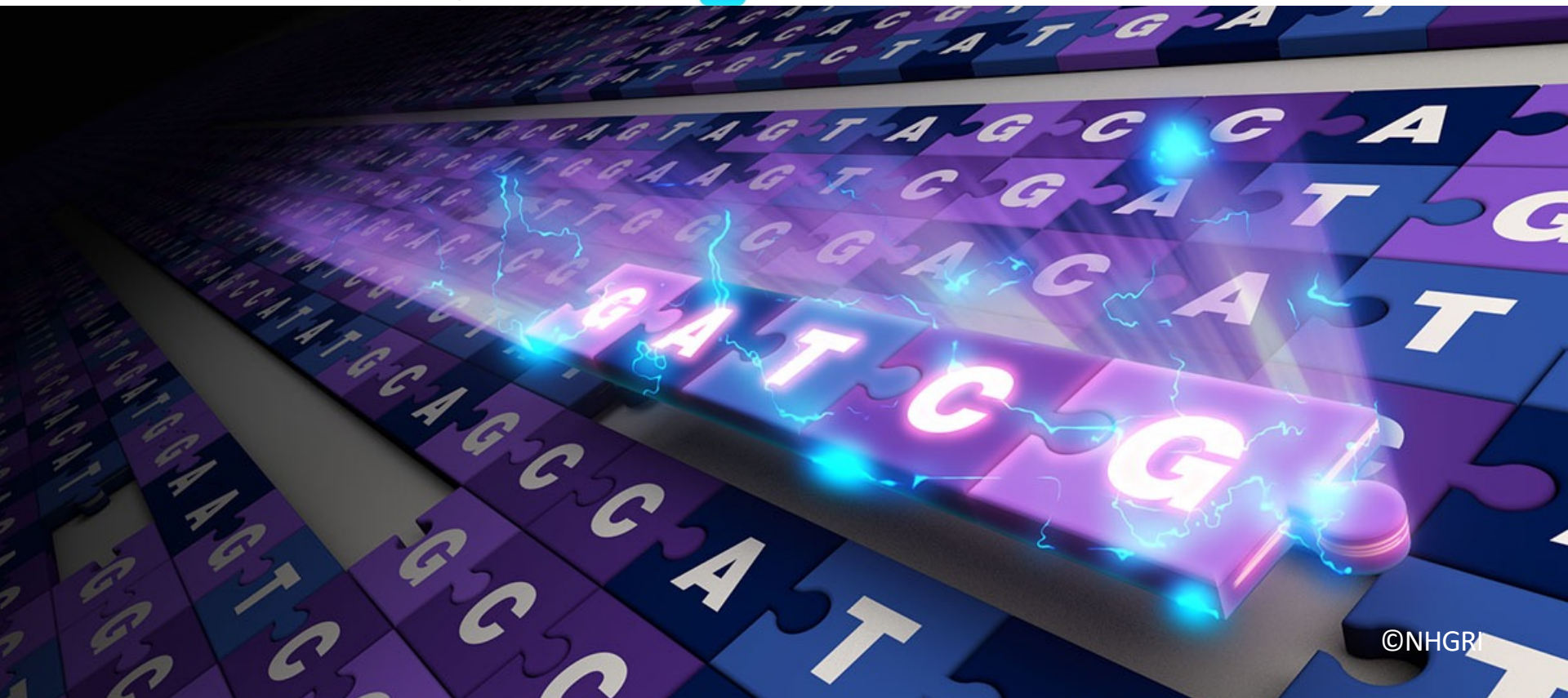
Genome Architecture in 3D Space

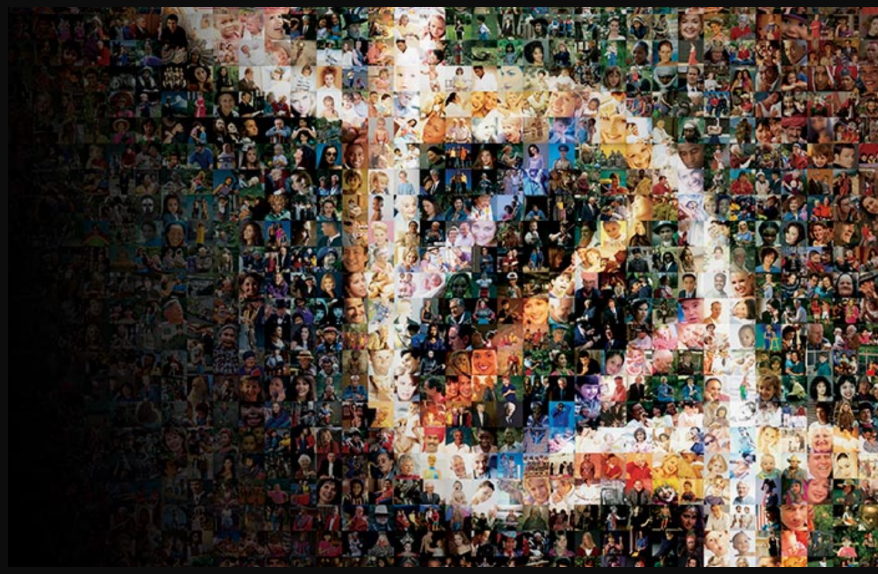
UniL, Nov. 3, 2023

Narcis Yousefi, University of Zürich narjes.yousefi@systbot.uzh.ch @YousefiNarcis



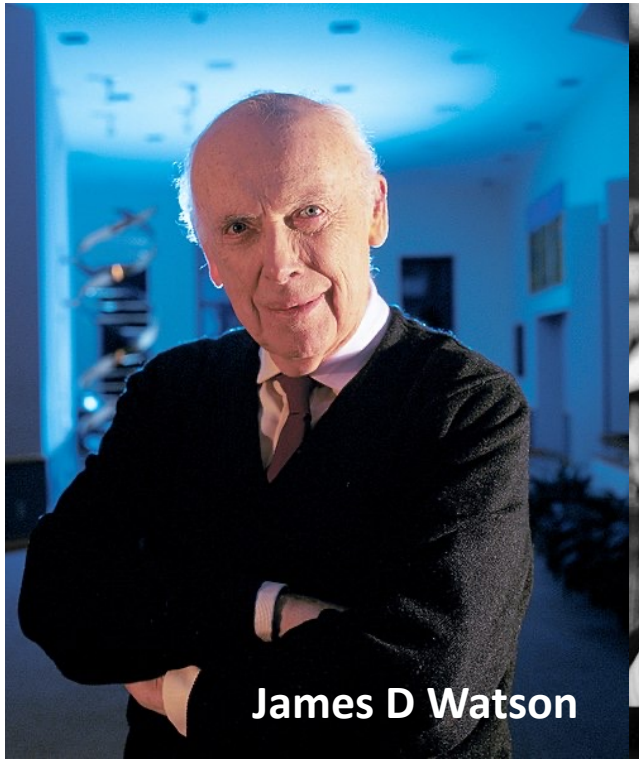
Outline

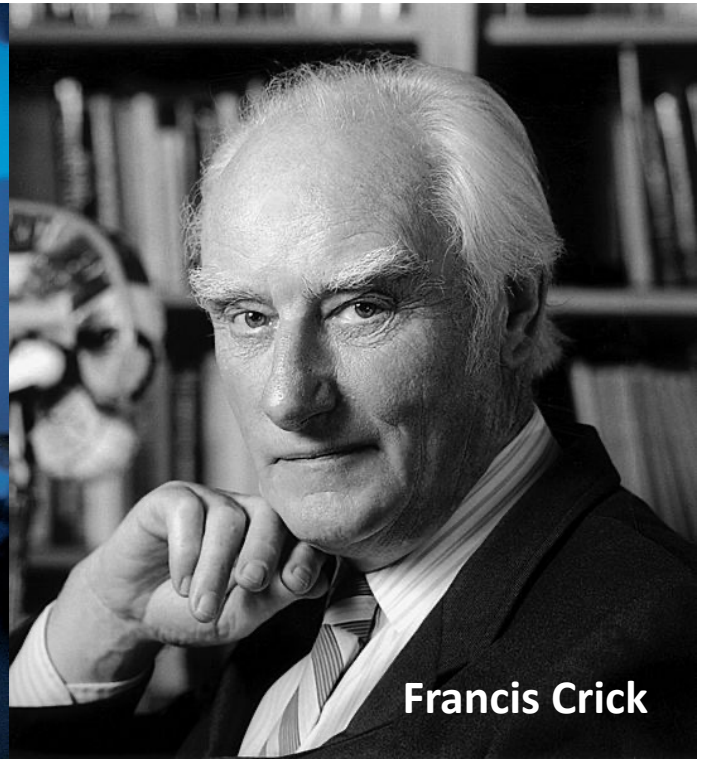
- History of DNA and human genome
- Complexity of the genome
- Supergenes
- 3D genome architecture
- Sequencing techniques

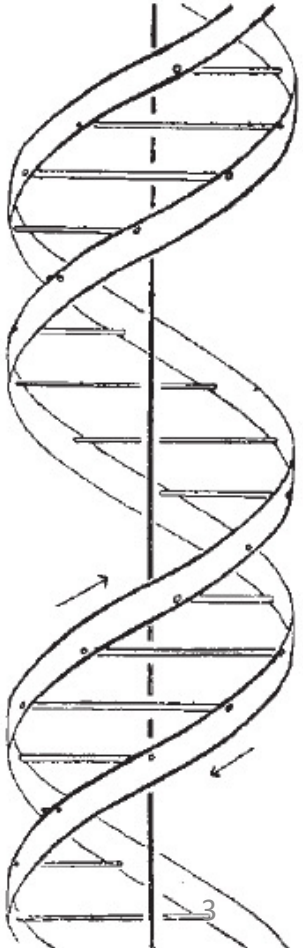


History of sequencing DNA

- 1953: Watson and Crick described the three-dimensional structure of DNA, based on crystallographic data produced by Rosalind Franklin and Maurice Wilkins

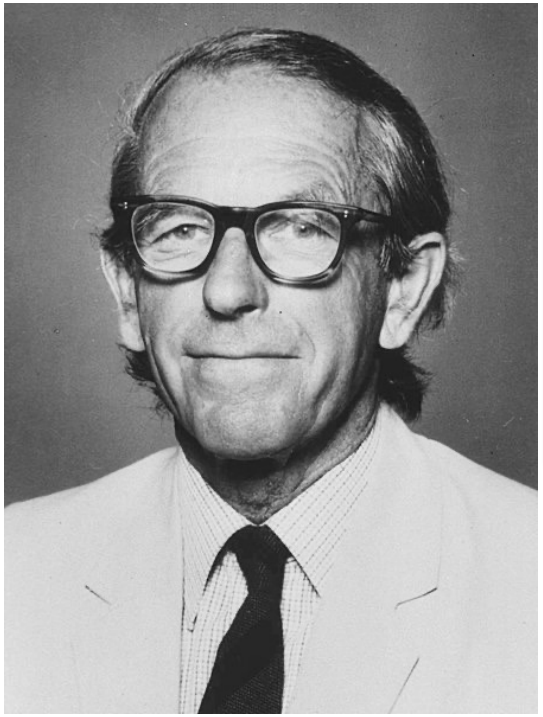






History of sequencing DNA

- 1977: Sanger sequencing method



Frederick Sanger 1918-2013

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

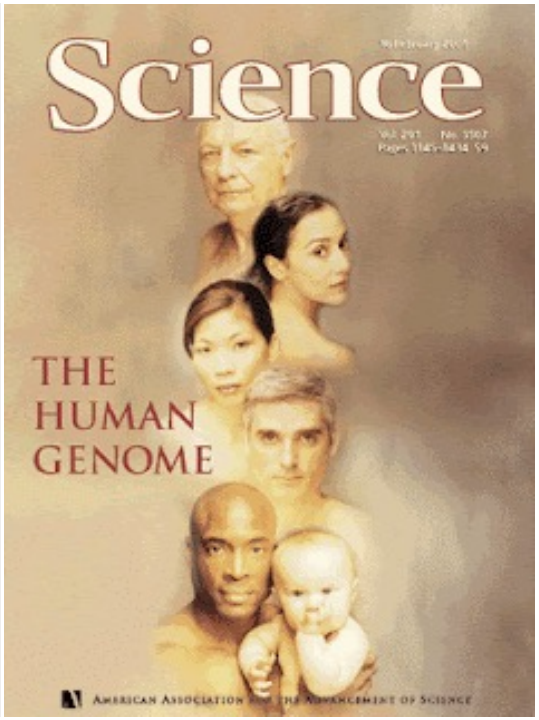
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

Sanger is one of the few scientists who was awarded **two Nobel prizes**, one for the sequencing of proteins, and the other for the sequencing of DNA!

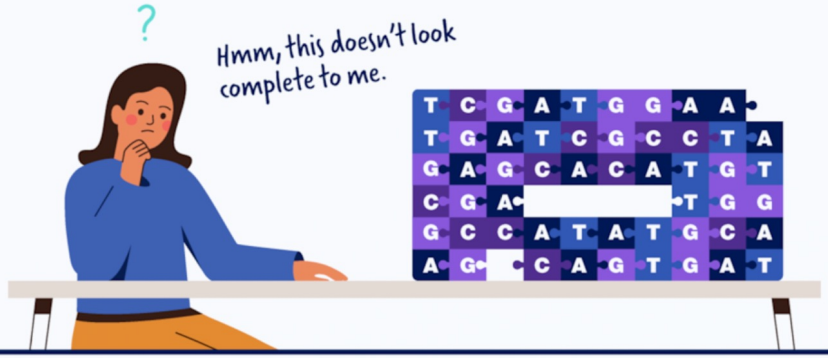
The human reference genome

- 2001 initial releases (>10 years and $3billion; drafts published in *Nature* and *Science*)
- 2003 complete (92%)
- 2022 complete (99%)
- 2023 complete (with last piece of the puzzle; chr. Y)



Why was it so difficult to fully complete the human genome sequence?

The Human Genome Project ended in 2003, but genomic researchers had not yet determined every last base (or letter) of the human genome sequence. Instead, they had only completed about 92% of the sequence at that time. Why did they stop there?

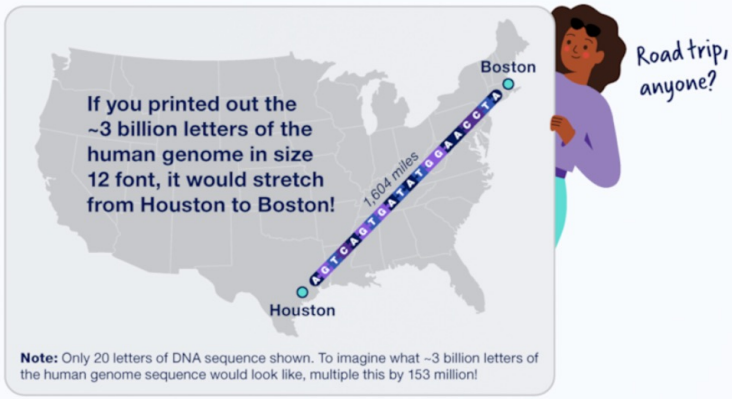


https://youtu.be/SZpcOy63lWw?feature=shared

Reason 1

The human genome contains a massive amount of DNA.

The human genome consists of about 3 billion bases in a precise order, each of which can be represented by a letter (G, A, T or C). A genome's sequence cannot be read out end-to-end. Rather, researchers must first determine the sequence of random pieces of DNA and then use those smaller sequences to put the whole genome sequence back together like a massive puzzle.



Note: Only 20 letters of DNA sequence shown. To imagine what ~3 billion letters of the human genome sequence would look like, multiple this by 153 million!

Reason 2

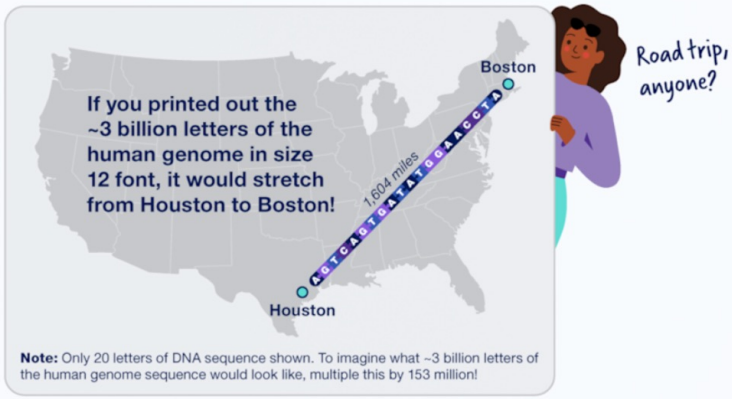
Some parts of our DNA are painfully repetitive.

Some sections of the human genome sequence consist of long, repetitive stretches of letters that are difficult to put in the right place. Over the past two decades, researchers developed new technologies to read longer stretches of DNA — from only about 500 to now over 100,000 letters at a time — which allowed them to assemble the full length of the most difficult repeats.

Reason 1

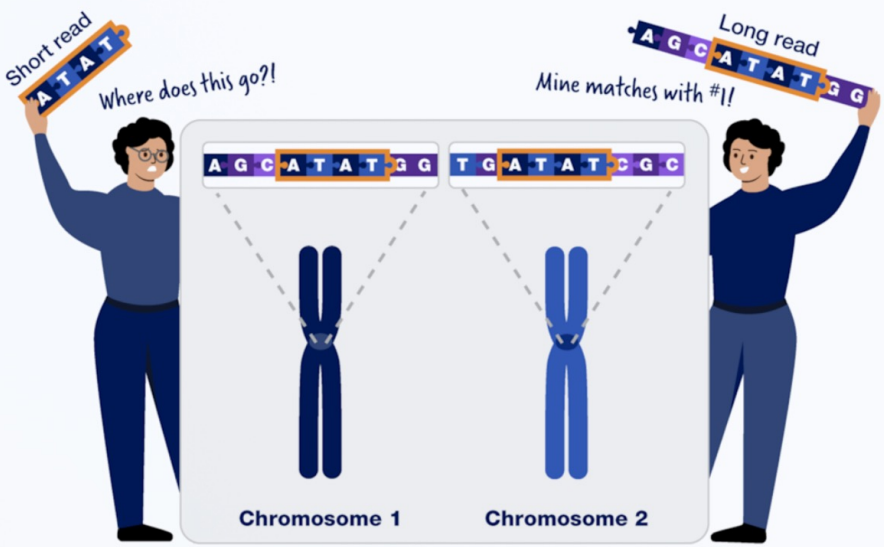
The human genome contains a massive amount of DNA.

The human genome consists of about 3 billion bases in a precise order, each of which can be represented by a letter (G, A, T or C). A genome's sequence cannot be read out end-to-end. Rather, researchers must first determine the sequence of random pieces of DNA and then use those smaller sequences to put the whole genome sequence back together like a massive puzzle.



Note: Only 20 letters of DNA sequence shown. To imagine what ~3 billion letters of the human genome sequence would look like, multiple this by 153 million!

Reason 2



Some parts of our DNA are painfully repetitive.

Some sections of the human genome sequence consist of long, repetitive stretches of letters that are difficult to put in the right place. Over the past two decades, researchers developed new technologies to read longer stretches of DNA — from only about 500 to now over 100,000 letters at a time — which allowed them to assemble the full length of the most difficult repeats.

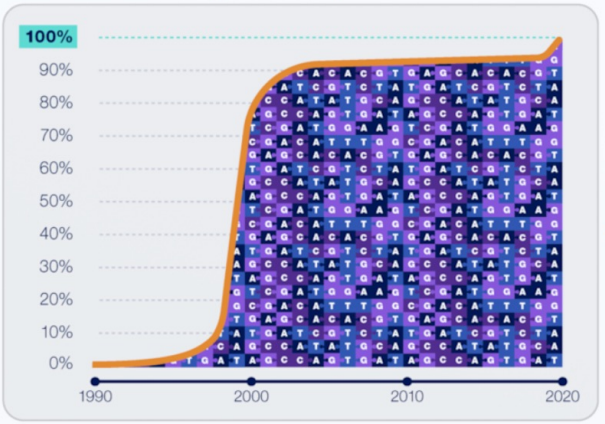
https://youtu.be/SZpcOy63lWw?feature=shared

Reason 3

The first 92% was hard. The last 8% was excruciating.

Those DNA repeats and other obstacles stood between the genomic researchers and the final 8% of the human genome sequence until new laboratory and computational technologies were developed. It took almost twice as long to finish the last 8% of the human genome as it did the first 92%!

Percent of human genome sequence released



Reason 4

The last 8% needed a generation of dedicated genomic researchers with a vision.

Even with new technologies, genome sequencing is still tough, time-consuming work that requires a lot of skill and dedication. The current generation of genomic researchers are true perfectionists and brought everything together to finally complete the human genome sequence.

genome.gov

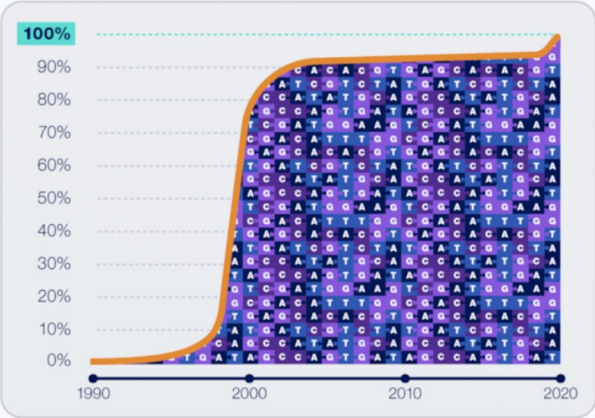
https://youtu.be/SZpcOy63lWw?feature=shared

Reason 3

The first 92% was hard. The last 8% was excruciating.

Those DNA repeats and other obstacles stood between the genomic researchers and the final 8% of the human genome sequence until new laboratory and computational technologies were developed. It took almost twice as long to finish the last 8% of the human genome as it did the first 92%!

Percent of human genome sequence released



Reason 4

The last 8% needed a generation of dedicated genomic researchers with a vision.

Even with new technologies, genome sequencing is still tough, time-consuming work that requires a lot of skill and dedication. The current generation of genomic researchers are true perfectionists and brought everything together to finally complete the human genome sequence.

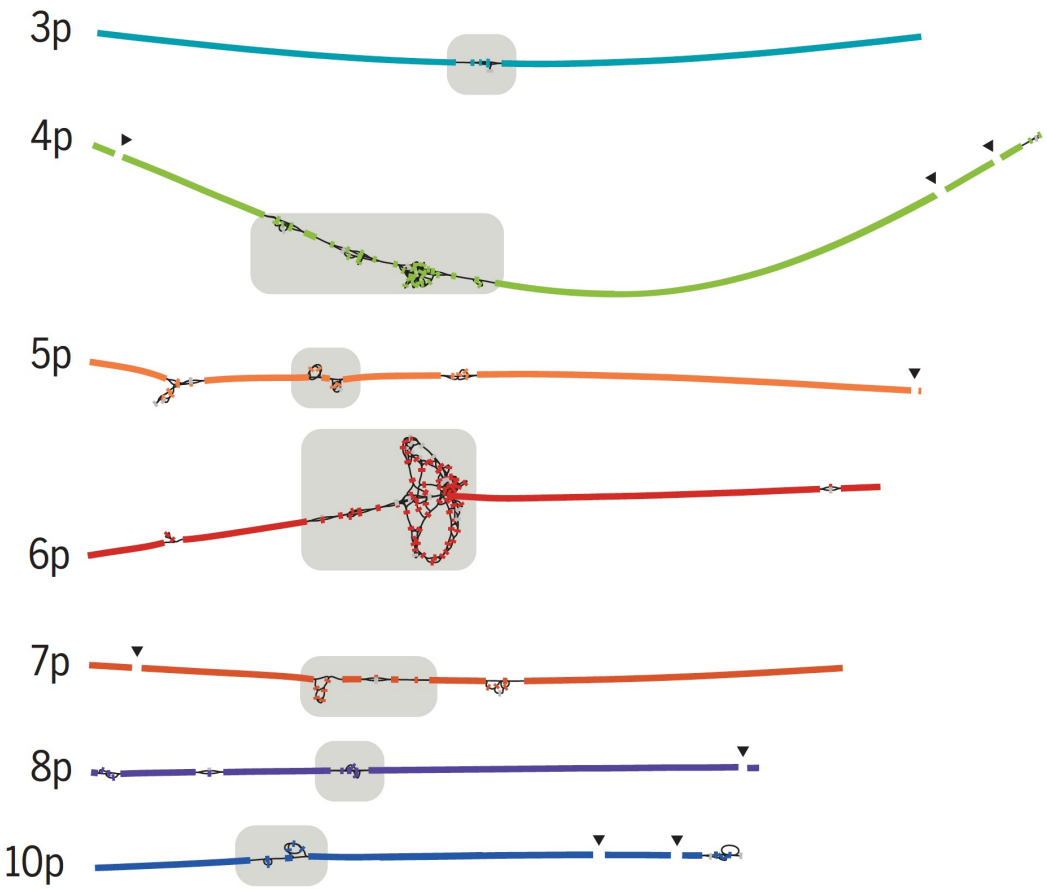
genome.gov

https://youtu.be/SZpcOy63lWw?feature=shared

https://youtu.be/SZpcOy63lWw?feature=shared

Human ref. genome graph for chr. 3-10

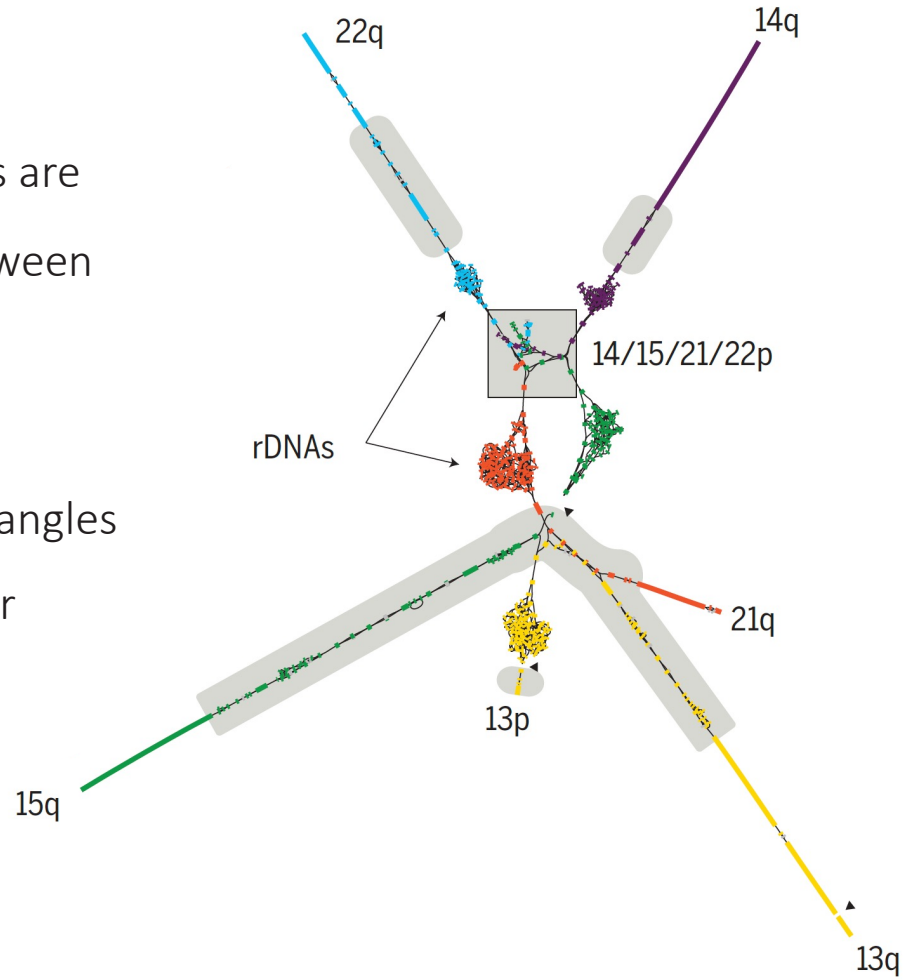
- Solid-color lines are unambigousely assembled regions (**euchromatic**)
- Centromeric satellites are the source of most ambiguity in the graph (gray highlights, **heterochromatic**).



Nurk et al. 2022, Science 376
Graph visualized in Bandage

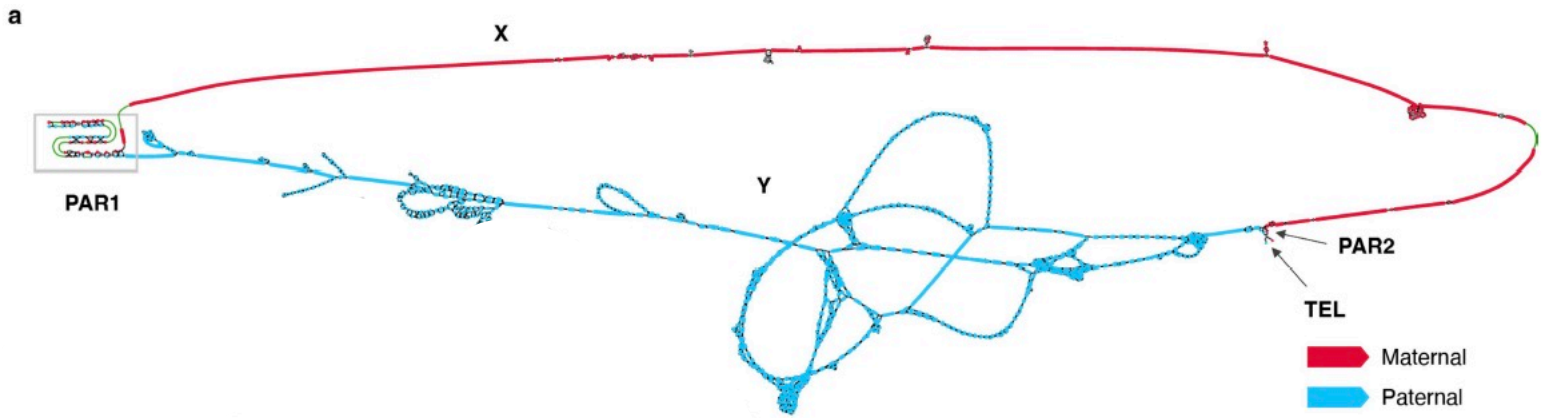
Human ref. genome graph for acrocentric chr. 13-15,21,22

- The five acrocentric chromosomes are connected owing to similarity between their short arms.
- The rDNA arrays form five dense tangles because of their high copy number



Nurk et al. 2022, Science 376
Graph visualized in Bandage

The human ref. genome **chr. Y**



Article

The complete sequence of a human Y chromosome

https://doi.org/10.1038/s41586-023-06457-y

Received: 2 December 2022

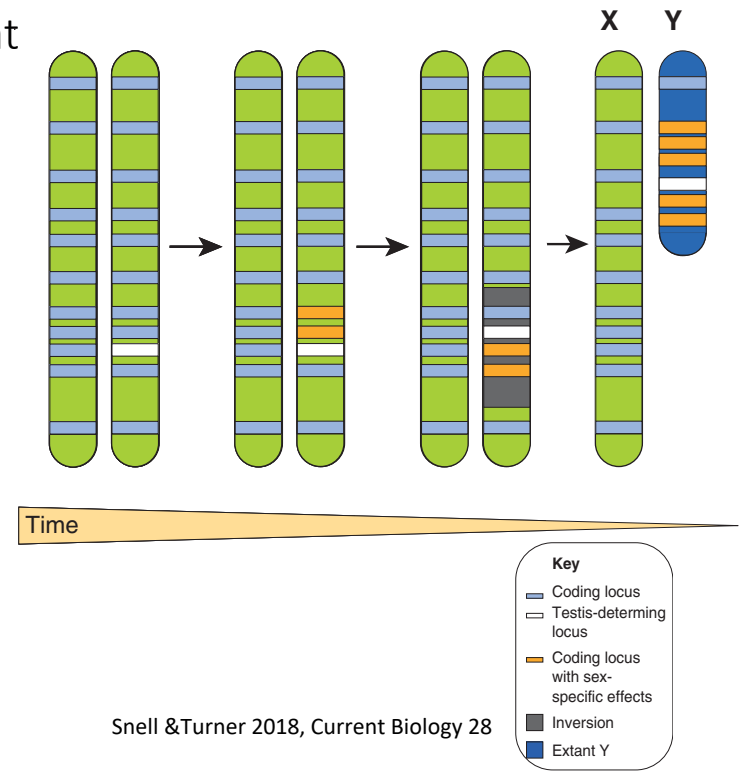
Accepted: 19 July 2023

Published online: 23 August 2023

Arang Rhie[1,53], Sergey Nurk[1,51,53], Monika Cechova[2,3,53], Savannah J. Hoyt[4,53], Dylan J. Taylor[5,53], Nicolas Altemose[6], Paul W. Hook[7], Sergey Koren[1], Mikko Rautiainen[1], Ivan A. Alexandrov[8,9,52], Jamie Allen[10], Mobin Asri[11], Andrey V. Bzikadze[12], Nae-Chyun Chen[13], Chen-Shan Chin[14,15], Mark Diekhans[11], Paul Flicek[10,16], Giulio Formenti[17], Arkarachai Fungtammasan[18], Carlos Garcia Giron[10], Erik Garrison[19], Ariel Gershman[7], Jennifer L. Gerton[20,21],

Evolution of mammalian sex chromosome (a supergene)

- A testis-determining locus was acquired on an autosome around 148–166 million years ago
- Sexually antagonistic alleles (orange) then evolved at nearby loci, selected for in males due to their **tight linkage** to testis-determining locus
- **Recombination suppression** likely followed on from chromosomal inversions
- Short-term expansion: **lack of sexual recombination led to the appearance of repetitive DNA sequences**
- In the longer term, large deletions took place. The outcome of this process is the **small, relatively gene poor Y chromosome** observed in most eutherian mammals today

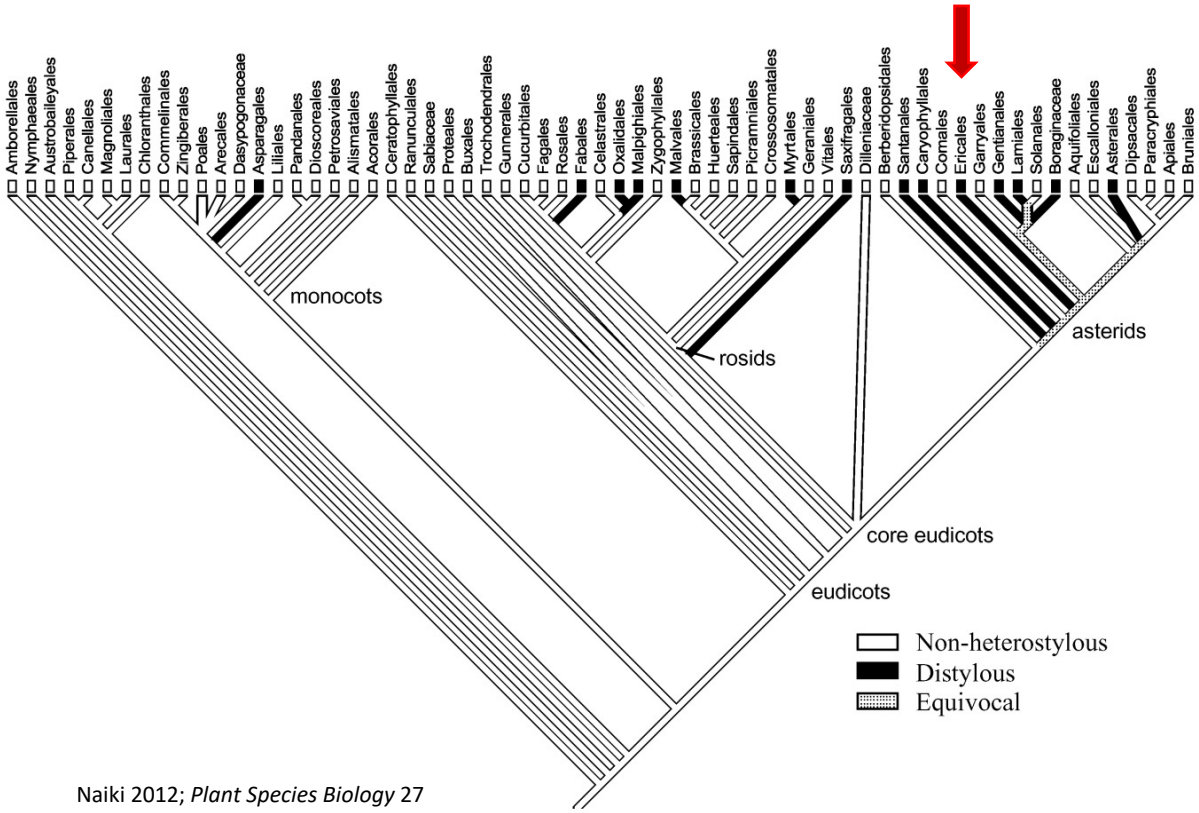


Snell &Turner 2018, Current Biology 28

Case study: Heterostyly supergene in *Primula*

Heterostyly in flowering plants

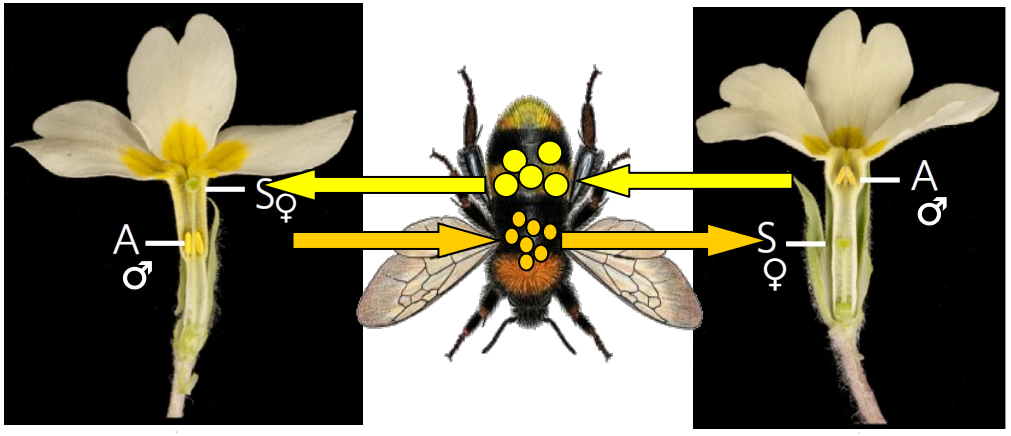
Heterostyly is found in 15 orders, 28 families, 199 genera



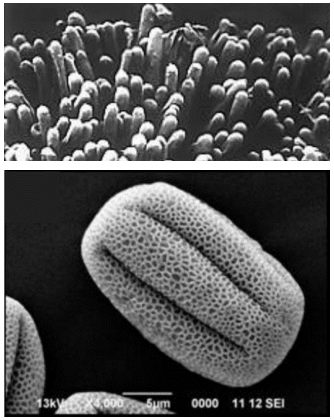
Naiki 2012; *Plant Species Biology* 27

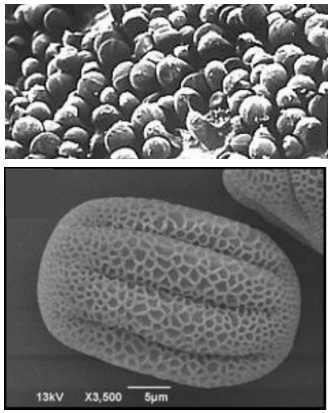
Heterostyly (distyly)

- Reciprocal herkogamy
- Secondary features
  - Stigma papillae size and shape
  - Pollen size and production
- Self-incompatibility
- Promotes outcrossing



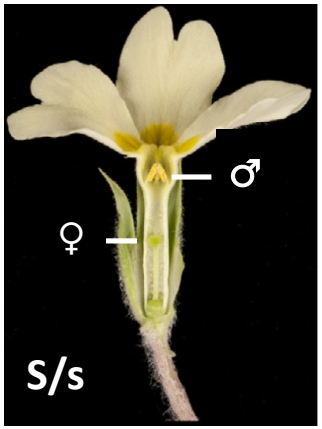
Pin Thrum



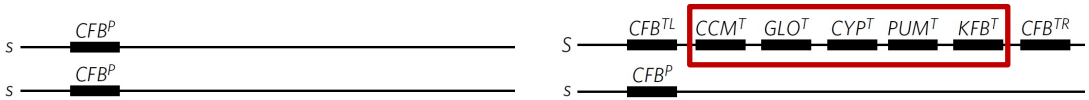


Heterostyly is controlled by S-locus supergene

- ✦ Hemizygous region ca. 280 kb
- ✦ Present in thrums
- ✦ Absent in pins

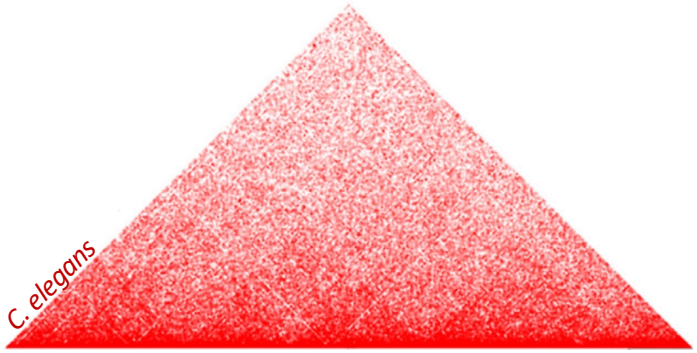


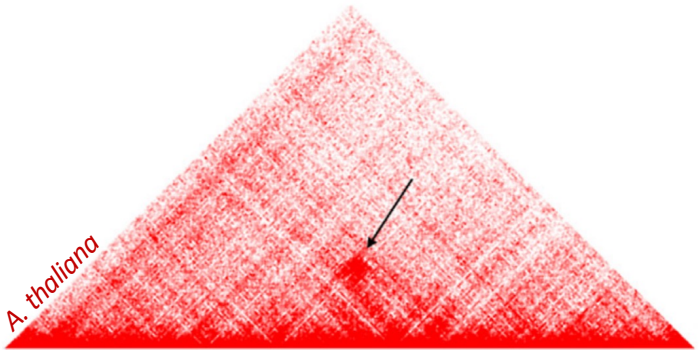
Pin **Thrum**



CYP^T → controlling style length
GLO^T → determining anther position

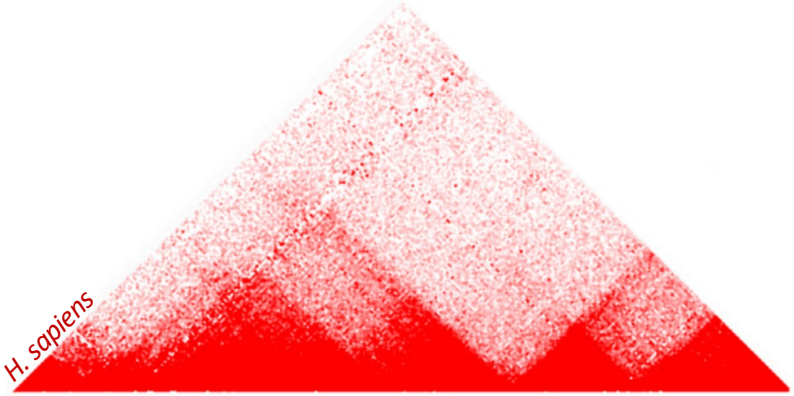
Li et al. 2016; Nowak et al. 2015; Huu et al. 2020





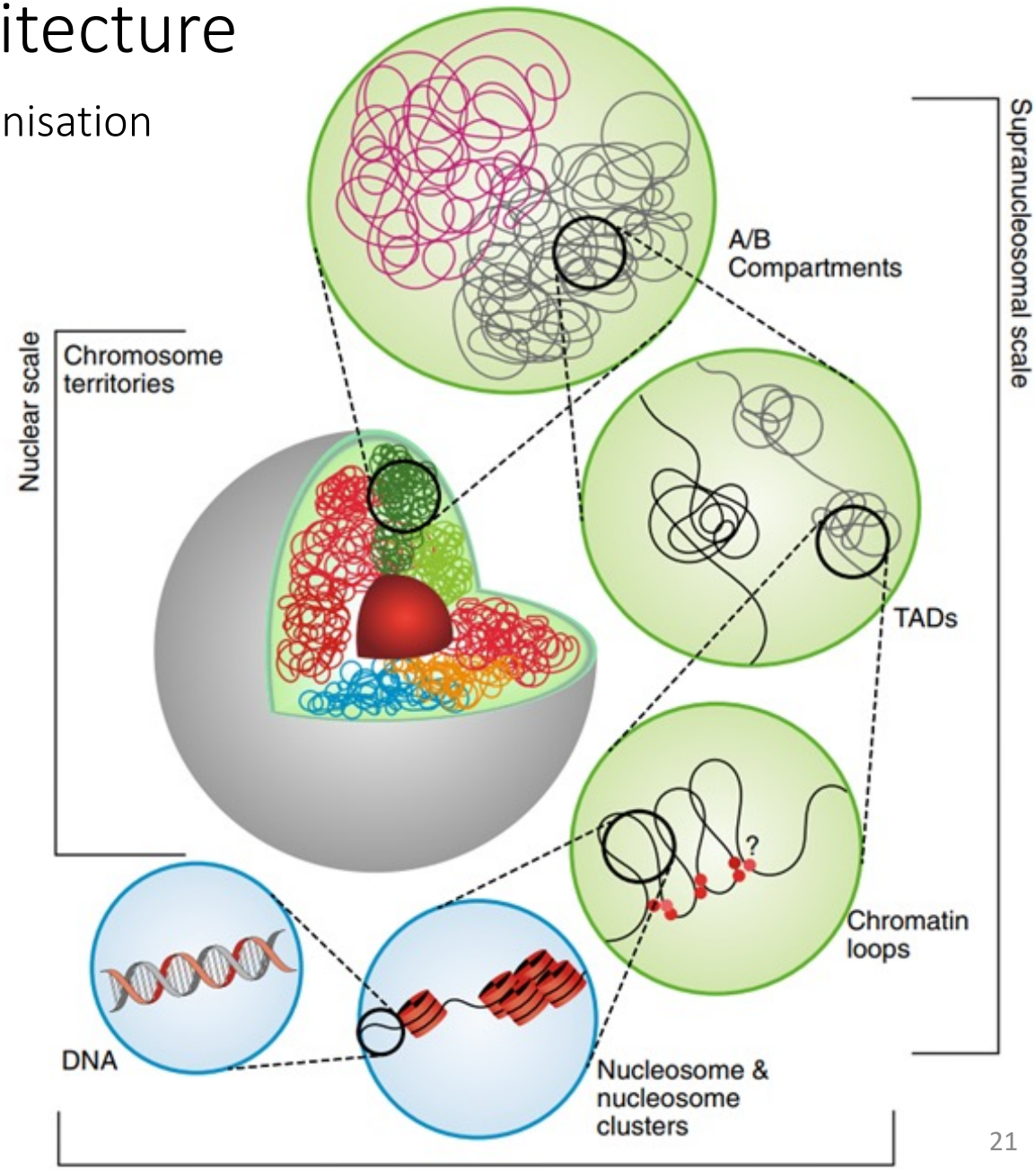
3-D genome architecture

Photos: Rowley & Corces 2016



3-D genome architecture

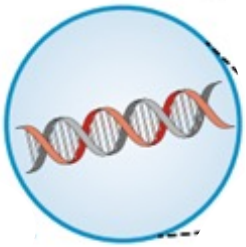
Hierarchical chromatin organisation



Doğan and Liu, 2018. Nature Plants, 4, 2018

3-D genome architecture

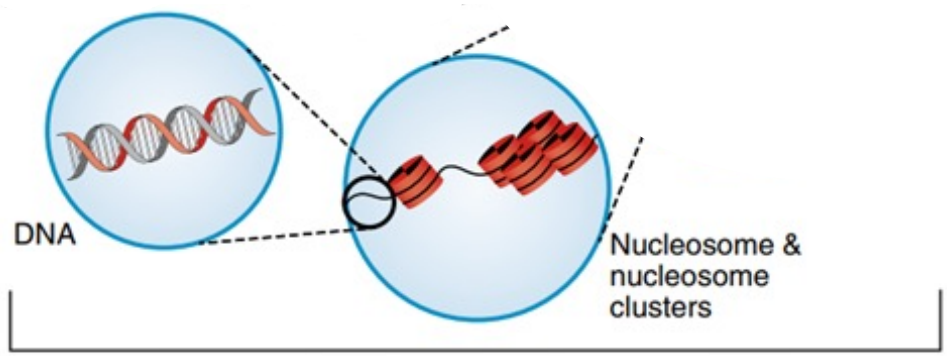
Hierarchical chromatin organisation



Doğan and Liu, 2018. Nature Plants, 4, 2018

3-D genome architecture

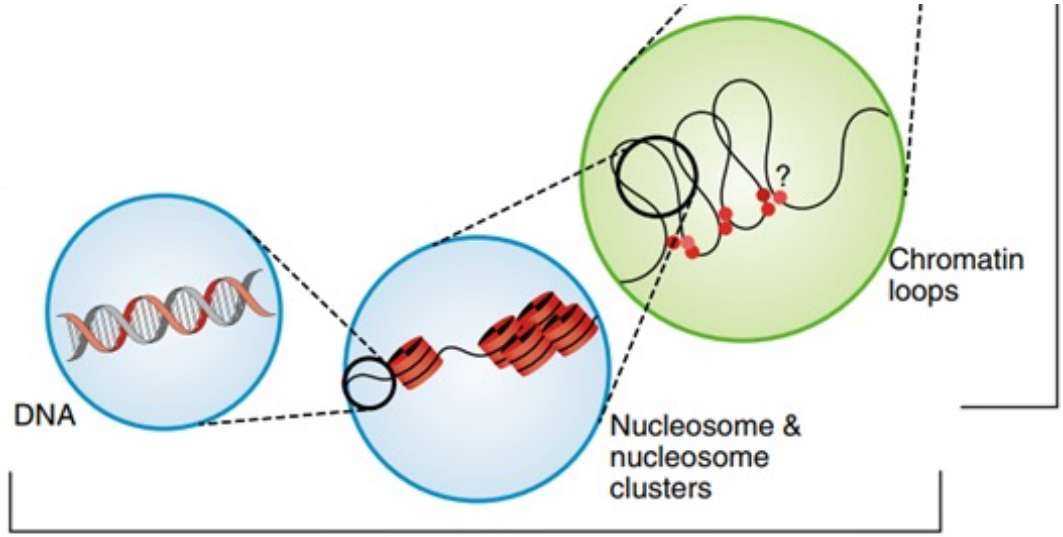
Hierarchical chromatin organisation



Doğan and Liu, 2018. Nature Plants, 4, 2018

3-D genome architecture

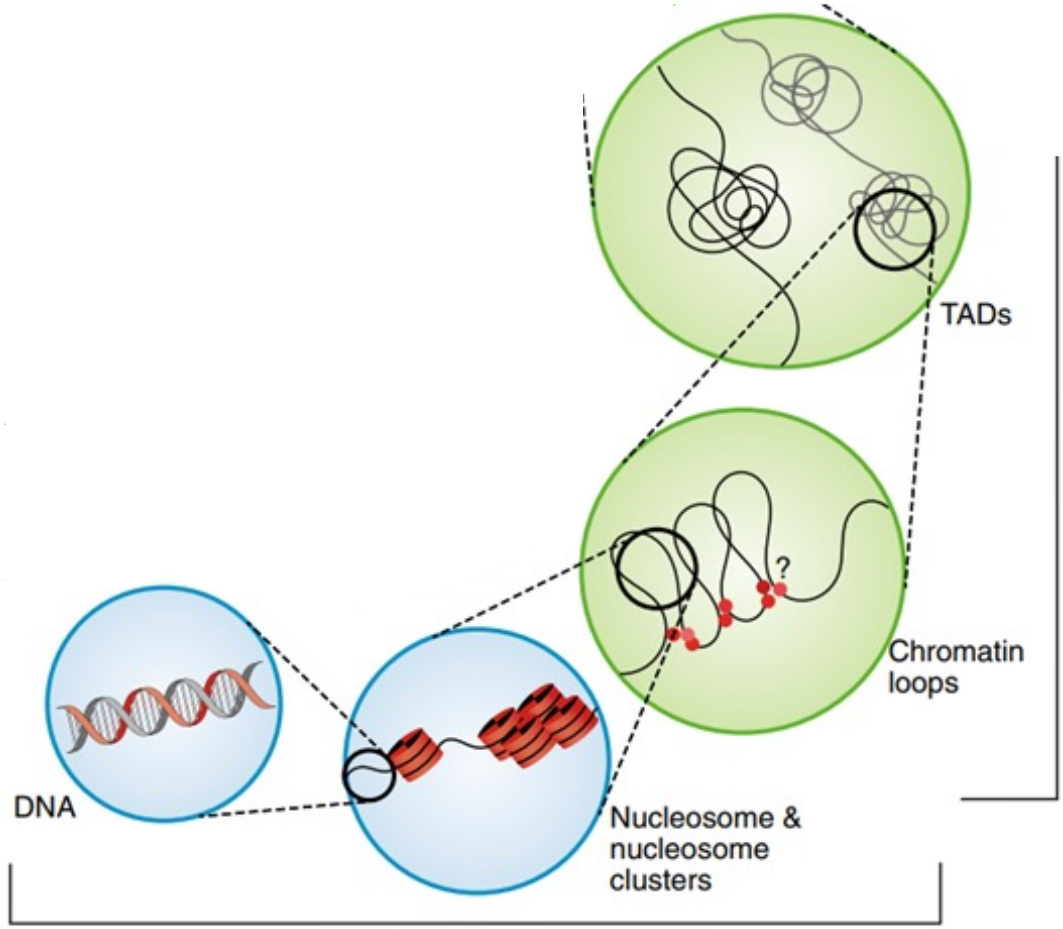
Hierarchical chromatin organisation



Doğan and Liu, 2018. Nature Plants, 4, 2018

3-D genome architecture

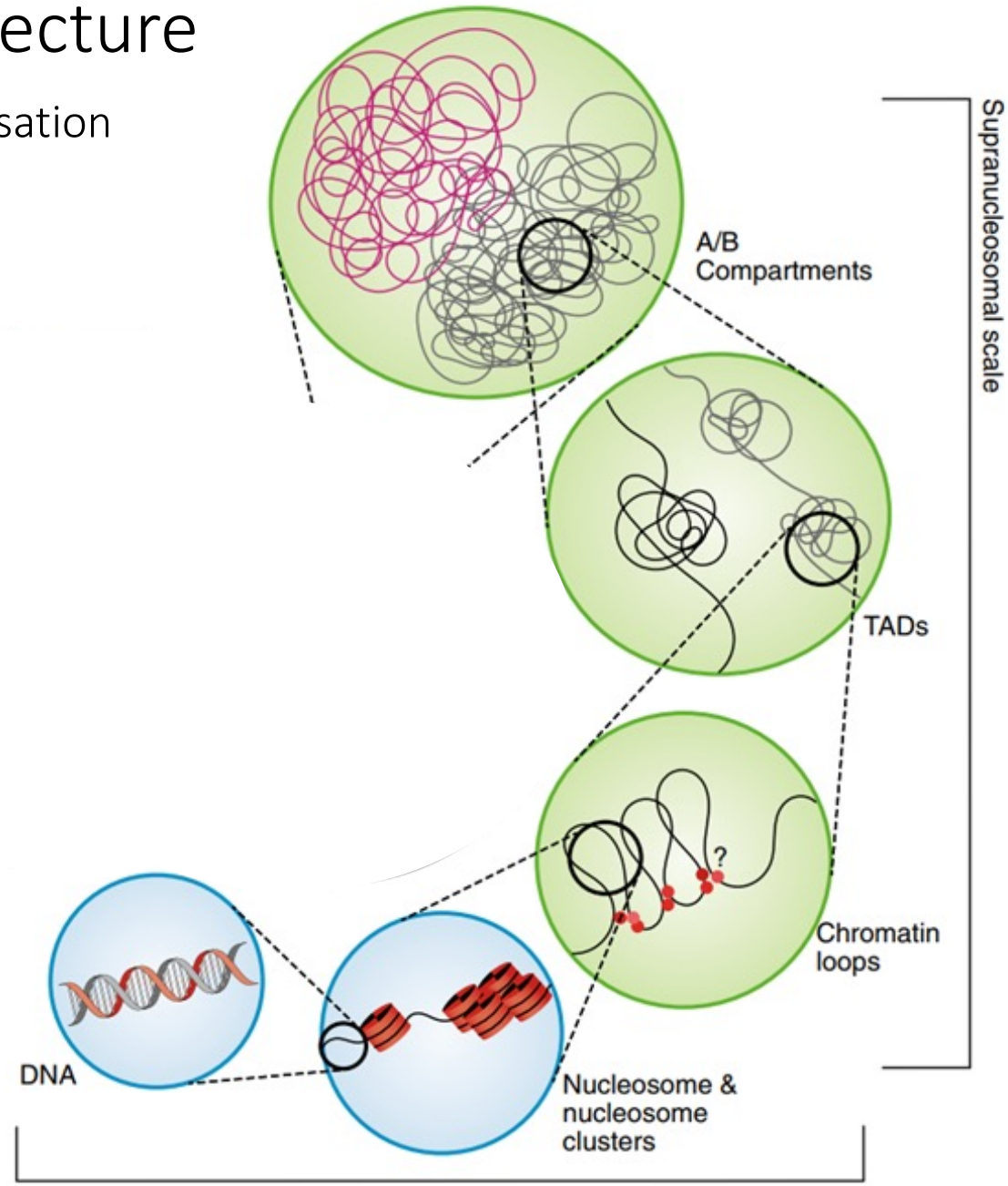
Hierarchical chromatin organisation



Doğan and Liu, 2018. Nature Plants, 4, 2018

3-D genome architecture

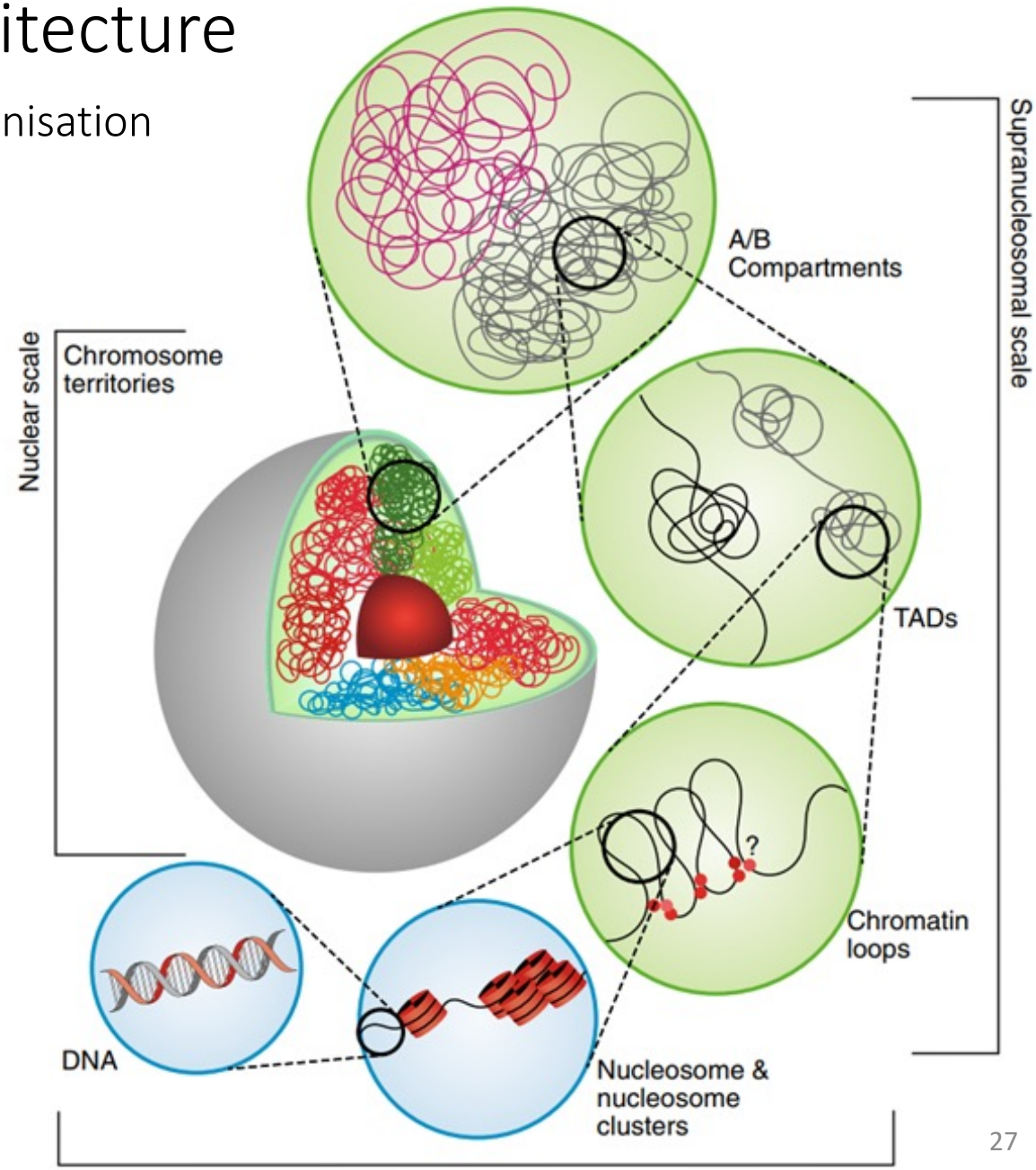
Hierarchical chromatin organisation



Doğan and Liu, 2018. Nature Plants, 4, 2018

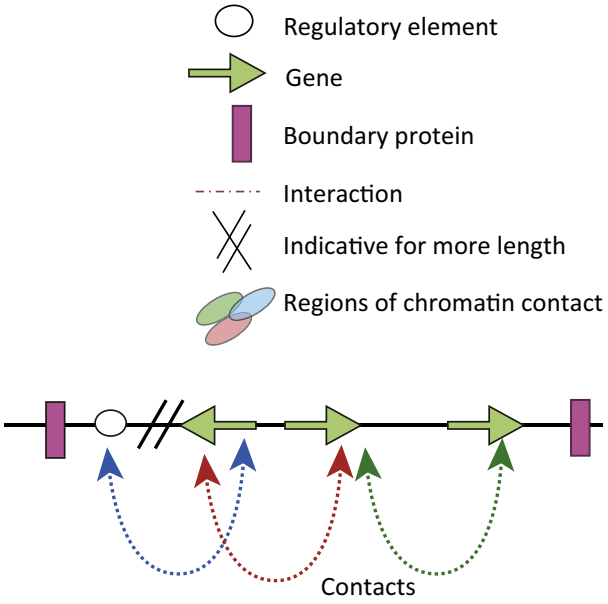
3-D genome architecture

Hierarchical chromatin organisation

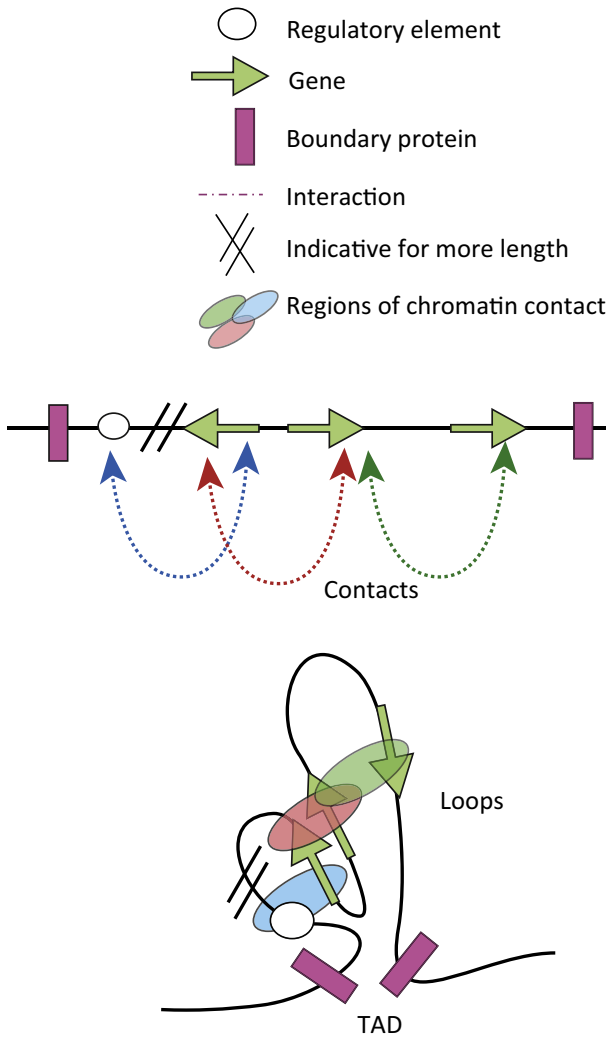


Doğan and Liu, 2018. Nature Plants, 4, 2018

Hierarchical chromatin organisation

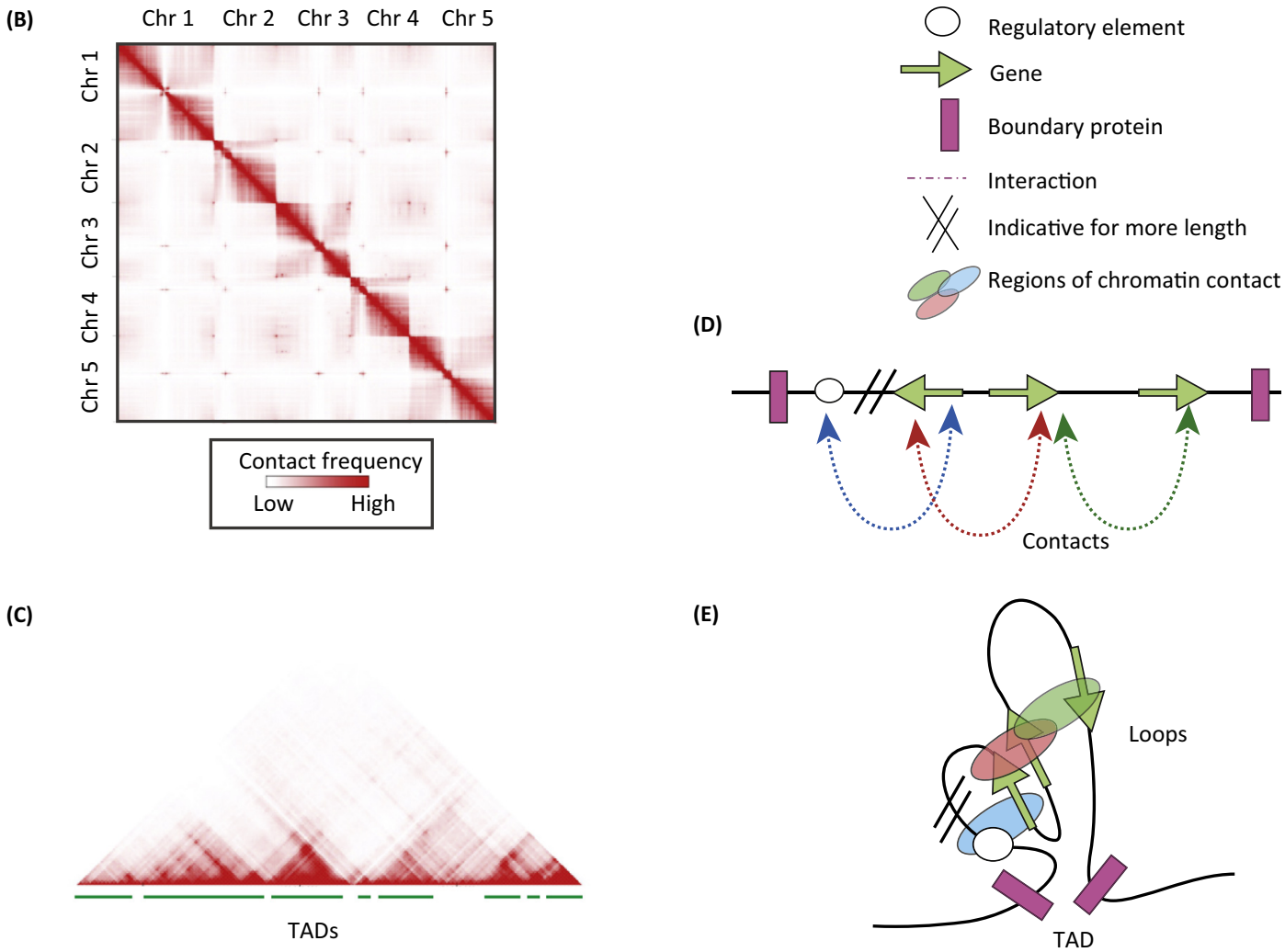


Hierarchical chromatin organisation



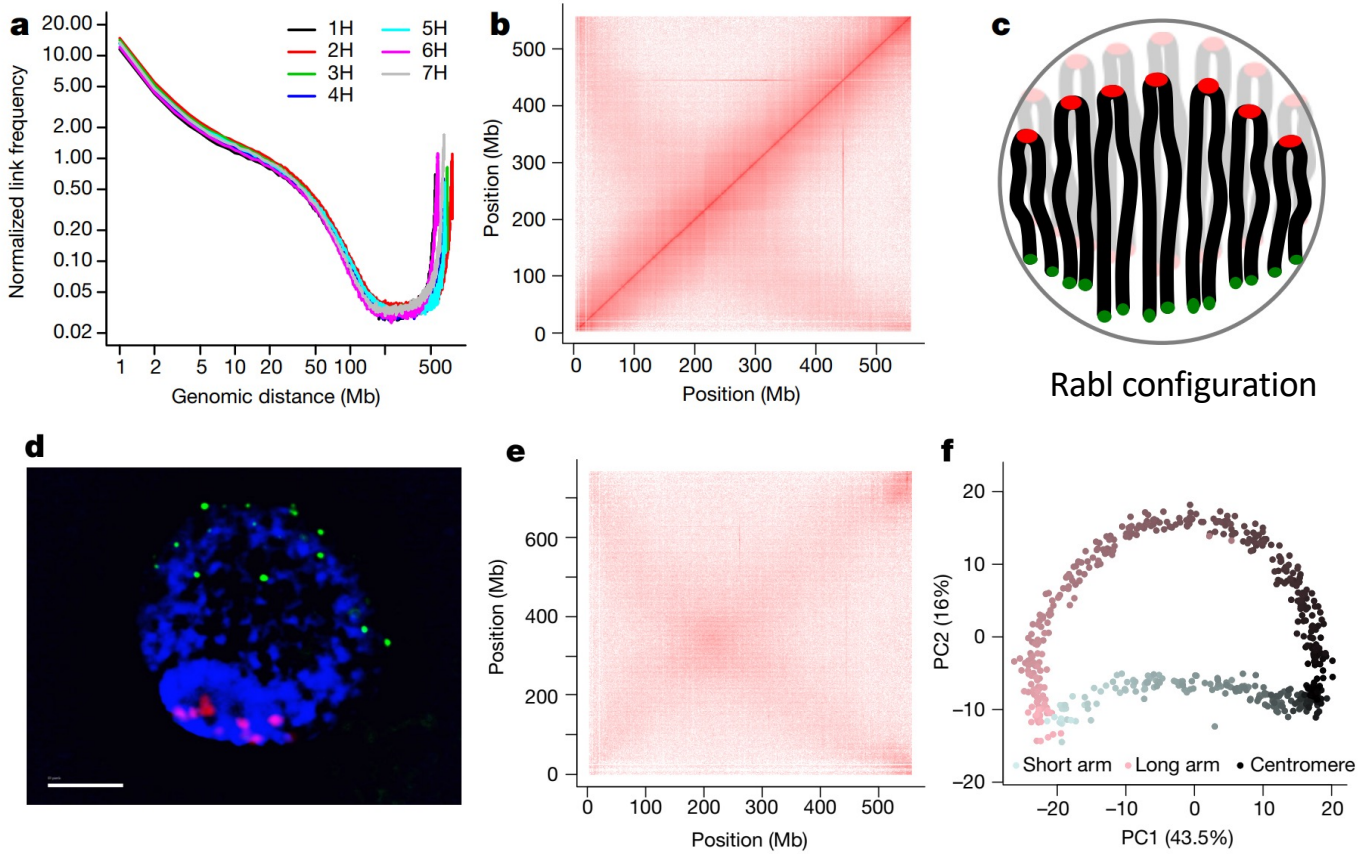
Trends in Plant Science

Hierarchical chromatin organisation



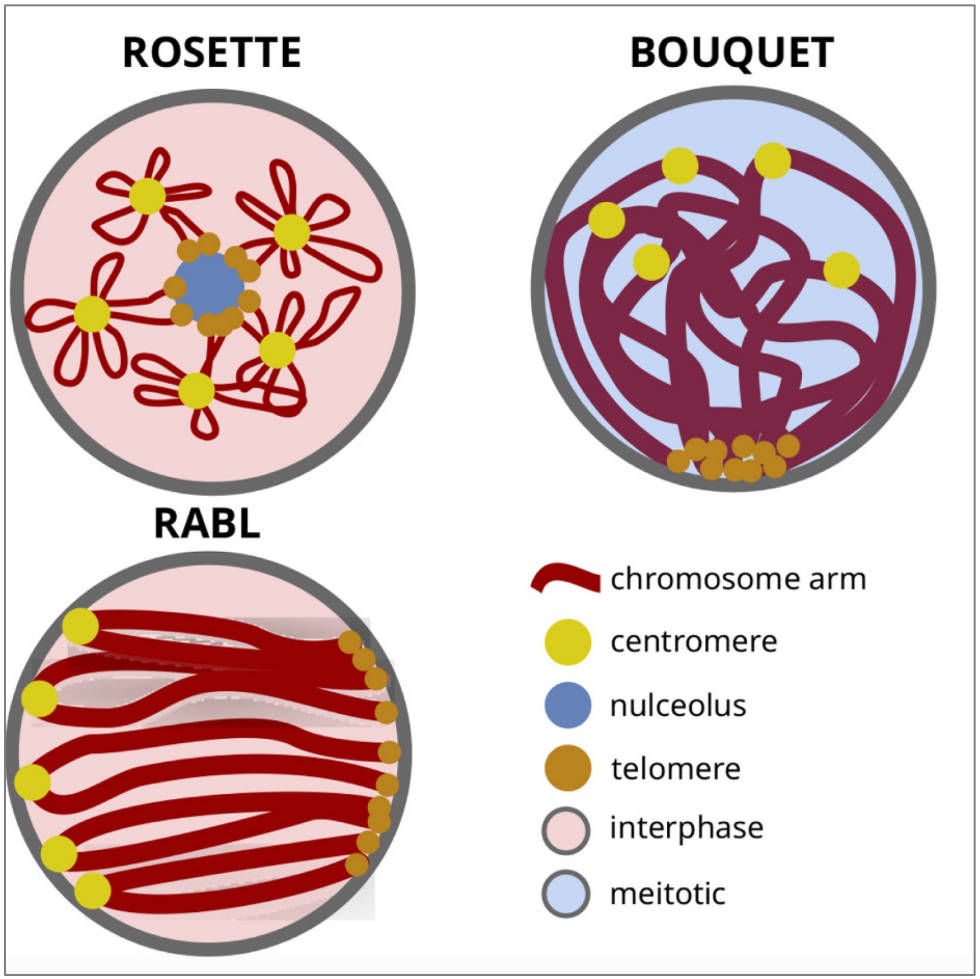
Trends in Plant Science

Organization of the chromosomes in the nucleus



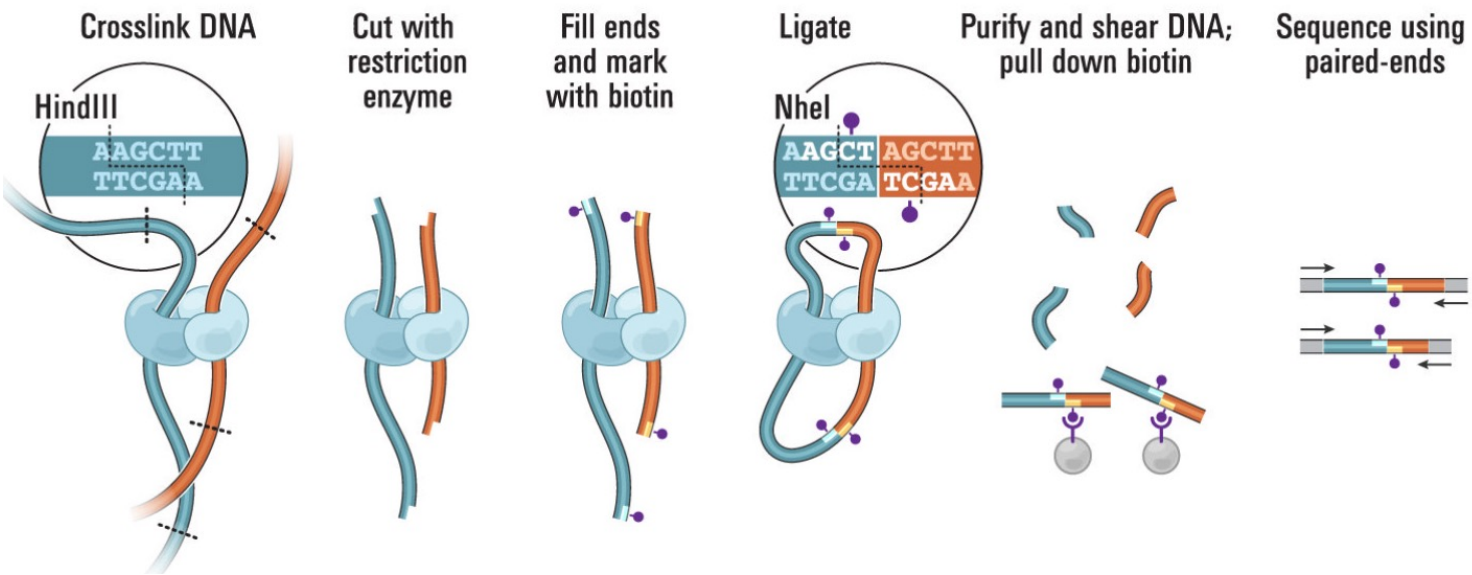
Mascher et al. 2017

Chromosome configuration in plants



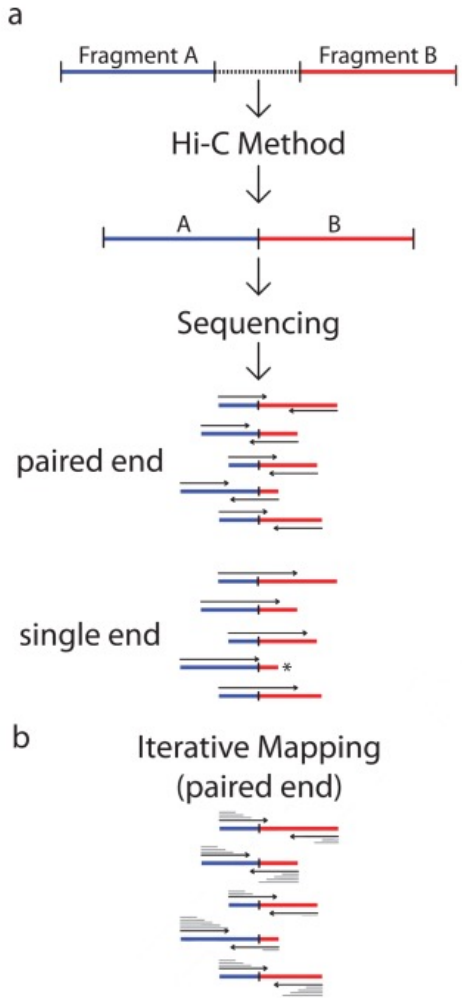
Tourdot & Grob 2023

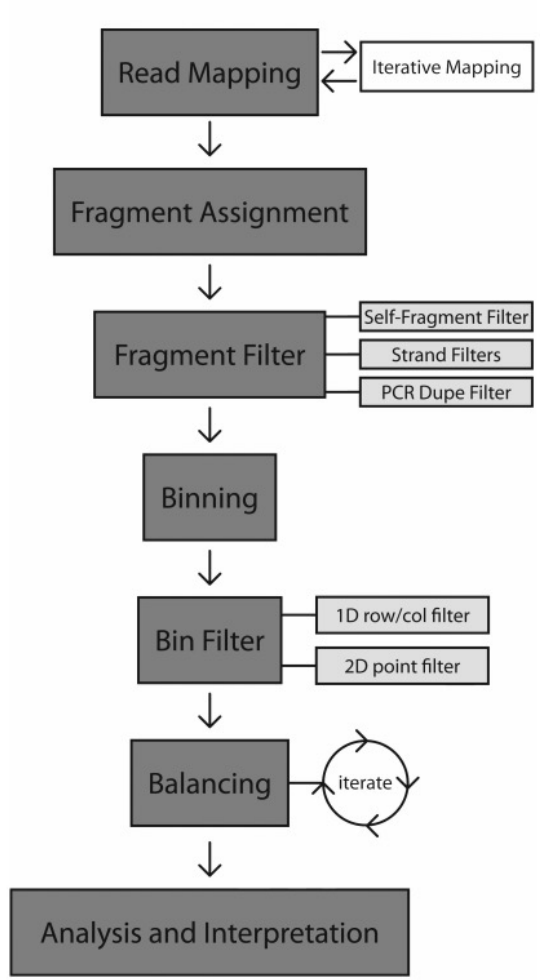
3-D genome architecture: Hi-C technique, library preparation

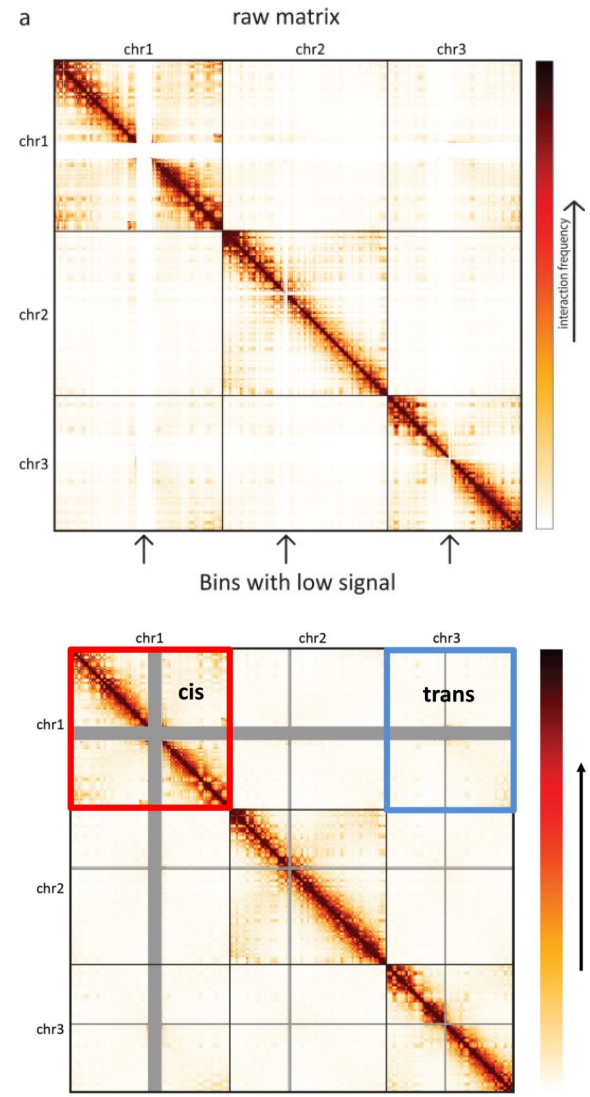


Lieberman-Aiden et al. 2009

3-D genome architecture: Hi-C technique data analysis

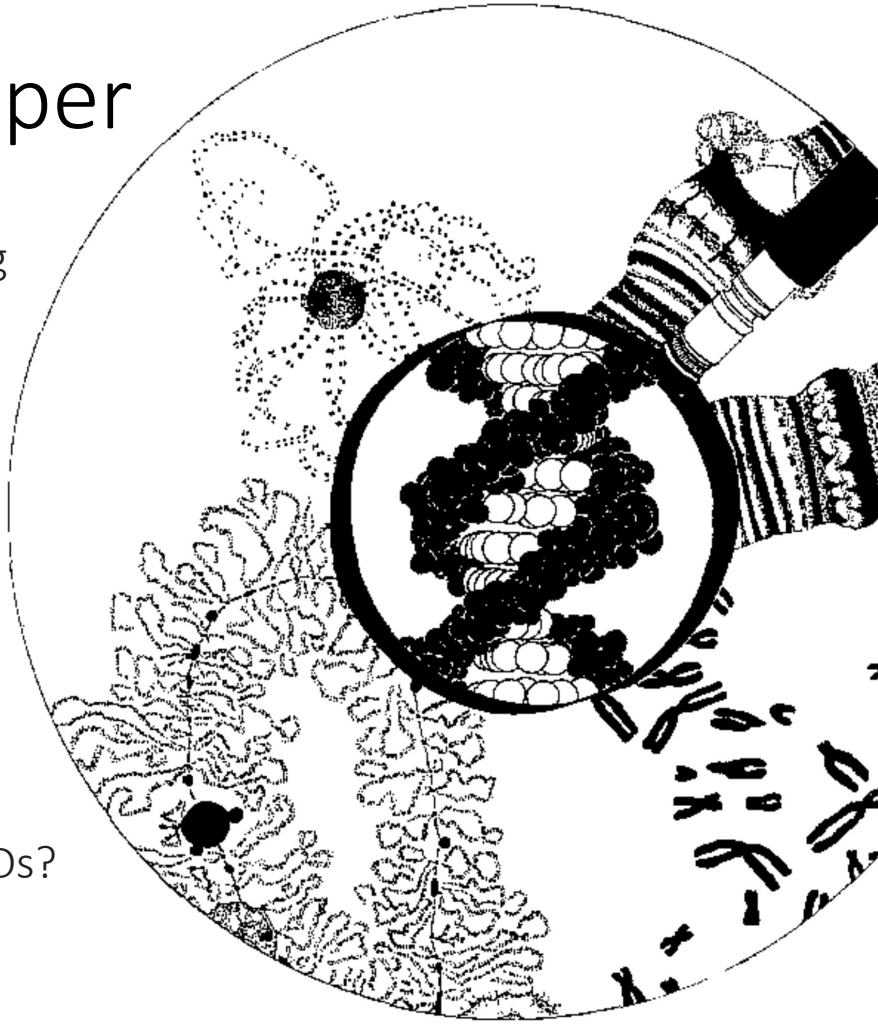






Lajoie et al. 2015

Points for the review paper



- Can we model the 3D map of the genome using only the linear DNA (with or without methylome)?

- How frequent are supergenes in the genome (using relative physical linkage and recombination maps, and heatmaps)

- Are supergenes always making one or more TADs? (verifying using Hi-C heatmaps)

- Are genes in TADs putative supergenes?

- How are supergenes regulated? (all genes together, separate, or mixed model)

NGS technologies

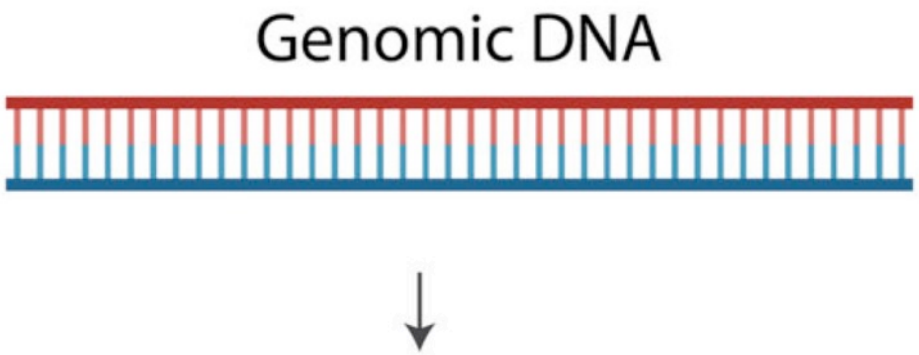
-Short reads:

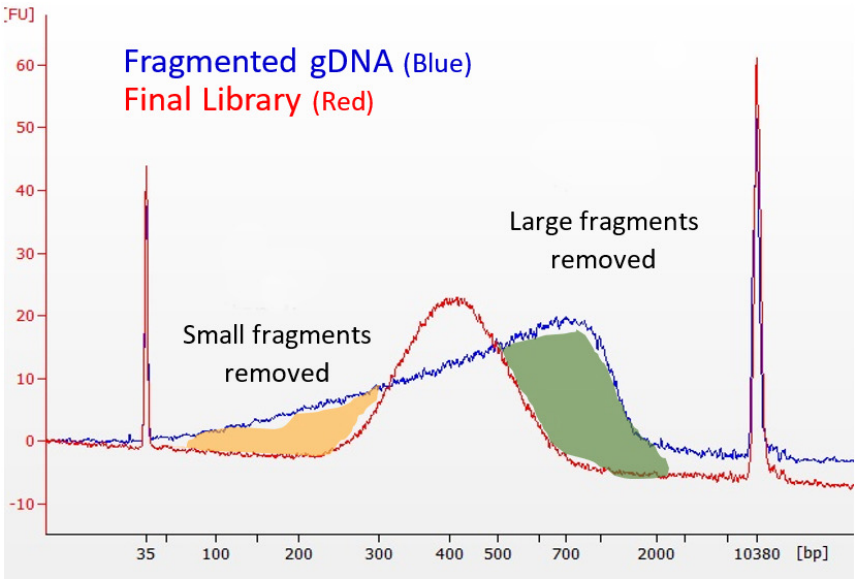
- Illumina

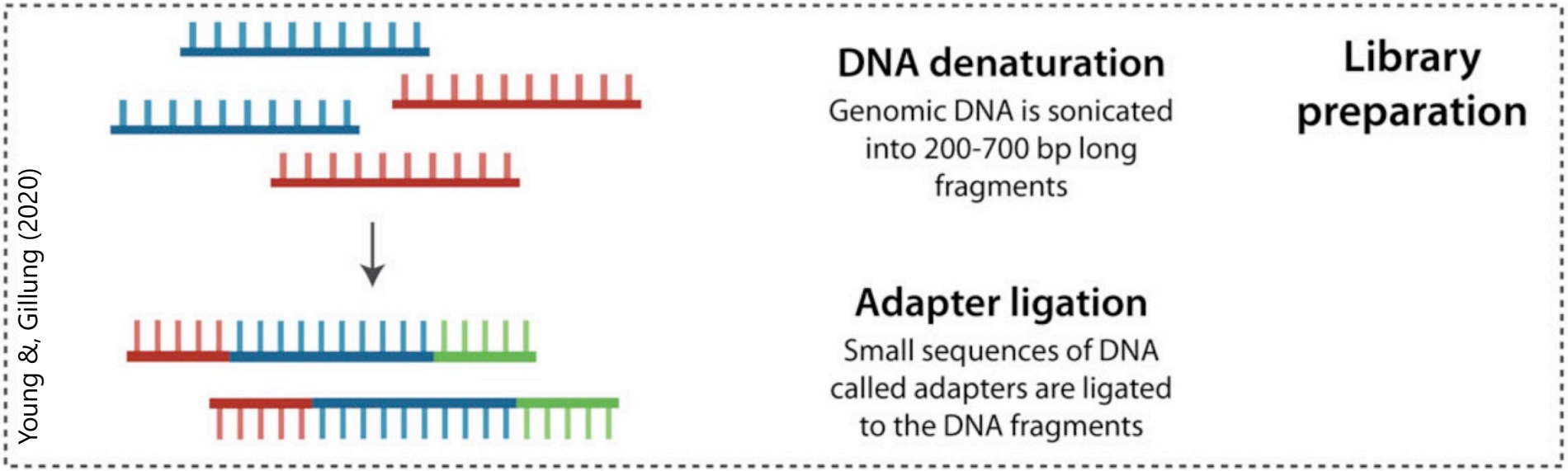
-Long reads:

- Pacific Biosciences RS
- Oxford Nanopore Sequencing

Short-Read Sequencing: Cyclic Reversible Termination (CRT) - Illumina

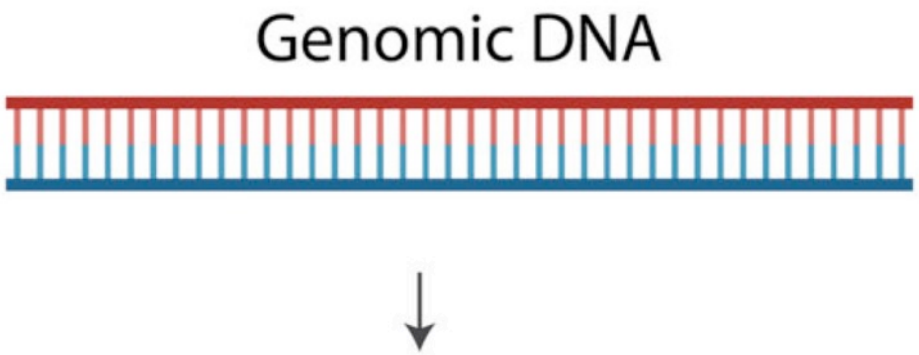


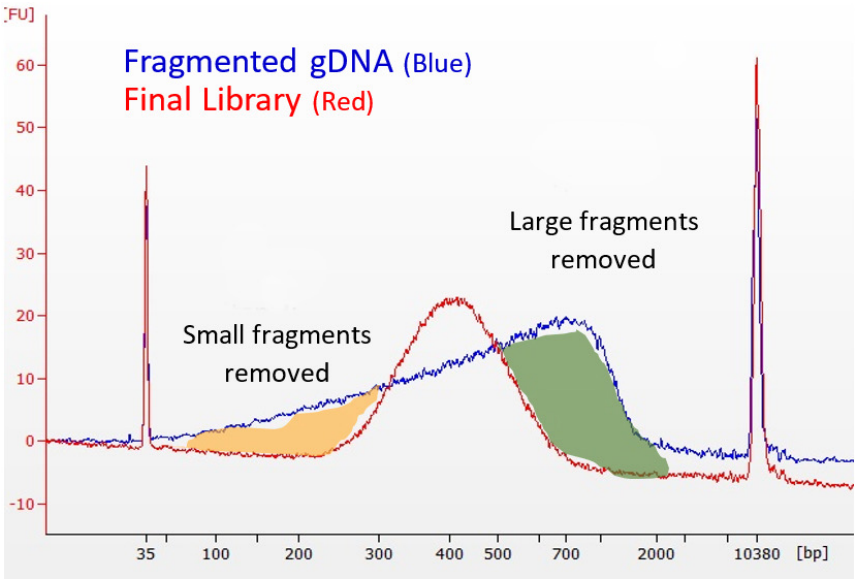


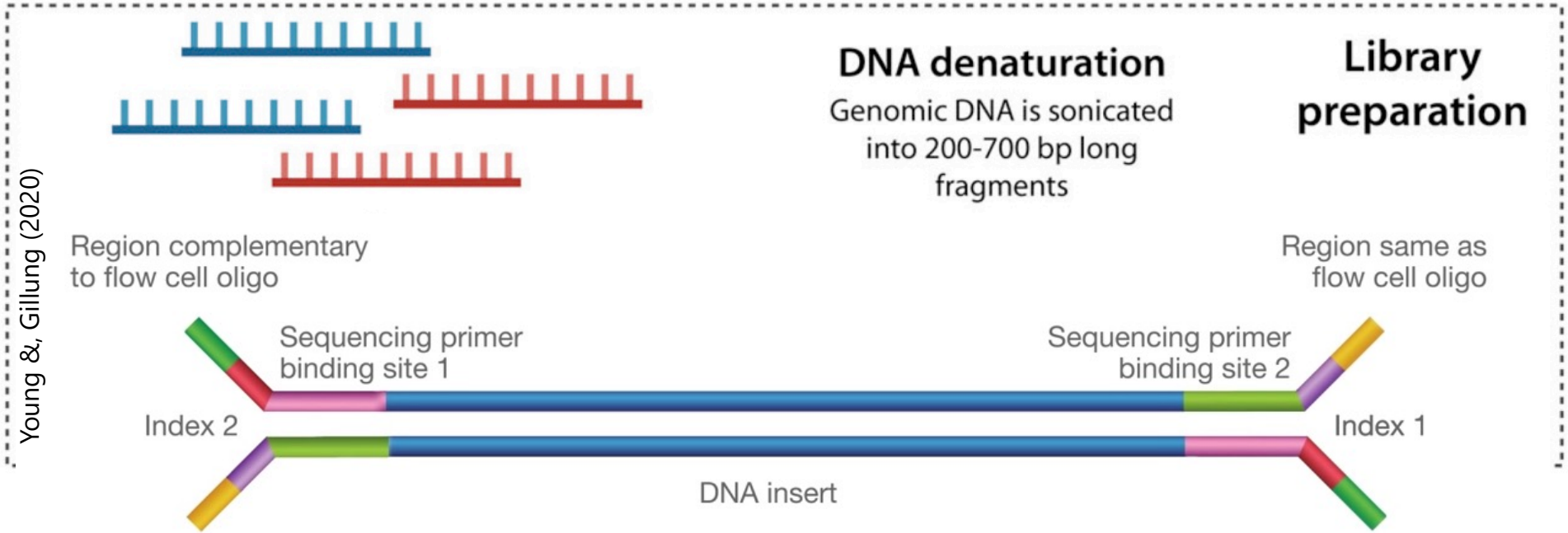


Short-Read Sequencing: Cyclic Reversible Termination (CRT) - Illumina

1. Library preparation

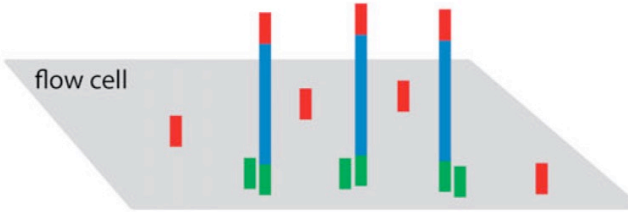






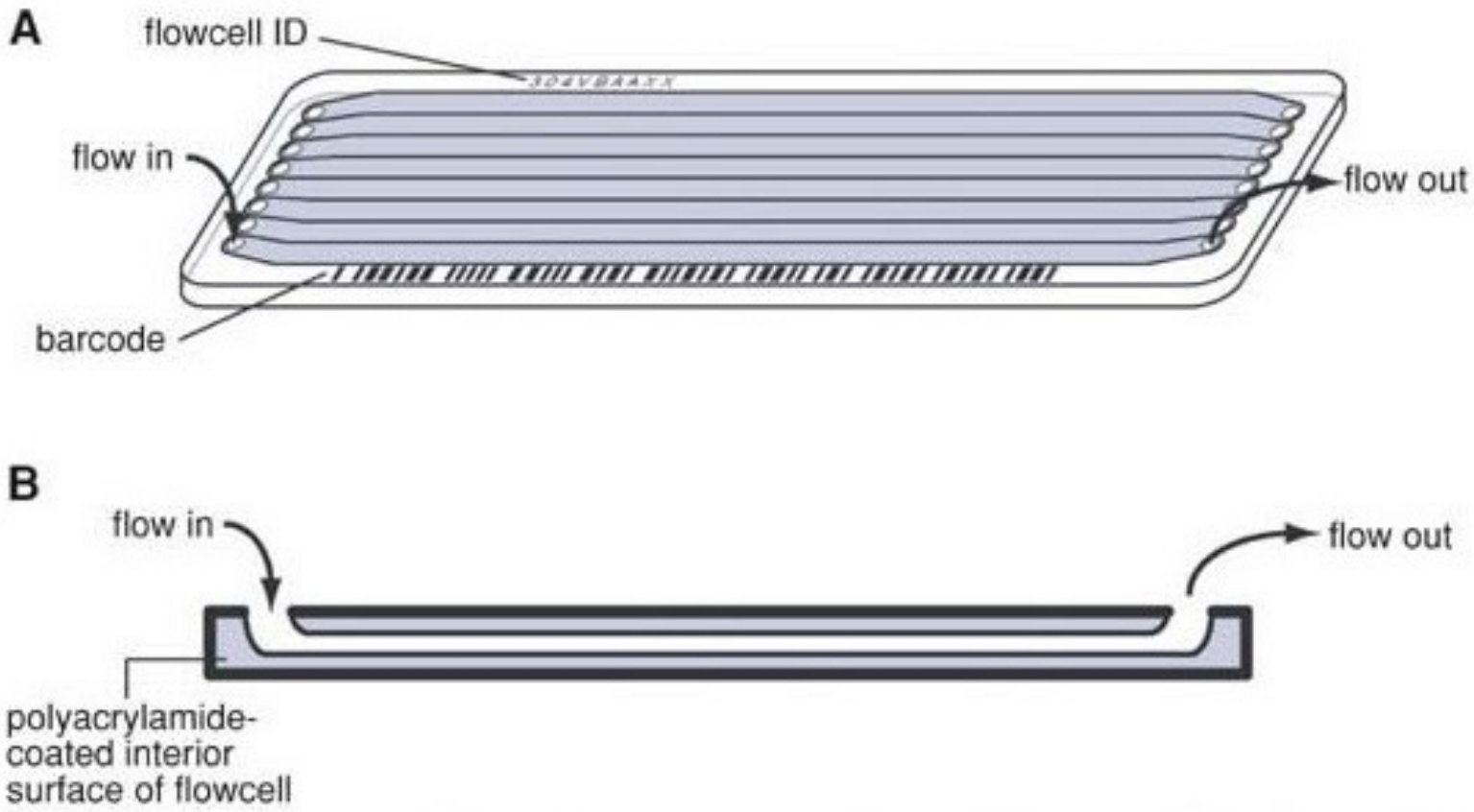
Young &, Gillung (2020)

Short-Read Sequencing: Cyclic Reversible Termination (CRT) - Illumina



Region complementary to the flow cell oligo binds to the flow cell

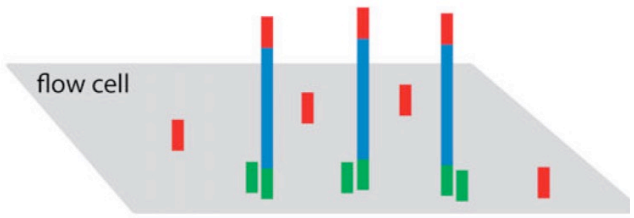
2. Cluster amplification



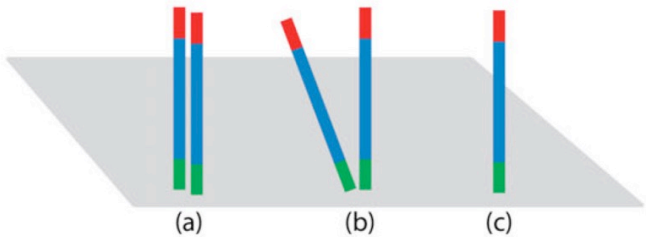
Young &, Gillung (2020)

Short-Read Sequencing: Cyclic Reversible Termination (CRT) - Illumina

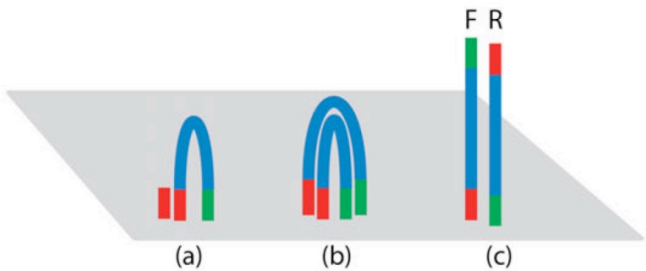
2. Cluster amplification



Region complementary to the flow cell oligo binds to the flow cell

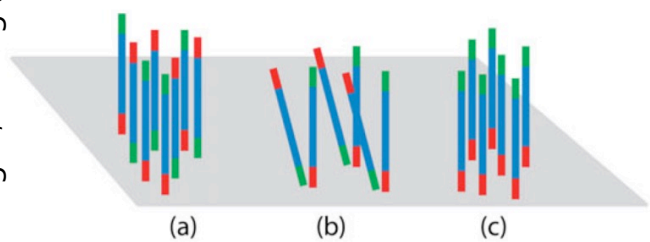


(a) A polymerase creates complement of original fragment
(b) Denaturation of the double stranded molecule
(c) Original template is washed away



Bridge amplification
(a) Strands fold over
(b) Complementary strand is synthesized forming a double-stranded bridge
(c) Denaturation results in two singles stranded fragments

This process is repeated over and over, and occurs simultaneously for millions of clusters
→ **Clonal amplification** of the original fragments



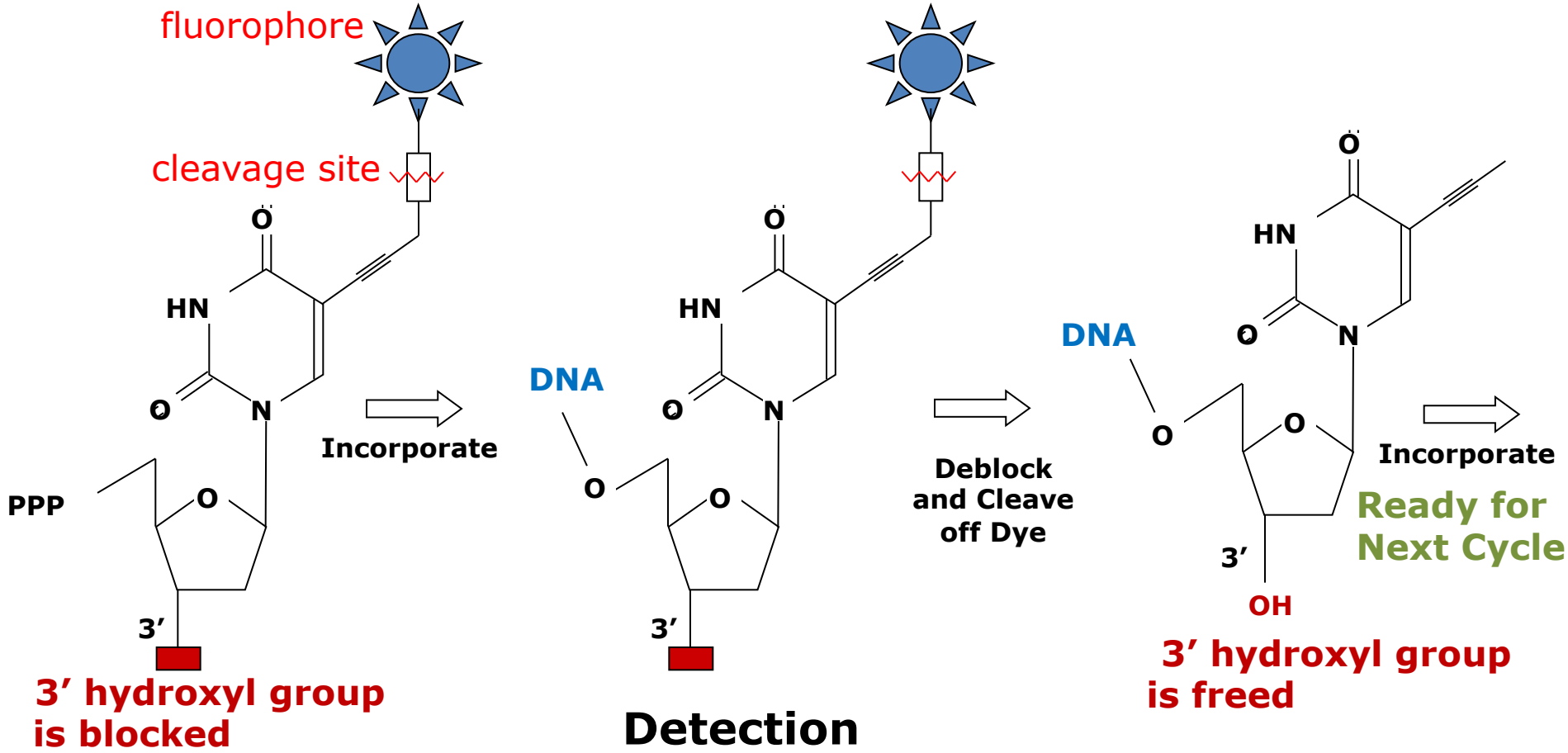
(a) Denaturation of the double stranded molecule
(b) all **R**everse strands are cleaved and washed away
(c) Leaving clusters of **F**orward strands on the flow cell

Young &, Gillung (2020)

Short-Read Sequencing:

Cyclic Reversible Termination (CRT) - Illumina

3. Sequencing using Reversible Terminators

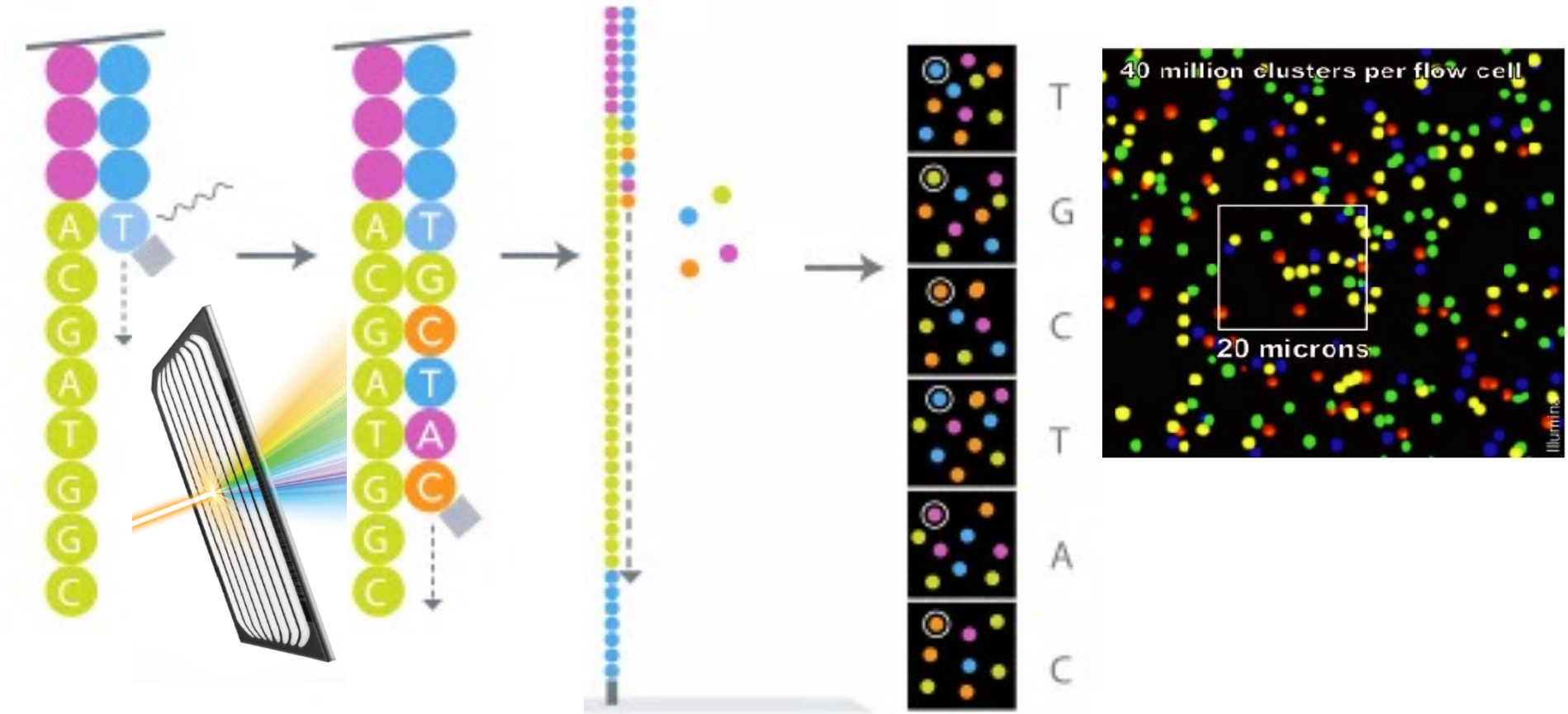


A **fluorescently labeled reversible terminator is imaged as each nucleotide is added**, hence this sequencing technology is also called **sequencing by synthesis**

Short-Read Sequencing:

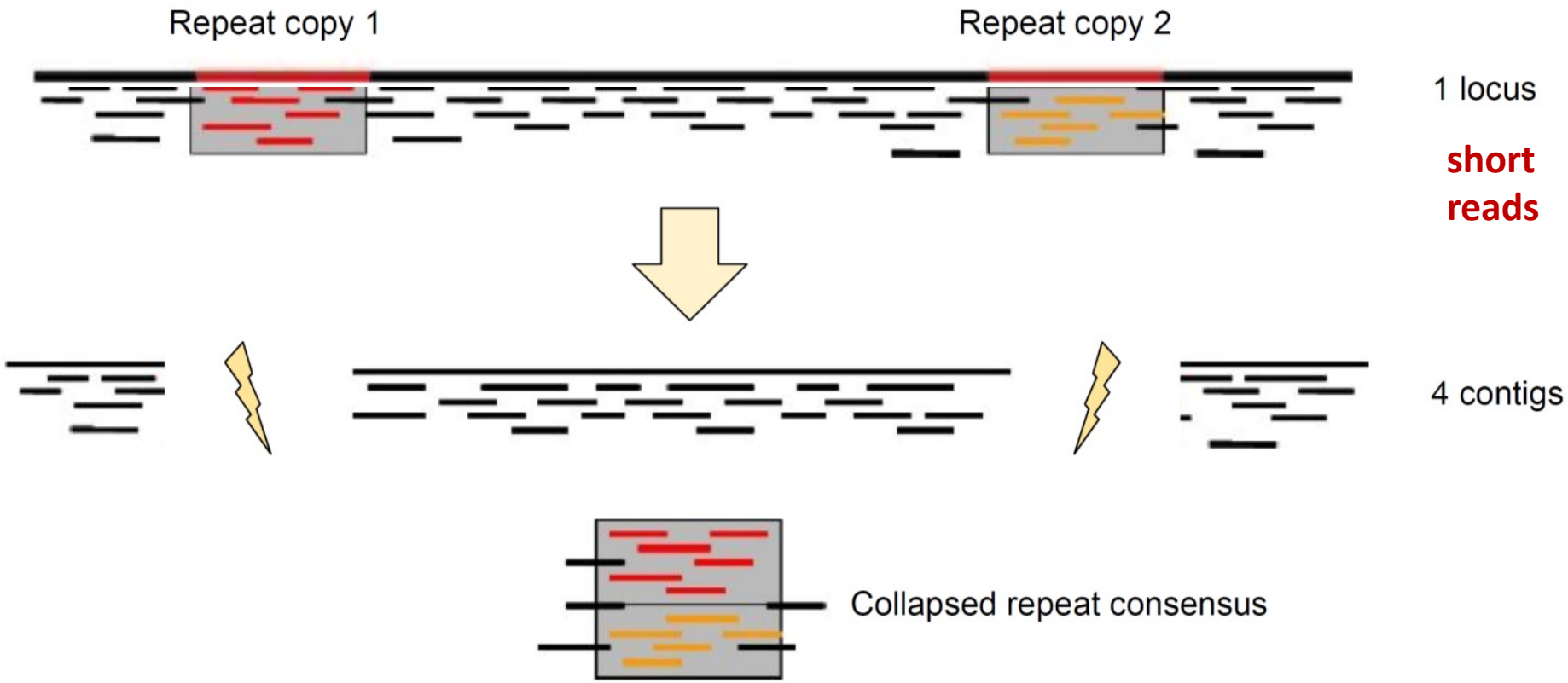
Cyclic Reversible Termination (CRT) - Illumina

4. Signal Detection



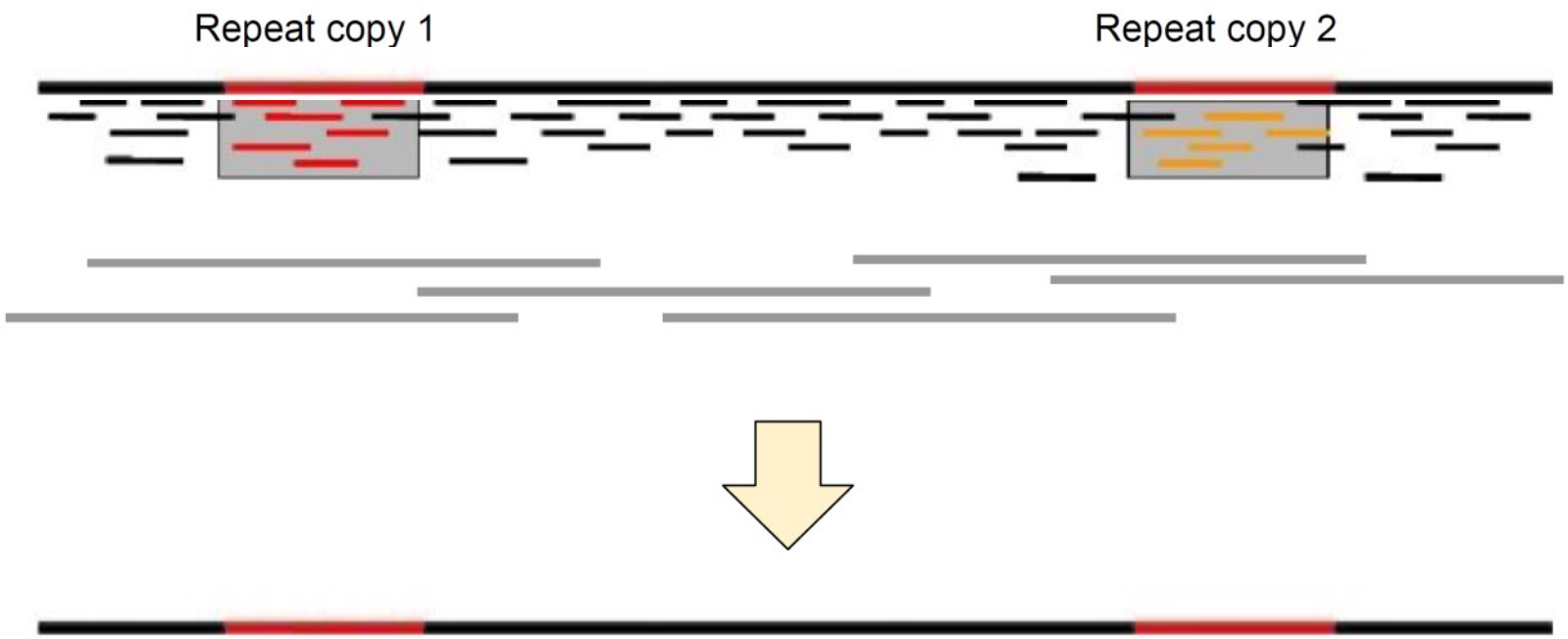
A **fluorescently labeled reversible terminator is imaged as each nucleotide is added**, hence this sequencing technology is also called **sequencing by synthesis**

Why do we need long reads?



Slide from Torsten Seemann

Why do we need long reads?



short reads

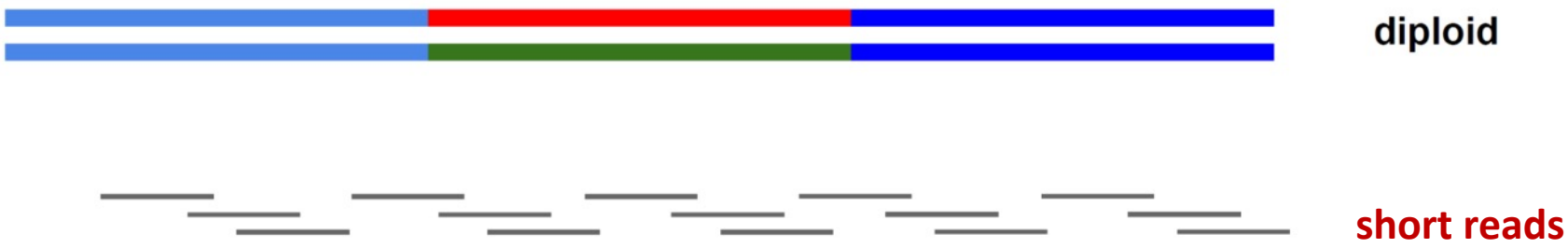
long reads

→ **Long reads can span repeats**

Slide from Torsten Seemann

Why do we need long reads?

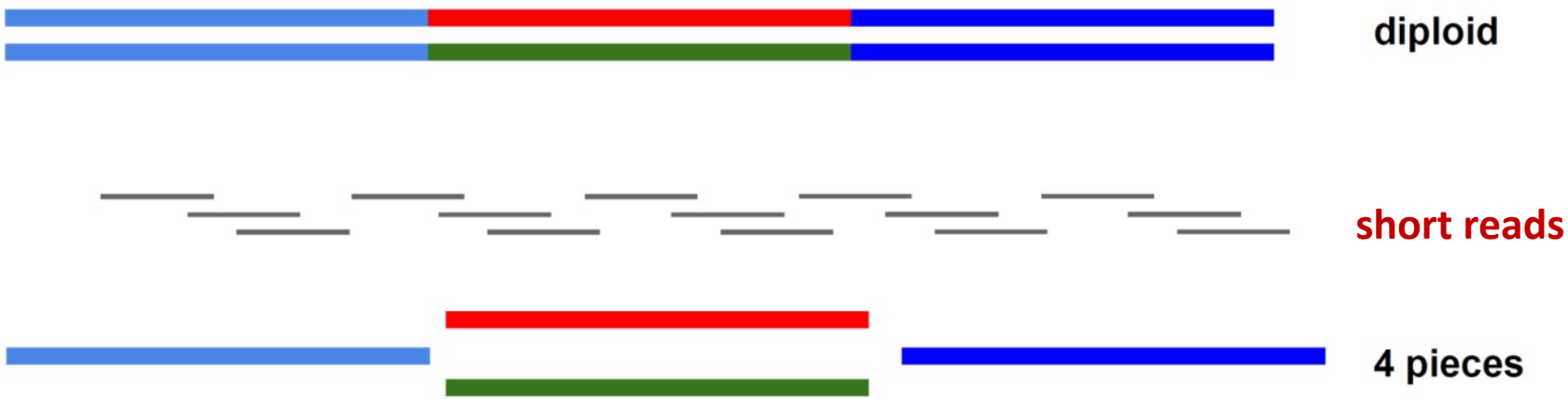
Heterozygosity:



Slide from Torsten Seemann

Why do we need long reads?

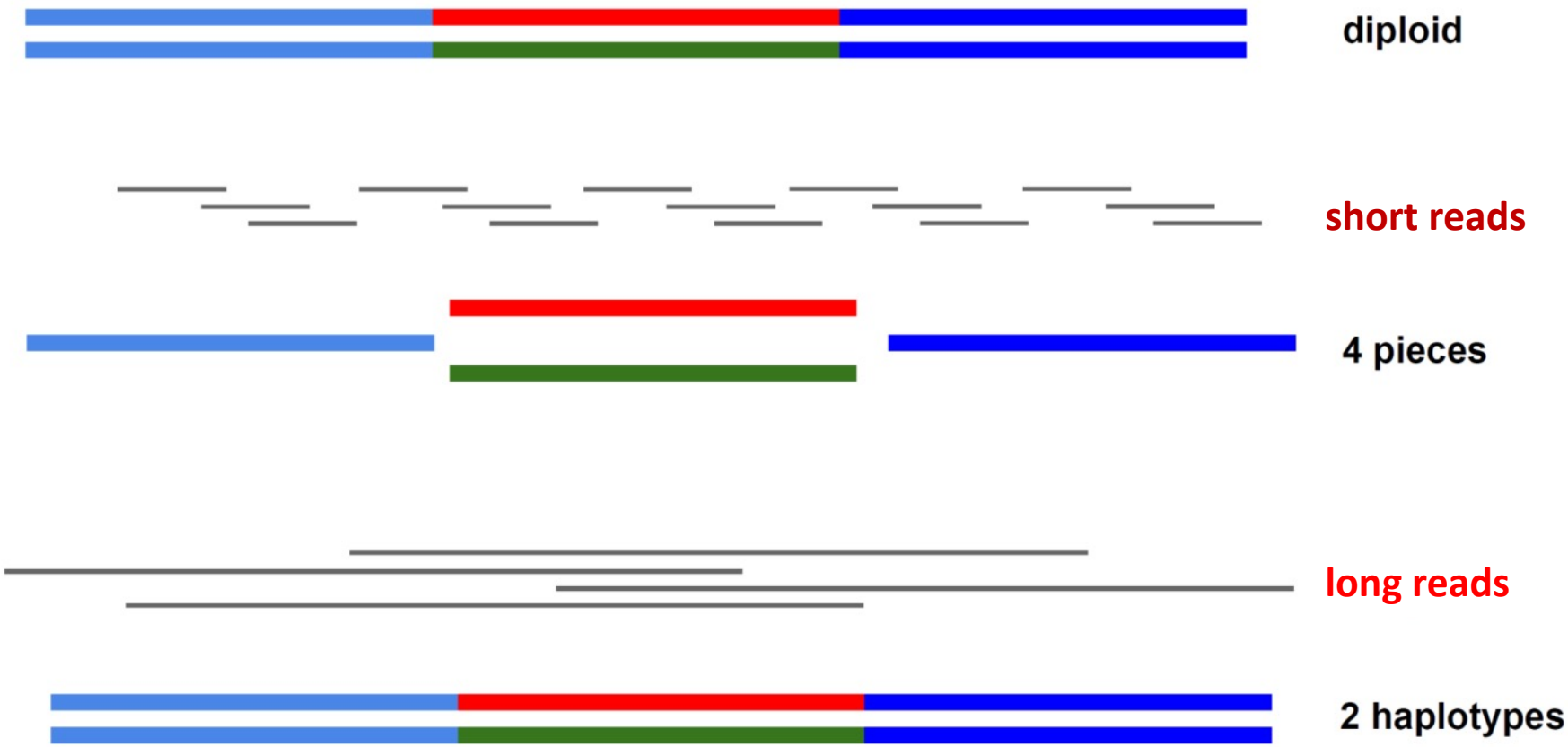
Heterozygosity:



Slide from Torsten Seemann

Why do we need long reads?

Heterozygosity:



Slide from Torsten Seemann

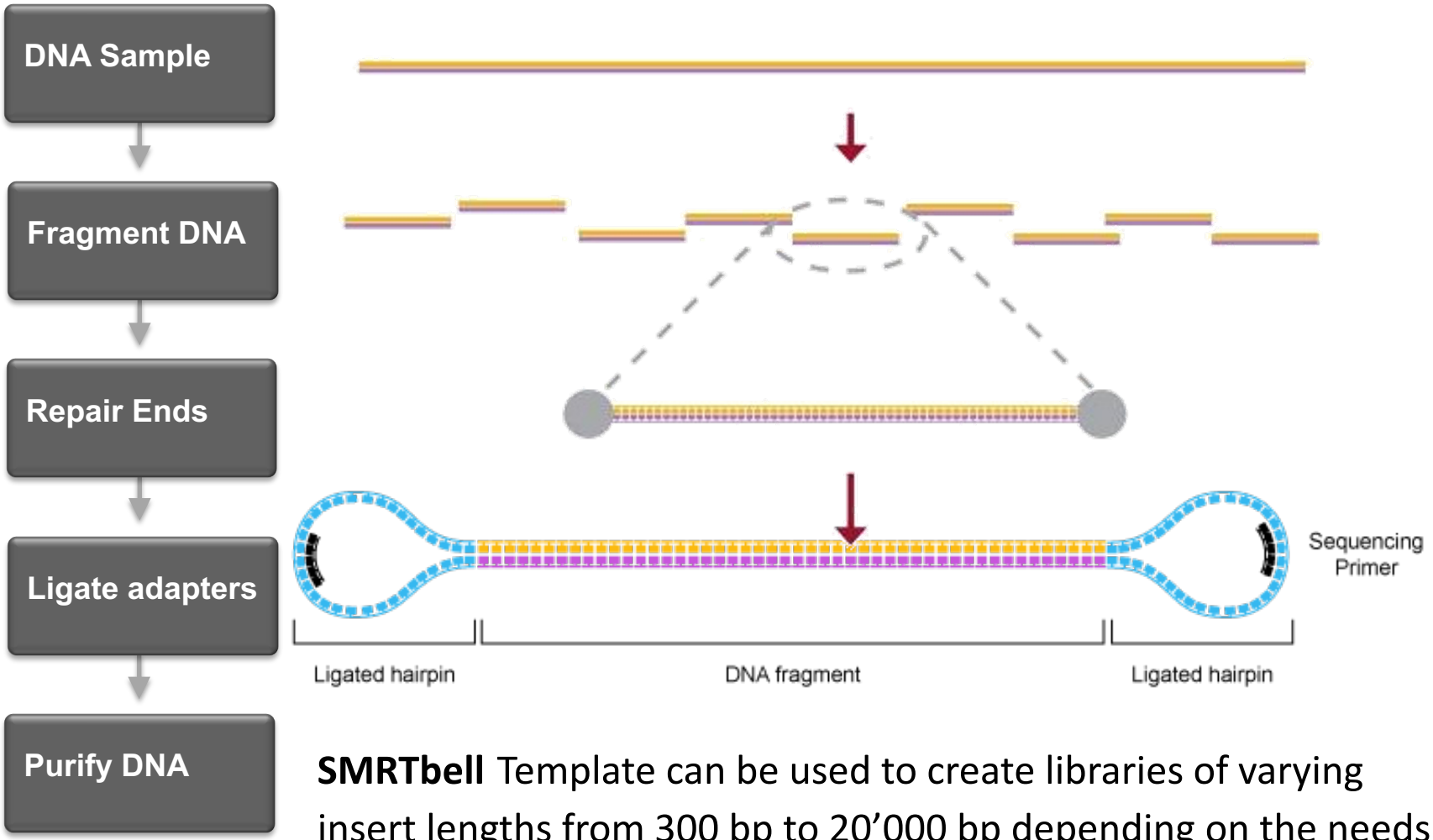
Long-Read Sequencing: Platforms

- **No** PCR amplification needed!
- **No** 'wash-and-scan' step required → Faster!

✓ **Single molecules** are immobilized on a solid surface

- **Pacific Biosciences (PacBio RS II, Sequel, Sequel II)**
- **Oxford Nanopore Technologies (ONT: MinION, GridION & PromethION)**

Long-Read Sequencing: PacBio

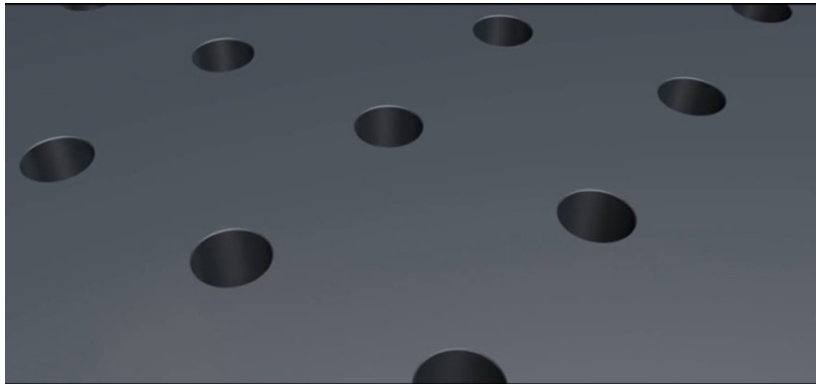


SMRTbell Template can be used to create libraries of varying insert lengths from 300 bp to 20’000 bp depending on the needs of the application.

→ The same insert can be sequenced multiple times

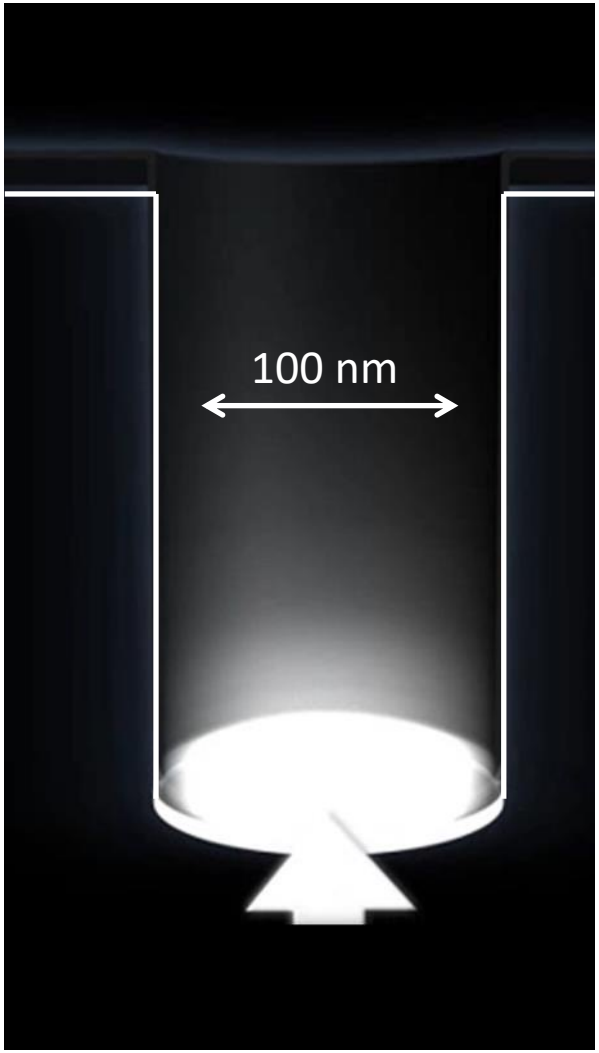
Long-Read Sequencing: PacBio

SMRT = Single Molecule Real Time Technology



http://www.pacificbiosciences.com/

Long-Read Sequencing: PacBio

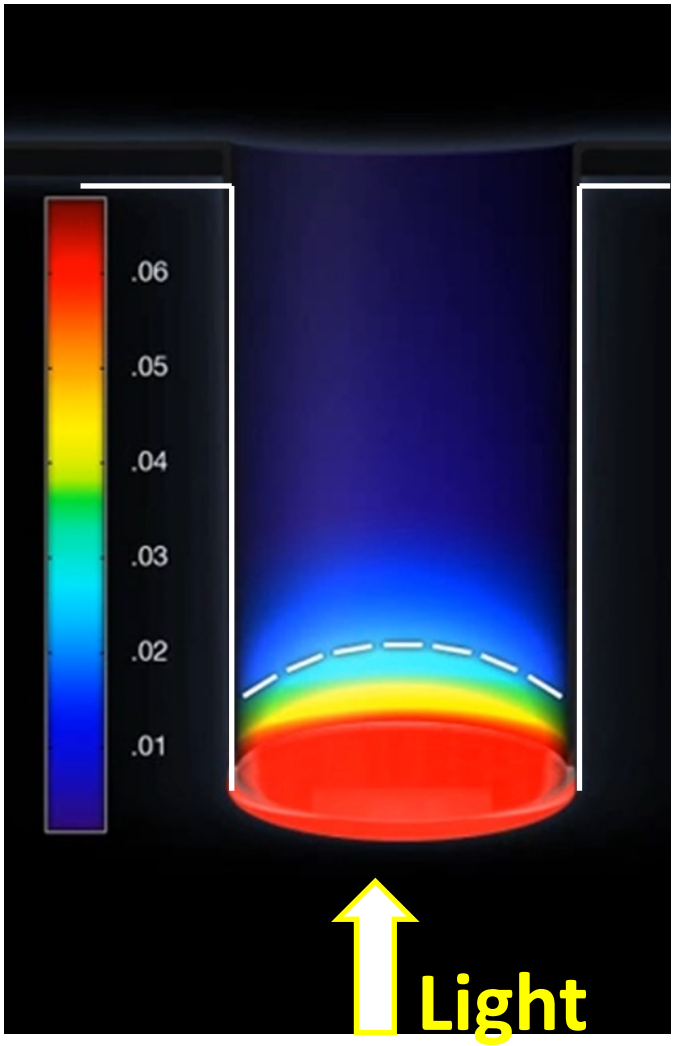


Zero-mode waveguide (ZMW) detector:

- Nanostructure device
- Diameter << wavelength of laser (532/ 643nm)
- Light can‘t efficiently pass through

http://www.pacificbiosciences.com/

Long-Read Sequencing: PacBio



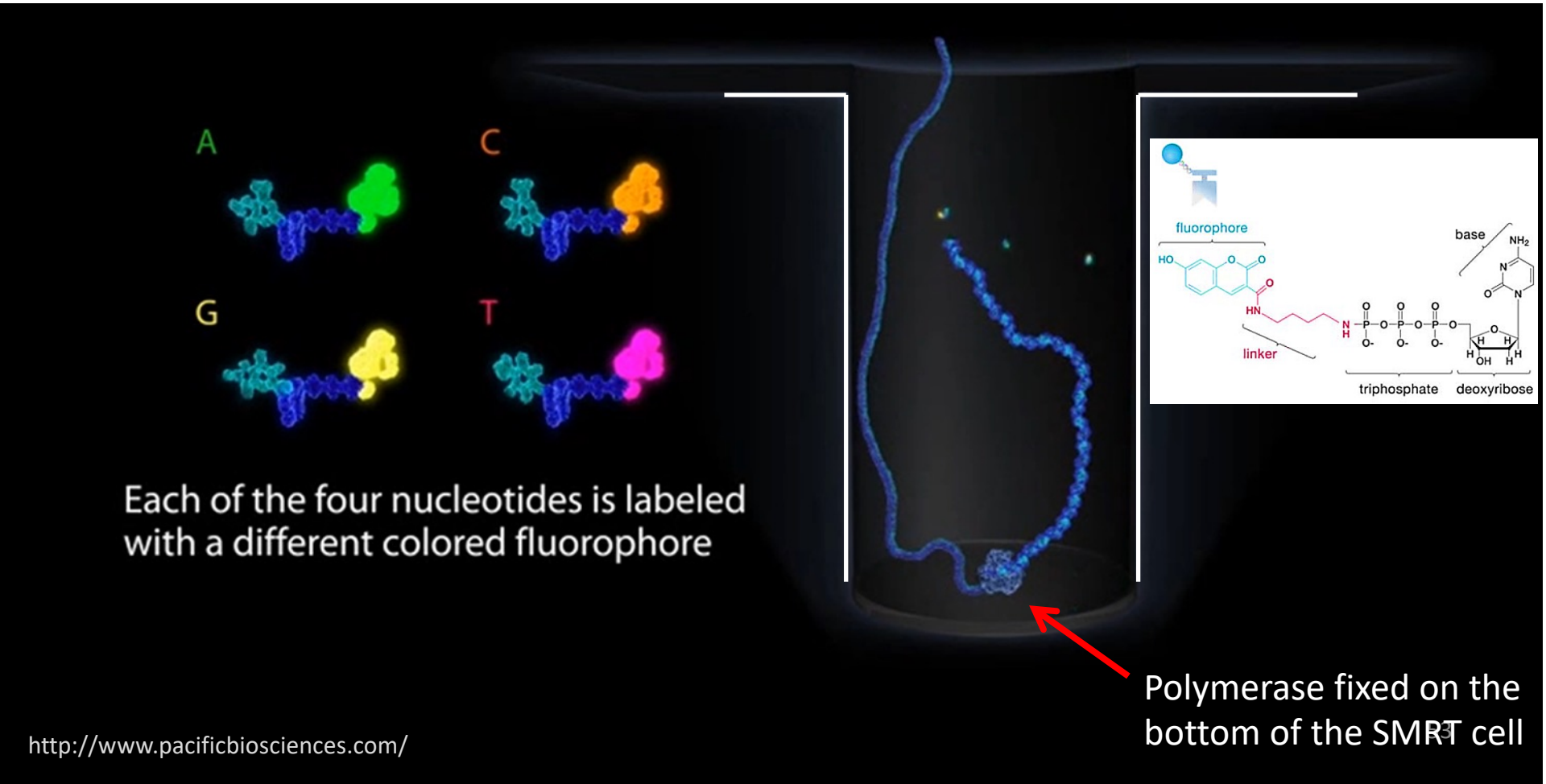
Attenuated light from the excitation beam penetrates the lower 20-30 nm of each ZMW...

...creating the world's most **powerful light microscope** with a detection volume of 20 zeptoliters (10^{-21} liters)

http://www.pacificbiosciences.com/

Long-Read Sequencing: PacBio

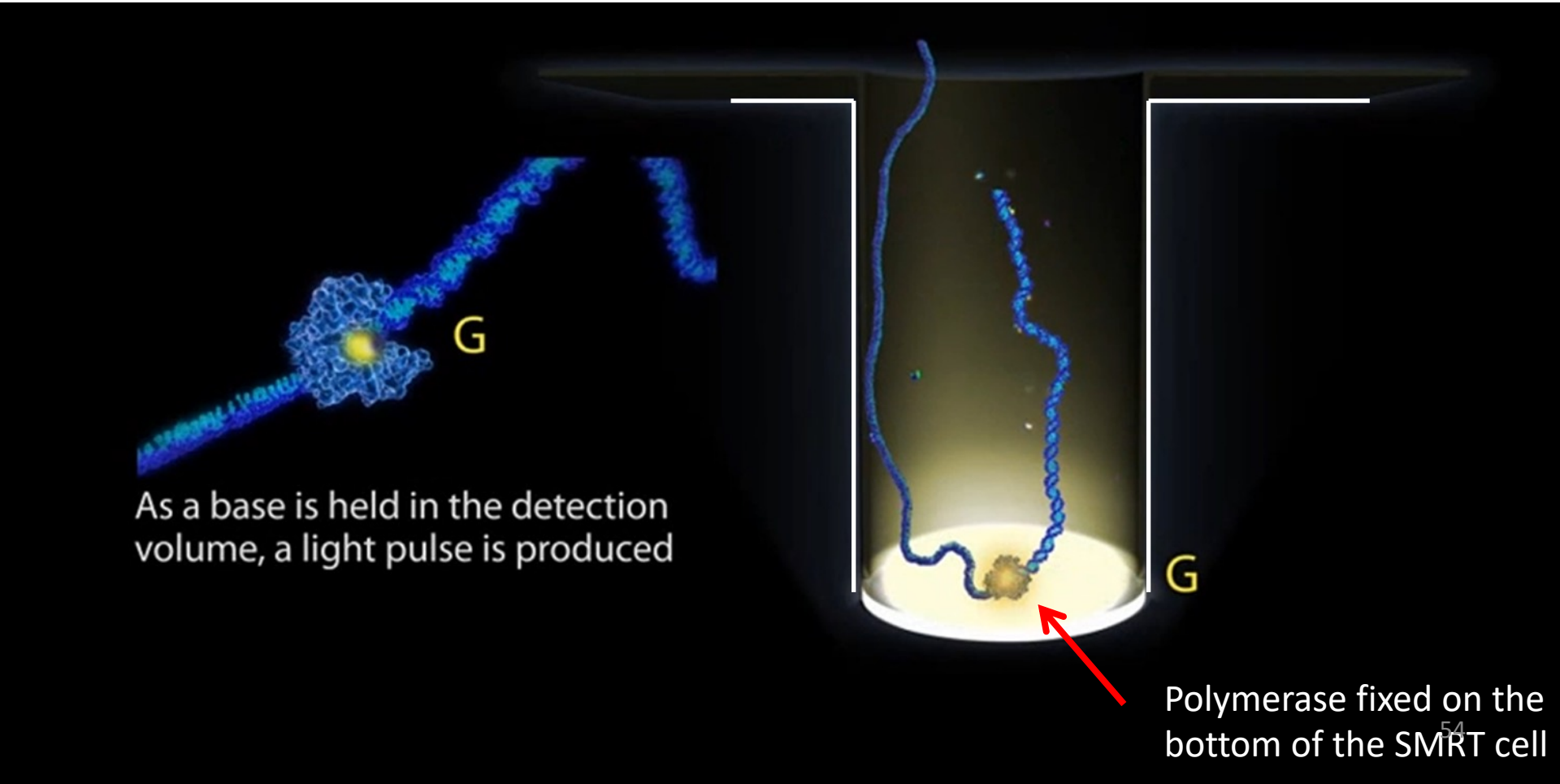
- Single polymerase molecule immobilized in each ZMW
- DNA sequence is read in real time of nucleotide incorporation
- Significant larger DNA molecules can be used (up to tens of 1000 bp)



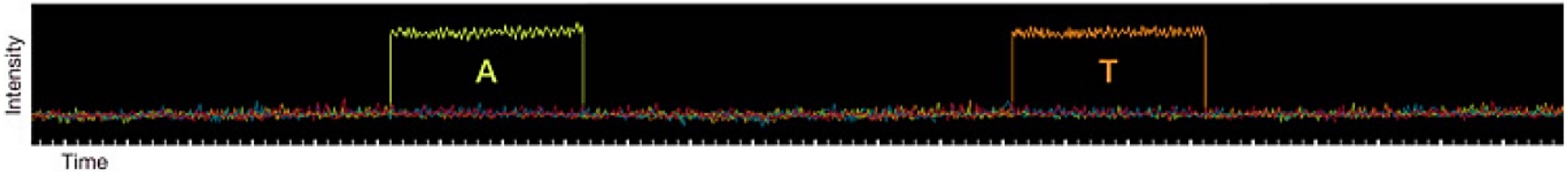
http://www.pacificbiosciences.com/

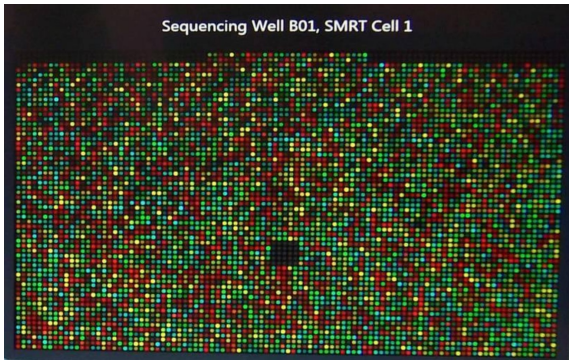
Long-Read Sequencing: PacBio

- Nucleotides diffuse in and out of detection volume (background noise)
- Fluorescent signal of the correct base remains longer (until base is incorporated and fluorophor released when phosphate chain is cut)

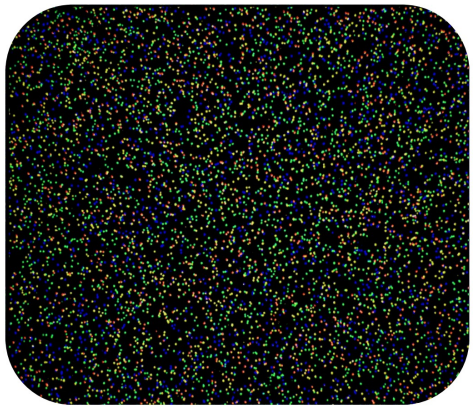


Long-Read Sequencing: PacBio



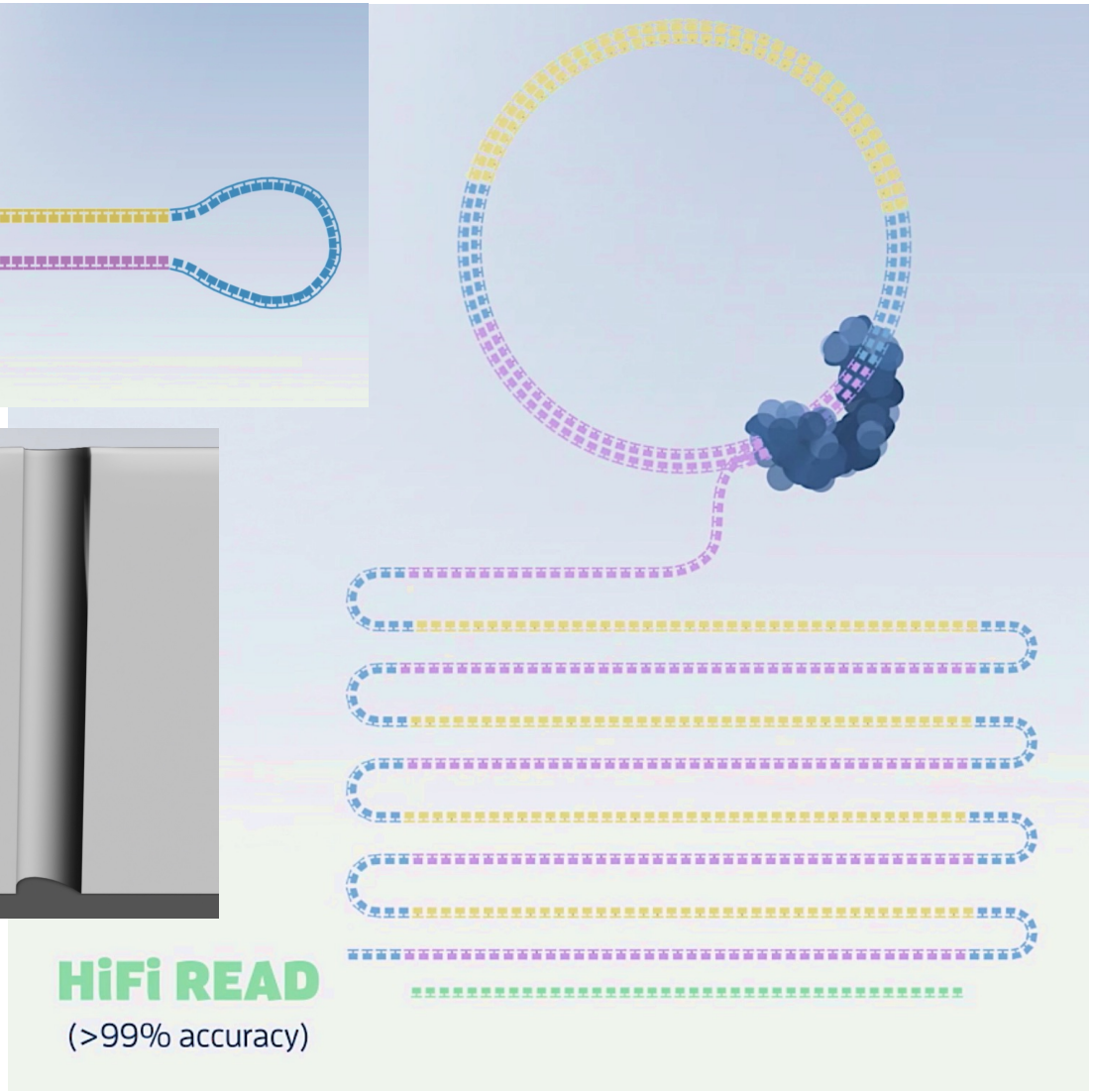


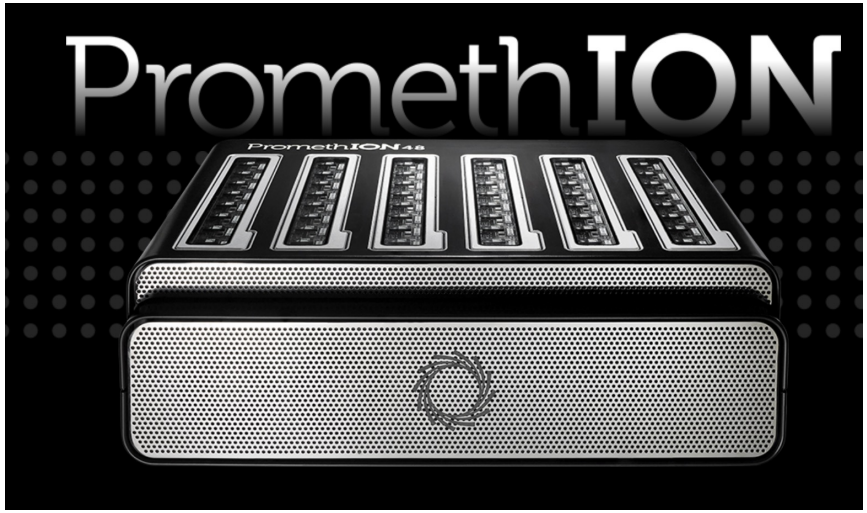
PacBio ZMWs with single DNA strand
Ordered



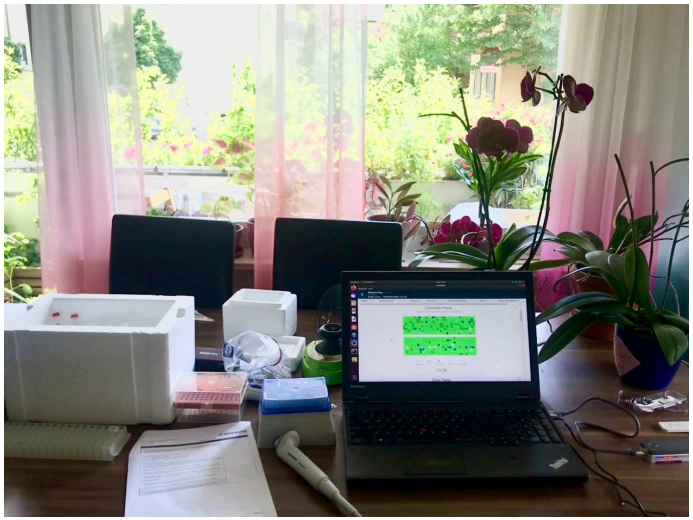
Illumina DNA mono-colonal clusters
unordered

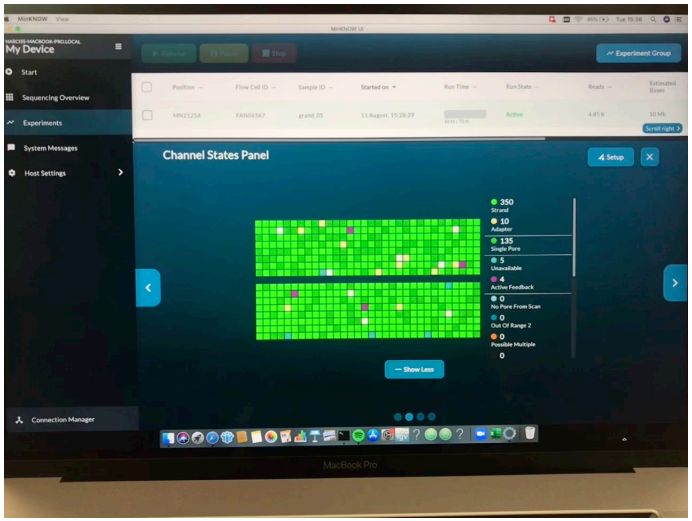
Long-Read Sequencing: PacBio

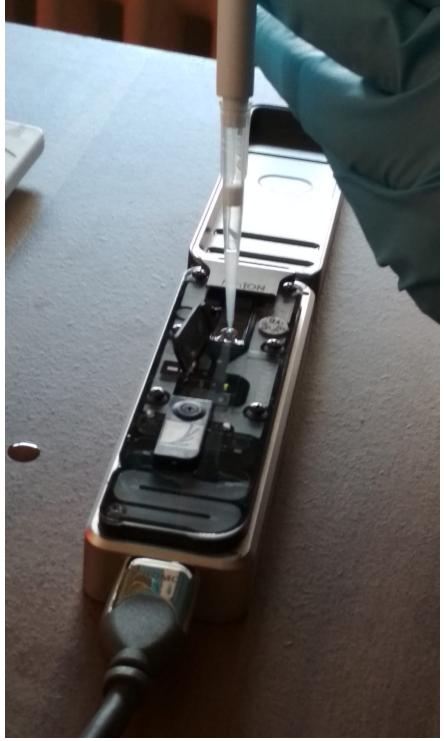




https://nanoporetech.com/products/minion

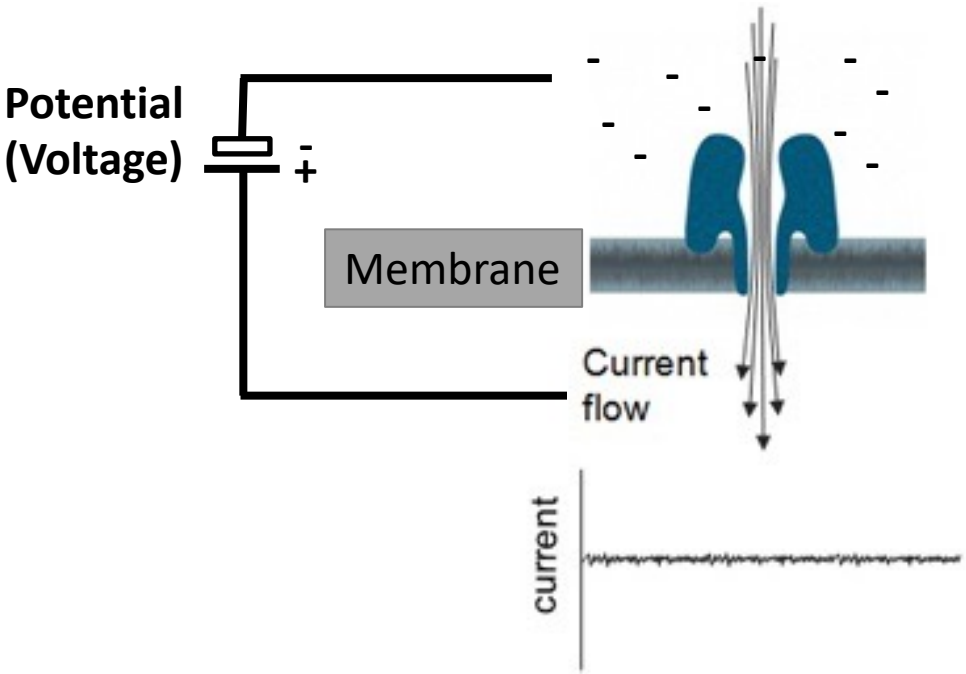






https://nanoporetech.com/products/minion
N. Yousefi

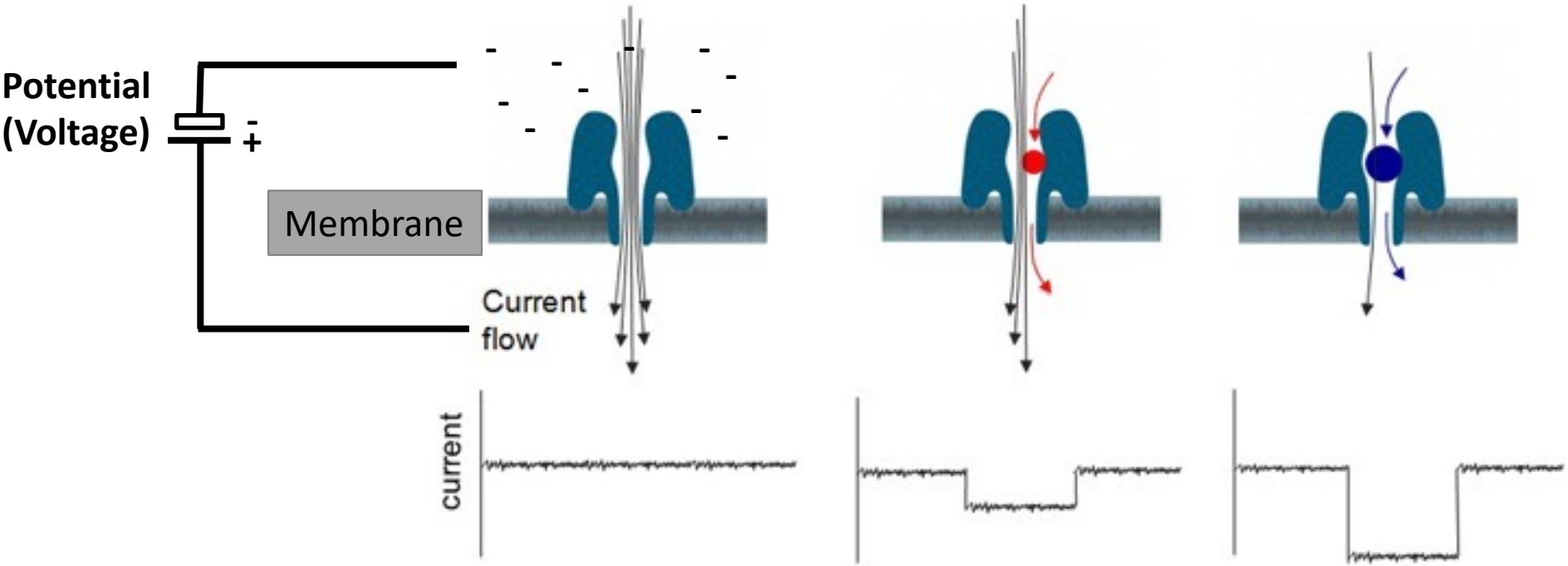
Long-Read Sequencing: Nanopore



http://cdn.phys.org/newman/gfx/news/hires/2014/oxfordnanopo.jpg

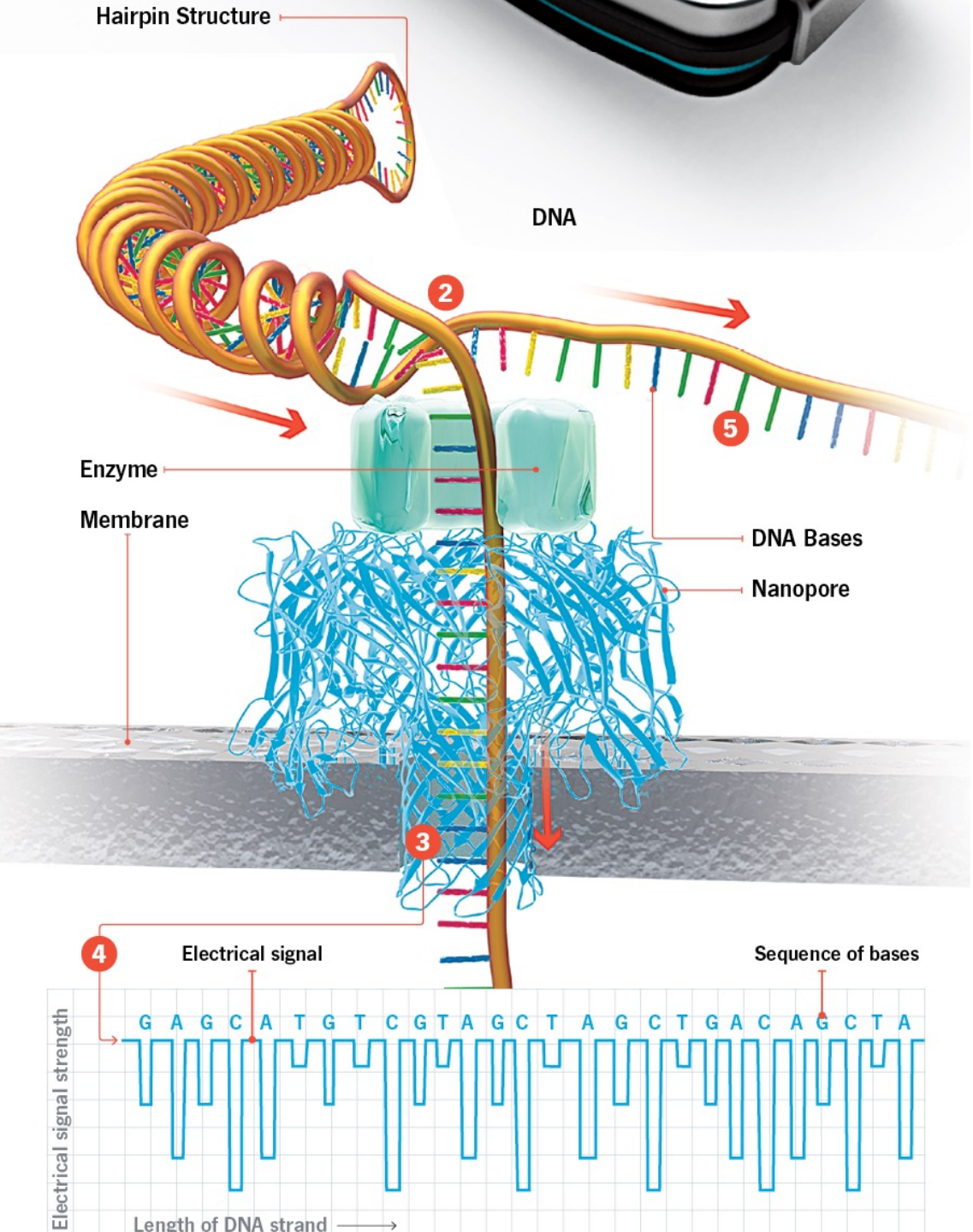
- The transmembrane protein (nanopore) is immersed in a conductive fluid to which a potential is applied
- The applied potential causes ions to flow trough the nanopore
  → we measure a current
- If something is inside the nanopore the amplitude of this current changes

Long-Read Sequencing: Nanopore



http://cdn.phys.org/newman/gfx/news/hires/2014/oxfordnanopo.jpg

- The transmembrane protein (nanopore) is immersed in a conductive fluid to which a potential is applied
- The applied potential causes ions to flow trough the nanopore
  → we measure a current
- If something is inside the nanopore the amplitude of this current changes



Long-Read Sequencing: ONT

- DNA strand is fed through a nanopore by a **processive enzyme**

- Hairpin structure: sequence **both strands**

Break?

Questions?