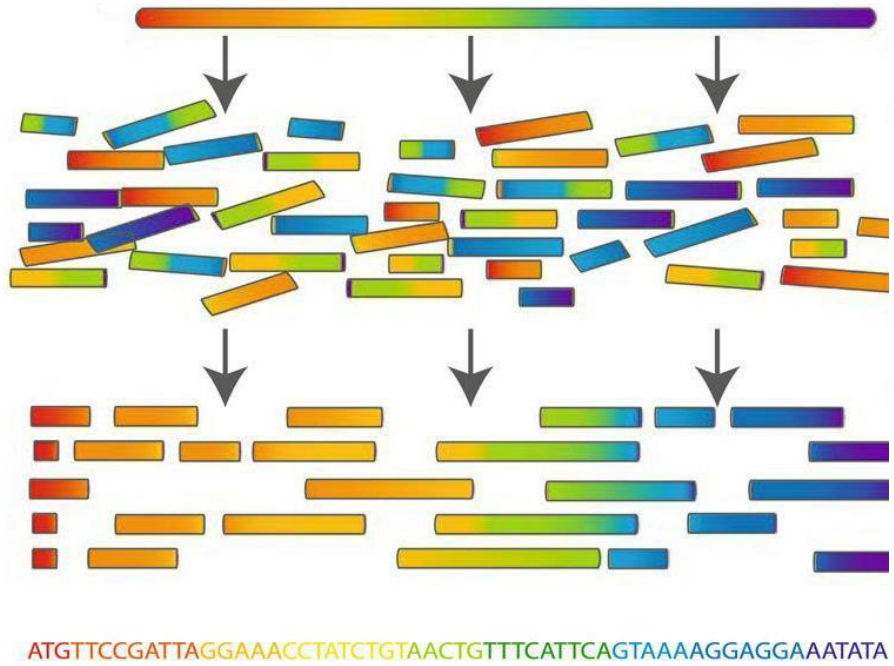
WHOLE CITY IS ABLAZE

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

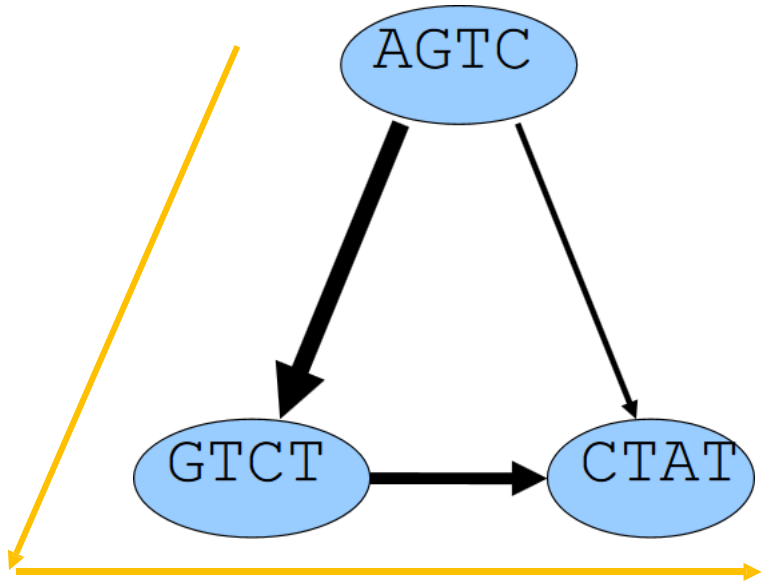
Overlaps:

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

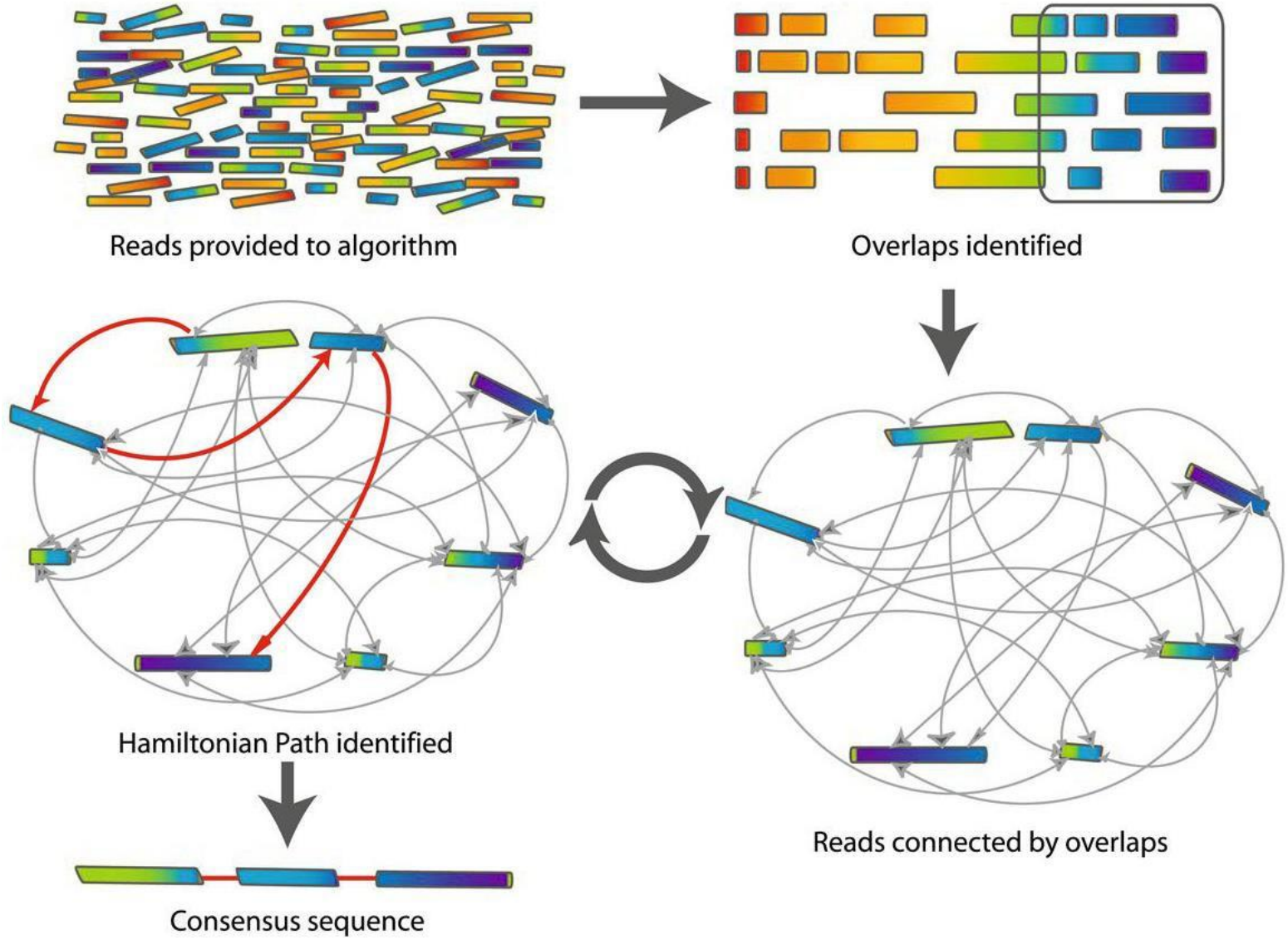
Steps:

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

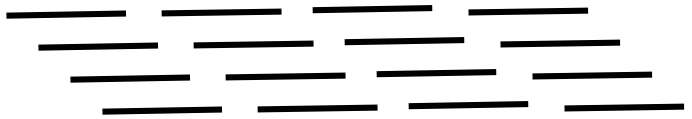
Every base of genome covered by ~100 sequence reads

	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