FRST302: Forest Genetics

Lecture 1.4: DNA Sequencing & Forest Genomics

Lecture 2 - Recap

- Chromosome structure
- Genetic linkage
- Genetic mapping
- DNA structure
- Mutation

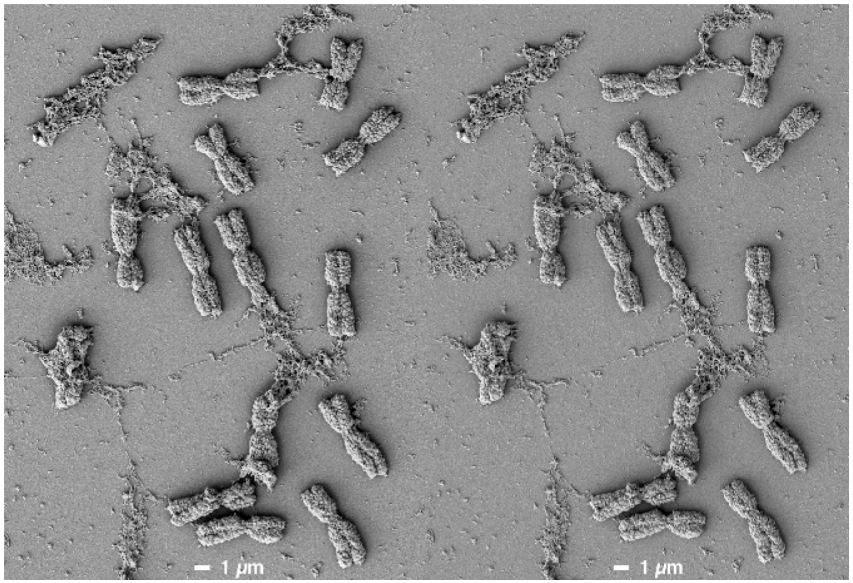


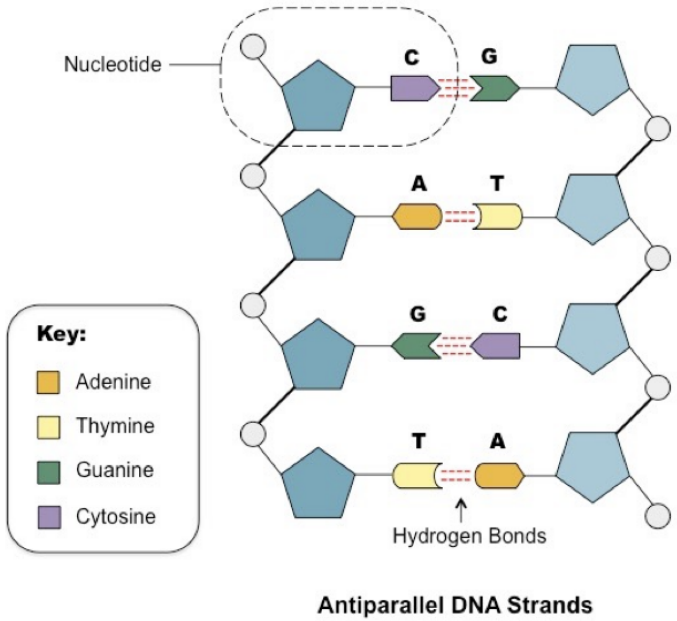
Figure from Punnett 1911

What is Sequencing?

A genome is the complete DNA present in an individual cell or organism

DNA sequencing is the process of decoding the sequence of As, Ts, Cs and Gs in a sample of DNA

Structure of DNA



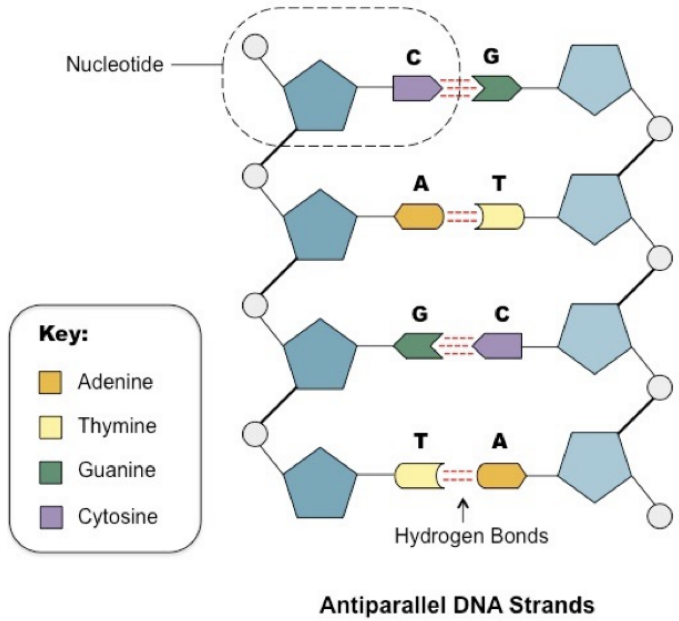
Antiparallel DNA Strands

What is Sequencing?

A genome is the complete DNA present in an individual cell or organism

DNA sequencing is the process of decoding the sequence of As, Ts, Cs and Gs in a sample of DNA

Structure of DNA



Antiparallel DNA Strands

How we measure DNA sequence data

Base Pairs	Unit	Example
1	1bp	Single Nucleotide
1,000	1kbp	The average human gene is 10-15kbp long
1,000,000	1Mbp	The human X-chromosome is 154Mbp long
1,000,000,000	1Gbp	The human is genome 3.05Gbp long

What are some uses for genomics in forestry?

Some Uses for Genomics in UBC Forestry

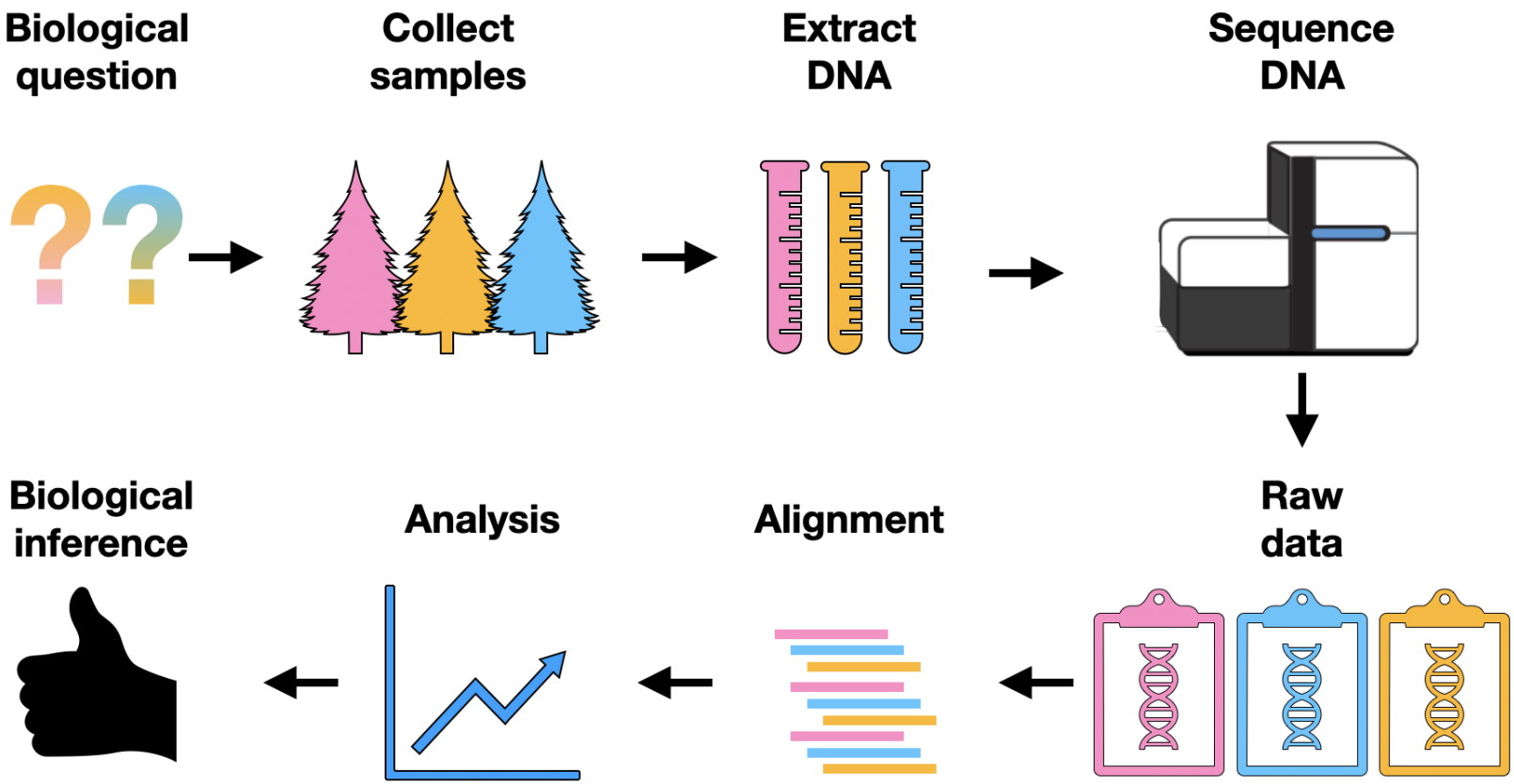
Developing breeding programs

- Identification of important/useful genetic variation (e.g. targets for transformation)
- Genome guided selection (e.g. breeding trees with specific attributes)

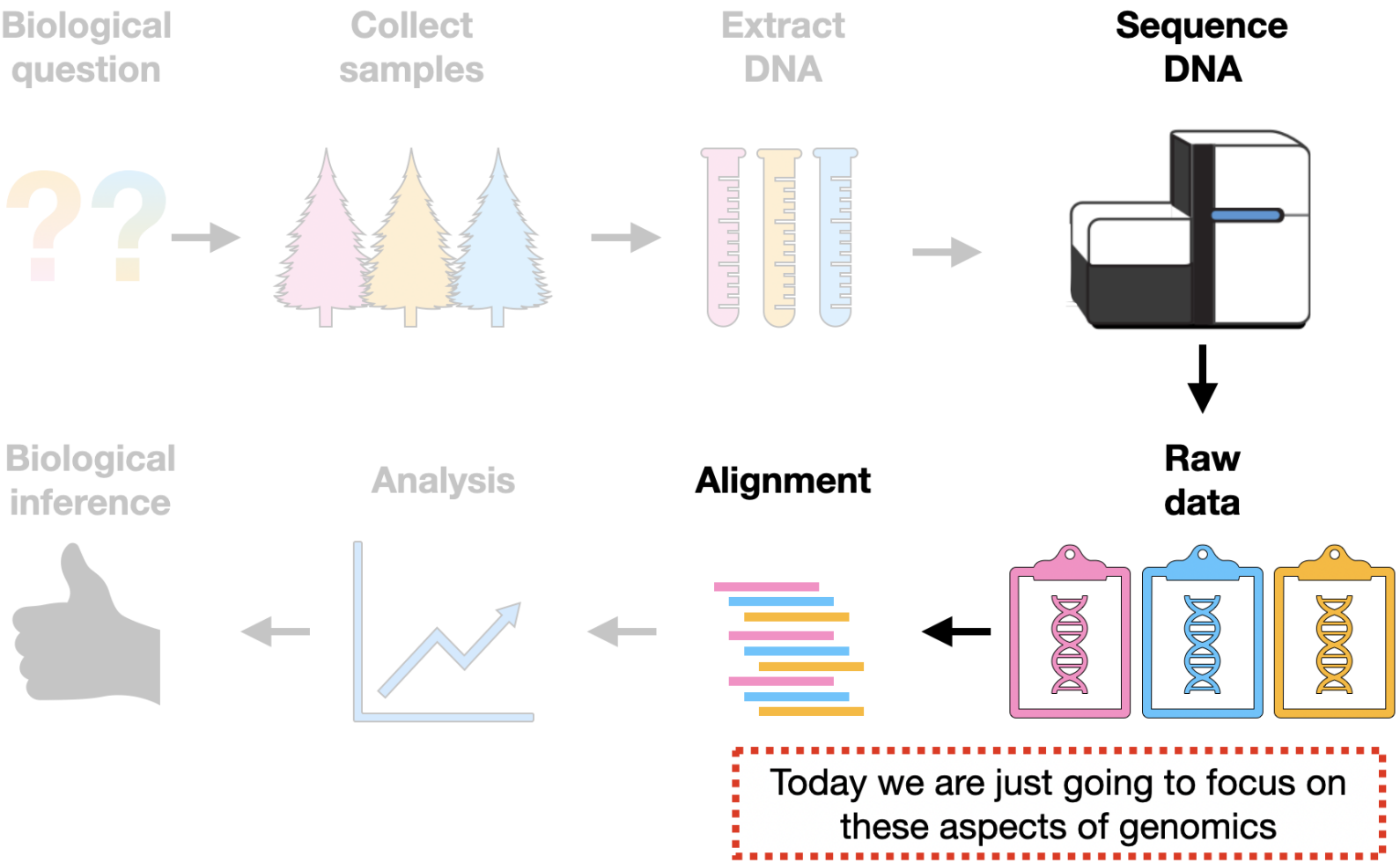
Identifying populations that are:

- Sensitive to climate change (e.g. given climate change projections)
- Particularly distinct (e.g. culturally or otherwise important to stakeholders)

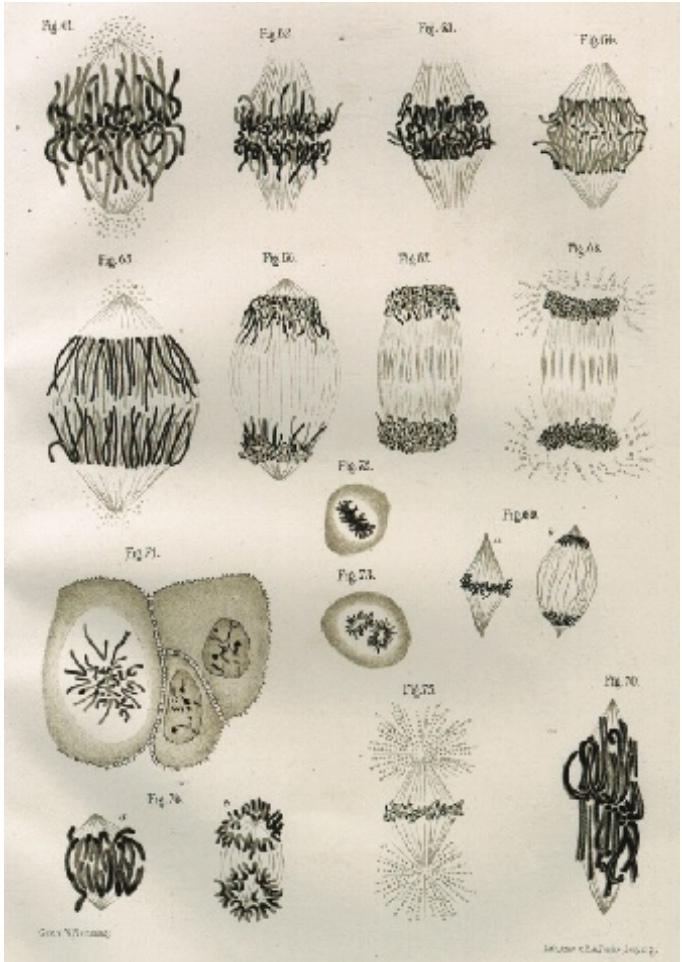
An Example Genomics Project



An Example Genomics Project



A Timeline of Some Discoveries



1915 Morgan and Sturtevant constructed their genetic map for *D. melanogaster*

1932 Barbara McClintock confirms that genes are exchanged during crossing-over

1940s DNA is determined to be the material within chromosomes that carry heritable information

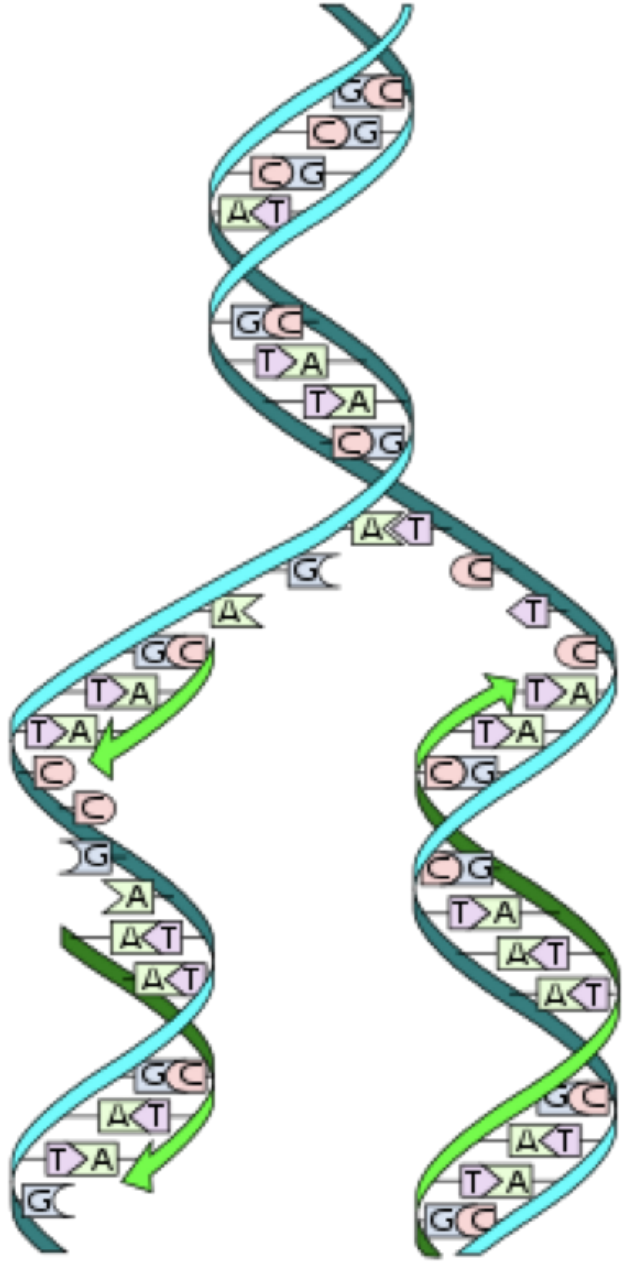
1950 The composition of DNA is determined - including Chargaff's rules

1953 The structure of DNA is determined

Drawing of mitosis by Walther Flemming 1882

DNA Replication

- Both mitosis and meiosis involve DNA replication
- DNA replication produces two identical replicas of DNA from one original DNA molecule.
- During replication, the double-stranded DNAs are separated. Each strand of the original DNA molecule serves as a template for the production of its counterpart.
- This process occurs in all living organisms and is the basis for growth and inheritance. [a]



[a] Image from: *Wikipedia*

The Polymerase Chain Reaction - PCR

PCR is a technique to amplify a single copy or a few copies of a piece of DNA to millions of copies of a particular DNA sequence.

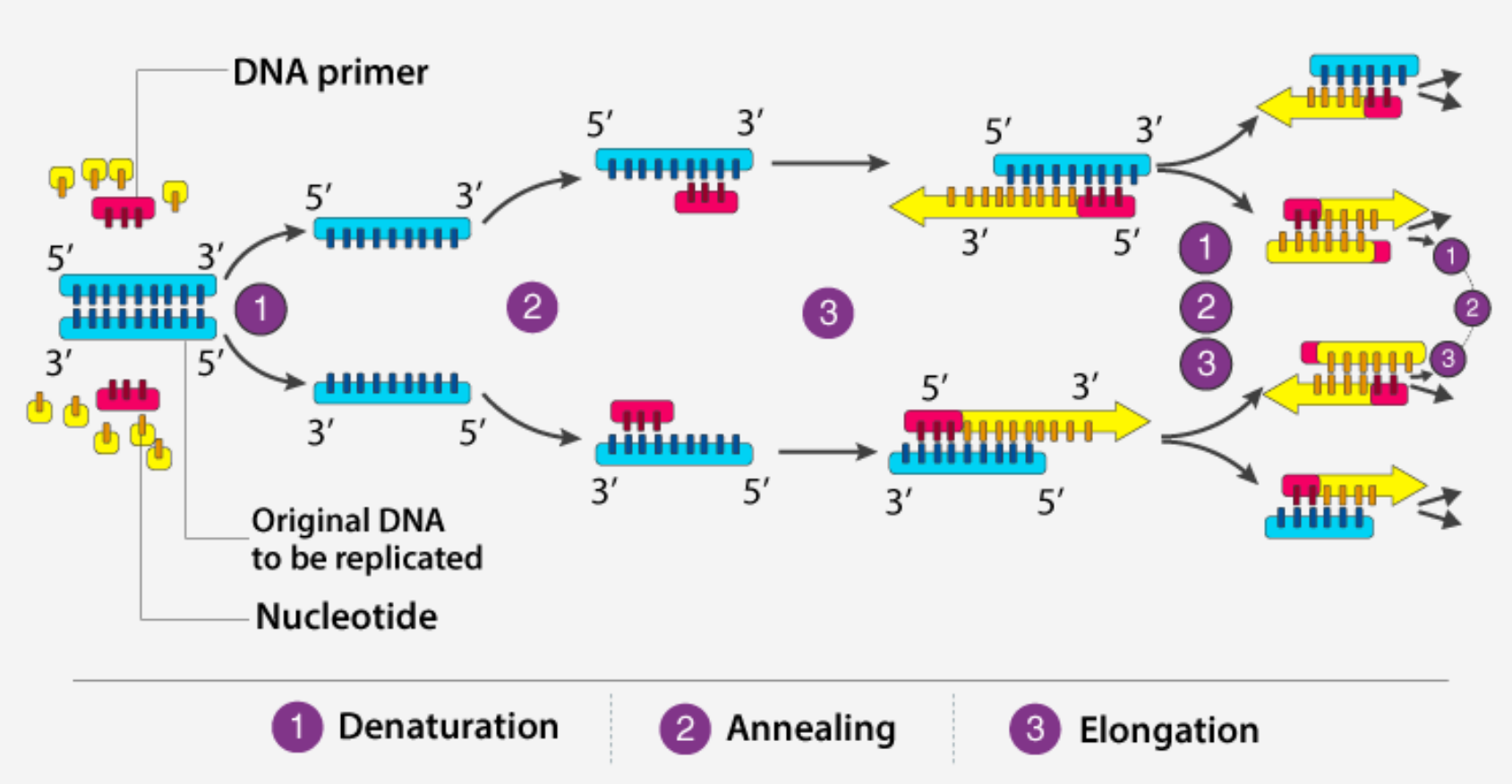


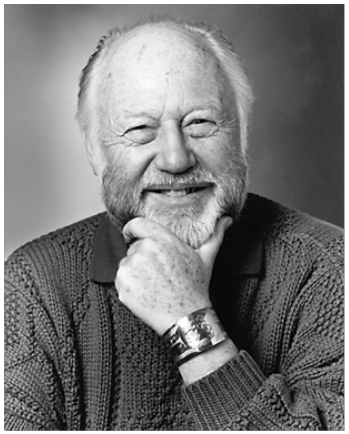
Image modifed from: `https://byjus.com/biology/pcr/`

Applications of PCR

Developed in the early 1980s, PCR is fundamental in modern biology and is widely used in clinical and research settings for:

- DNA cloning for sequencing
- Functional analysis of genes
- Diagnosis of hereditary diseases
- DNA fingerprinting
- Detection and diagnosis of infectious diseases.

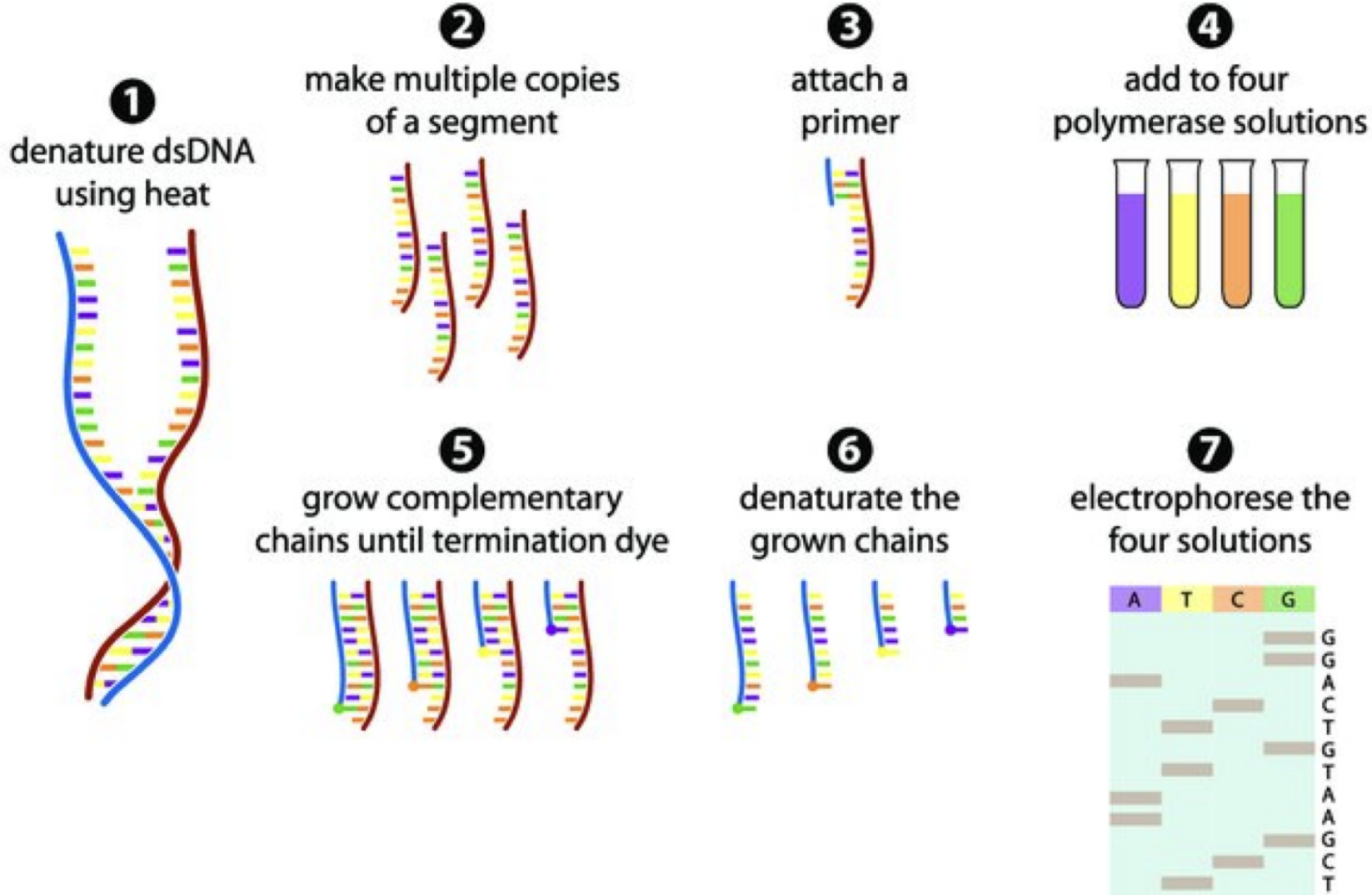
In 1993, Mullis and Michael Smith (UBC) was awarded the Nobel Prize in Chemistry



Technological Milestones

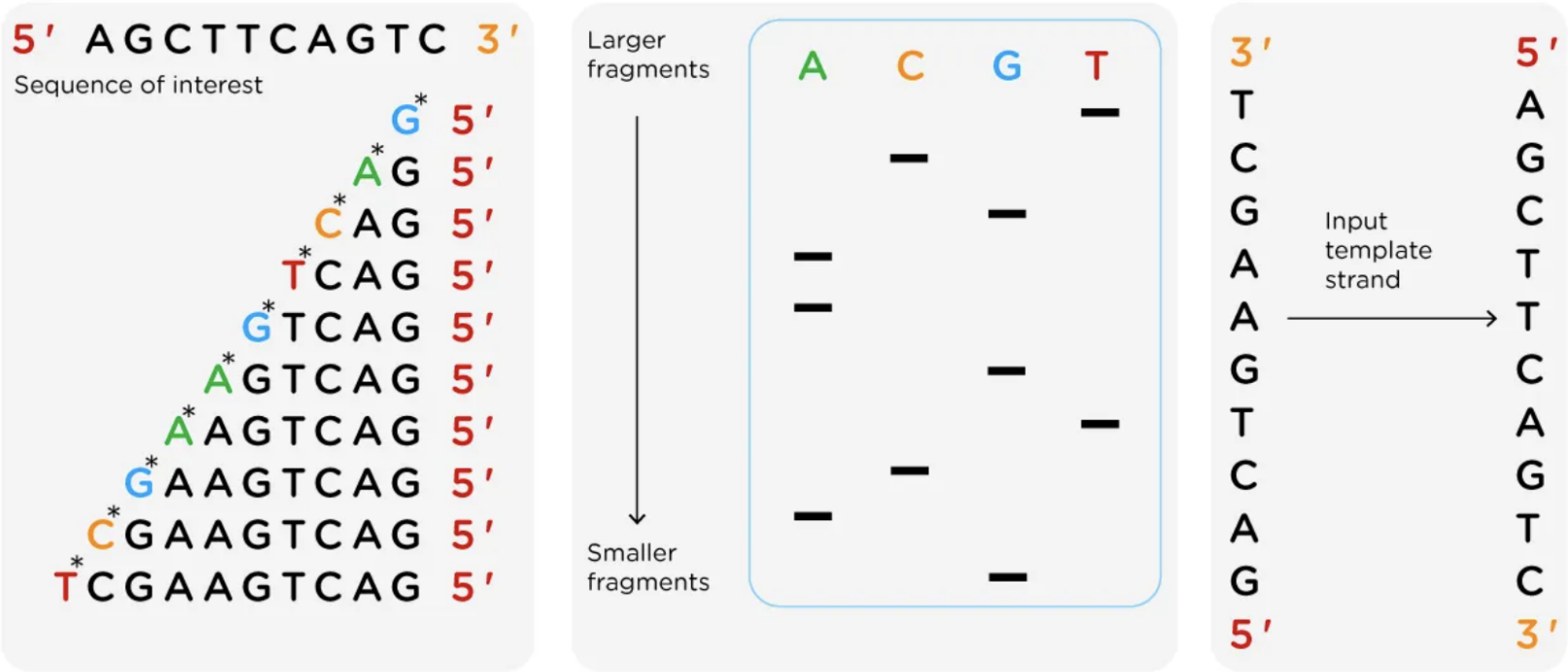
1953 **Sequencing of insulin protein**
1965 Sequencing of alanine tRNA
1977 Maxam–Gilbert sequencing
1977 **Sanger sequencing**
1990 Paired-end sequencing
2000 Massively parallel signature sequencing by ligation
2003 **Single-molecule massively parallel sequencing-by-synthesis**
2003 Sequencing by synthesis of in vitro DNA colonies in gels
2007 Large-scale targeted sequence capture
2010 Direct detection of DNA methylation during single-molecule sequencing
2010 Single-base resolution electron tunnelling through a solid state detector
2011 Semiconductor sequencing by proton detection
2012 **Reduction to practice of nanopore sequencing**
2012 Single-stranded library preparation method for ancient DNA

DNA Sequencing - Sanger



Developed by Frederick Sanger and colleagues in the 1970s
Figure from PhD thesis of Michel G. Gauthier

DNA Sequencing - Sanger



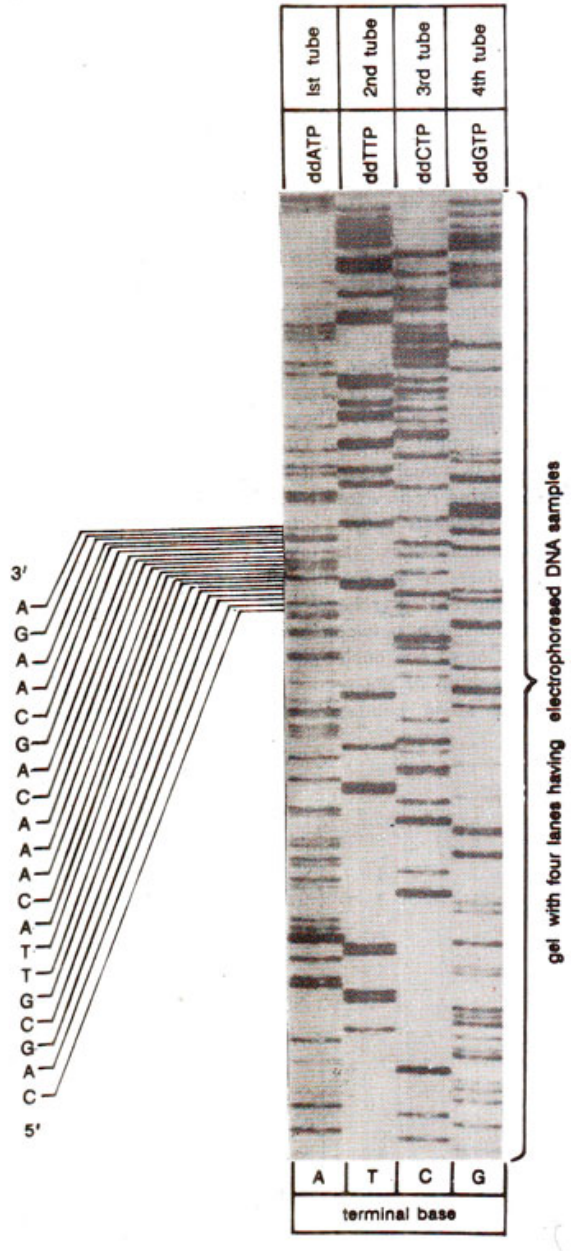
The sequence identified from the gel will be the reverse complement of the input sequence!

---

https://jackwestin.com/resources/mcat-content/recombinant-dna-and-biotechnology/dna-sequencing

DNA Sequencing - Sanger

- Gold standard for accuracy (99.99% accurate)
- Cheap equipment (low initial investment)
- Labour intensive
- Can only be used for short DNA strands (100 to 1000 base pairs)
- Time consuming & low throughput
- The basis of the human genome project



Genome Milestones

1977 Bacteriophage ϕX174

1982 Bacteriophage λ

1995 *Haemophilus influenzae*

1996 *Saccharomyces cerevisiae*

1998 *Caenorhabditis elegans*

2000 *Drosophila melanogaster*

2000 *Arabidopsis thaliana* - The first plant genome sequenced!

2001 *Homo sapiens*

2002 *Mus musculus*

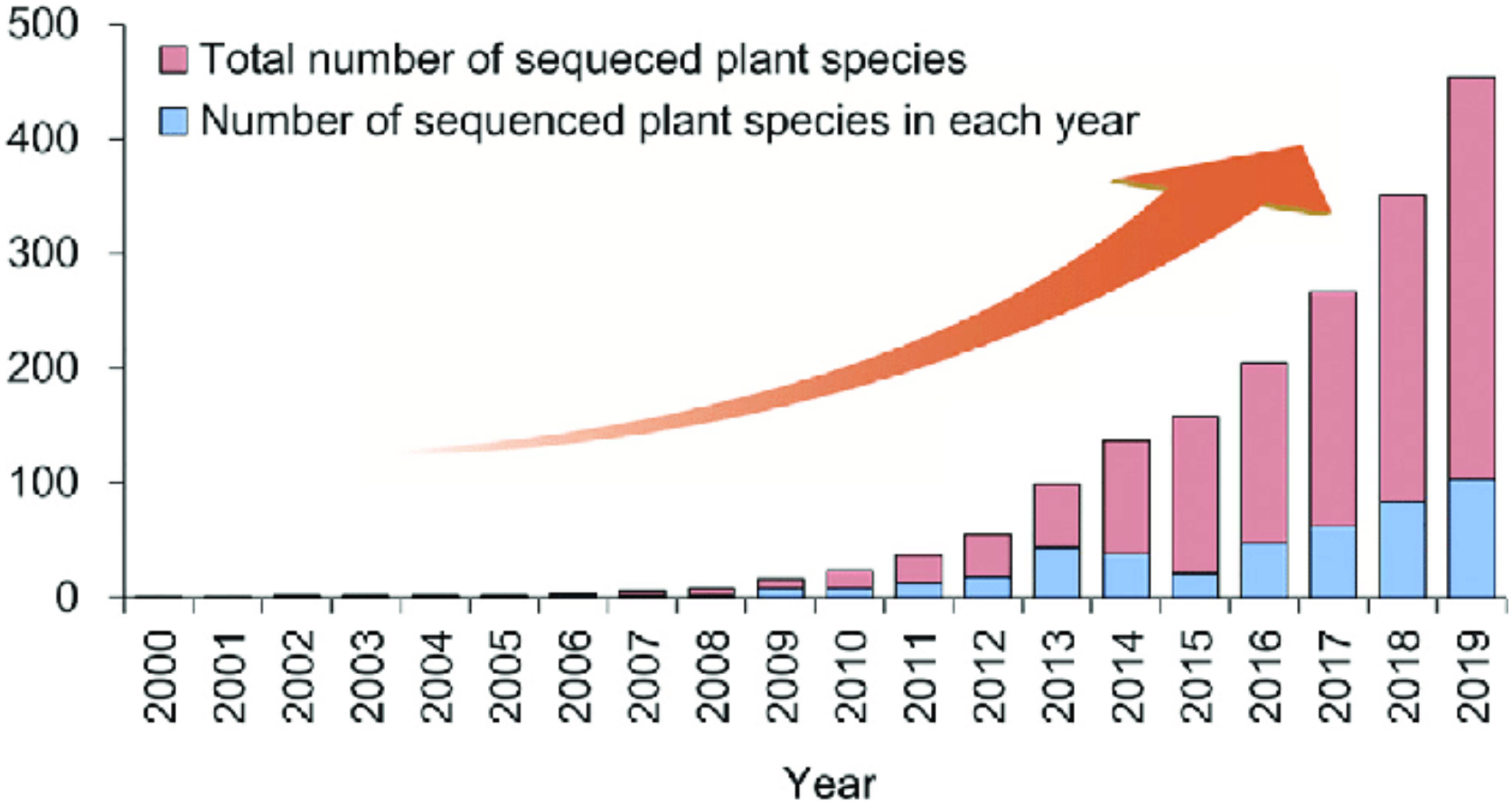
2004 *Rattus norvegicus*

2005 *Pan troglodytes*

2005 *Oryza sativa*

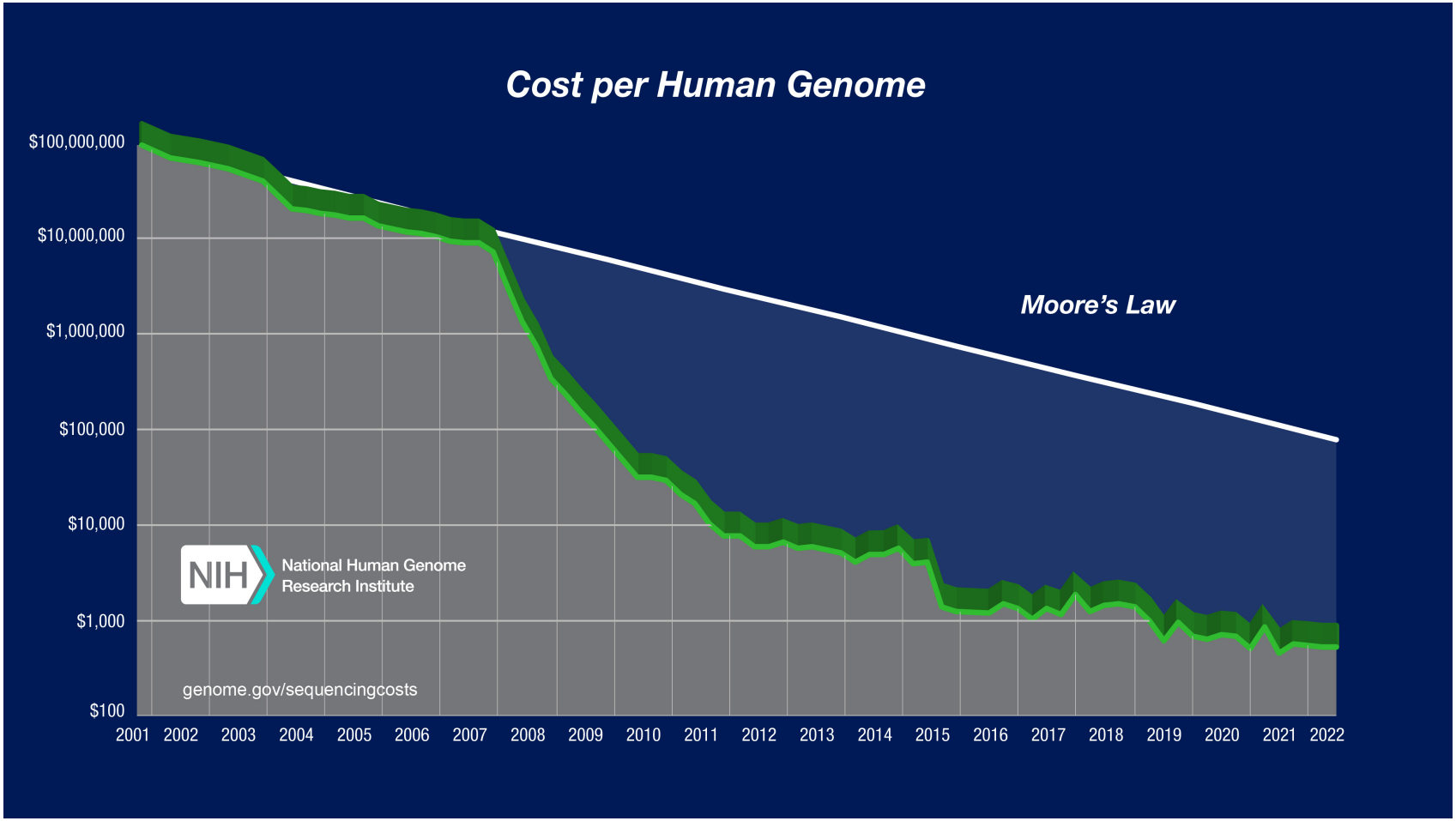
Genome Sequencing

Genome sequencing is now a routine part of genetic research



From: Tang et al. 2019 *Plant Communications*

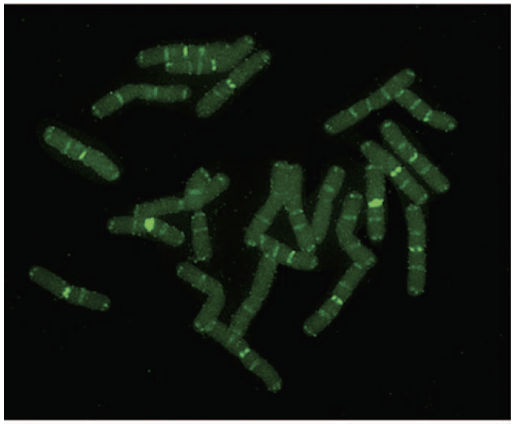
DNA Sequencing - Costs



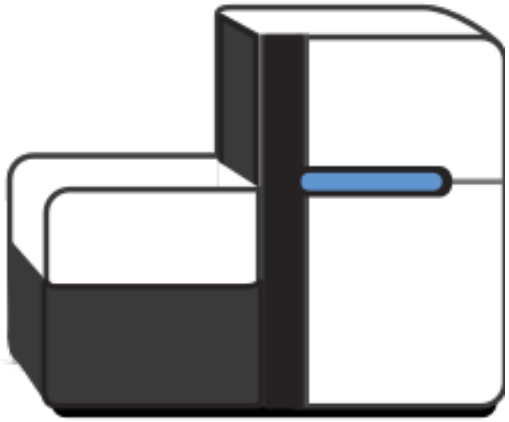
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

DNA Sequencing

The rapid accceleration in genome science has been facilitated by technological advances in sequencing technology



$\rightarrow$



DNA sequencing machines

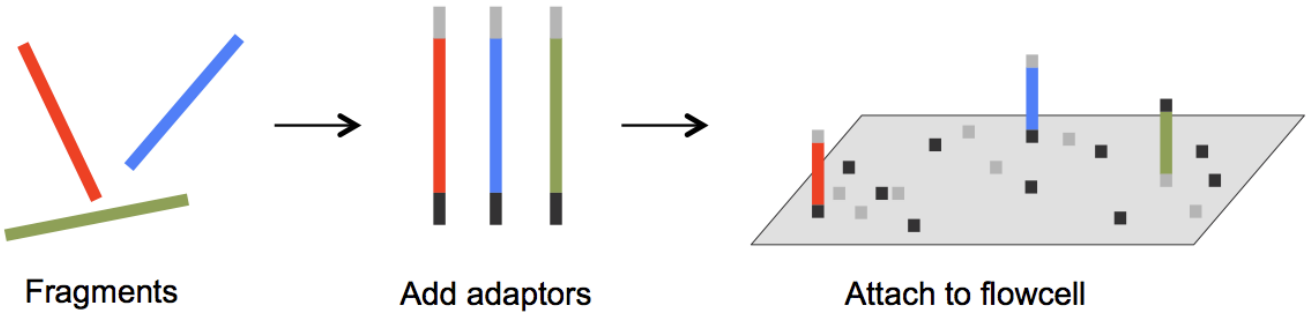
- There are numerous technologies available
- The various technologies have different attributes

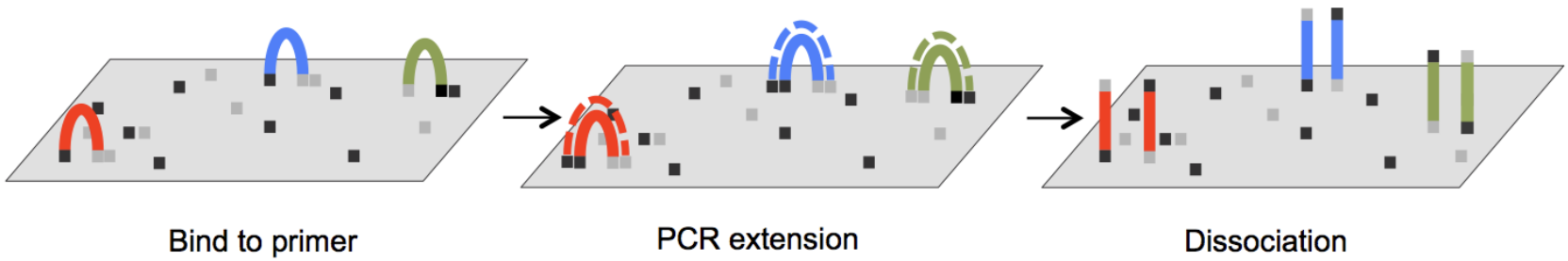
DNA Sequencing - Illumina technology

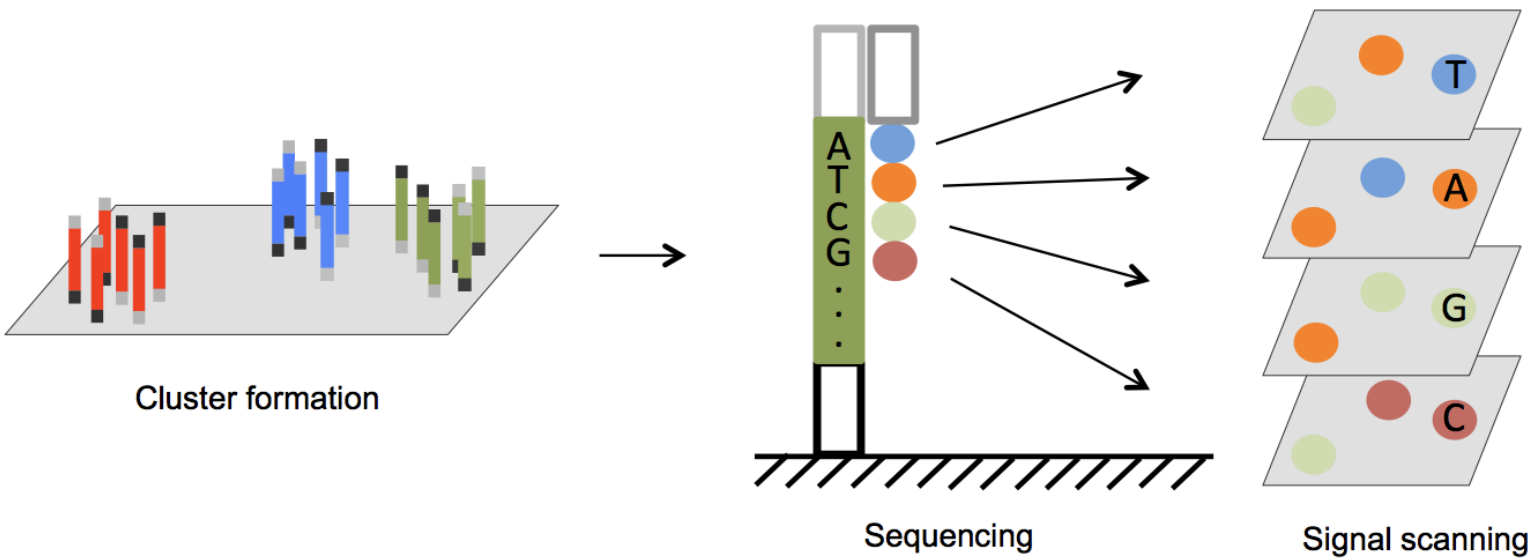
- Based on sequencing by synthesis - chain elongation without termination
- Solves the problem of throughput
- Sequence millions of DNA fragments at the same time
- MASSIVE parallelization

There are/were several other companies in this space, but they have mostly been eclipsed by Illumina

DNA Sequencing - Illumina technology

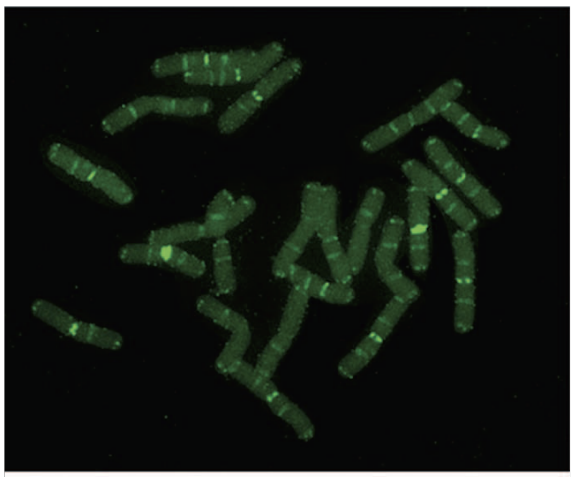




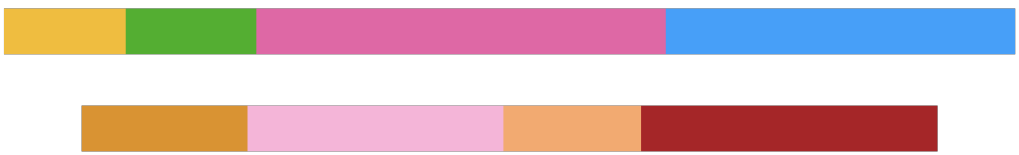


Genome Sequencing

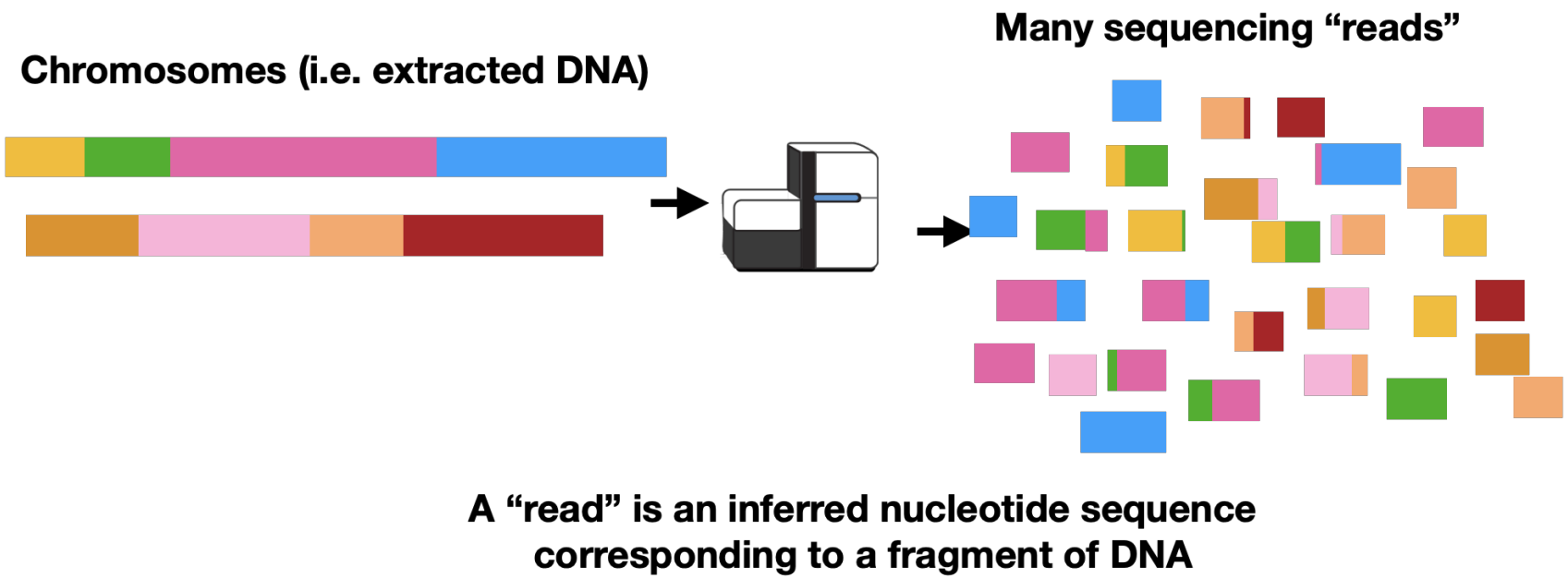
Loblolly pine chromosomes



Cartoon chromosomes

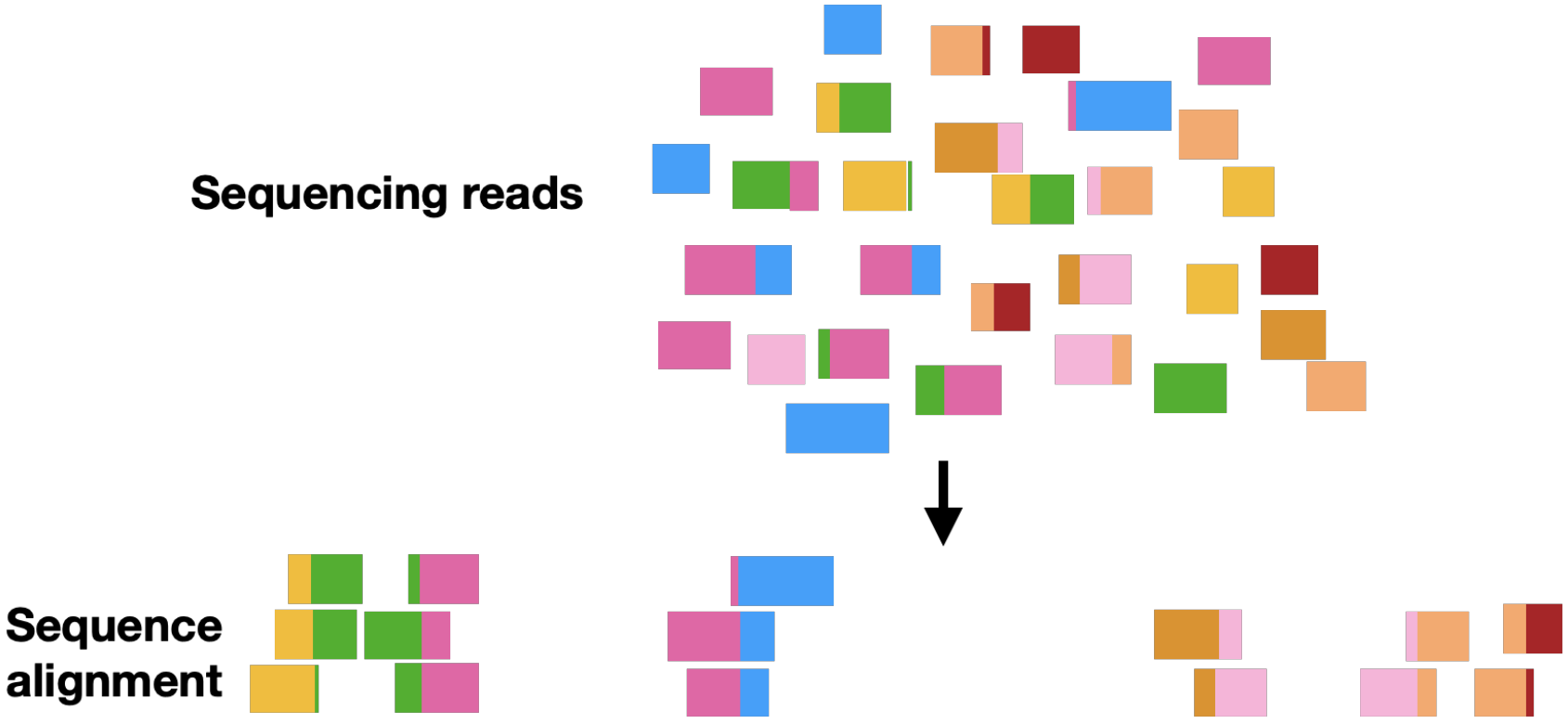


Genome Sequencing



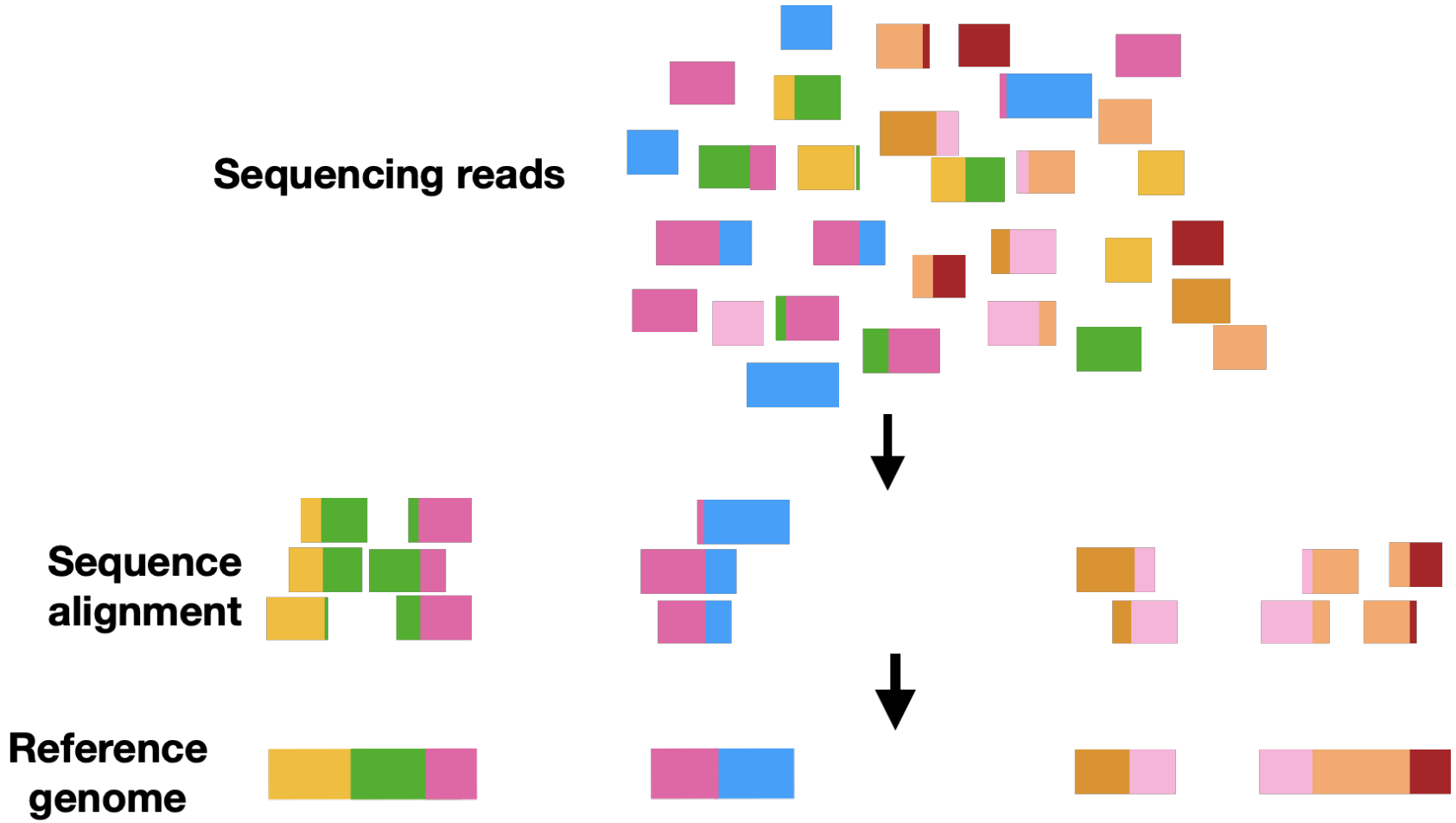
This basic idea has been around since the 1980s (*i.e. shotgun sequencing*), but represents the foundation of how we do genome sequencing today

Genome Sequencing



This step involves identifying overlaps between reads - it can involve *LOT* of computation

Genome Sequencing

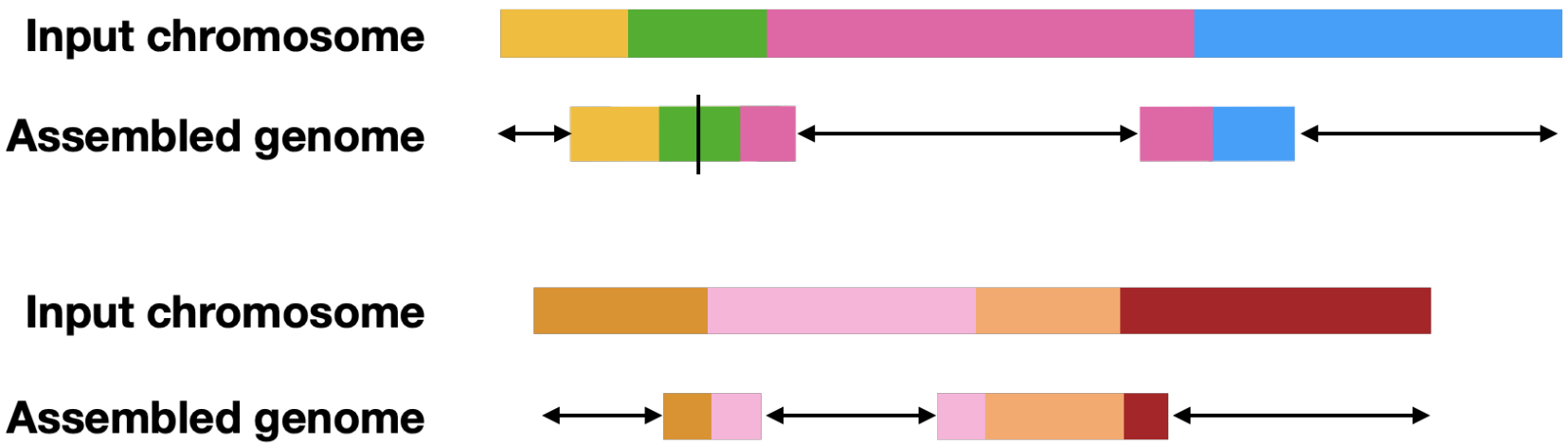


The alignment of the sequencing reads is used to reconstruct the input sequence

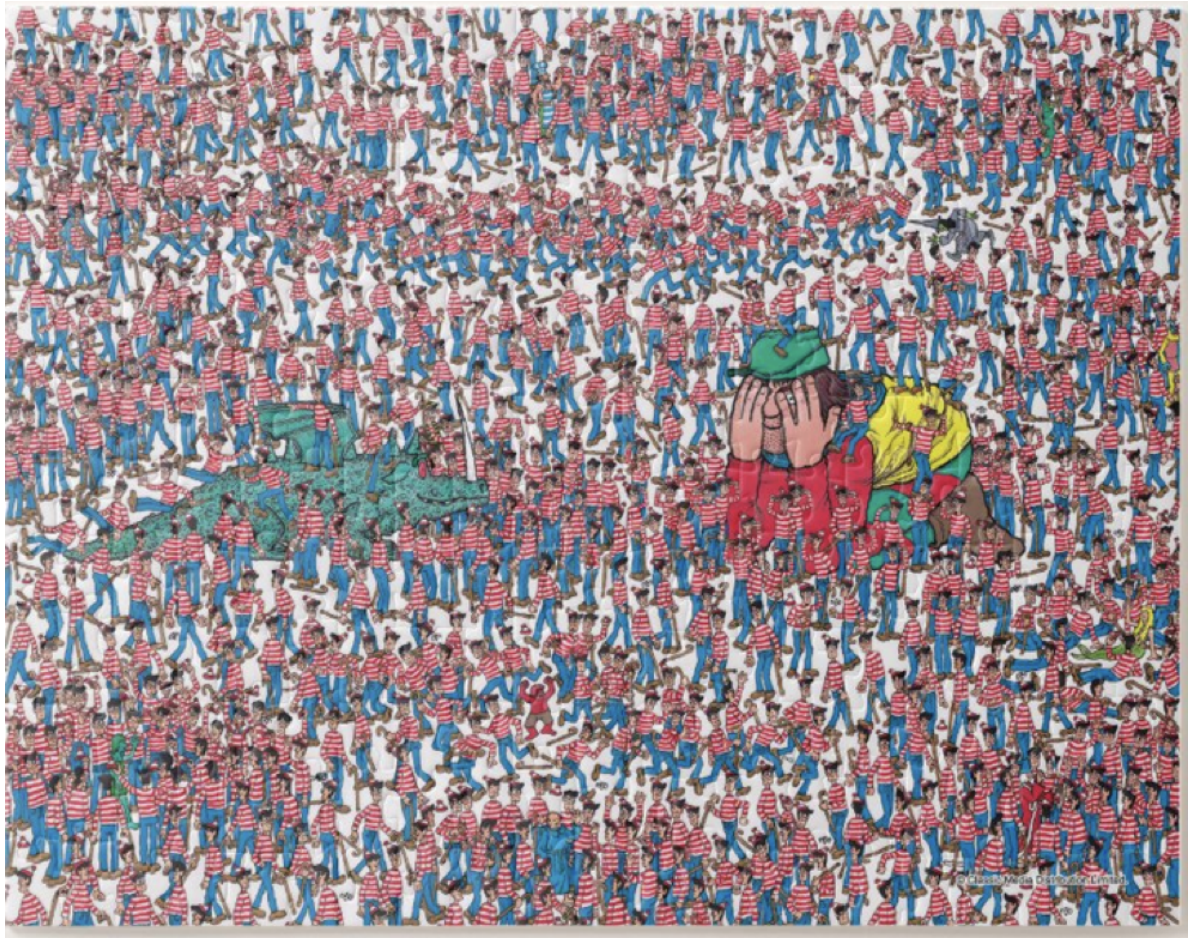
Genome Sequencing - Reference Genome

A reference genome is a representation of the average genome for a species/population

- Used as the template against which to evaluate genetic variation[see lecture on genetic variation]
- Reference genomes are incomplete pictures

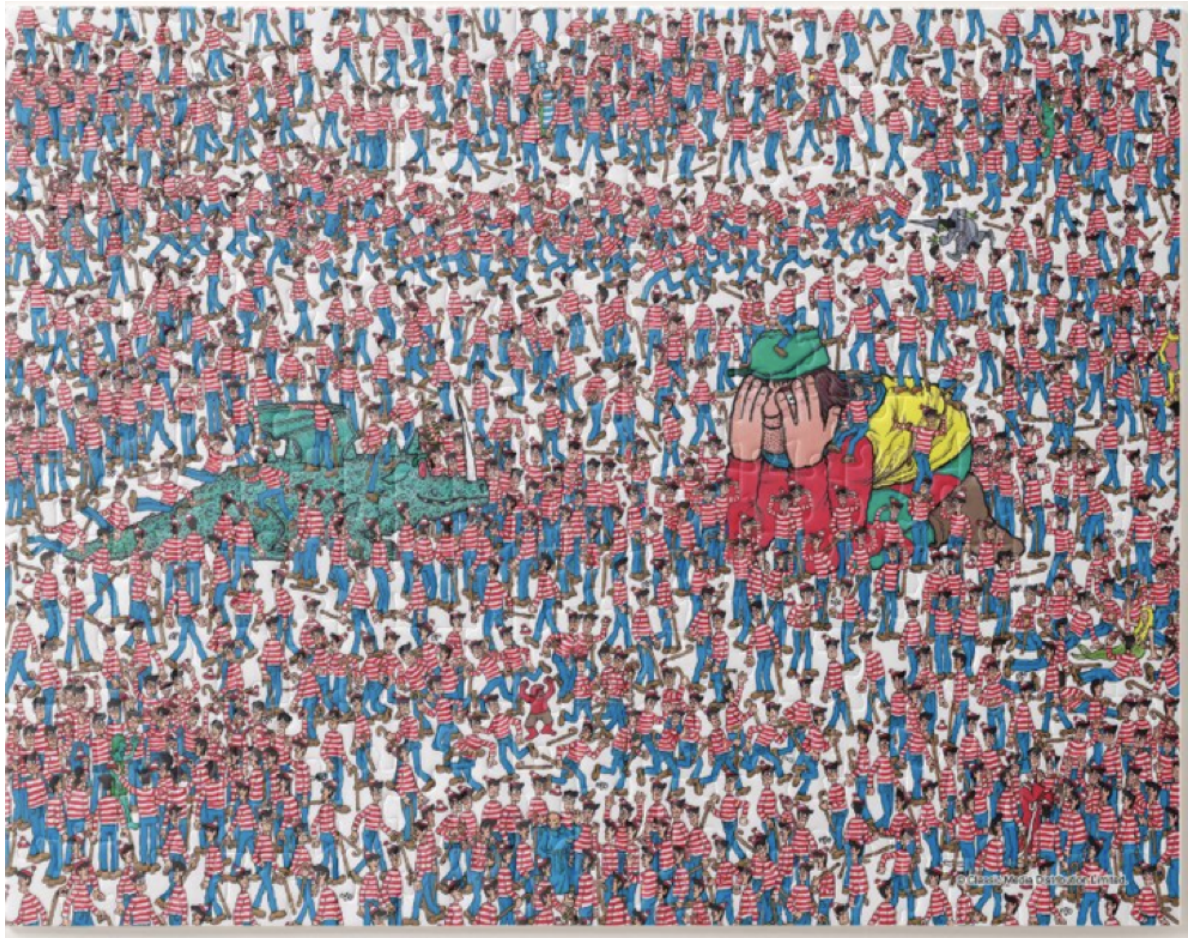


The Land of Waldos - Martin Handford

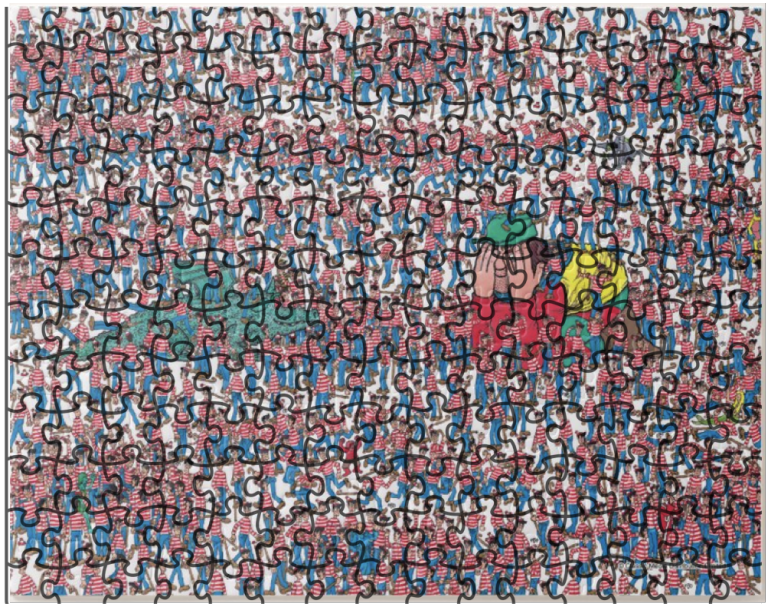


Any questions?

The Land of Waldos - Martin Handford



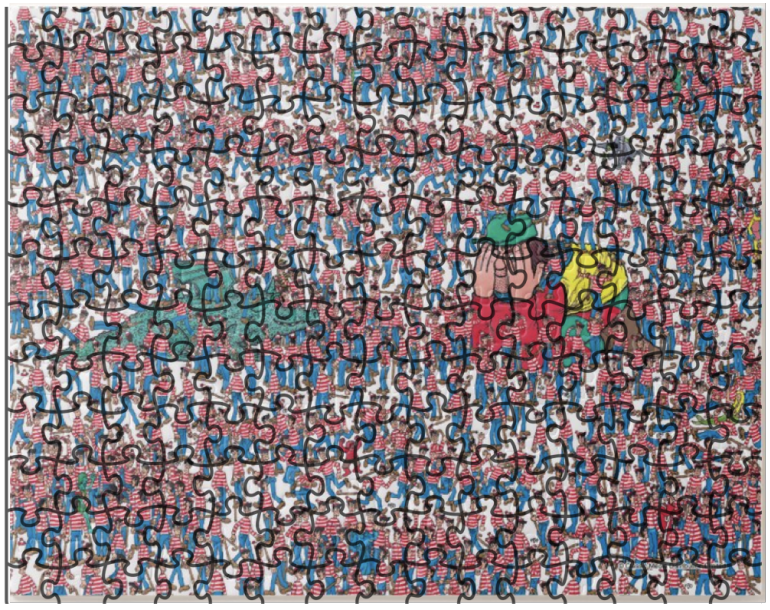
Any questions? Let's take a short break



***The Land of Waldos* - Martin Handford**

A bowl of popcorn

Which of these jigsaw puzzles would be harder to solve?

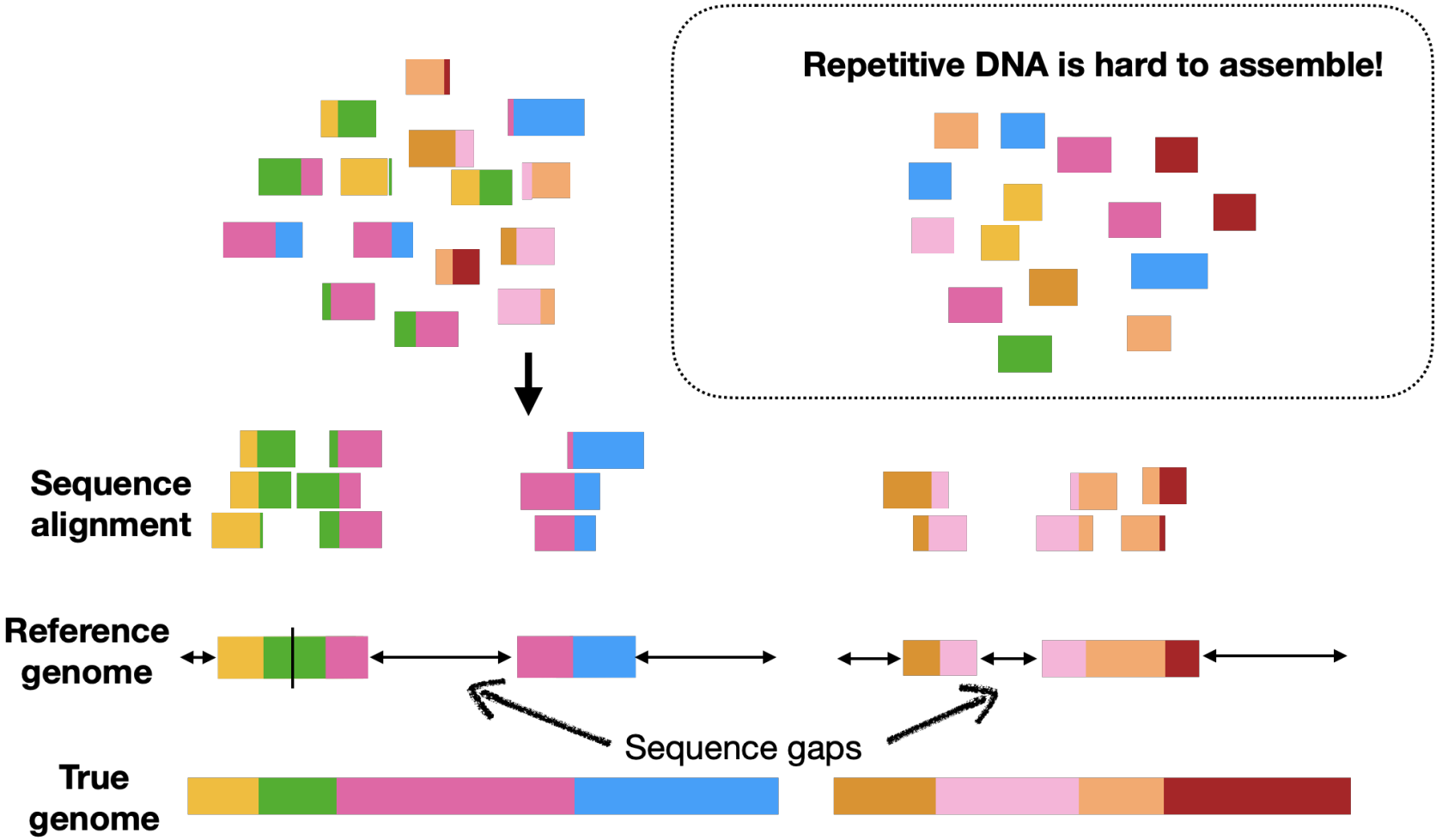


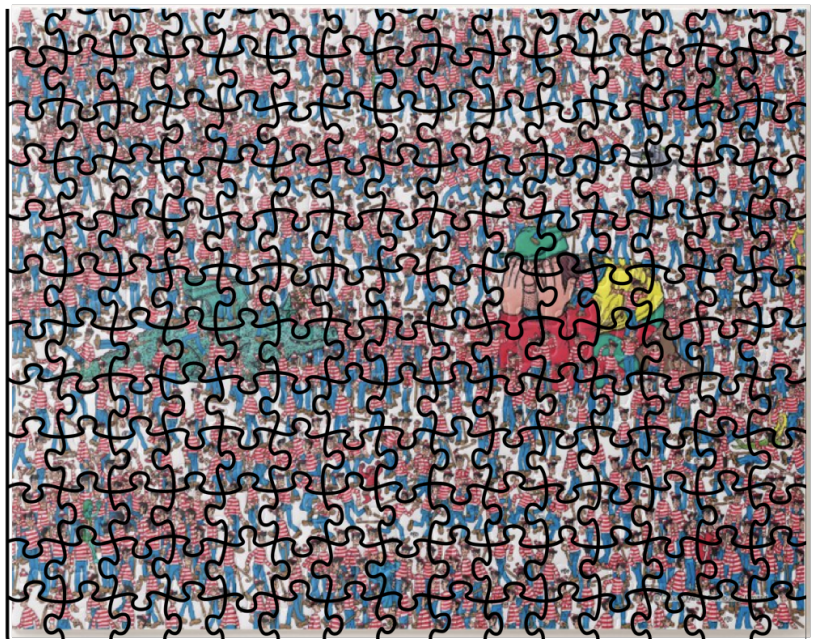
***The Land of Waldos* - Martin Handford**

A bowl of popcorn

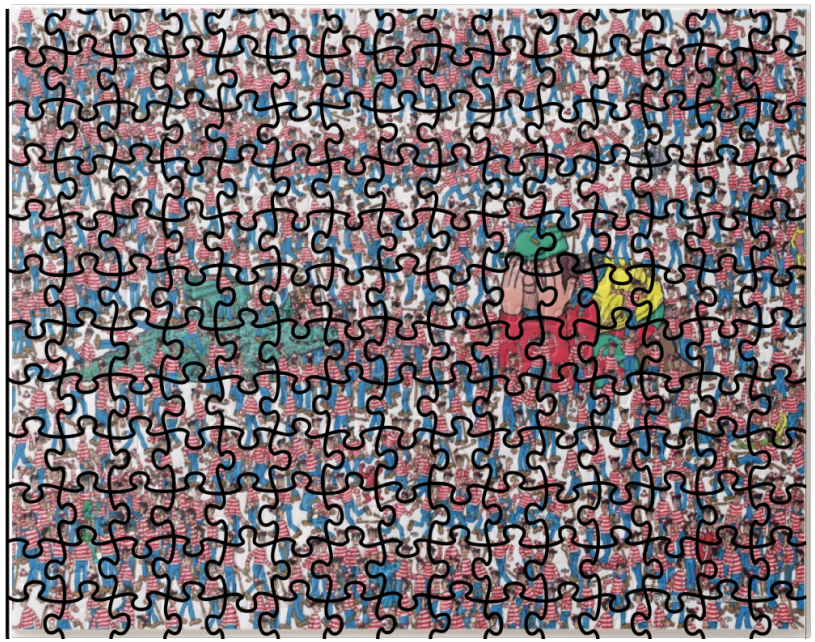
Which of these jigsaw puzzles would be harder to solve? Why?

Genome Sequencing



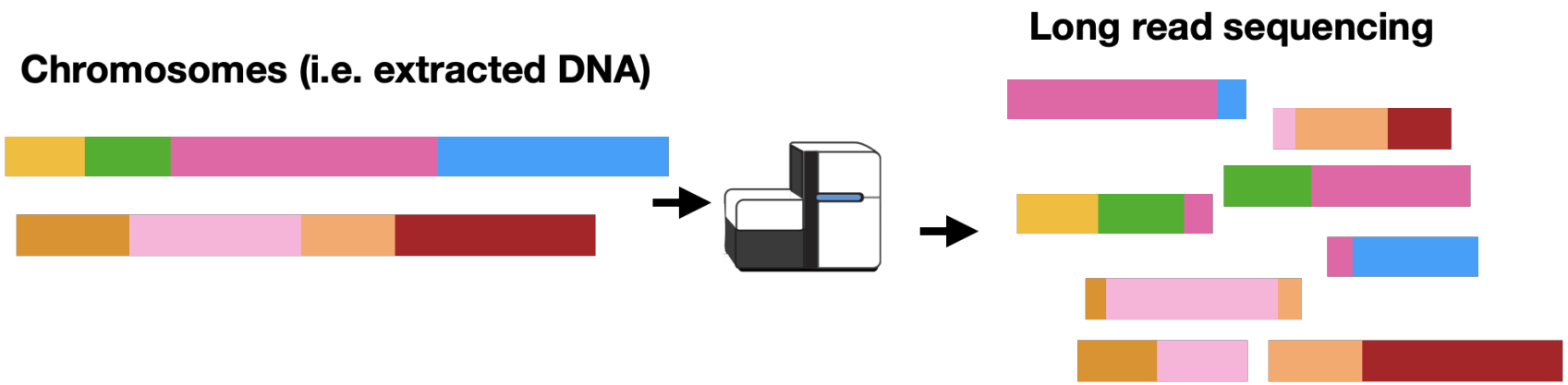


Why is it easier to solve the puzzle on the right?



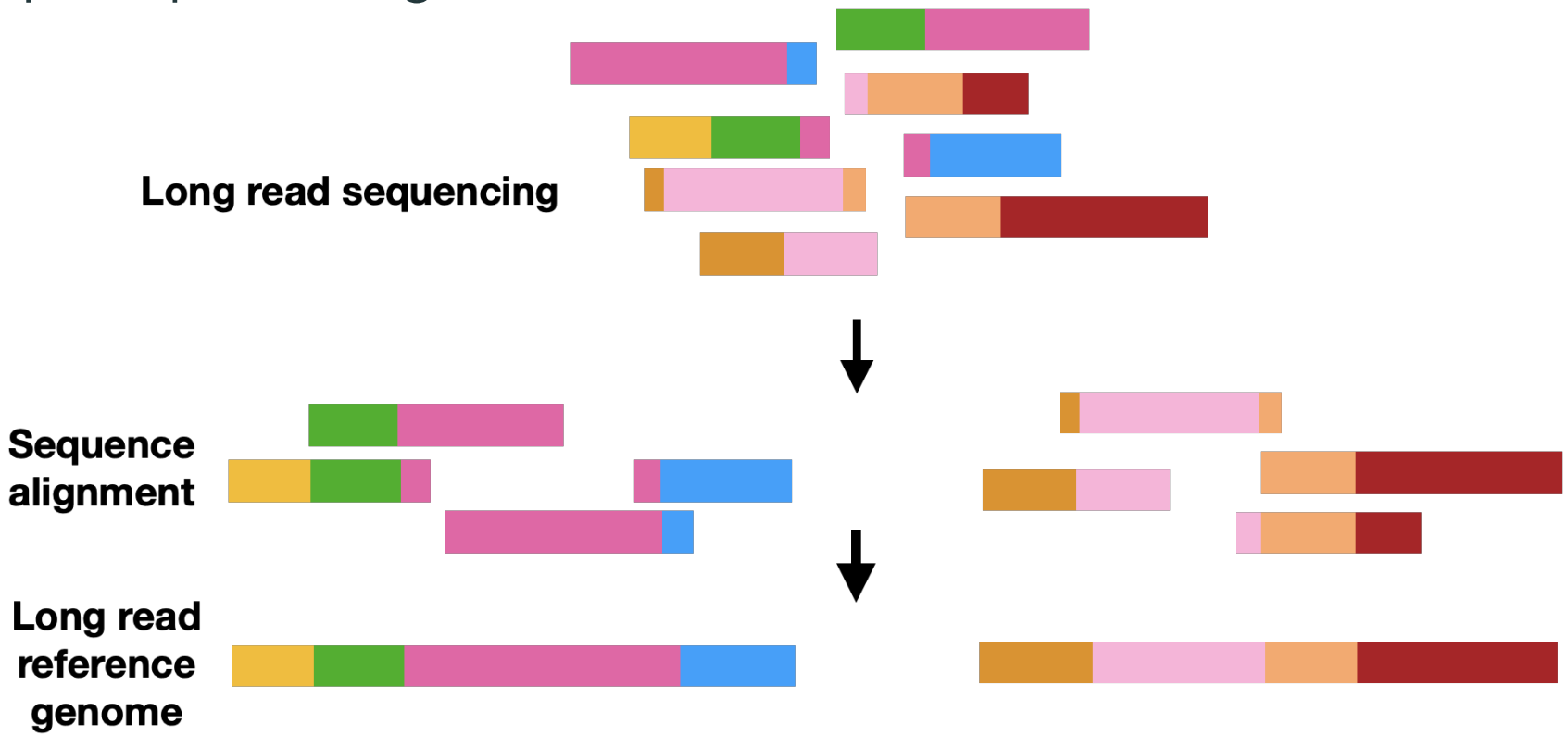
Why is it easier to solve the puzzle on the right?
Puzzles with fewer, larger pieces are easier to solve!

Genome Sequencing - Long Reads

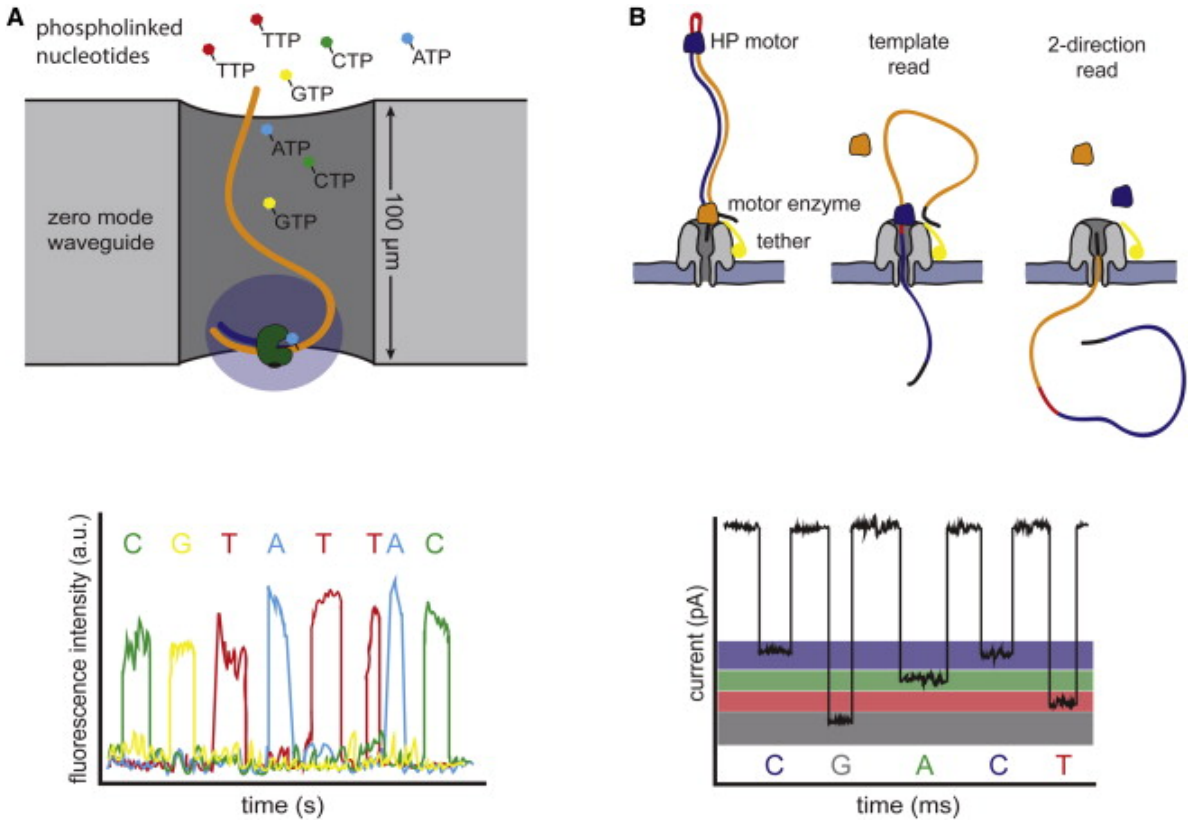


Genome Sequencing - Long Reads

Longer reads improve assembly quality as they are more likely to span repetitive regions!



Genome Sequencing - Long Read Technologies



A PacificBioscience's SMRT sequencing

B Oxford Nanopore sequencing

Image from: Reuter et al 2015 - *Molecular Cell*

Genome Sequencing - Sequencers

Short reads

MiSeq

NovaSeq X

Long reads

MinION

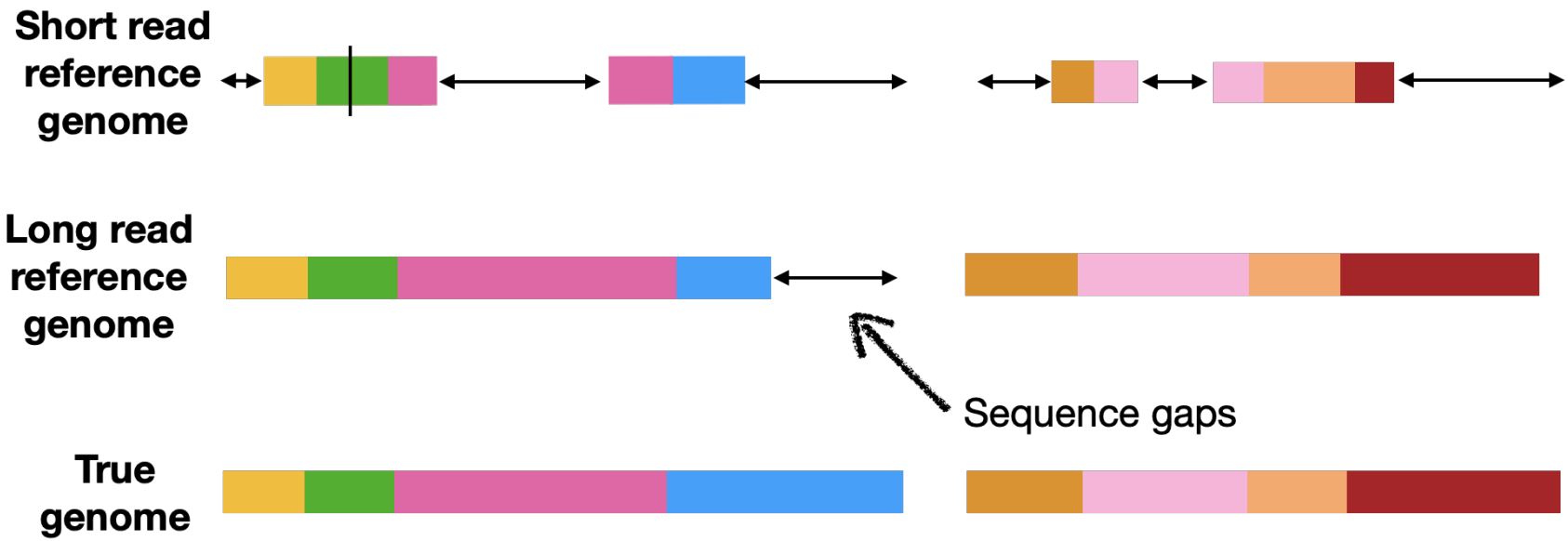
PromethION

Revio

You don't have to memorize the names!

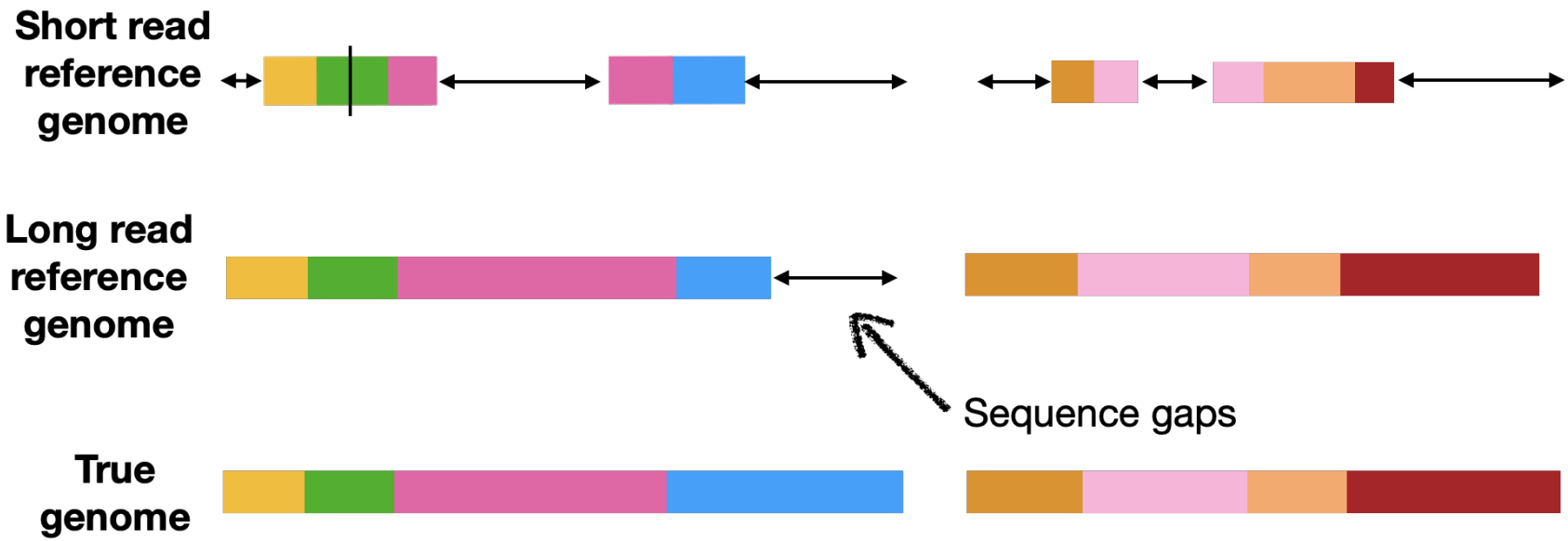
Genome Sequencing - Short vs Long Reads

Longer reads improve assembly quality as they are more likely to span repetitive regions!



Genome Sequencing - Short vs Long Reads

Longer reads improve assembly quality as they are more likely to span repetitive regions!



So why use short reads?

Genome Sequencing - Different approaches

Different sequencing methodologies have their uses in different settings

	Sanger sequencing	Illumina (Short-reads)	PacBio (long-reads)	ONT Nanopore (long-reads)
Cost	—	+++	-	+
Accuracy	+++	++	+	0
Assembly	-	-	+++	++
Computation	—	-	++	++

The above table related to the applicability of the different methods for genome assembly

Questions?

Questions?
Let's take a short break


Welcome to
Twin Peaks
WELCOMETOTWINPEAKS.COM
Man, smell those trees!
Smell those douglas fir genomes

Conifer genomes

Conifer species have clear ecological and cultural importance
Fundamental to forestry in Canada

Douglas Fir

lá:yelhp
(Halkomelem)

Pseudotsuga menziesii

White/Interior Spruce

k̓isiɬqaǂ
(Ktunaxa)

Picea glauca

Lodgepole Pine

apahtó'kii
(Káinai)

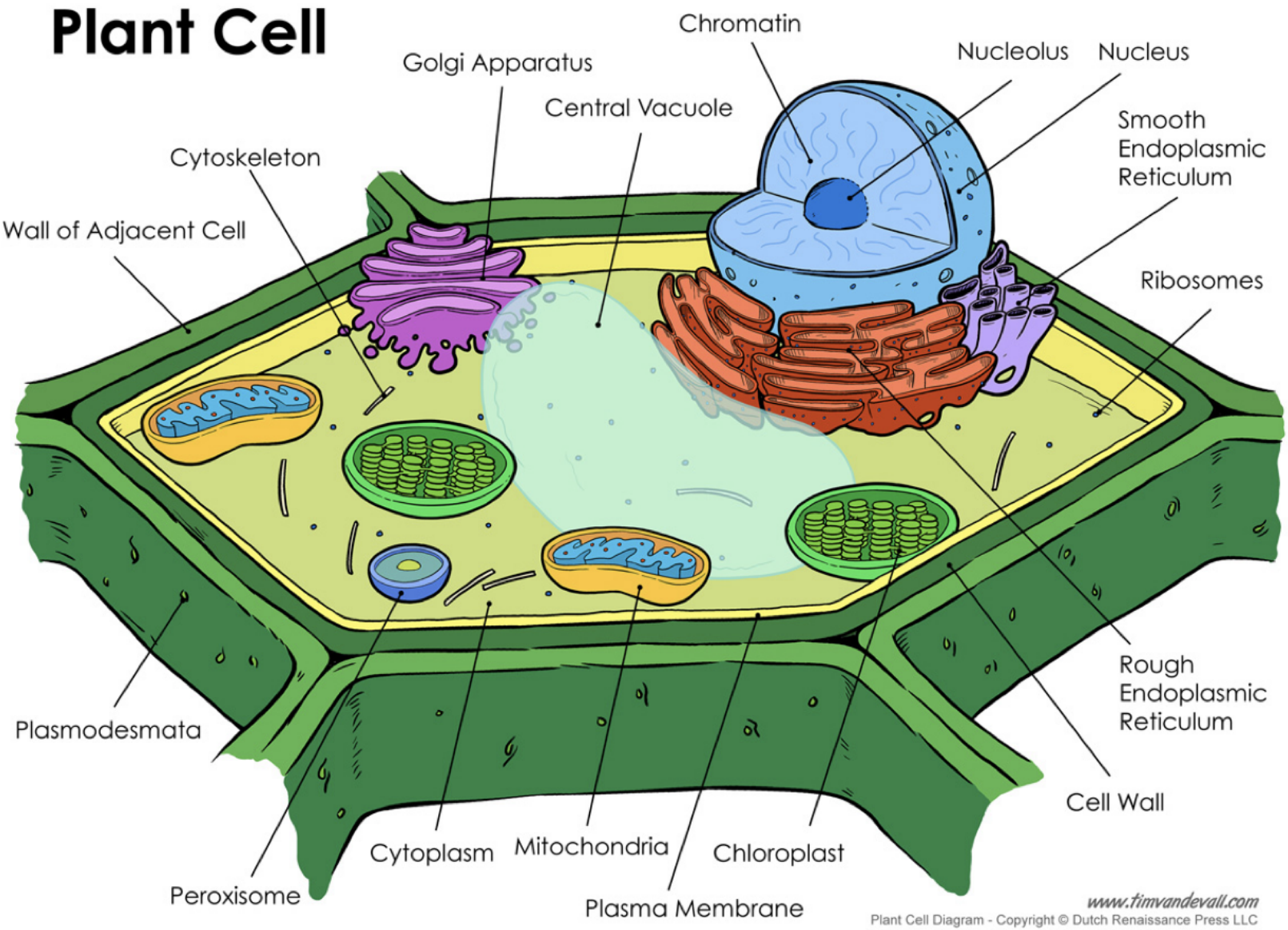
Pinus contorta

Western Red Cedar

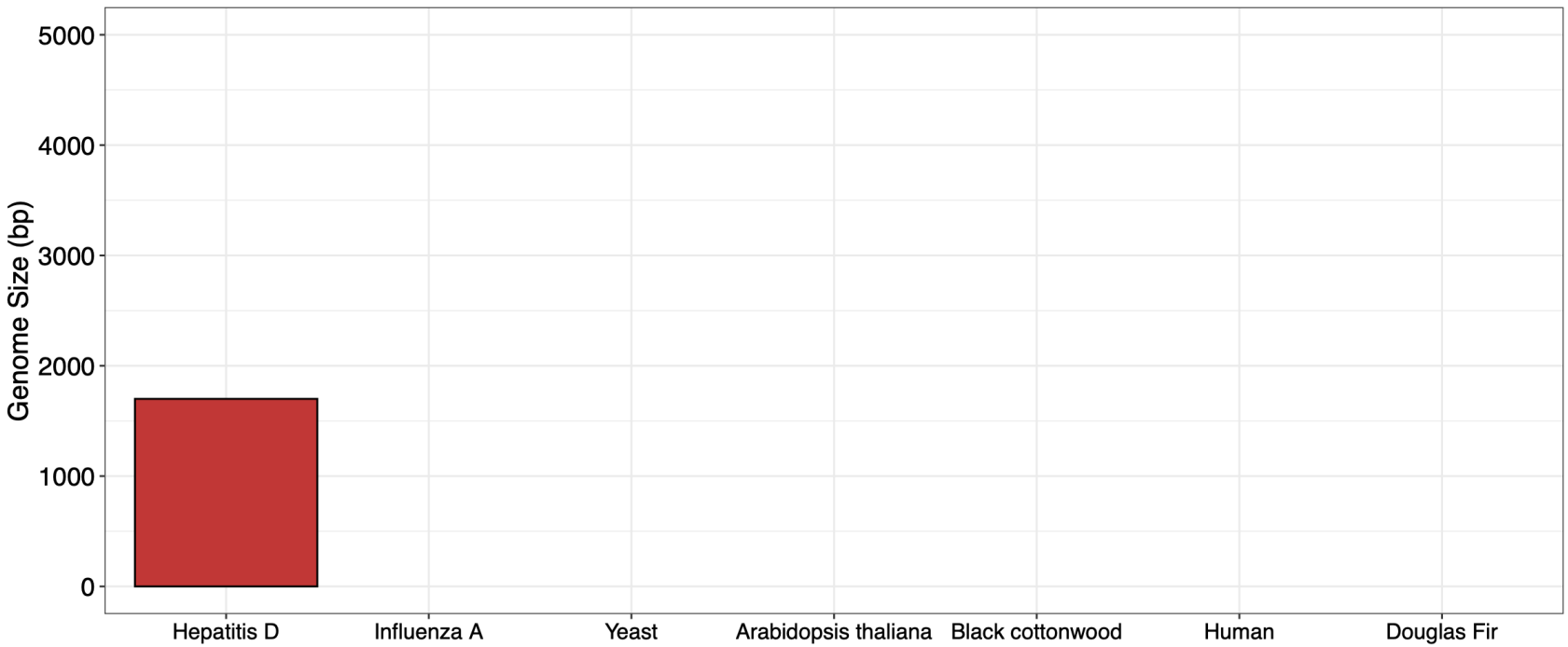
xápaẏay
(Squamish)

Thuja plicata

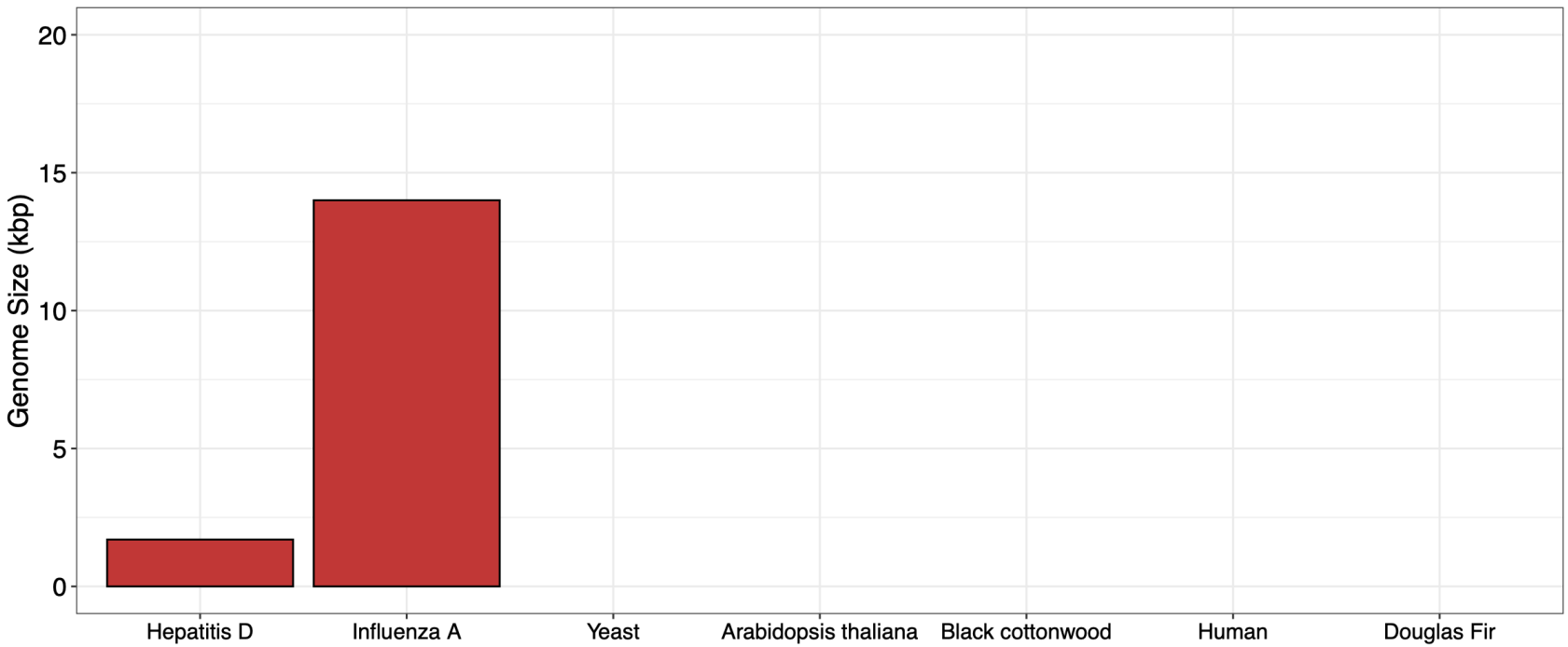
Genome Sizes



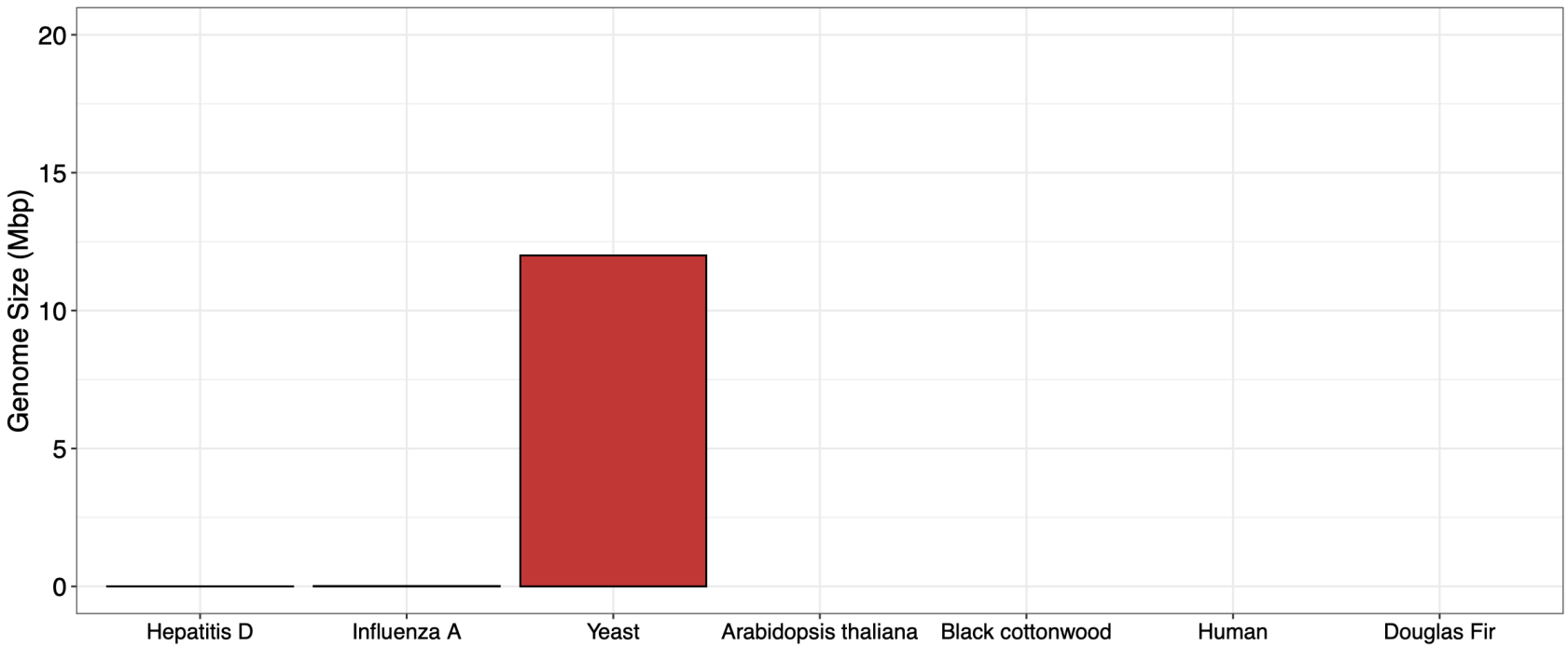
Genome Sizes



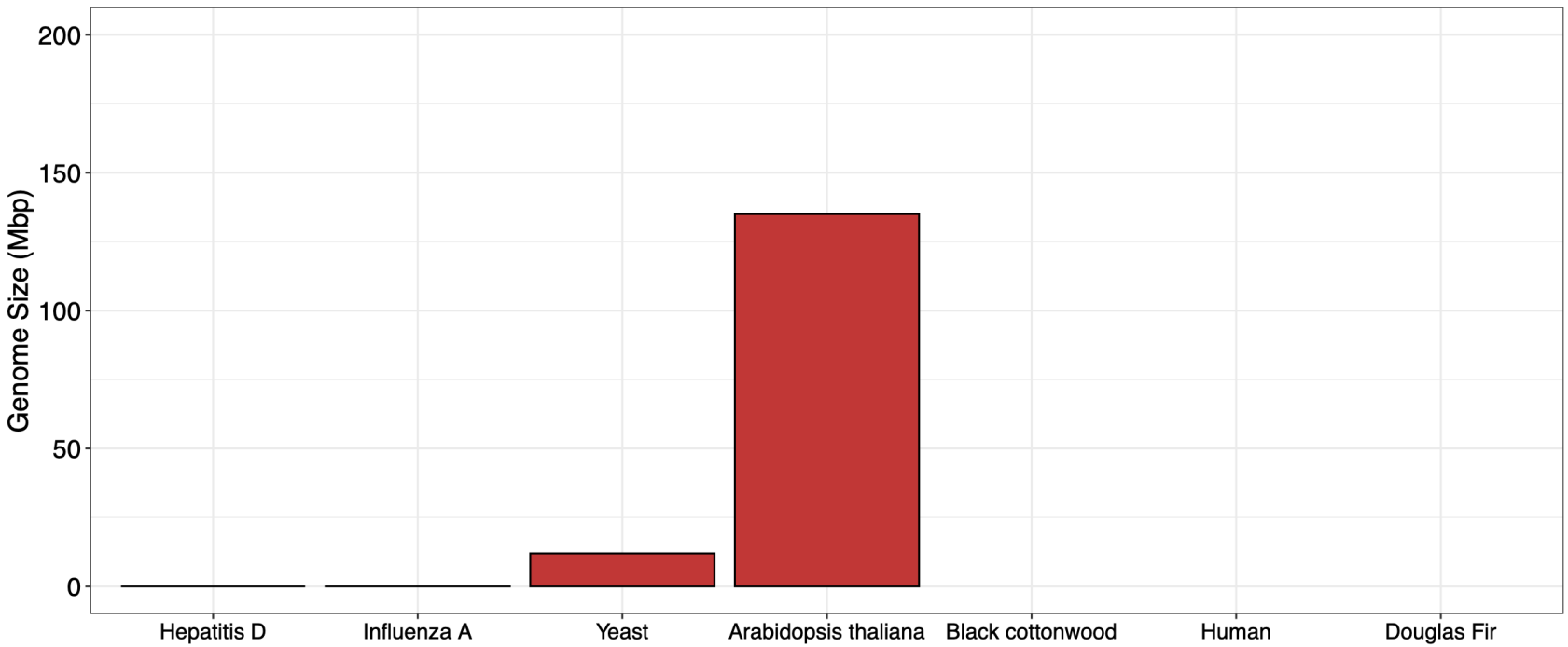
Genome Sizes



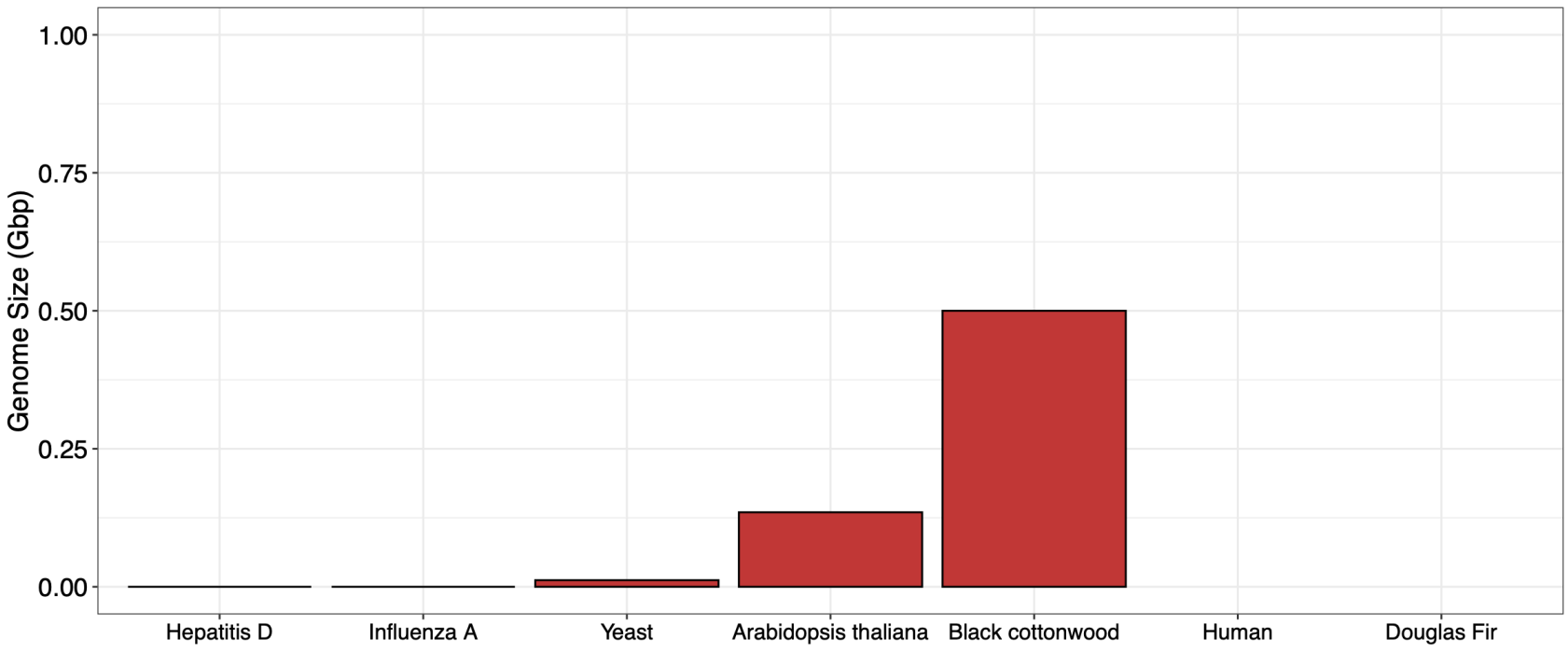
Genome Sizes



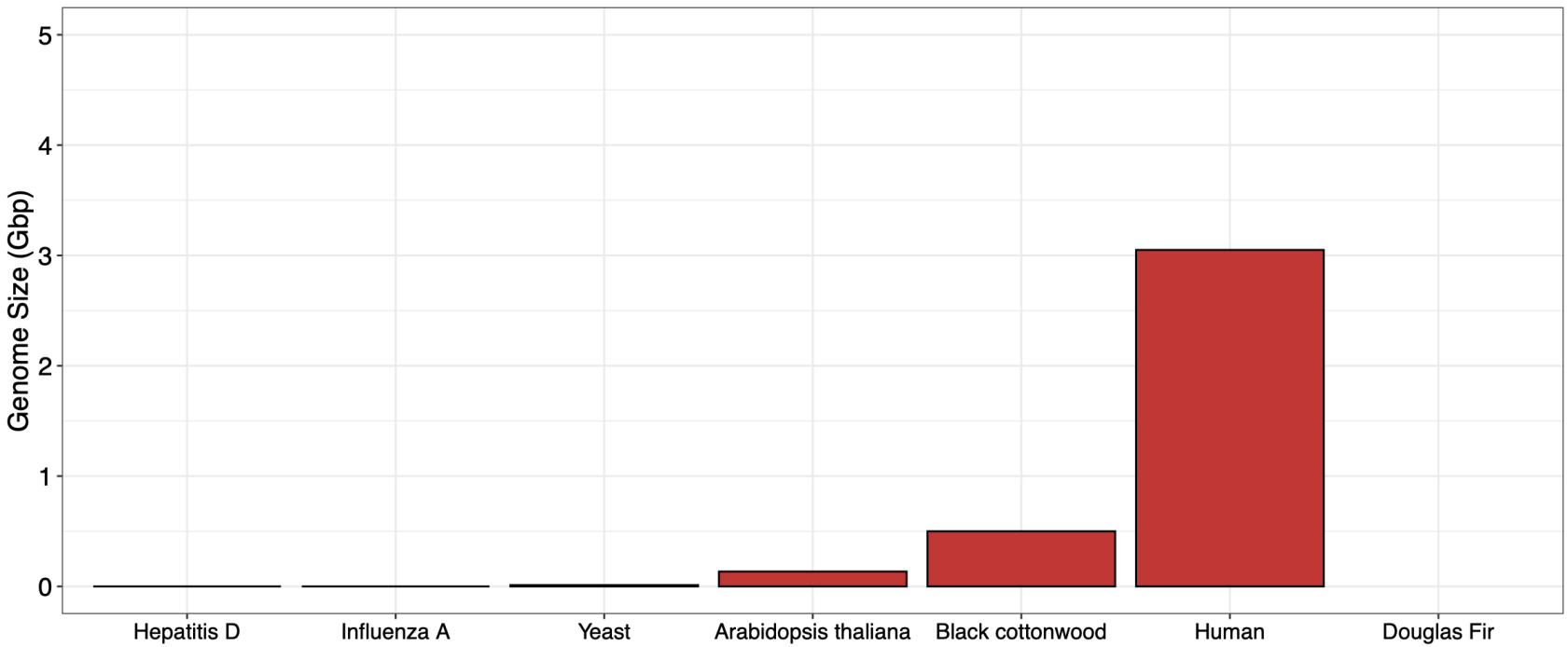
Genome Sizes



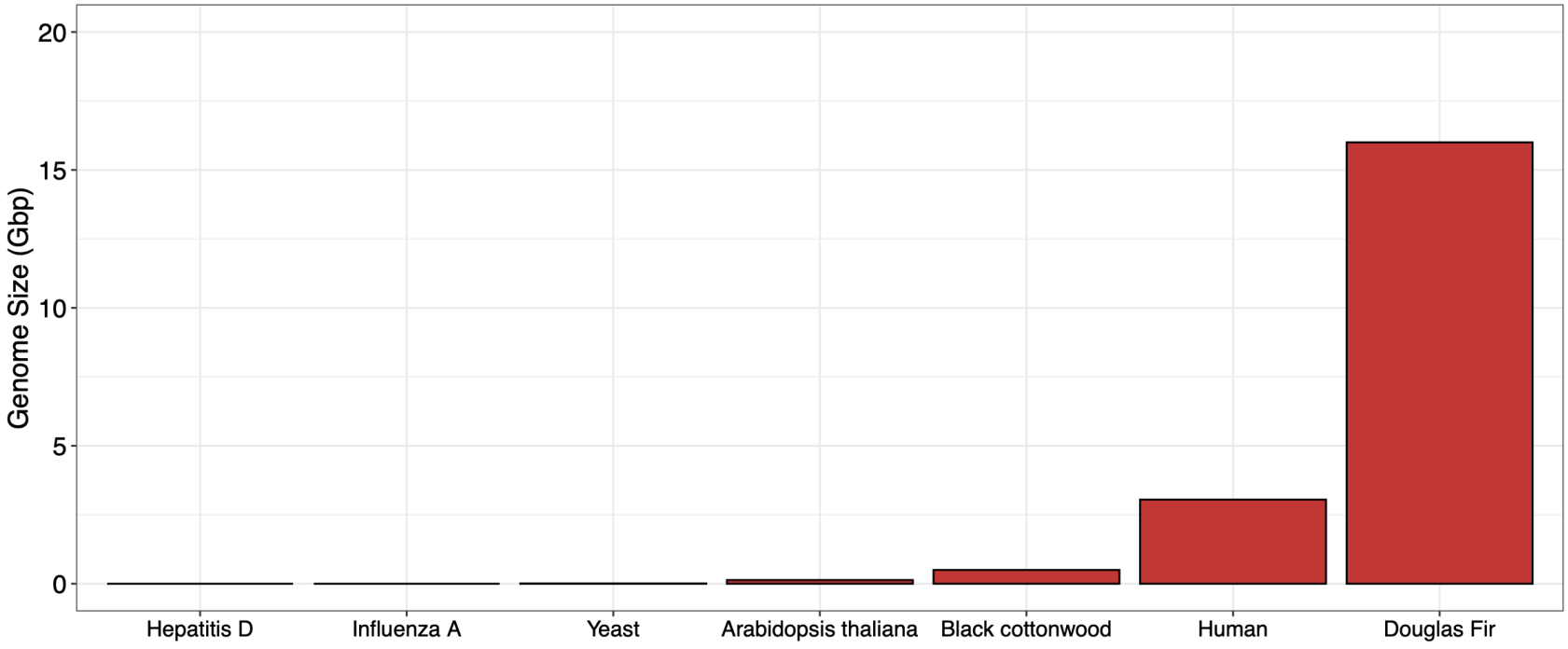
Genome Sizes



Genome Sizes



Genome Sizes



Every cell of a Douglas-fir contains more than 5x the DNA than any human cell!

As with ploidy and chromosome count, DNA volume is not a good predictor of an organism's complexity

Genome Sizes - Conifers

Douglas Fir

lá:yelhp
(Halkomelem)

Pseudotsuga menziesii

16Gbp

White/Interior Spruce

ǩisiȼ̓qaǂ
(Ktunaxa)

Picea glauca

20Gbp

Lodgepole Pine

apahtó'kii
(Káinai)

Pinus contorta

>20Gbp

Western Red Cedar

x̱ápay̓ay
(Squamish)

Thuja plicata

10Gbp

The smallest of these is still more than $3x$ the human genome!

Conifer genomes are big and repetitive

conifer genomes are ibg and repetitive

Confiiger genomes are big and repetitive

Cnifer genomes are big and repetitive

Conifer GENomes are big and repetitive

Conifer genomes are big and repetitive

Conifere genomes are big and repettiive

Print Out the Douglas-fir Genome

The Douglas fir genome is 16Gbp long!

How many trees would you need to cut down to print out the Douglas fir genome?

Single sided using MS Word default settings

Print Out the Douglas-fir Genome

- The Douglas fir genome is approximately 16Gbp in size
- With standard formatting a piece of paper (US Letter) would contain about 3,000 characters (As, Ts, Gs, Cs and Ns)
- That gives a total of $(16 \times 10^9)/3000 = 5,333,333$ million sheets of paper
- A single pine tree (45' trunk, 8" diameter) would give you about 10,000-20,000 sheets of paper

Print Out the Douglas-fir Genome

- The Douglas fir genome is approximately 16Gbp in size
- With standard formatting a piece of paper (US Letter) would contain about 3,000 characters (As, Ts, Gs, Cs and Ns)
- That gives a total of $(16 \times 10^9)/3000 = 5,333,333$ million sheets of paper
- A single pine tree (45' trunk, 8" diameter) would give you about 10,000-20,000 sheets of paper

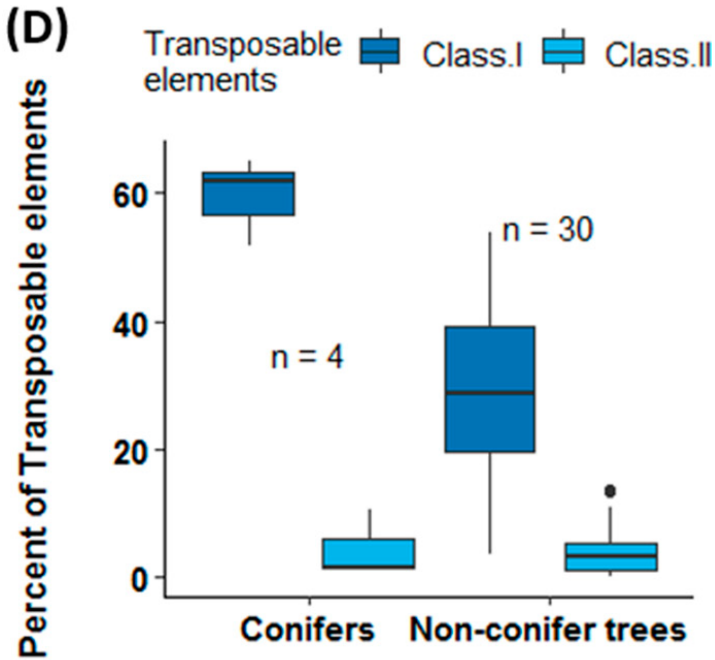
You would need to cut down about 360 trees to print it out

What Makes Conifer Genomes so Large?

Mobile DNA elements (aka jumping genes) can duplicate themselves in a host's genome

Similar life cycle to a virus

Responsible for huge proportion of the repetitive DNA in eukaryotic genomes



Transposable elements are truly fascinating! Figure from Liu and El-Kassaby 2019 *Genes*

The Douglas-fir genome

The first assembly of the Douglas-fir genome was published in 2017

- Primily based on short-read (151bp) technology
- Assembed genome size of 14.95Gbp
- N50 = 341kbp
- 2.8 million assembled chunks

N50 = at east half the assembled chunks are in chunks greater than or equal to this size. This number can be used to compare different genome assemblies

The Douglas-fir genome

The first assembly of the Douglas-fir genome was published in 2017

- Primily based on short-read (151bp) technology
- Assembed genome size of 14.95Gbp
- N50 = 341kbp
- 2.8 million assembled chunks

UBC MSc student Meg Smith is currently (i.e. upstairs right now) re-assembling the Douglas-fir genome

- Primily based on long-read technology (median read length 15kbp)
- Assembed genome size of 15.21Gbp
- N50 = 1.34Gbp
- 3,012 assembled chunks

Note: This is not a criticism of the 2017 assembly, it's a statement on how rapidly the technilogy has moved on!

N50 = at east half the assembled chunks are in chunks greater than or equal to this size. This number can be used to compare different genome assemblies

The Douglas-fir genome

The long-read based genome assembly of Douglas-fir is a vast improvement over the earlier, short-read based one. However,

- Repetitive DNA is still a **major** stumbling block for genomics
- Understanding the genome is now a managable for Douglas-fir, but this does not make application of genomics straightforward

Learning Outcomes

- Introduction to different sequencing methods
- An introduction to genomics in conifers
- The difficulty of repetitive DNA for genomic analysis
- The pros and cons of different sequencing methods

A Thought For Next Time...

What are some uses for genomics in forestry?

Think about how different aspects of conifer life history make the application of genetic and genomic technology difficult in conifers...