Plant Genomics

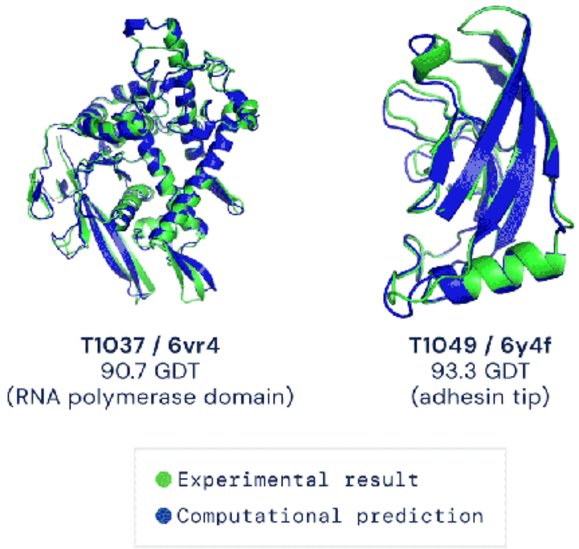
Alex Harkess

Omics, genomics, transcriptomics, so many -omics

What is computational biology?

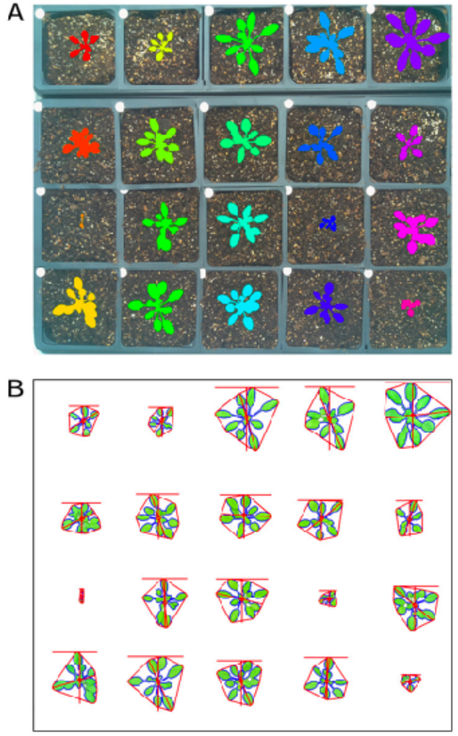
The development and application of large-scale datasets to study biological questions.

Protein modeling (AlphaFold)



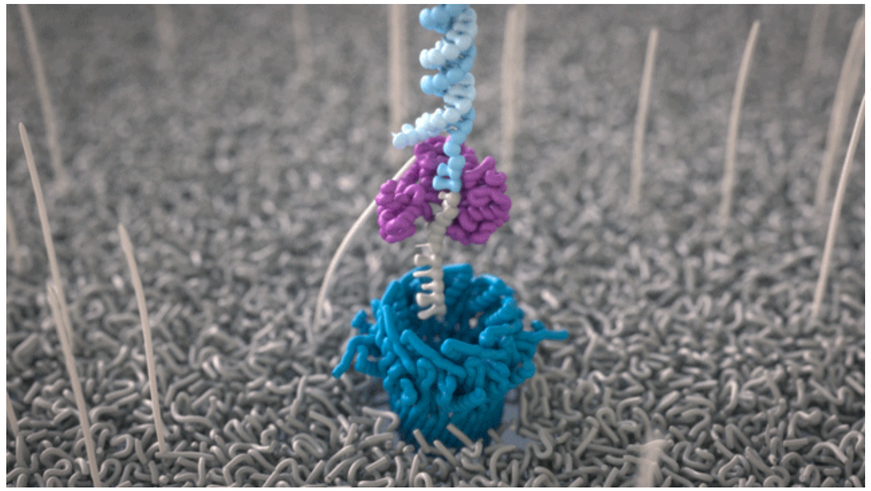
deepmind.com

Image analysis (PlantCV2)



Gehan et al. 2017

Nucleic acid sequencing (DNA/RNA)



Credit: Oxford Nanopore

You will be a stronger computational biologist
if you **understand the biology of plants**

Plant genomes are complicated
but **the technology is getting better**

How do we approach genome projects
because **plants are absurd organisms**

How do we apply computational genomics
to **assemble plant genomes**

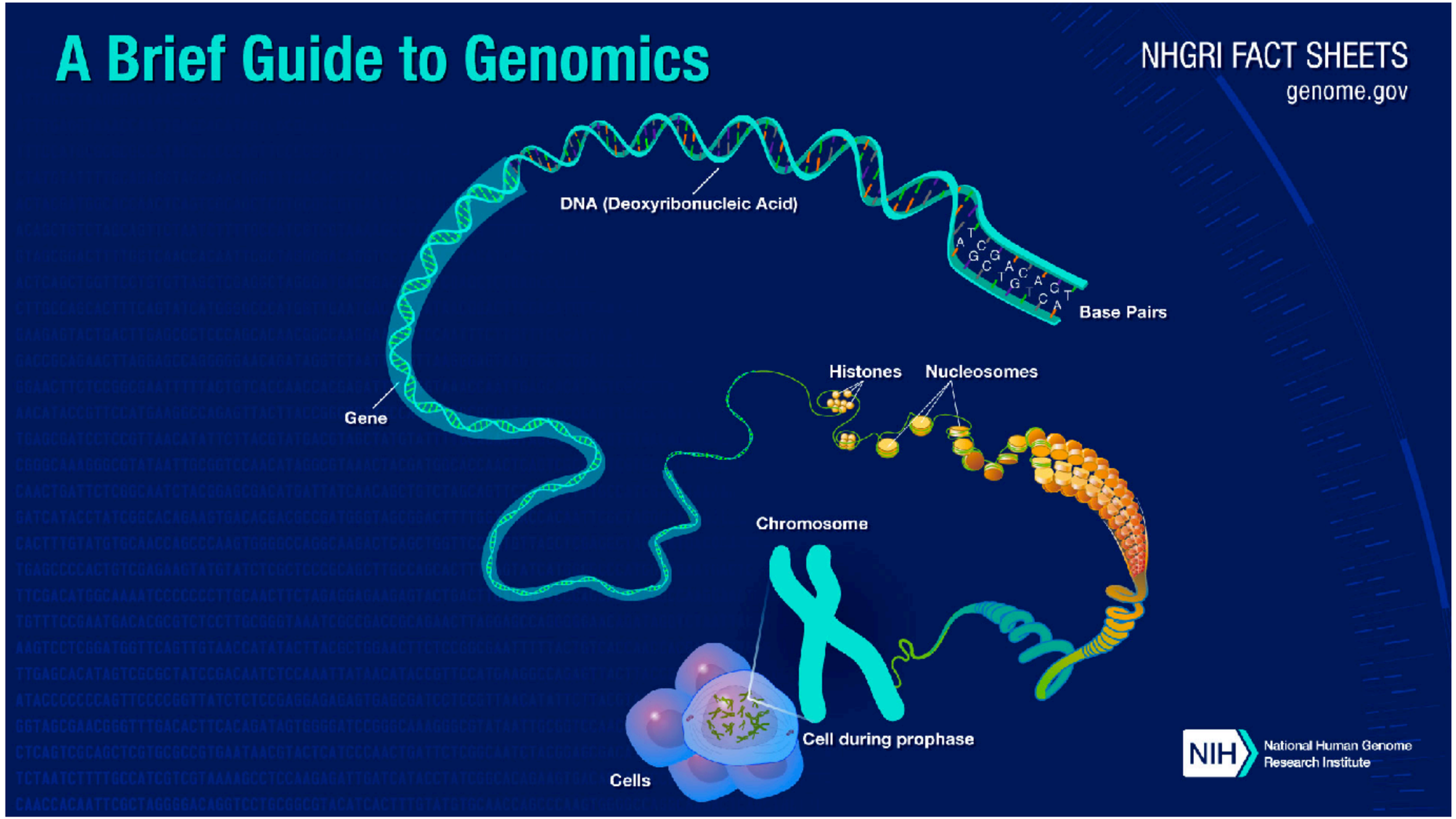
You will be a stronger computational biologist
if you **understand the biology of plants**

Plant genomes are complicated
but **the technology is getting better**

How do we approach genome projects
because **plants are absurd organisms**

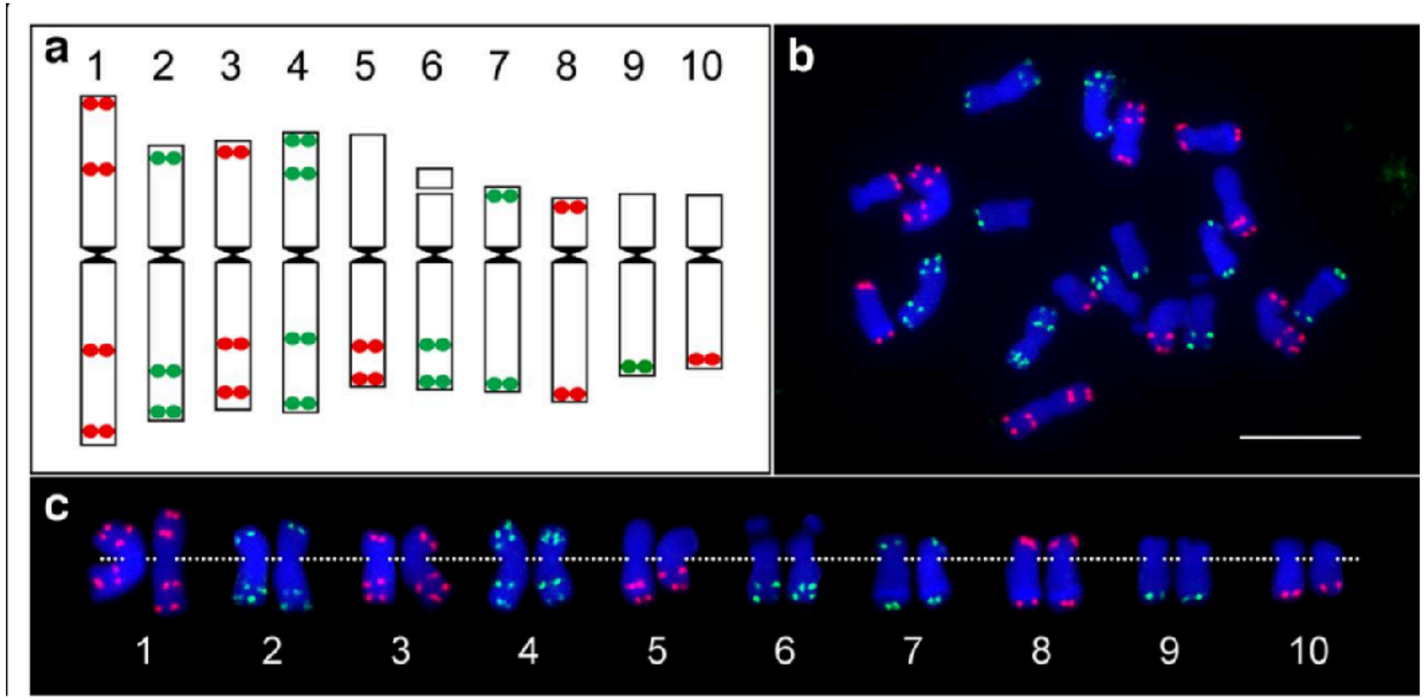
How do we apply computational genomics
to **assemble plant genomes**

What is genomics

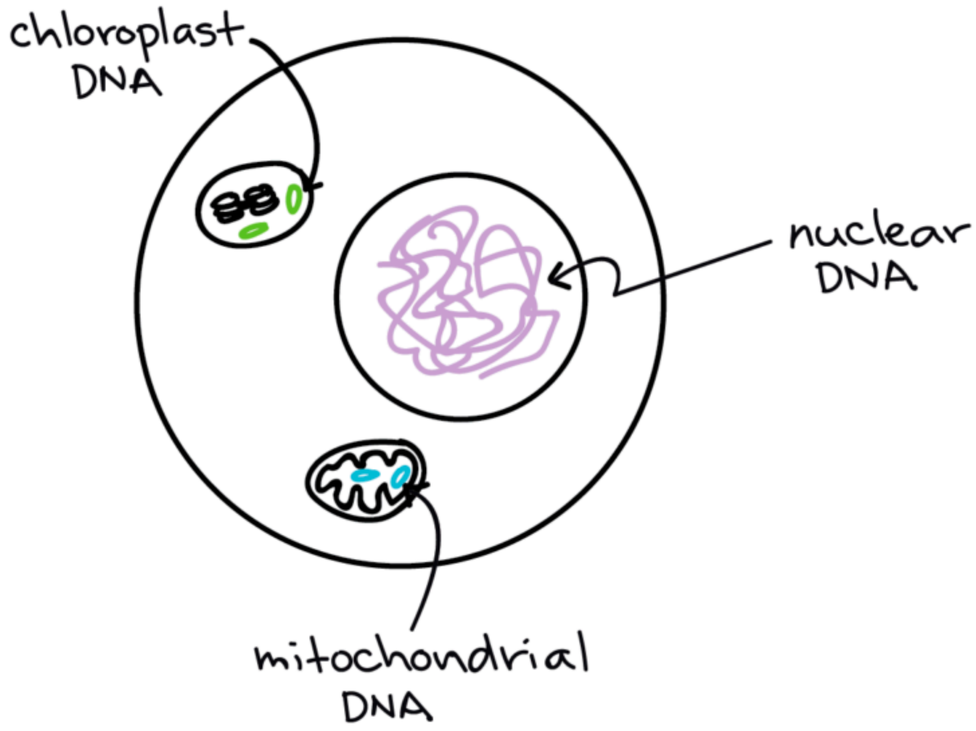


Plants have 3 distinct genomes

Nuclear genome



Braz et al. 2020



Nuclear genome sizes vary widely in plants

Utrichularia gibba
Creeping bladderwort
80 Mb

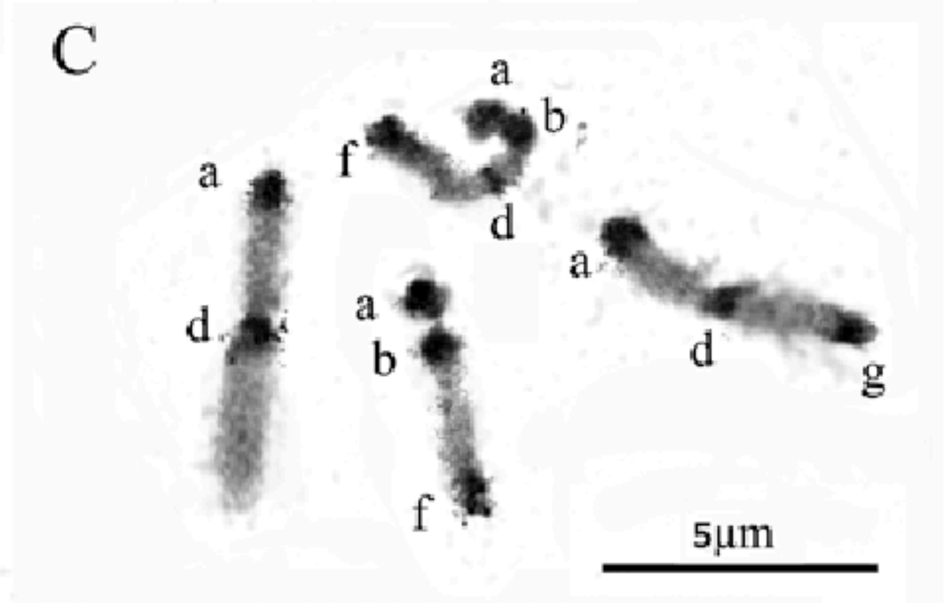
Zea mays
Corn
2.7 Gb

Paris japonica
149 Gb

...and in chromosome numbers

N = 2 chromosome pairs

Credit: Wikipedia



Haplopappus gracillis
2N = 4
(2 pairs of chromosomes)
Castiglione et al. 2008

...and in chromosome numbers

N = 630 chromosome pairs

http://cookislands.bishopmuseum.org/

Ophioglossum reticulatum
2N = 1260
(630 pairs of chromosomes)
Abraham et al. 1954

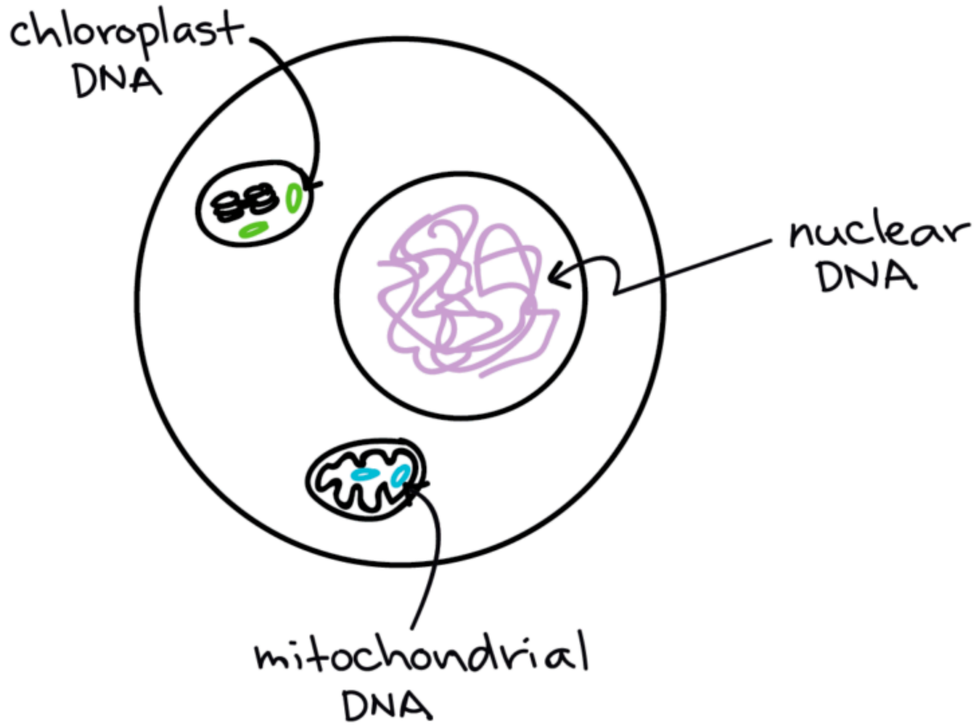
Plants have 3 distinct genomes

Chloroplast genome

- Typically ~150 kilobase circular genome
- 80-90 protein coding genes
- ~30 transfer RNA (tRNA) genes
- ~4 ribosomal RNA (rRNA) genes

Useful for phylogenetics

- maternally inherited
- low substitution rate compared to nuclear genes
- and do not recombine



The first chloroplast genome (1986)

Marchantia polymorpha (a liverwort)

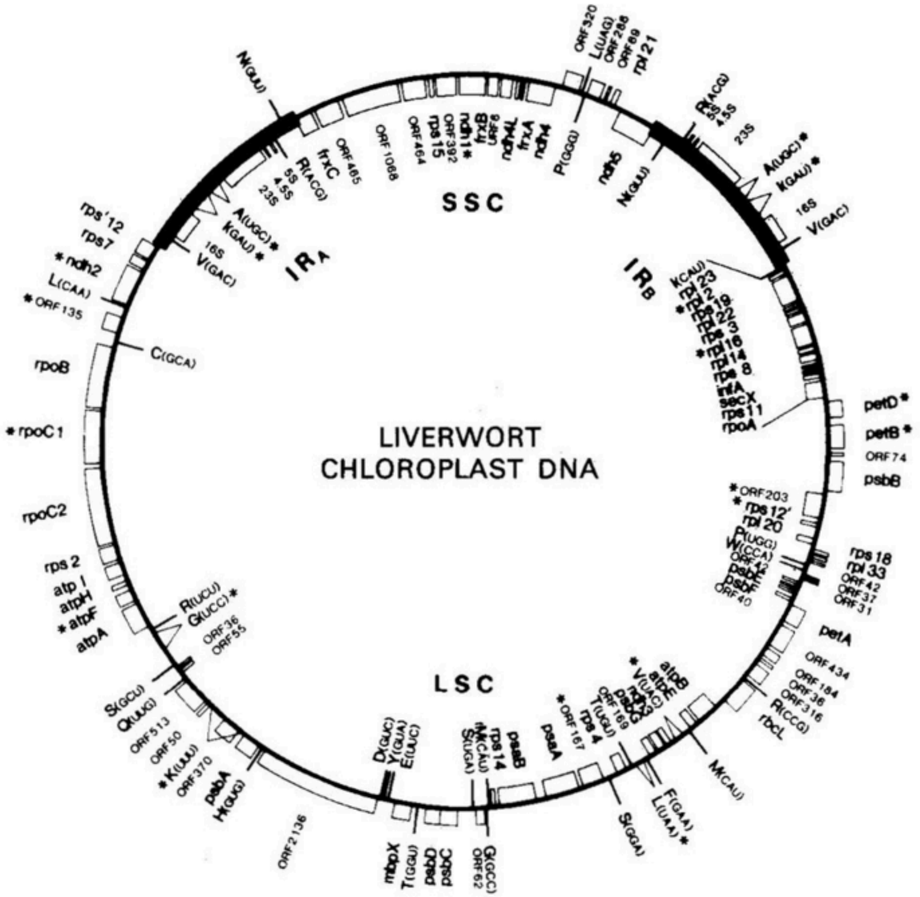


Table 1 Genes encoded by the chloroplast DNA from a liverwort, *M. polymorpha*

rRNA genes (IR_A and IR_B)
16S, 23S, 4.5S and 5S

tRNA genes
37 tRNAs (see Fig. 1)

RNA polymerase genes
rpoA: homologous to *E. coli* α-subunit
rpoB: homologous to *E. coli* β-subunit
*rpoC1**: homologous to *E. coli* β′-subunit
rpoC2: homologous to *E. coli* β′-subunit

Ribosomal protein genes and related genes

50S subunit	30S subunit
*rpl2**, *rpl14*, *rpl16**, *rpl20*, *rpl21*, *rpl22*, *rpl23*, *rpl33*	*rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12**, *rps14*, *rps15*, *rps18*, *rps19*

Others
infA, *secX*

Genes for photosynthesis
rbcL, large subunit of ribulose bisphosphate carboxylase
psaA, photosystem I P_{700} chlorophyll *a* apoprotein
psaB (same as above)
psbA, photosystem II 32K protein
psbB, photosystem II P_{680} chlorophyll *a* apoprotein
psbC (same as above)
psbD, photosystem II D2 protein
psbE, cytochrome b_{559}
psbF (same as above)
psbG, photosystem II G protein
atpA, ATPase F_1 subunit α
atpB, ATPase F_1 subunit β
atpE, ATPase F_1 subunit ε
*atpF**, ATPase F_0 subunit I
atpH ATPase F_0 subunit III
atpI, ATPase F_0 subunit IV
petA, cytochrome *f*
*petB**, cytochrome b_6
*petD**, subunit 4 of cytochrome b_6/f complex

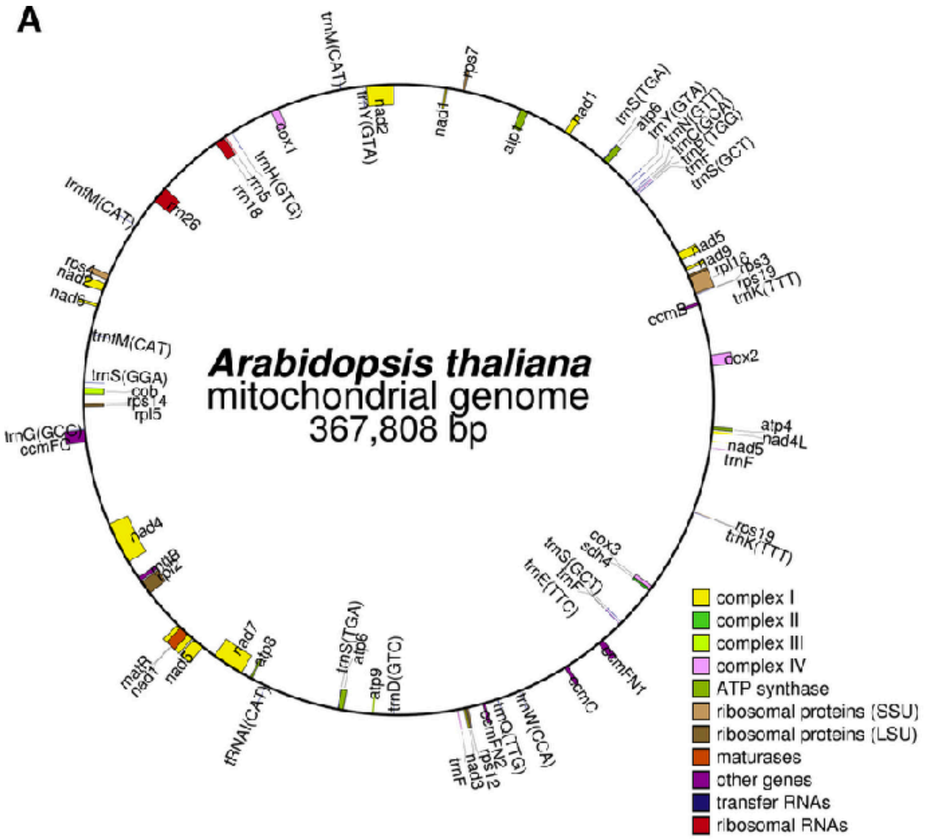
Genes predicted by amino-acid sequence homology
*ndh1**, homologous to mammalian mitochondrial URF1
*ndh2**, URF2 (same as above)
ndh3, URF3 (same as above)
ndh4, URF4 (same as above)
ndh4L, URF4L (same as above)
ndh5, URF5 (same as above)
URF6, homologous to *Aspergillus nidulans* mitochondrial URFC (human mitochondrial URF6)
frxA, homologous to 4Fe-4S type ferredoxin
frxB (same as above)
frxC, homologous to 4Fe-4S protein found in *R. capsulata*
mbpX, homologous to ATP-binding subunit of inner membrane permease in *E. coli* (*malK*) or *Salmonella typhymurium* (*hisP*)[29]

Unidentified genes
More than 28 ORFs

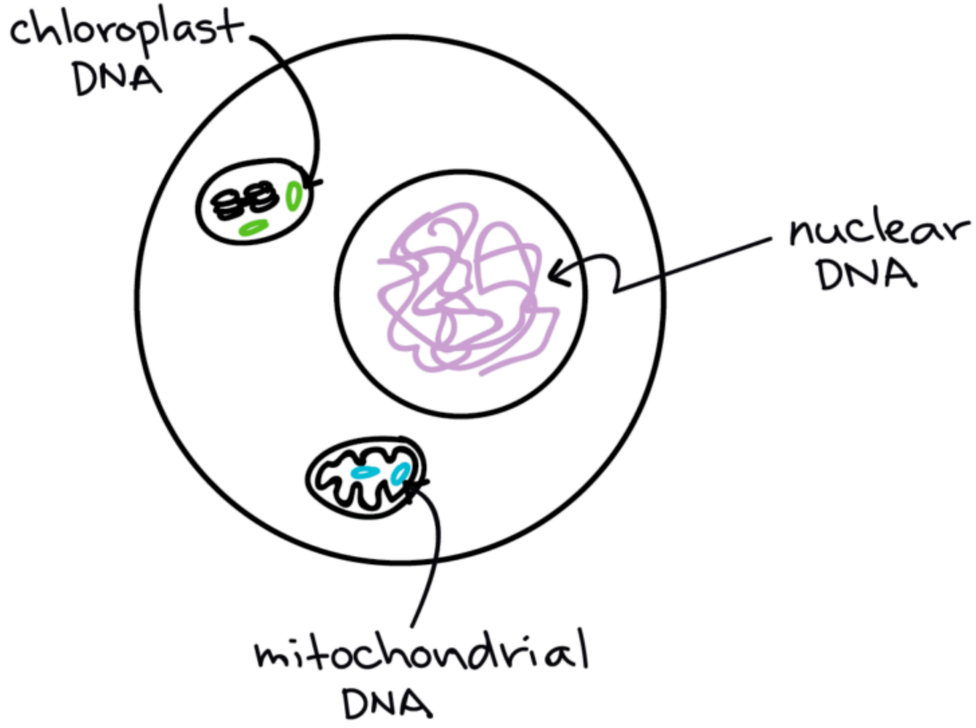
Ohyama et al. 1986 *Nature*

Plants have 3 distinct genomes

Mitochondrial genome



Johnston 2019



Plants vary in ploidy

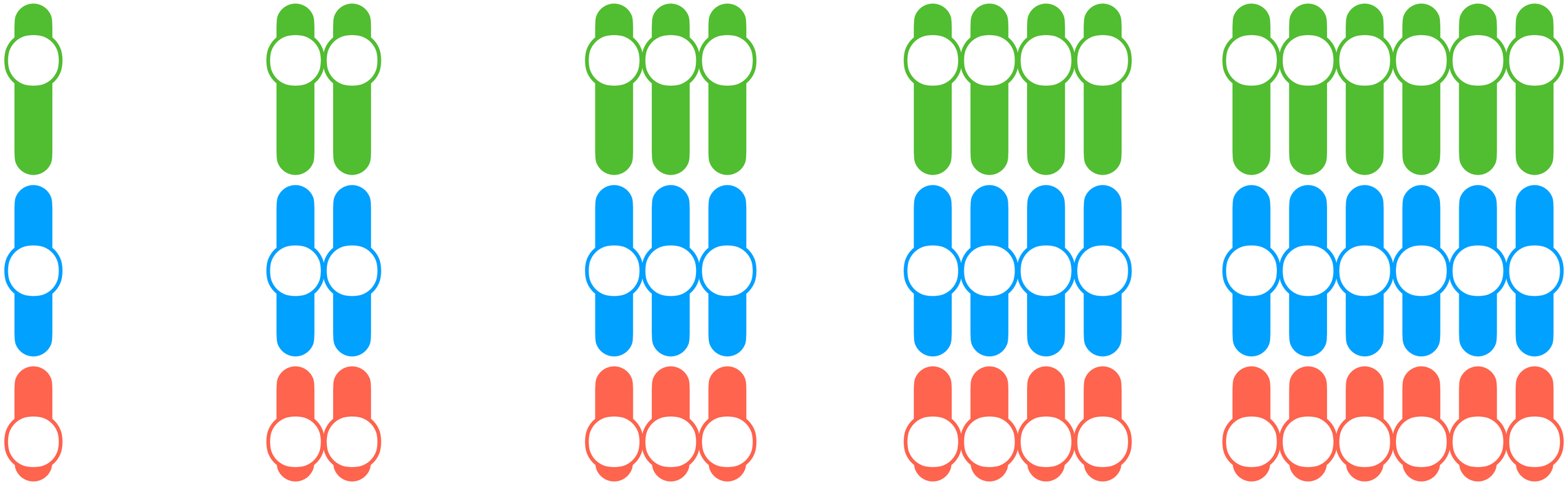
Haploid (n)

Diploid (2n)

Triploid (3n)

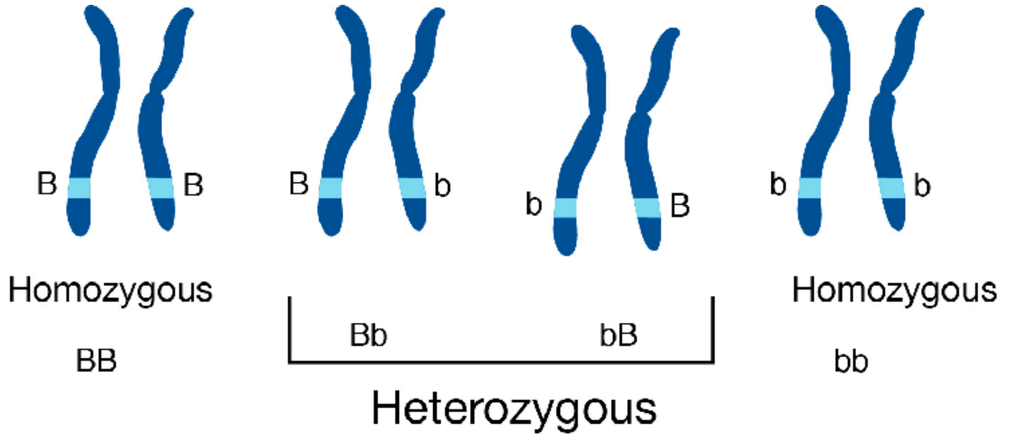
Tetraploid (4n)

Hexaploid (6n)



Plants vary in heterozygosity

- **Zygosity** = the degree to which both copies of a chromosome have the same allele
- **Homozygous** = one locus has the same allele on both chromosomes
- **Heterozygous** = one locus has different alleles
- **Hemizygous** = one locus has presence/ absence variation of an allele

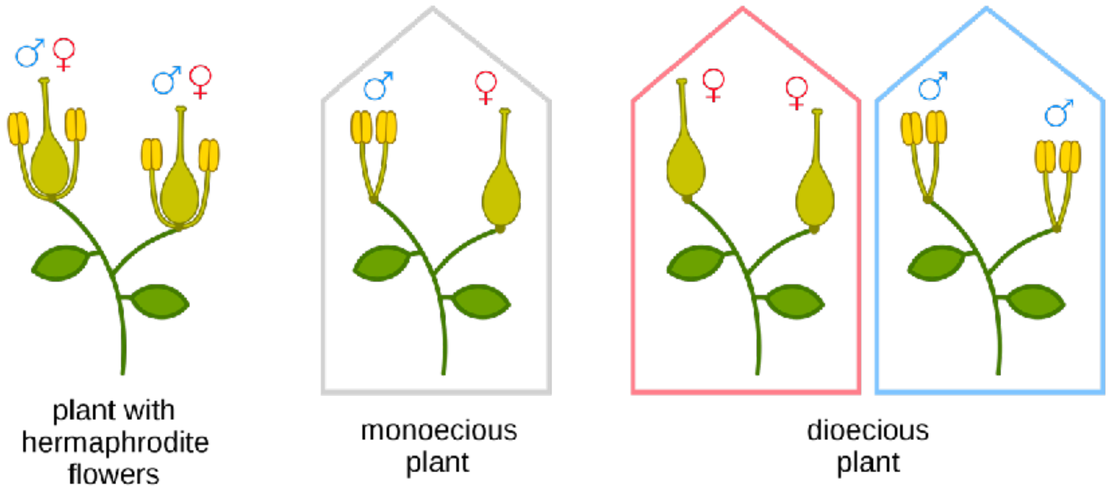


Credit: Wikipedia

- **B** and **b** are two alleles of the same gene on homologous chromosomes
- **Allele** = a form of a gene; a variant

Heterozygosity relates to reproductive strategies

- Flowering plants (angiosperms) reproduce in a huge diversity of ways
- Hermaphrodite flowers
- Self-compatible, self-incompatible
- Obligate outcrossers (dioecious)
- And all combinations within!



Credit: Wikipedia

You will be a stronger computational biologist
if you **understand the biology of plants**

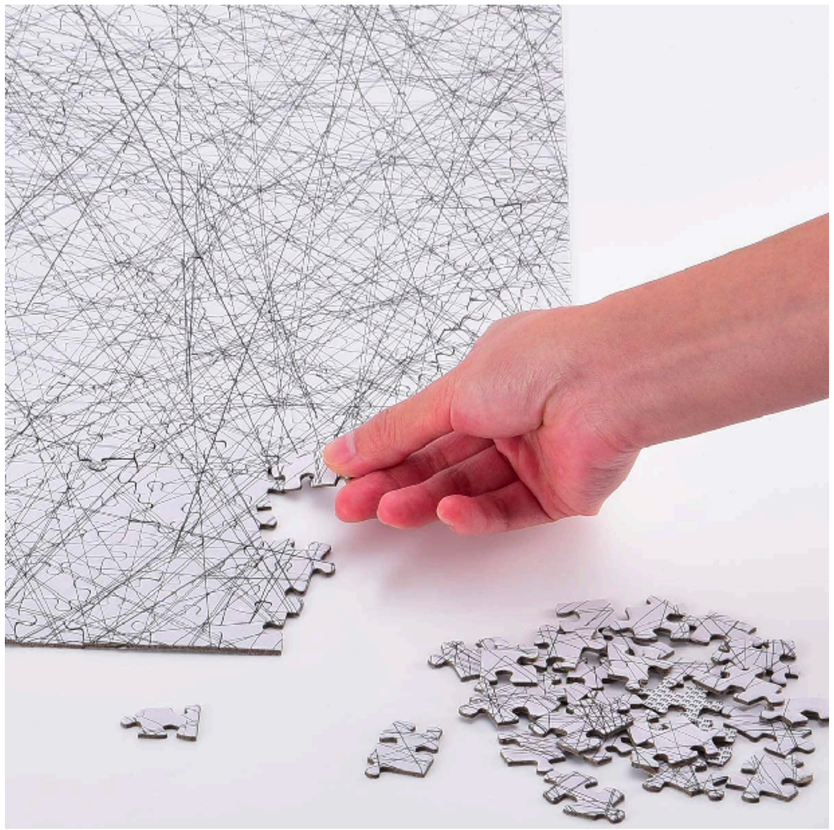
Plant genomes are complicated
but **the technology is getting better**

How do we approach genome projects
because **plants are absurd organisms**

How do we apply computational genomics
to **assemble plant genomes**

De novo genome assembly

The process of reconstructing the original DNA sequence with only the sequence **read** data



Like a jigsaw puzzle…kind of.
Lots of pieces
Easy to find the edges and corners

But…
Many pieces look the same
Pieces have 2 sides (forward/reverse strandedness)
Diploid organisms = 2 puzzles in same box
(But 95%+ of the pieces are identical)
There might not be a cover
Some of the pieces are wrong
Some of the pieces are from another puzzle

Modern data generation is massive

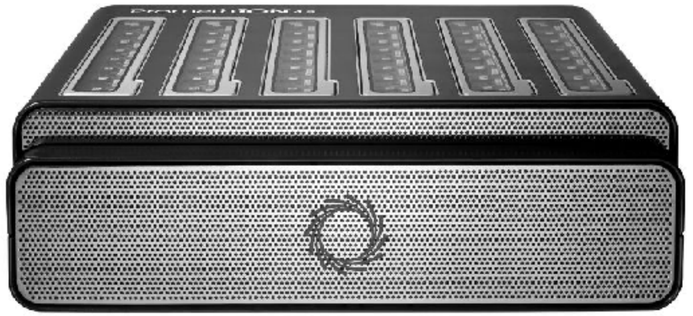
As of 2021...

Illumina NovaSeq6000

Around $1M instrument cost
20 billion reads in 48 hours
$20,000 per flow cell run
48 human genomes

PacBio Sequel IIe

Around $500k instrument cost
4 million reads in 30 hours
$3,500 per flow cell
0.5 human genomes



Oxford Nanopore PromethION

Around $500k instrument cost
4 million reads in 30 hours
$2,000 per flow cell
2 human genomes

Modern data generation is massive

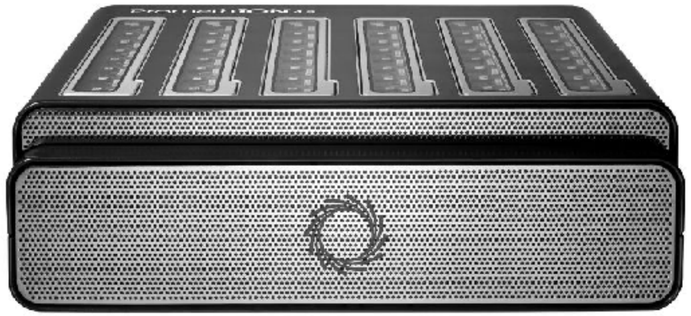
As of 2021...

Illumina NovaSeq6000

Tiny reads
accurate
billions of reads
super cheap ($)

PacBio Sequel IIe

Huge reads
super accurate
millions of reads
and expensive ($$$)

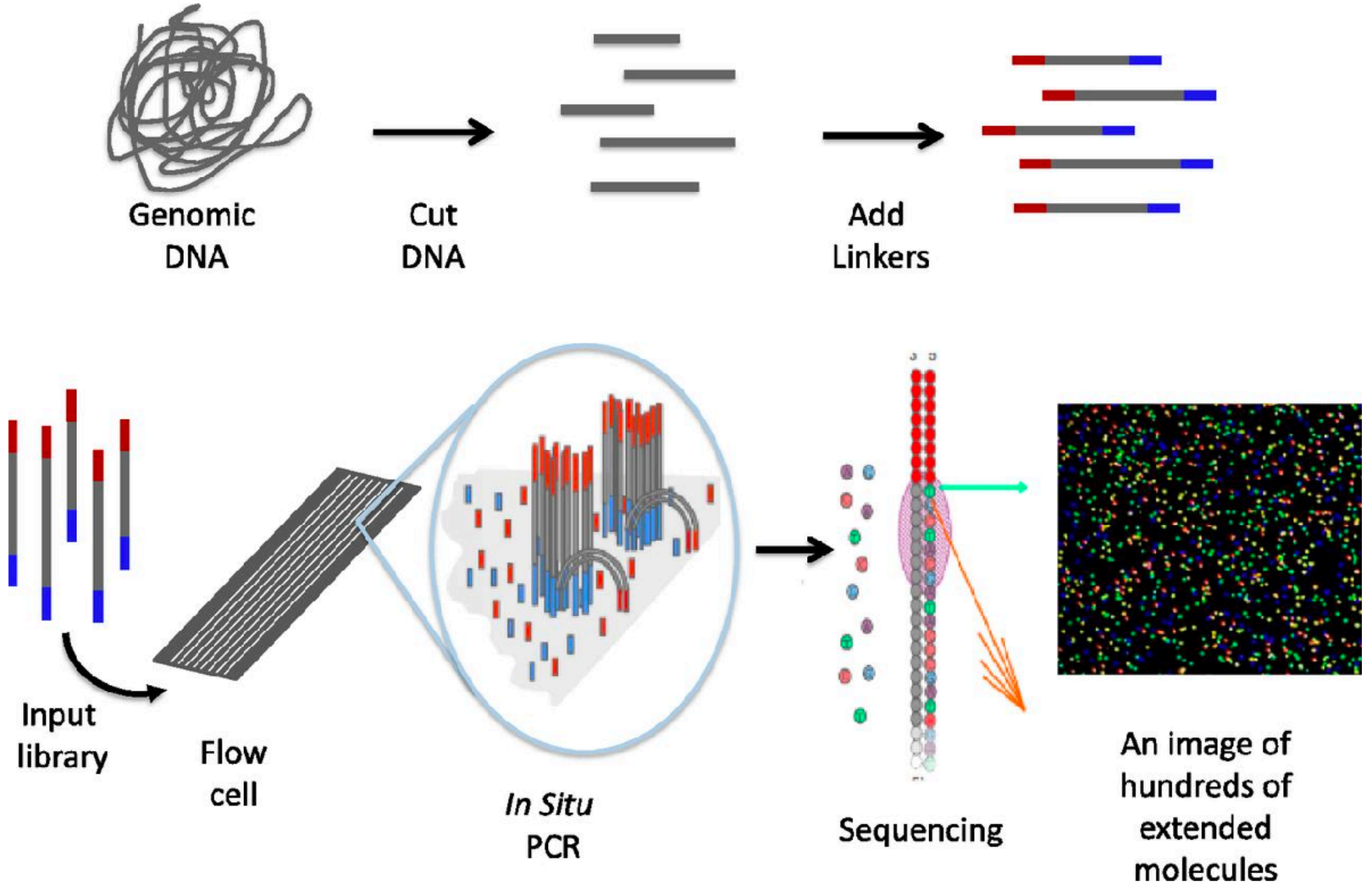


Oxford Nanopore PromethION

Huge reads
accurate-ish…
millions of reads
inbetween ($$)

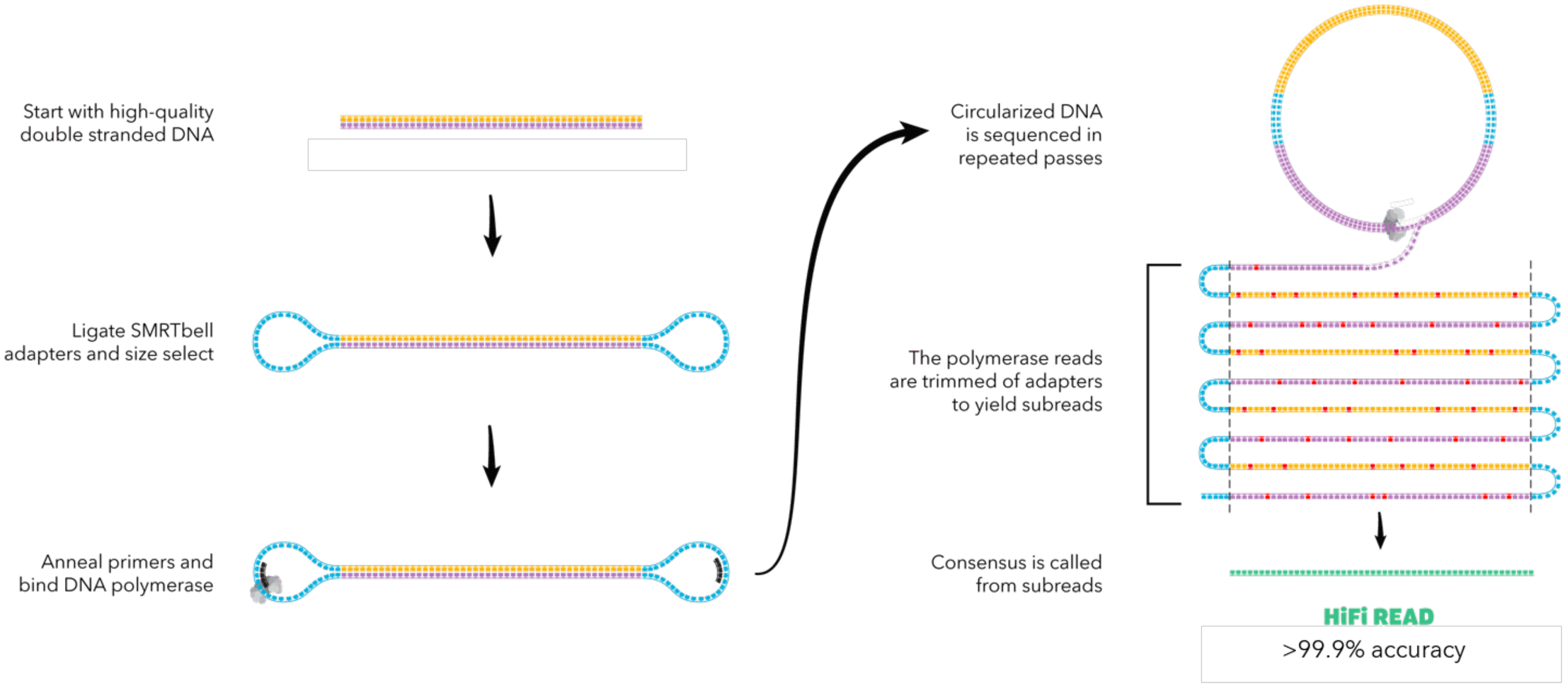
Illumina sequencing

Short reads (~150 nucleotides)



PacBio HiFi sequencing

Make bigger pieces

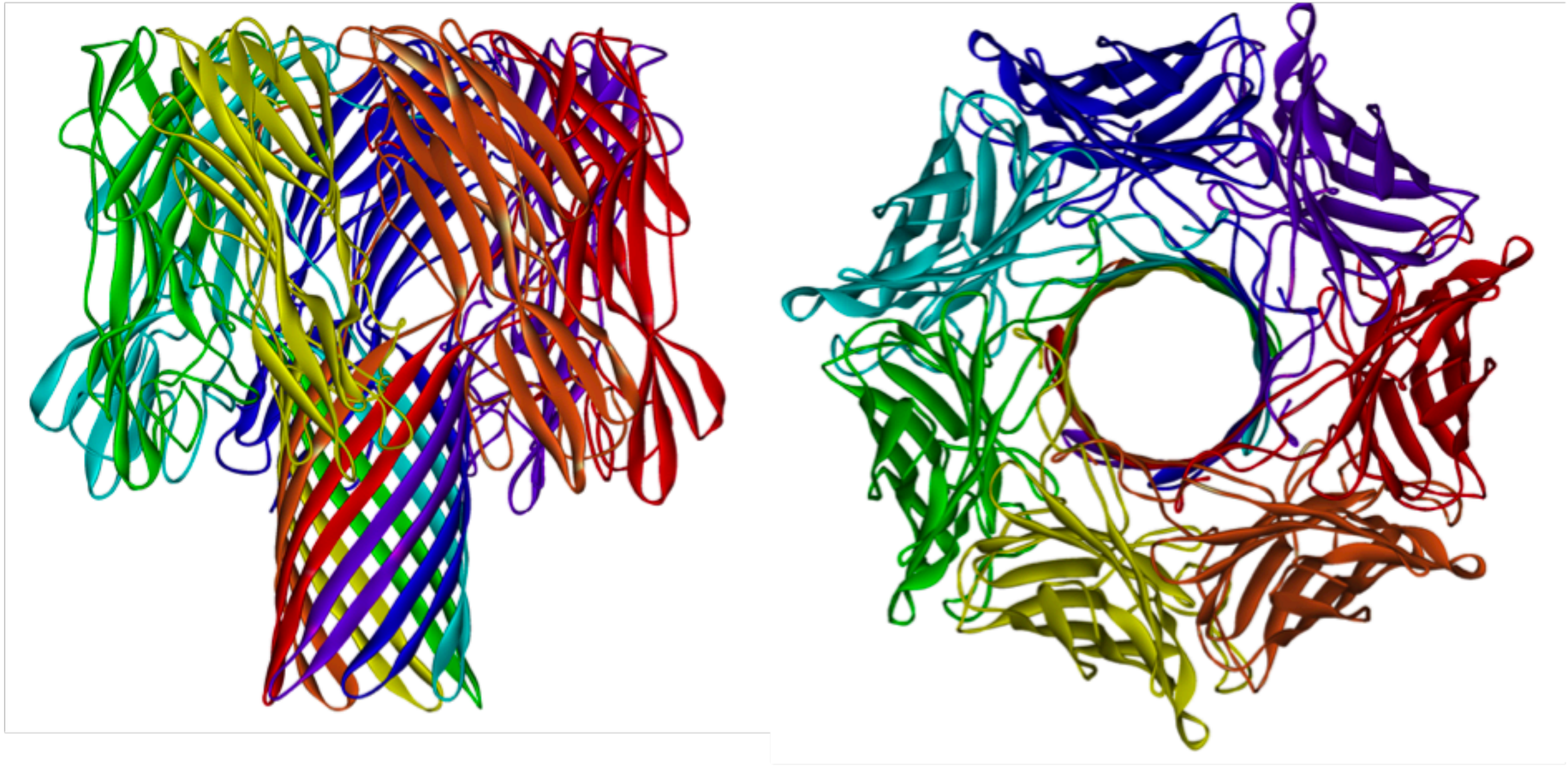


Oxford Nanopore sequencing

Make bigger pieces

Side view

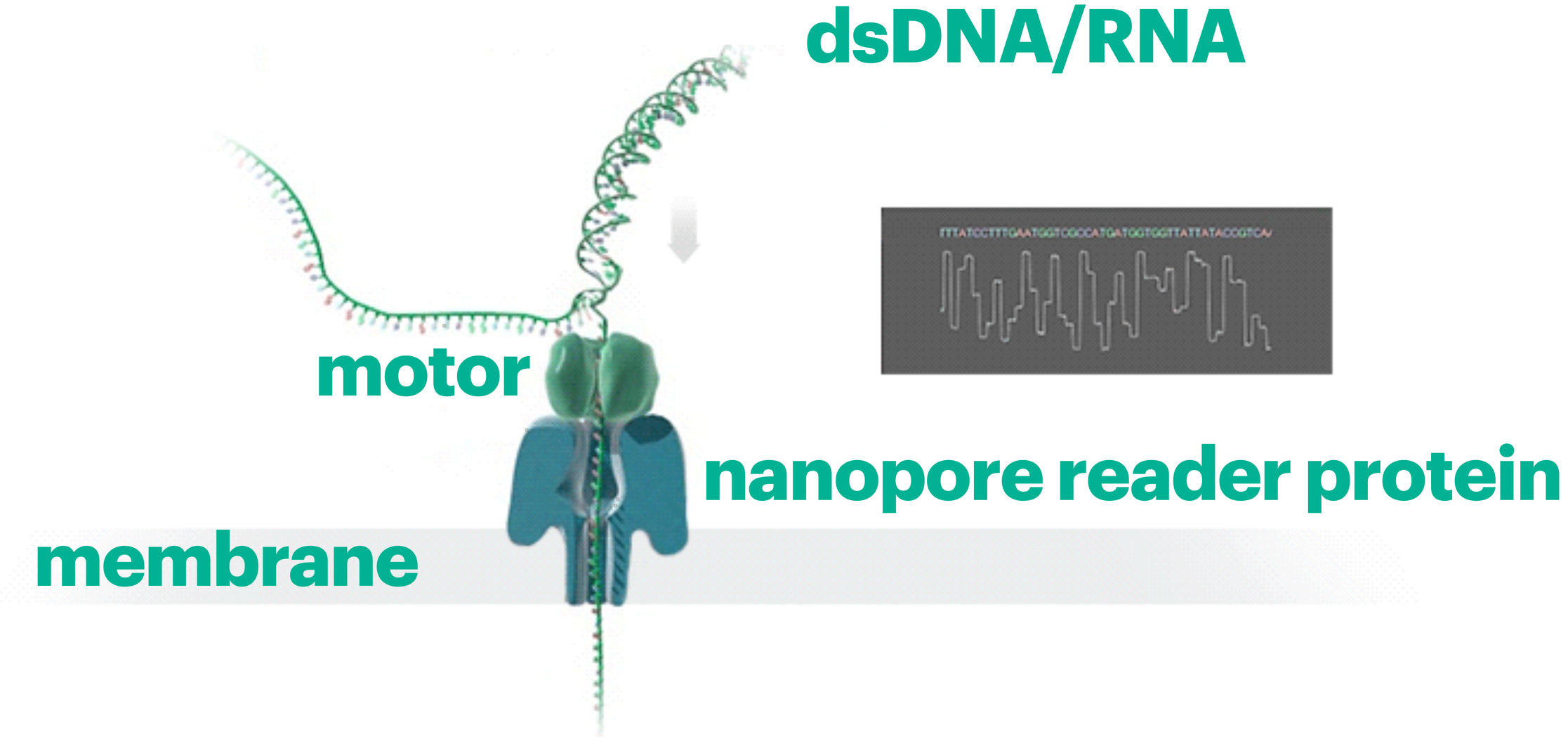
Top view



alpha hemolysin

Oxford Nanopore sequencing

Make bigger pieces



You will be a stronger computational biologist
if you **understand the biology of plants**

Plant genomes are complicated
but **the technology is getting better**

How do we approach genome projects
because **plants are absurd organisms**

How do we apply computational genomics
to **assemble plant genomes**

Major steps to start a genome project

Identifying the accession you want to sequence

Isolate ultra HMW DNA

Isolate RNA from a variety of tissues and conditions

Get “small DNA” for Illumina shotgun
Genome size, ploidy, heterozygosity

Get “big DNA” for PacBio sequencing

Get RNA to annotate genes,
or do an experiment
(e.g. differential expression)

Identifying the accession you want to sequence

Before you start:

- **Questions you should ask first:**
  - Do I know the genome size (flow cytometry or kmer-based)
  - Is this a selfer or an outcrosser?
    - Has this specific individual been inbred for several generations?
    - Is it wild collected?
  - Annual, or perennial? Can I return to this one plant?
  - Do I have adequate plant material (e.g. young leaves)?
  - Can I get a diversity of tissues for RNA sequencing?

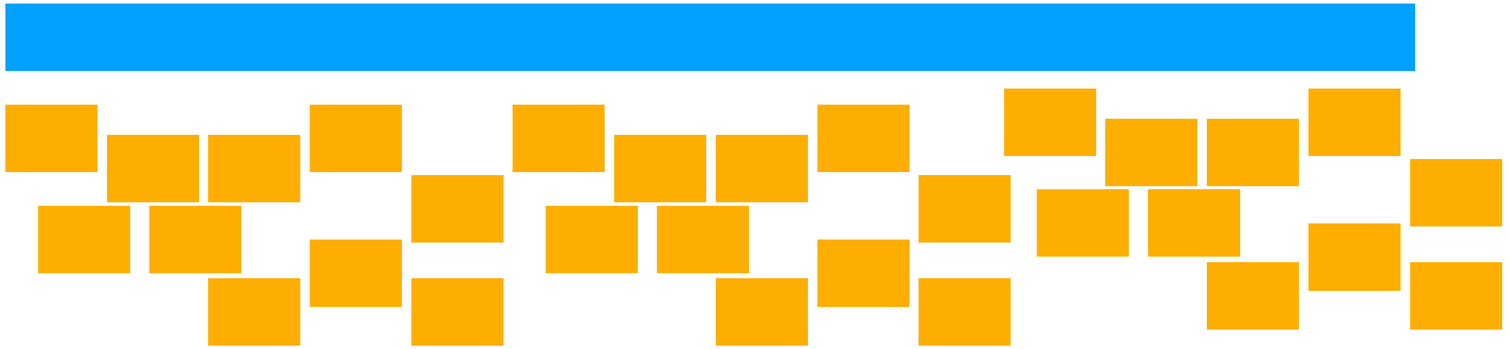
You will be a stronger computational biologist
if you **understand the biology of plants**

Plant genomes are complicated
but **the technology is getting better**

How do we approach genome projects
because **plants are absurd organisms**

How do we apply computational genomics
to **assemble plant genomes**

Long-read shotgun sequencing



Extracted “big” DNA

Fragmented DNA (20-30 kb+)

PacBio SMRTbell library

or Nanopore library

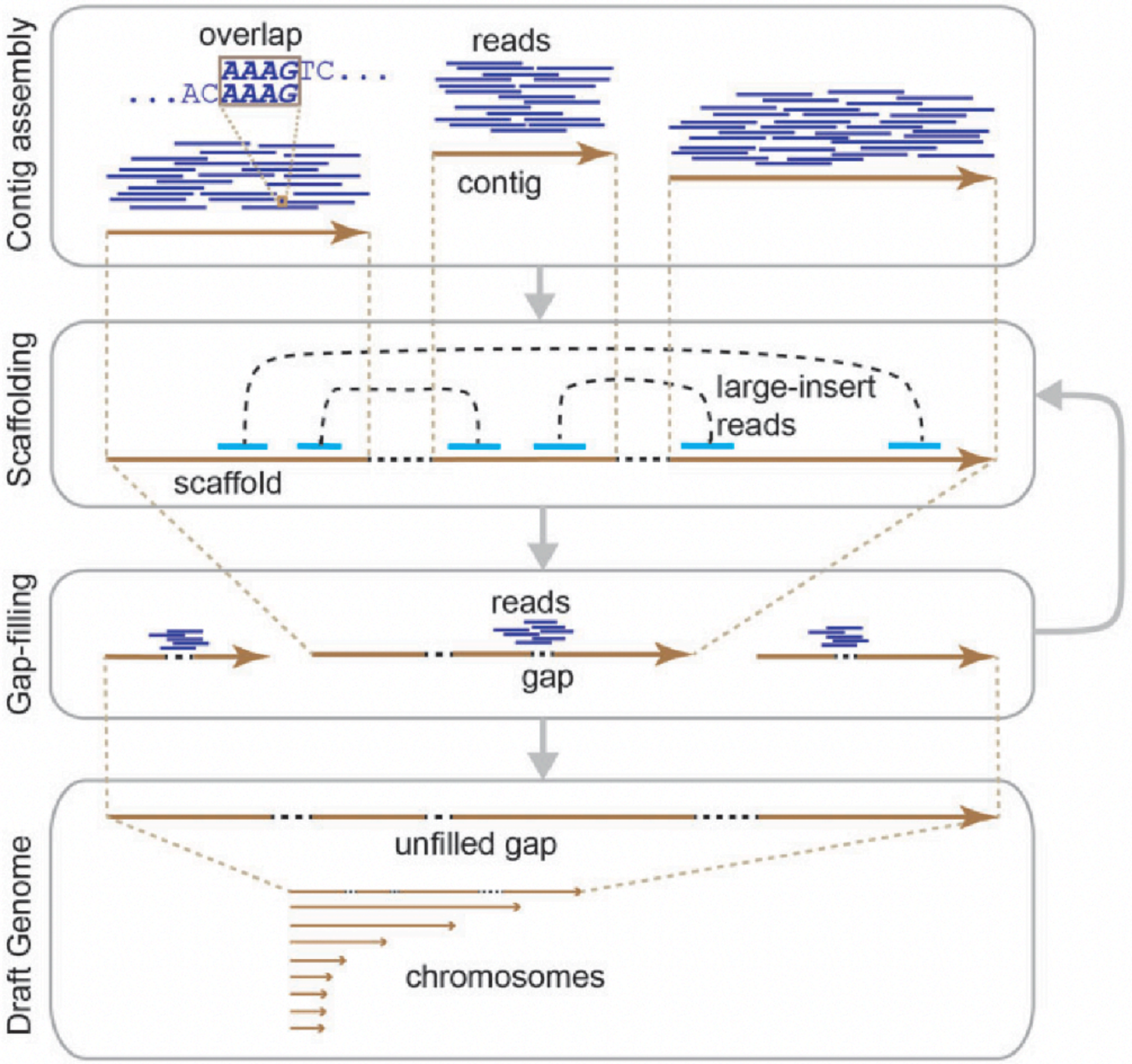
- Produces DNA sequence **reads**
- Based on the haploid genome size, calculate coverage of reads
    -

Long-read shotgun sequencing

- Fragment DNA randomly into pieces
- Then sequence on long-read (e.g. PacBio, Nanopore)
    - or short-read platform (e.g. Illumina)

Basic recipe for genome assembly

- **Find overlaps between the reads**
    - Depending how many reads we sequence, this adds up
- **Build a graph**
    - A picture of the connections between reads
- **Simplify the graph**
    - Sequencing errors complicate this
- **Traverse the graph**
    - Find the best path through the graph to build a consensus assembly


Contig assembly
overlap
AAAGTC...
...ACAAAG
reads
contig
Scaffolding
large-insert
reads
scaffold
Gap-filling
reads
gap
Draft Genome
unfilled gap
chromosomes

de Bruijn vs OLC

OLC

Overlap-Layout-Consensus

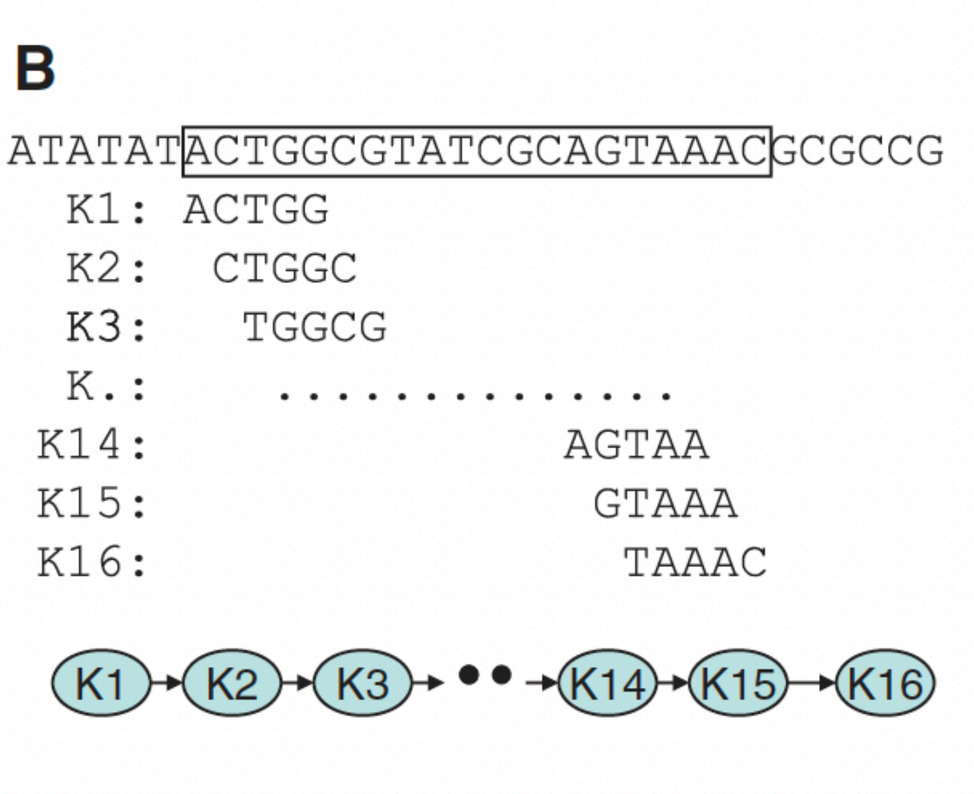
All-by-all pairwise read alignment to find overlaps

Construct a read graph where…
Nodes = reads
Links = two reads overlap larger than a cutoff length

Number of nodes = Number of reads

If requiring at least 5 nt overlaps….
R1 (read 1) overlaps with R2, R3, R4, but NOT R5

DBG
de Bruijn Graph

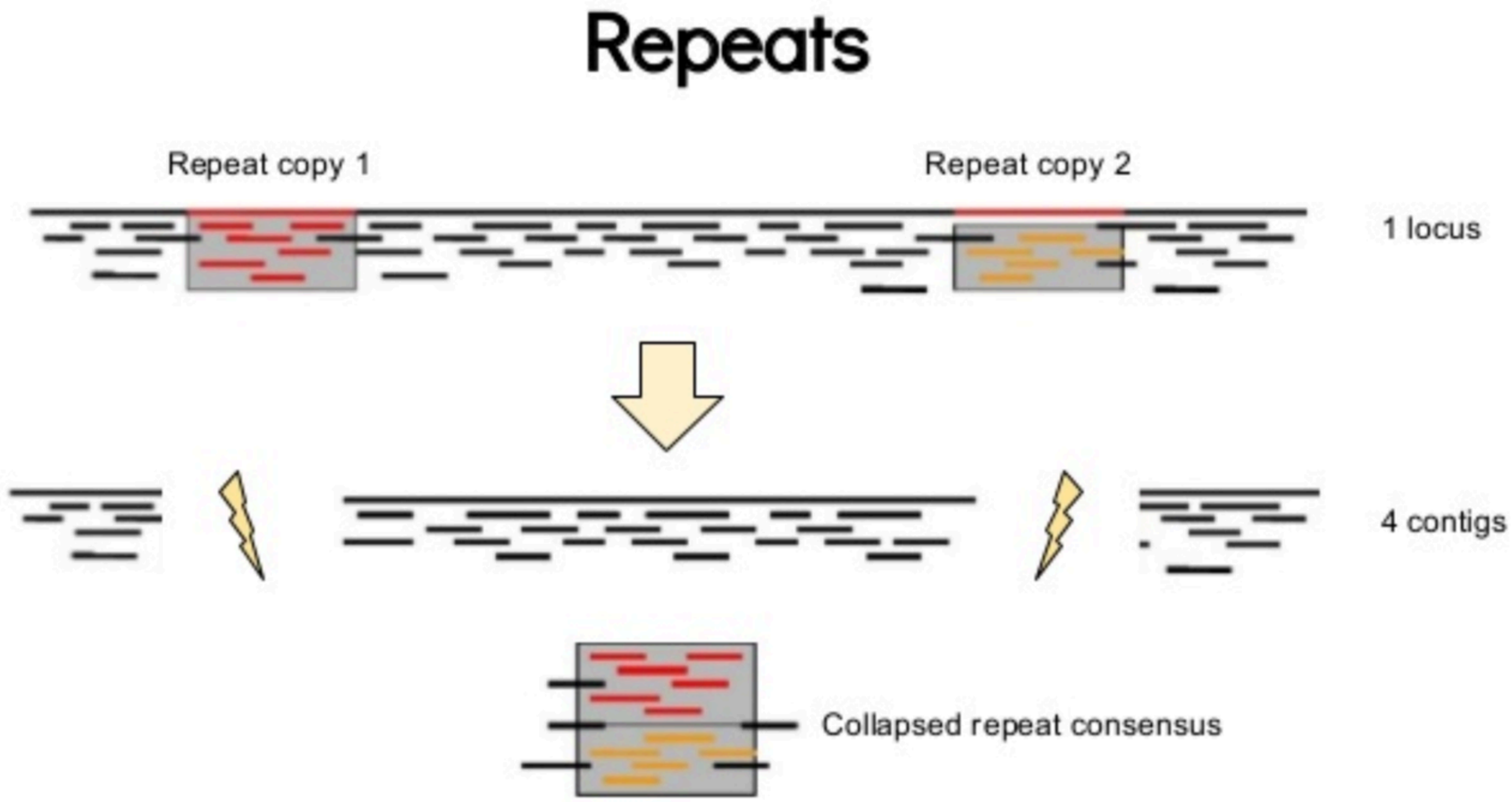


Break reads into **kmers** (e.g. k=5)
kmer = a sequence with length *k*

Construct a read graph where…
Nodes = kmers
Links = two kmers overlap by k-1

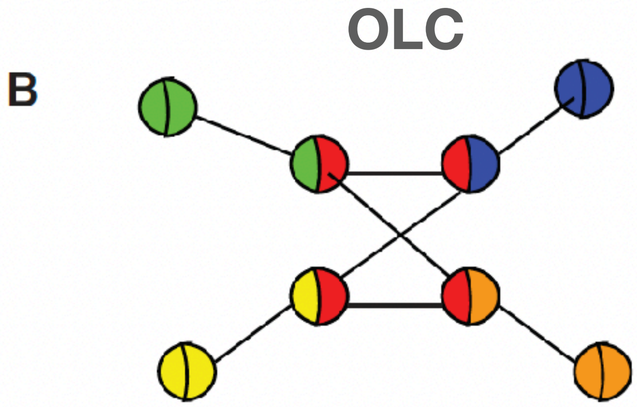
Number of nodes = genome size + errors

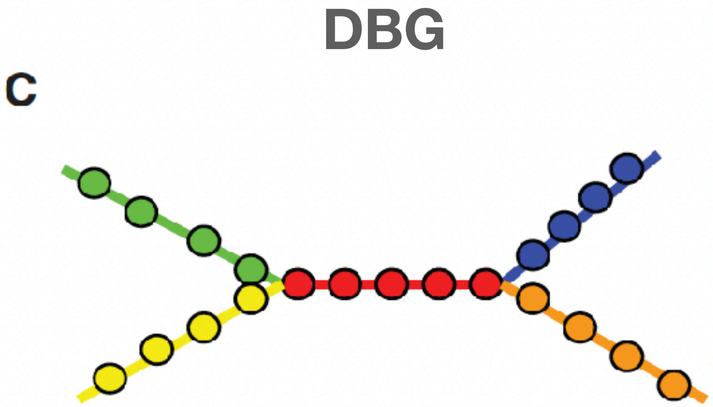
Assembling repeats



OLC vs de Bruijn graph assembly

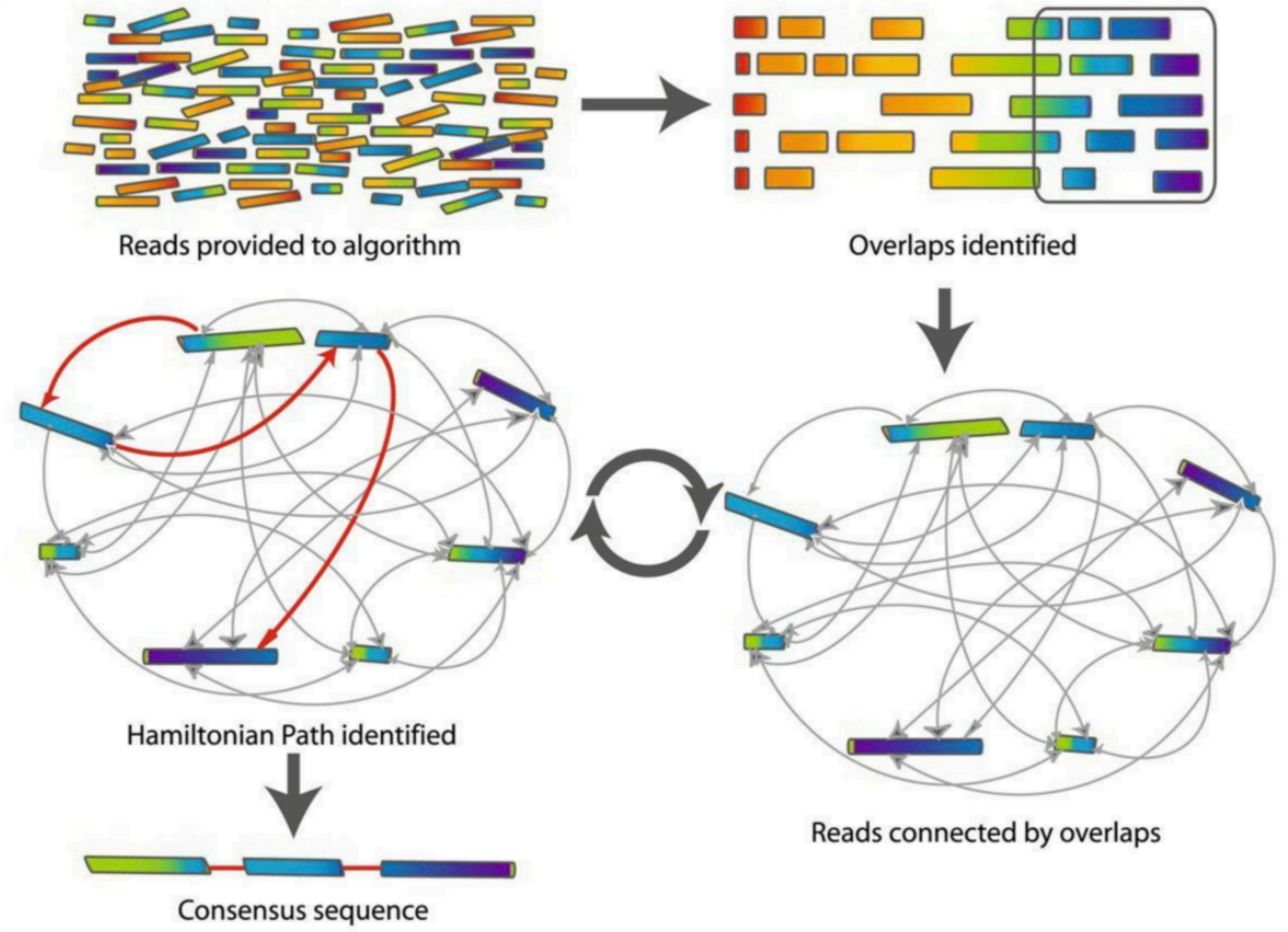
Two regions of the genome, same repeat





Keep increasing read length!

Can try to increase kmer size…


Reads provided to algorithm
Overlaps identified
Hamiltonian Path identified
Reads connected by overlaps
Consensus sequence

What complicates assembly

Biology can be annoying; data isn't perfect

- Raw data error rates
  - Illumina <<< 1%
  - PacBio ~1% or less (for HiFi)
  - Nanopore 3-15% depending
- Repetitive elements
- How much coverage did you sequence?
- Heterozygosity
- Ploidy and flavor of ploidy
- Genome size

What complicates assembly

Repetitive Elements