



# Designing a genome sequencing project

Fay-Wei Li  
Boyce Thompson Institute &  
Cornell University

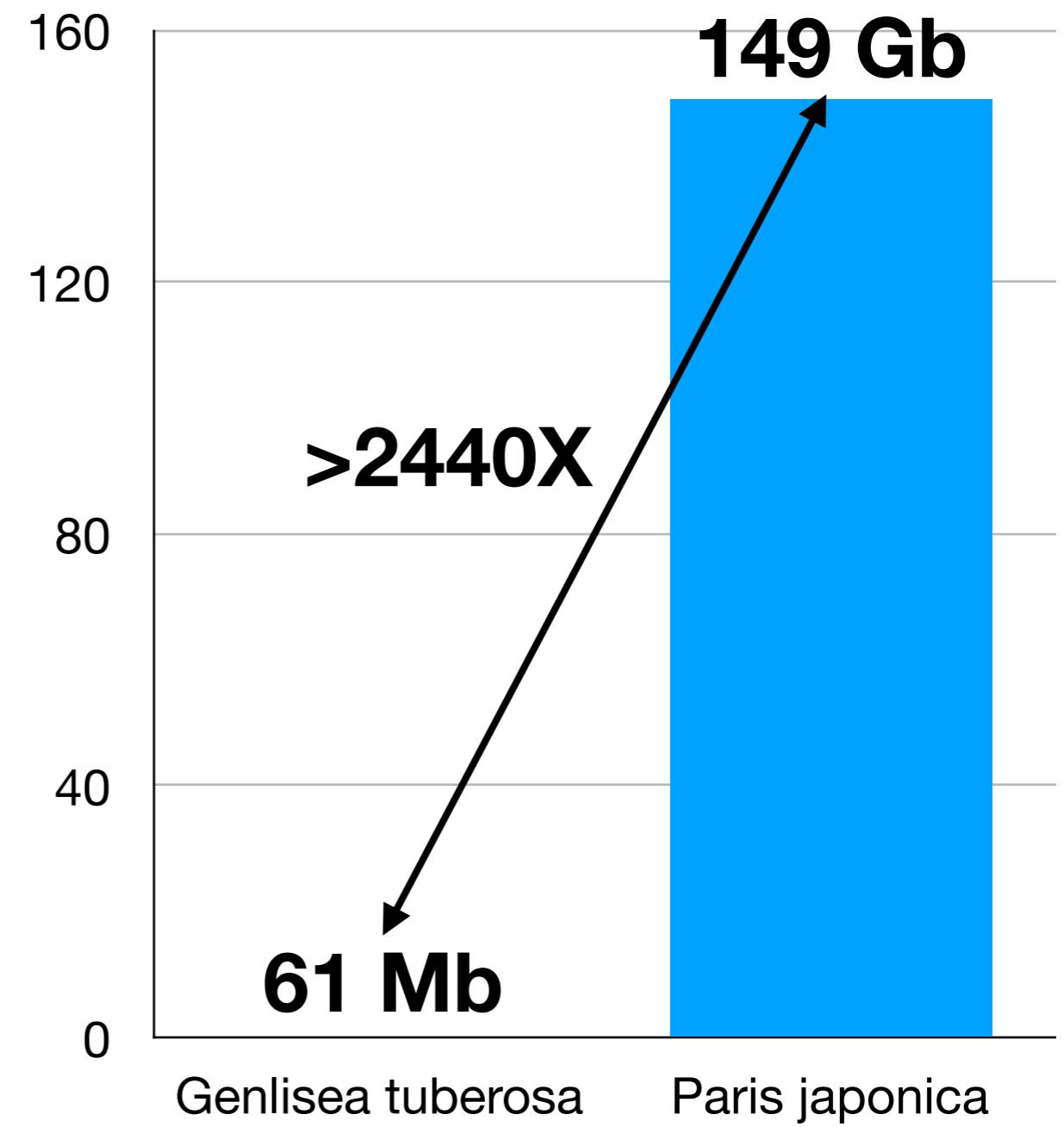
# Outline

- Select your species and individual
- Choose the sequencing platforms
- Get DNA and sequence it
- Example

# Outline

- **Select your species and individual**
- Choose the sequencing platforms
- Get DNA and sequence it
- Example

# Not all plants are equally sequenceable (yet)





*Ophioglossum reticulatum*  
 $2n = 1440$



*Xanthisma gracile*  
 $2n = 4$

**KNOW THE  
GENOME SIZE AND  
PLODY!!**

# Plant DNA C-values Database

Home

Home

Introduction

Search

All Plant C-values

Angiosperm C-values

Gymnosperm C-values

Pteridophyte C-values

Bryophyte C-values

Algal C-values

Release History

Related Databases

Contacts

## Plant DNA C-values Database

Release 7.1, April 2019. Leitch IJ, Johnston E, Pellicer J, Hidalgo O, Bennett MD

<https://cvalues.science.kew.org/>

All Plants



Angiosperm



Gymnosperm



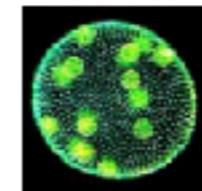
Pteridophyte



Bryophyte



Algae



The DNA amount in the unreplicated gametic nucleus of an organism is referred to as its C-value, irrespective of the ploidy level of the taxon. The Plant DNA C-values Database currently contains C-value data for 12,273 species comprising 10,770 angiosperms, 421 gymnosperms, 303 pteridophytes (246 ferns and fern allies and 57 lycophytes), 334 bryophytes, and 445 algae.

If you have comments and/or suggestions contact [dnac-value@kew.org](mailto:dnac-value@kew.org)

Search

Results

## Search Results (showing results 1 to 9 of 9)

1

## Summary Statistics

	Mean	Min	Max	Std Dev
1C (pg)	59.51	31.21	152.23	33.72
Genus	Species	Subspecies	DNA Amount	Original Reference

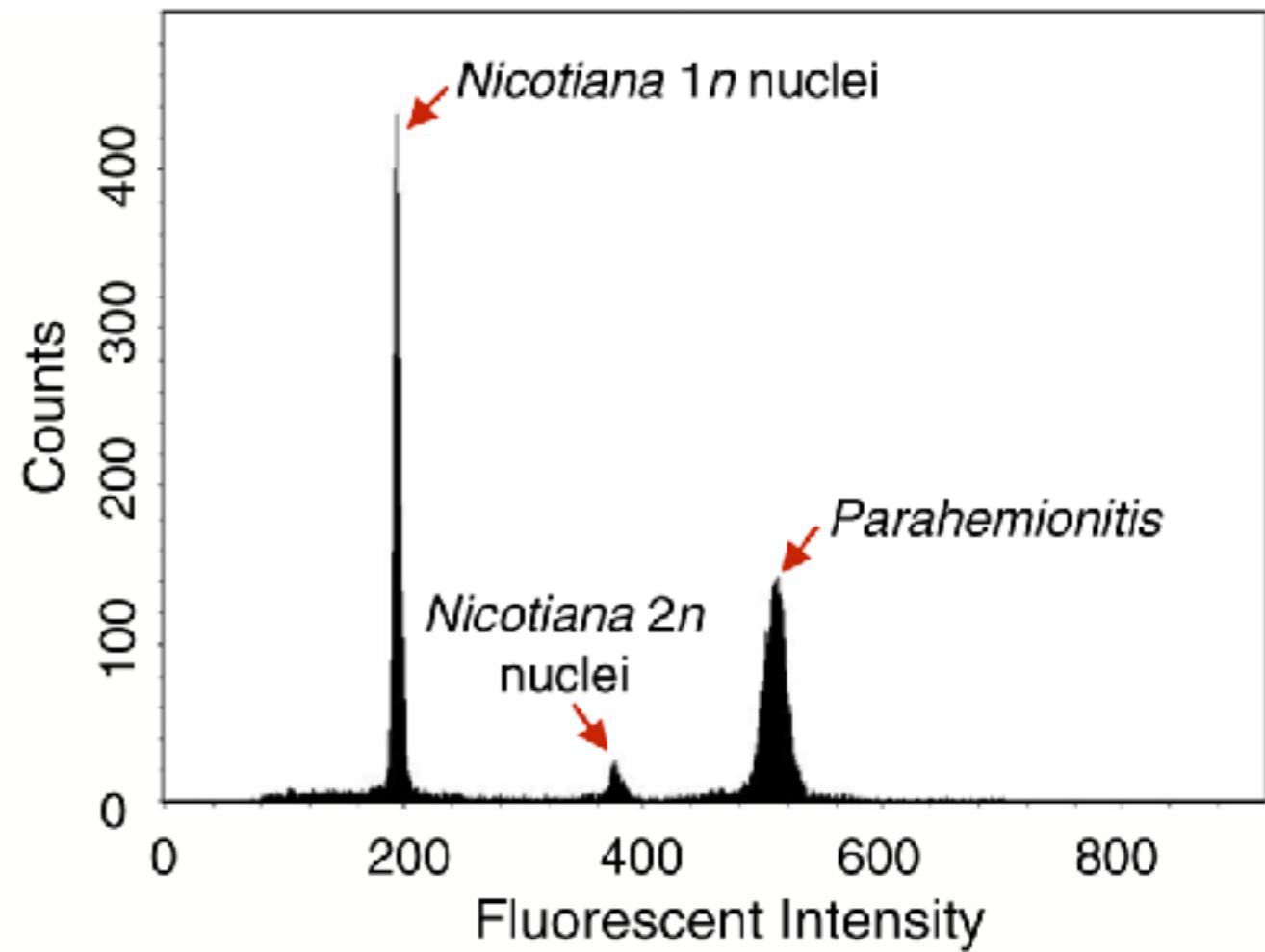
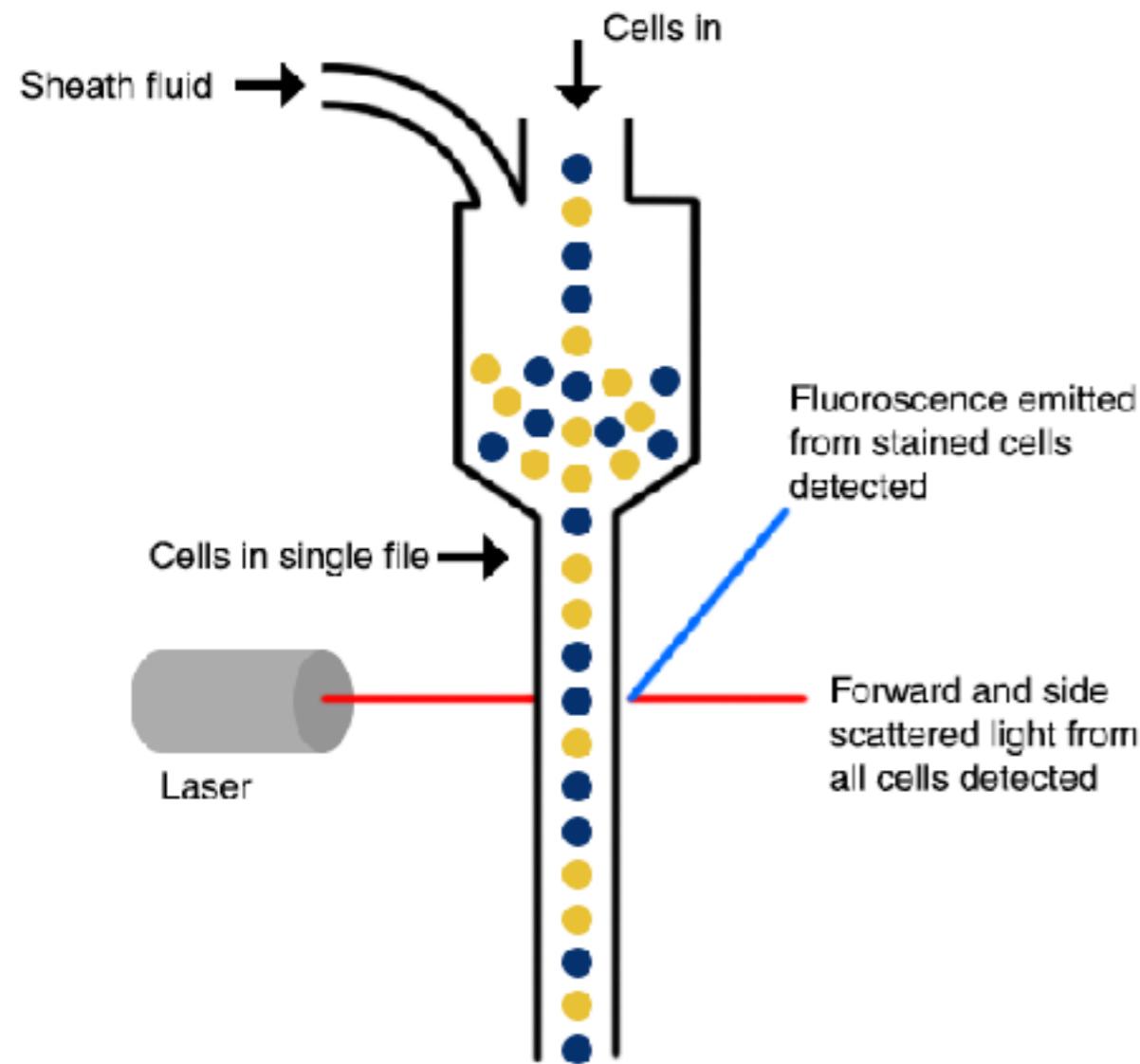
Paris	verticillata		31.21	Pellicer et al.,2014 
Paris	tetraphylla		40.75	Pellicer et al.,2014 
Paris	incompleta		42.25	Pellicer et al.,2014 
Paris	quadrifolia		50.52	Pellicer et al.,2014 
Paris	thibetica	var. thibetica	52.52	Pellicer et al.,2014 
Paris	polyphylla		53.61	Pellicer et al.,2014 
Paris	mairei		55.91	Pellicer et al.,2014 
Paris	forrestii		56.59	Pellicer et al.,2014 
Paris	japonica		152.23	Pellicer et al.,2010 

1C = haploid genome size

1pg = 978 Mbp

# Flow cytometry

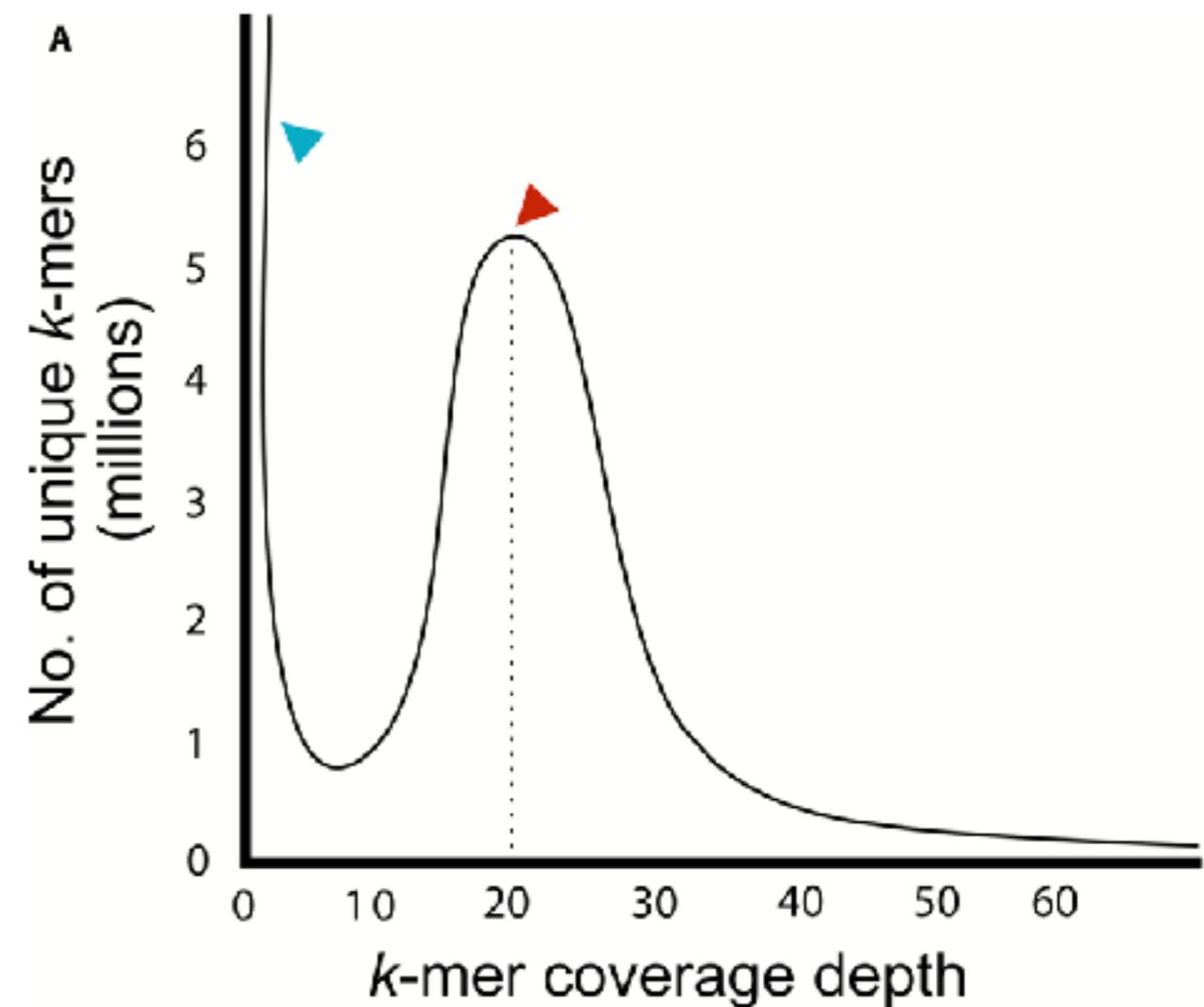
- Isolate nuclei and stain with propidium iodide
- The higher the intensity, the larger the genome



# K-mer frequency plot

**ATGCTAGCTAACTAGACTACTAAGCTAGCAT**  
ATGCTAGCTAACTAGACTAC  
TGCTAGCTAACTAGACTACT  
GCTAGCTAACTAGACTACTA  
CTAGCTAACTAGACTACTAA  
TAGCTAACTAGACTACTAAG  
AGCTAACTAGACTACTAAC  
GCTAACTAGACTACTAACG  
CTAACTAGACTACTAACGTA  
TAACTAGACTACTAACGCTAG  
AACTAGACTACTAACGCTAGC  
ACTAGACTACTAACGCTAGCA  
CTAGACTACTAACGCTAGCAT

$$\text{Coverage} = \frac{\text{Sequencing volume}}{\text{Genome size}}$$



# GenomeScope

Estimate genome heterozygosity, repeat content, and size from sequencing reads using a kmer-based statistical approach.

## Run GenomeScope

Click or drop .histo file here to upload

Description my sample

Kmer length 21

Read length 100

Max kmer coverage 1000

Submit

## Instructions

Upload results from running Jellyfish. Example: inputk21.hist

Instructions for running Jellyfish:

1. Download and install jellyfish from: <http://www.genome.umd.edu/jellyfish.html#Release>
2. Count kmers using jellyfish:

```
$ jellyfish count -C -m 21 -s 1000000000 -t 10 *.fastq -o reads.jf
```

Note you should adjust the memory (-s) and threads (-t) parameter according to your server. This example will use 10 threads and 1GB of RAM. The kmer length (-m) may need to be scaled if you have low coverage or a high error rate. You should always use "canonical kmers" (-C)

3. Export the kmer count histogram

```
$ jellyfish histo -t 10 reads.jf > reads.histo
```

Again the thread count (-t) should be scaled according to your server.

4. Upload reads.histo to GenomeScope

Note: High copy-number DNA such as chloroplasts can confuse the model. Set a max kmer coverage to avoid this. Default is -1 meaning no filter.

# CCDB CHROMOSOME COUNTS DATABASE

Enter a genus or genus and species

SEARCH

[Home](#) [About](#) [Browse](#) [Services](#) [Add new counts](#) [Contact](#)

Prof. Itay Mayrose Lab - Plant Evolution, bioinformatics, & comparative genomics

The **Chromosome Counts Database (CCDB, version 1.47)** is a comprehensive community resource for plant chromosome numbers.

CCDB aims to combine existing data **resources** into an extensive central database that will be updated regularly by the community.

Users and researchers are encouraged to contribute to the accuracy and completeness of the data in CCDB by **submitting new counts**, or **reporting erroneous counts**.

To start browsing for chromosome numbers, use the **Browse** page, or the search box above.

#### Recommended citation:

CCDB is built upon many individual data collection efforts. In addition to the main citation detailed below we recommend citing the major resources where the downloaded data were first assembled.

Main citation: Rice et al. 2015. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytol.* 206(1): 19-26.

#### *Abstract.*

## Browse



News

12/18 8 New counts from [Kemal2009](#)

12/18 13 New counts from [Yildiz2009](#)

12/18 6 New counts from [Minareci2009](#)

12/18 16 New counts from [Yildiz2006](#)

[▼](#) [▲](#)

CCDB → Angiosperms → Melanthiaceae → Paris

46 species in "Paris":

[Show Statistics](#) [Export to CSV](#) [Submit new counts](#) [Report counts](#)

Taxon name	Status	Median (n)
<i>Paris axialis</i> H. Li	Accepted	5
<i>Paris bashanensis</i> F. T. Wang & Tang	Accepted	5
<i>Paris cronquistii</i> (Takht.) H. Li	Accepted	5
<i>Paris daliensis</i> H. Li & V. G. Soukup	Accepted	5
<i>Paris delavayi</i> Franch.	Accepted	5
<i>Paris dulongensis</i> H. Li & Kurita	Accepted	5
<i>Paris dunniana</i> H. Lév.	Accepted	5
<i>Paris fargesii</i> Franch.	Accepted	5
<i>Paris forrestii</i> (Takht.) H. Li	Accepted	5
<i>Paris incompleta</i> M. Bieb.	Accepted	5
<i>Paris japonica</i> (Franch. & Sav.) Franch.	Accepted	20



**INVITED SPECIAL ARTICLE**

*For the Special Issue: Conducting Botanical Research with Limited Resources: Low-Cost Methods in the Plant Sciences*

# A step-by-step protocol for meiotic chromosome counts in flowering plants: A powerful and economical technique revisited

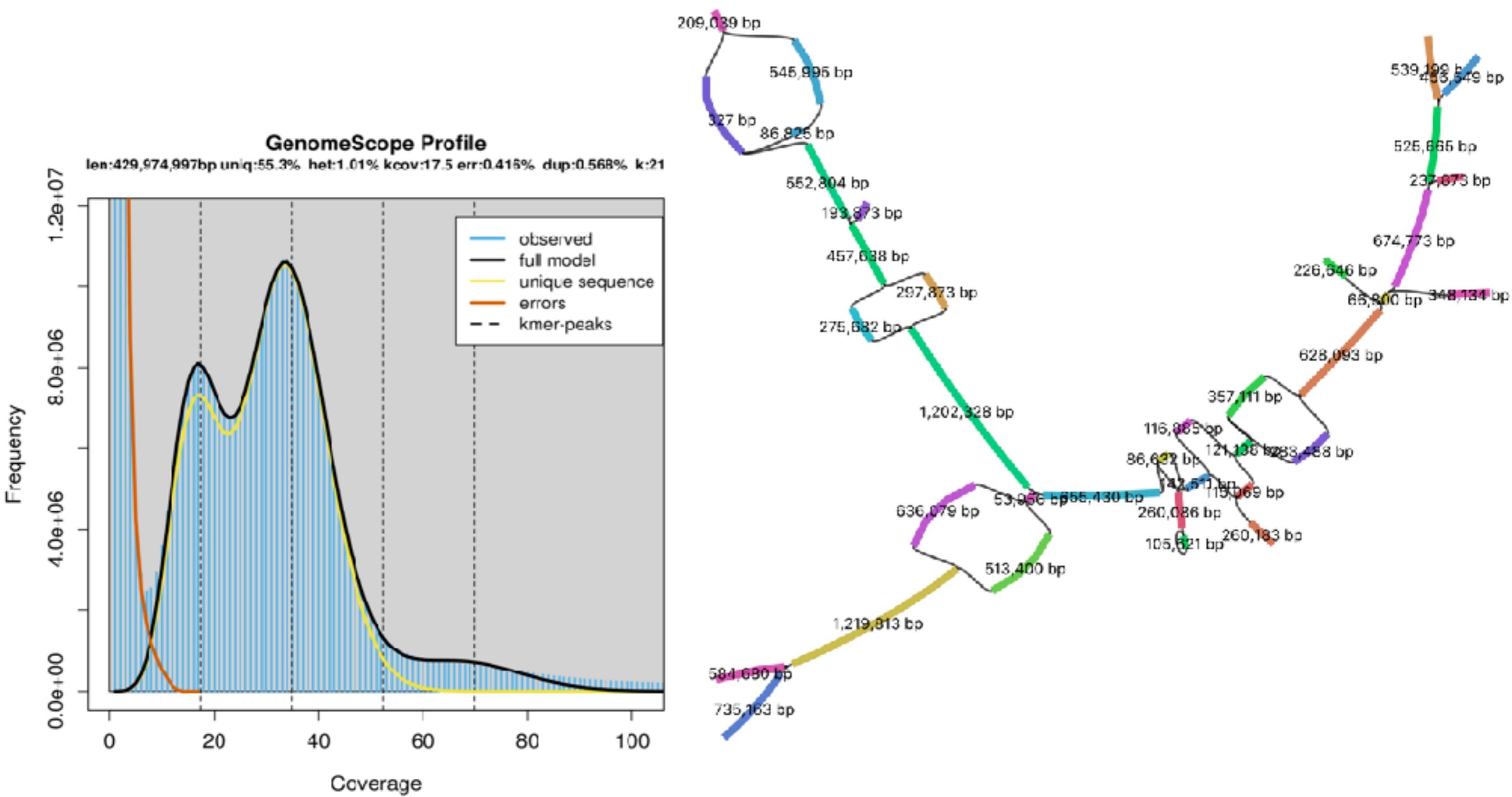
Michael D. Windham<sup>1</sup>, Kathleen M. Pryer<sup>1</sup>, Derick B. Poindexter<sup>2</sup>, Fay-Wei Li<sup>3</sup>, Carl J. Rothfels<sup>4</sup>, and James B. Beck<sup>5,6,7</sup> 

**KNOW THE  
GENOME SIZE AND  
PLODY!!**

# An ideal plant to sequence

- With lots fresh materials
  - Culturable in greenhouse / tissue culture
  - High molecular weight DNA
  - RNA from multiple tissues / developmental stages
  - Fresh tissue for Hi-C library
- Low heterozygosity
  - Self the plant if you have time
- “Clean” - not tangled up with other organisms

# Heterozygosity not good





*Agrostis* NN13

*Agrostis* NN13

*Agrostis* NN13

8102, 220 97256

# Genome survey

- >20X Illumina coverage

- Genome size

- Heterozygosity

- Repeat content

- Contamination

**GenomeScope**  
<http://qb.cshl.edu/genomescope/>

**Centrifuge**  
<https://ccb.jhu.edu/software/centrifuge/manual.shtml>

- Useful for polishing genome assembly

# Outline

- Select your species and individual
- **Choose the sequencing platforms**
- Get DNA and sequence it
- Example

# Illumina short reads cannot give good assemblies

If repeats are longer than read length:

**ATGAGAGAGAGAGA**

# GAGAGAGAGAGA

# GAGAGAGAGAGA

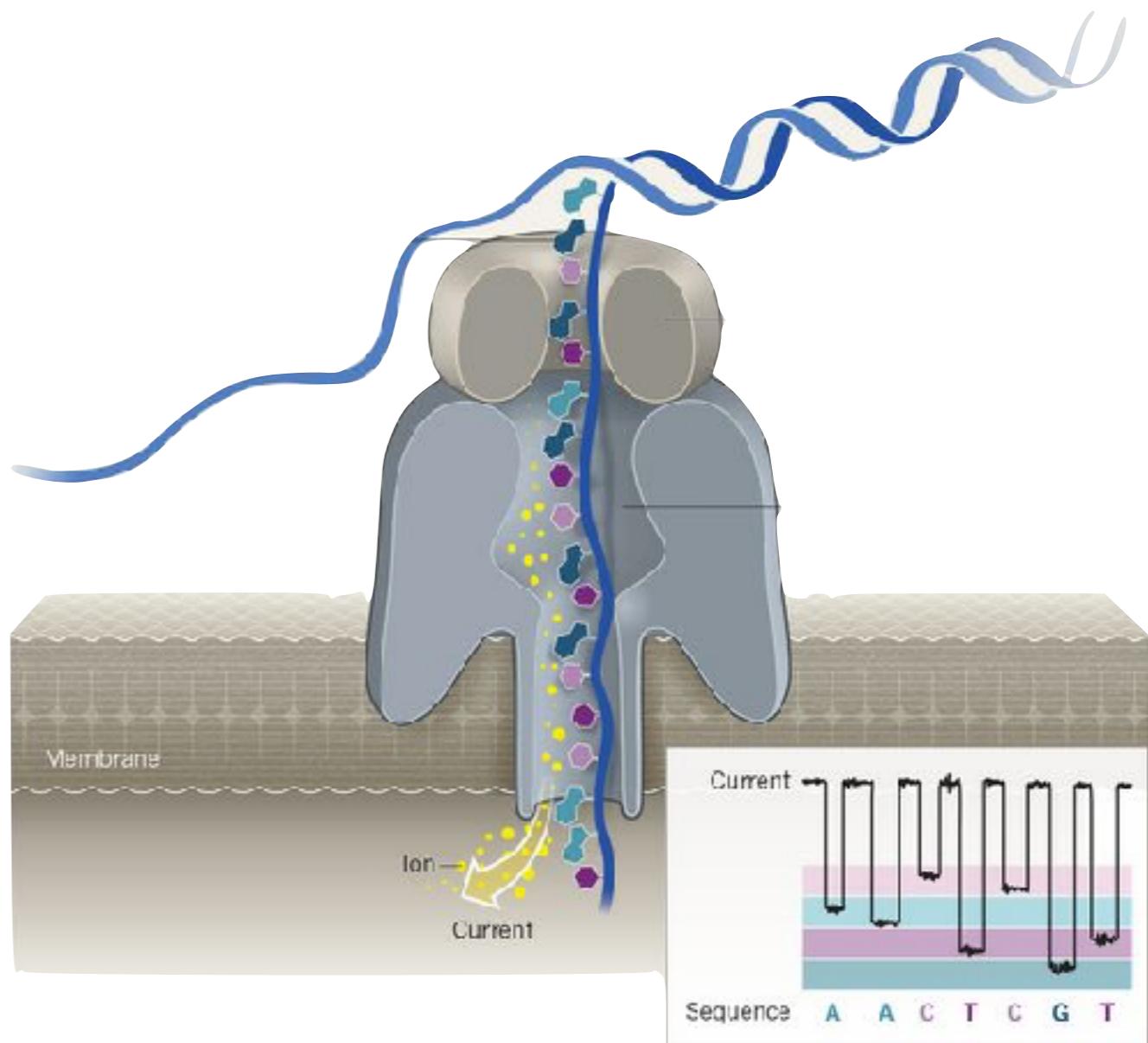
# GAGAGAGAGAGA

GAGAGAGAGAGACTT

# GAGAGAGAGAGA

# Single molecule sequencing

- PacBio, Oxford Nanopore
- Long read length
  - >10,000 bp (vs 150 bp)
  - up to 1-2 Mb
- High error rate
  - 10% (vs <1%)



**MinION**



**GridION**



**SmidgION**

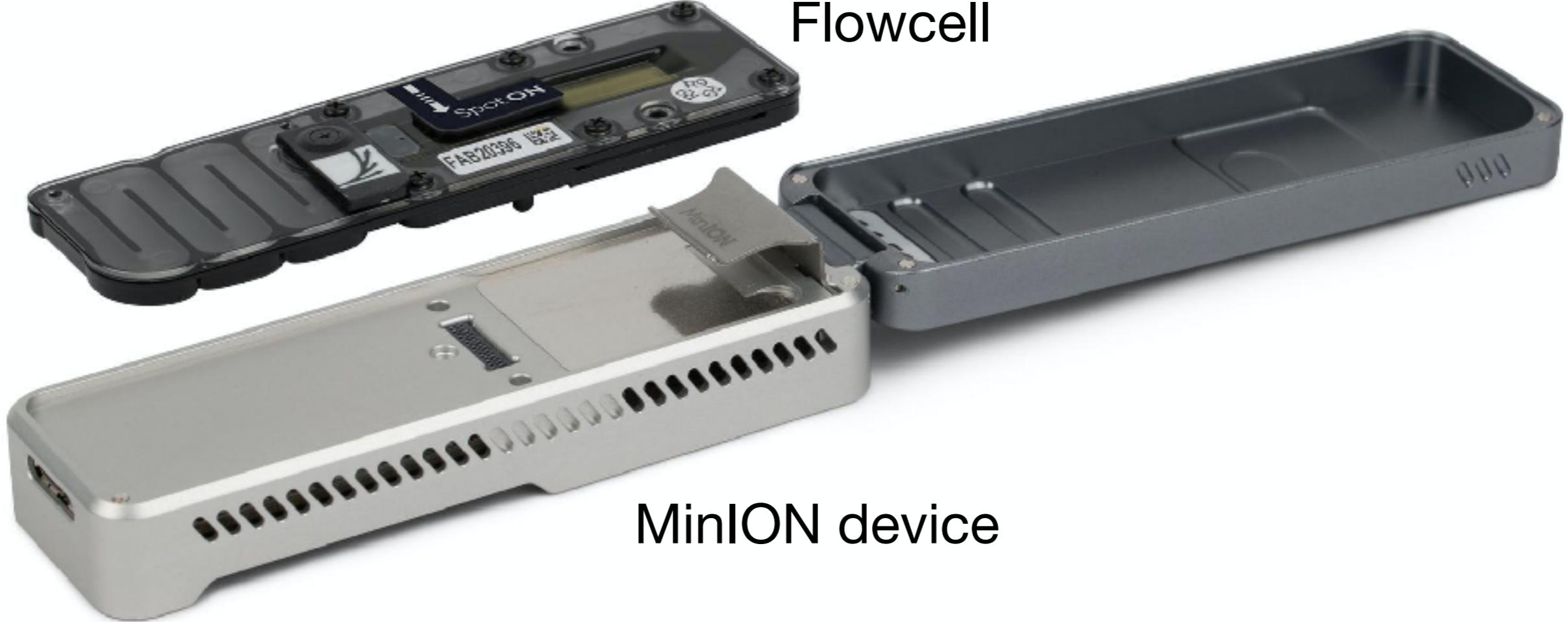


**PromethION**





Flowcell



MinION device

# MinION

- R9 flowcell
  - \$500
  - 10-20 Gb
- Flongle flowcell
  - \$90
  - 1-2 Gb
- Library prep
  - \$150
  - 1hr
  - 1ug DNA
  - (RNA, cDNA, amplicon, ...)
- Starter pack
  - \$1000
  - 2 R9 flowcells
  - 1 library prep kit



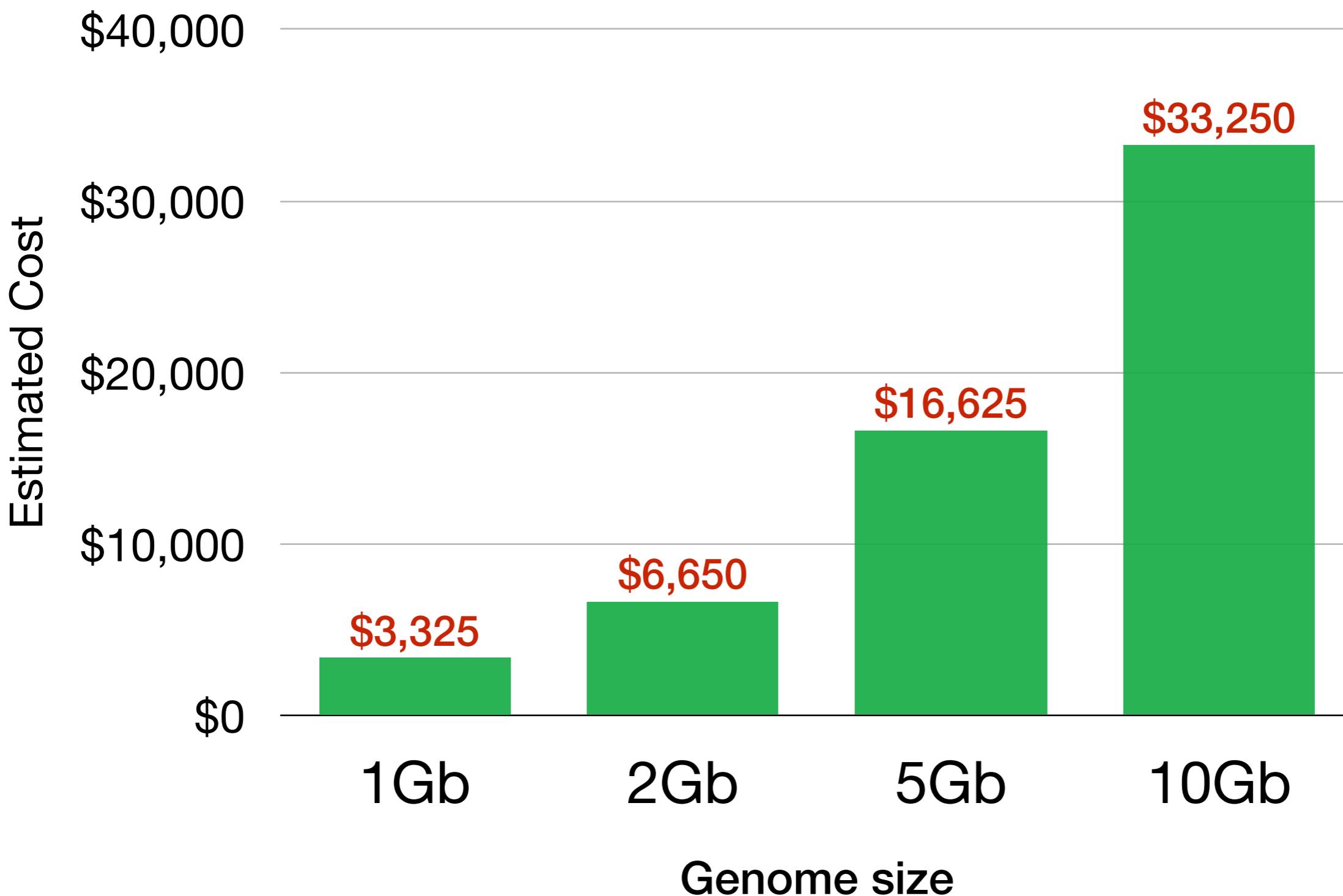
# How much \$\$ for an 1 Gb genome?

- 50X Illumina:
  - $50\text{Gb} \times \$26.5/\text{Gb} = \$1,325$
- 50X nanopore:
  - $50\text{Gb} \times \$40/\text{Gb} = \$2,000$

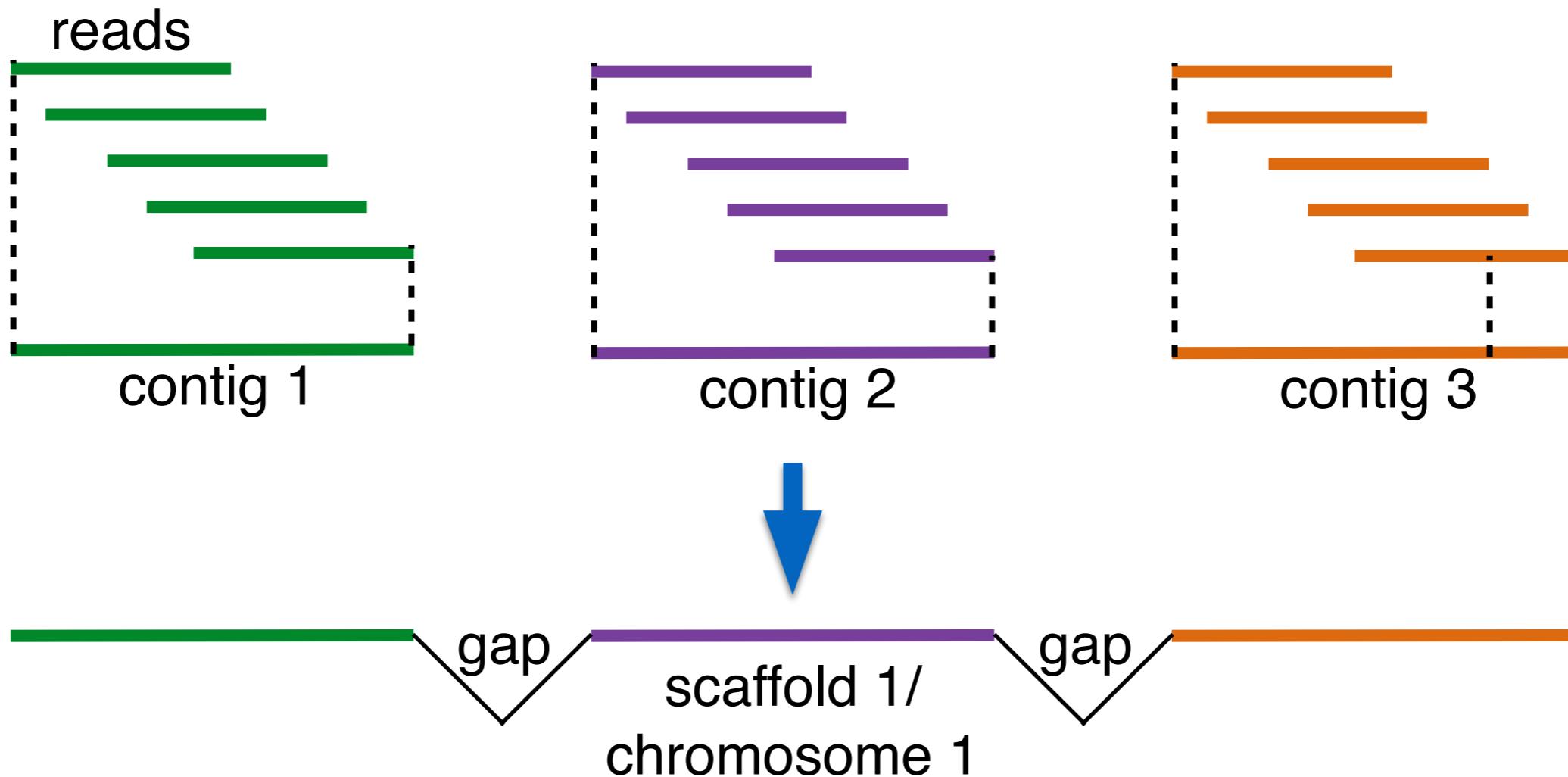
---

**\$3,325**

# How much \$\$?

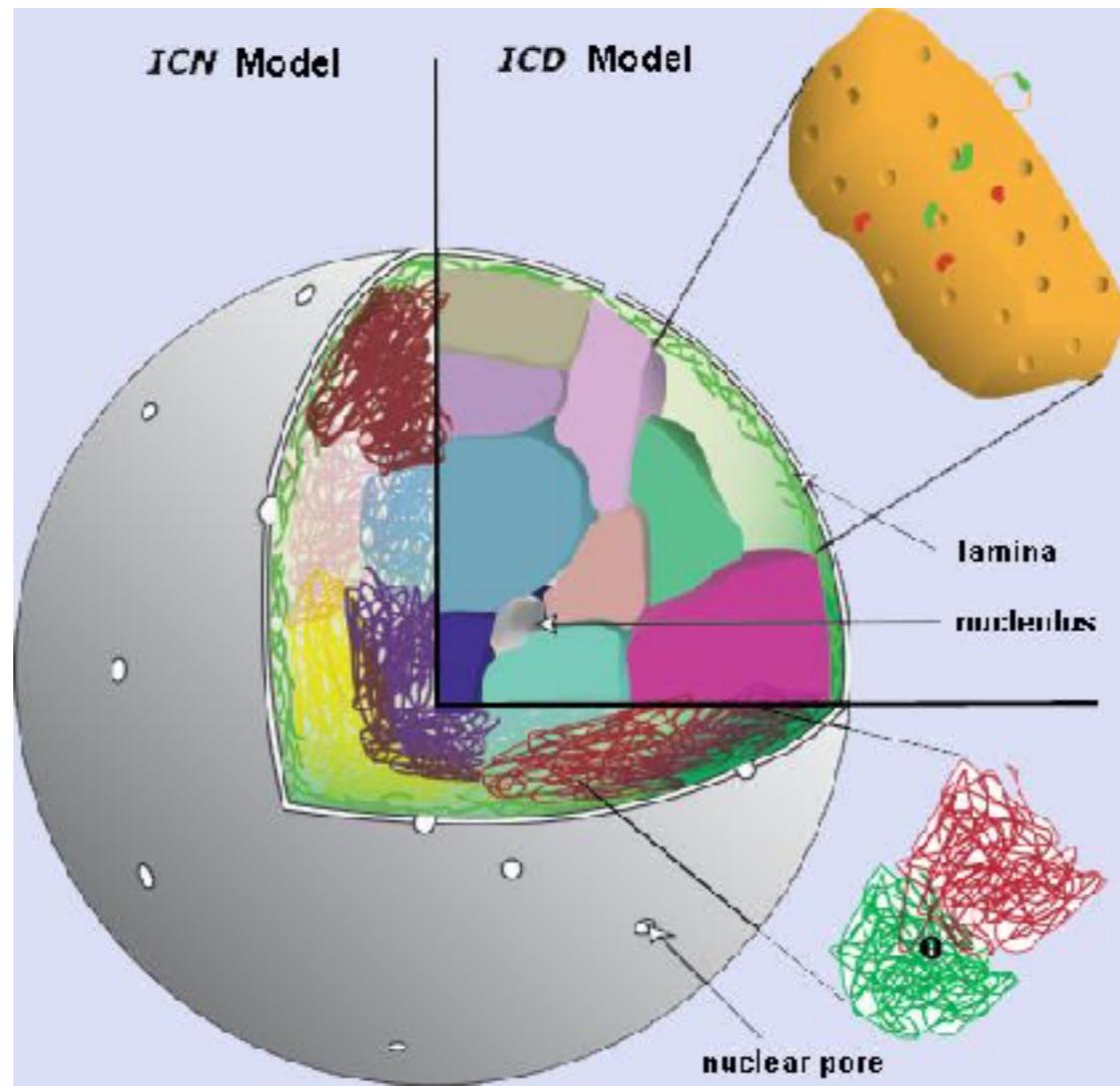


# Reads → Contigs → Scaffolds



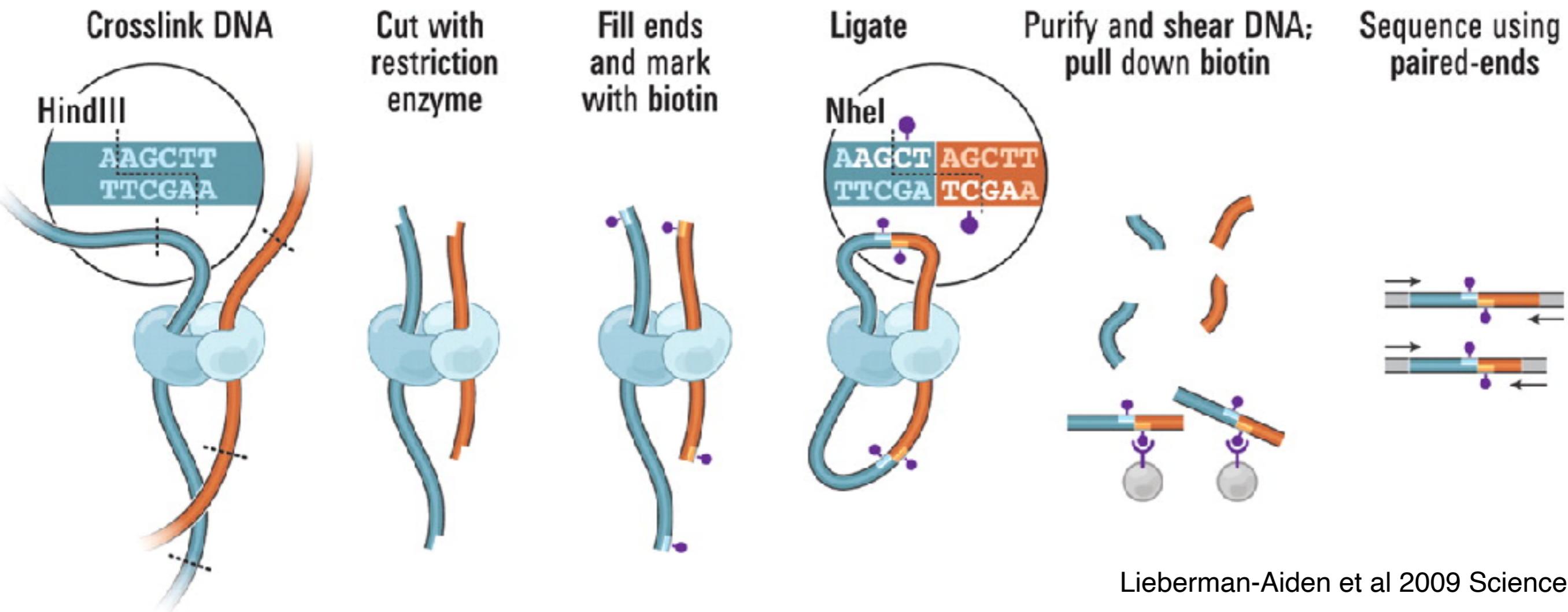
# Hi-C

- Hi-C = high throughput chromatin conformation capture
- DNA of the same chromosome will be ***spatially*** close



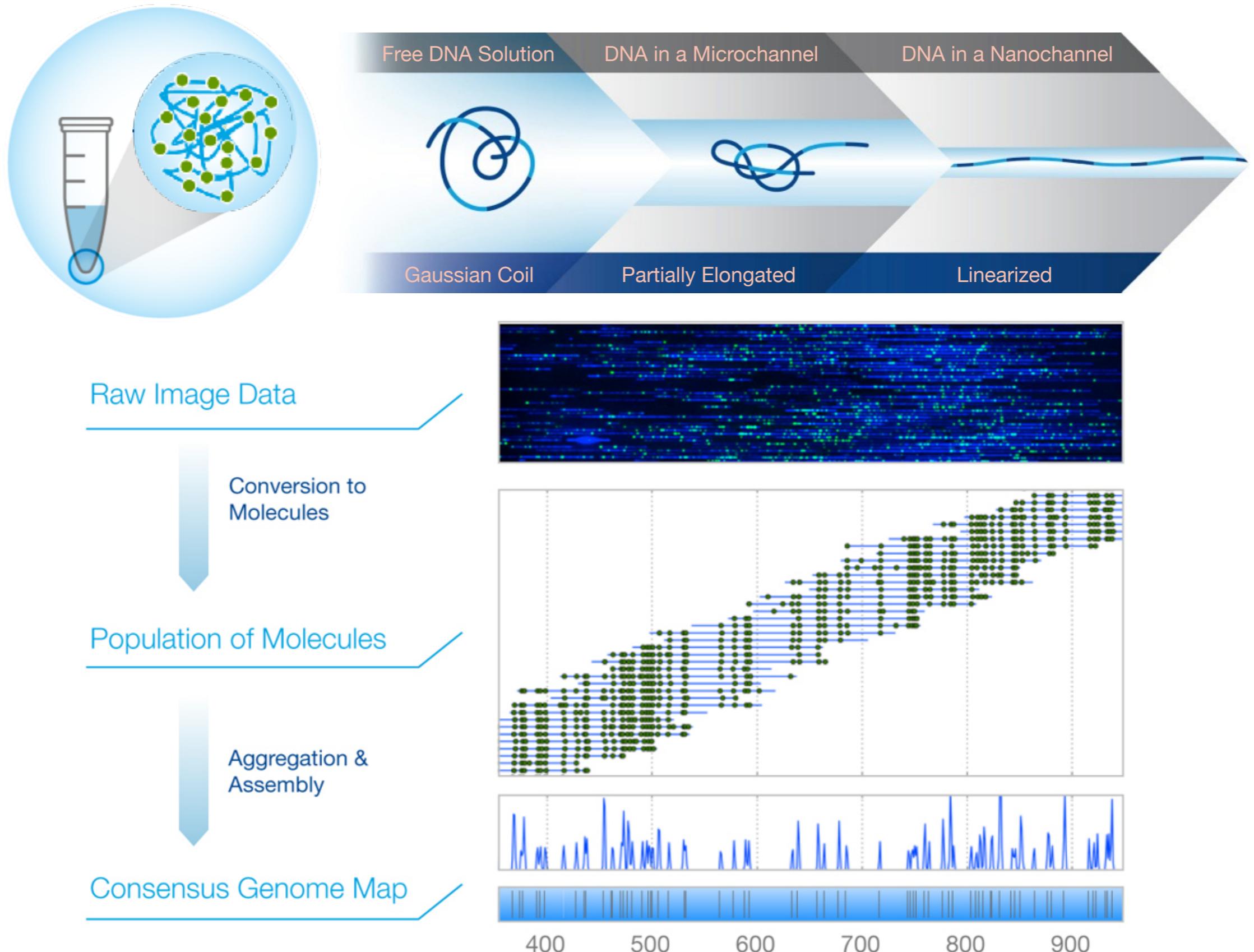
# Hi-C

- Hi-C = high throughput chromatin conformation capture
- DNA of the same chromosome will be ***spatially*** close



Lieberman-Aiden et al 2009 Science

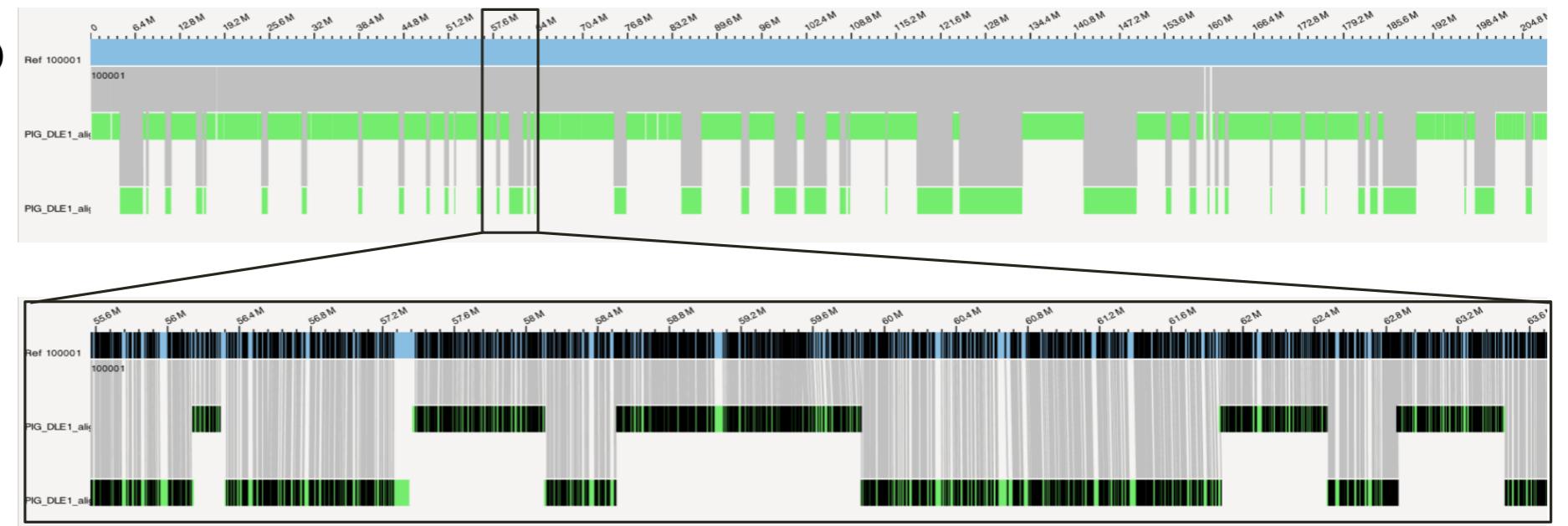
# Optical mapping



# Optical mapping

Use optical map to scaffold contigs

Genome map  
contigs



# How much \$\$ for scaffolding an 1 Gb genome?

- Hi-C
  - Lib prep kit **\$500** + 50X Illumina **\$1,325** = **\$1,825**
- Optical mapping by Bionano
  - Full service (HMW DNA extraction + Saphyr chip + data analysis) = **~\$3,000**

# Outline

- Select your species and individual
- Choose the sequencing platforms
- **Get DNA and sequence it on nanopore**
- Example

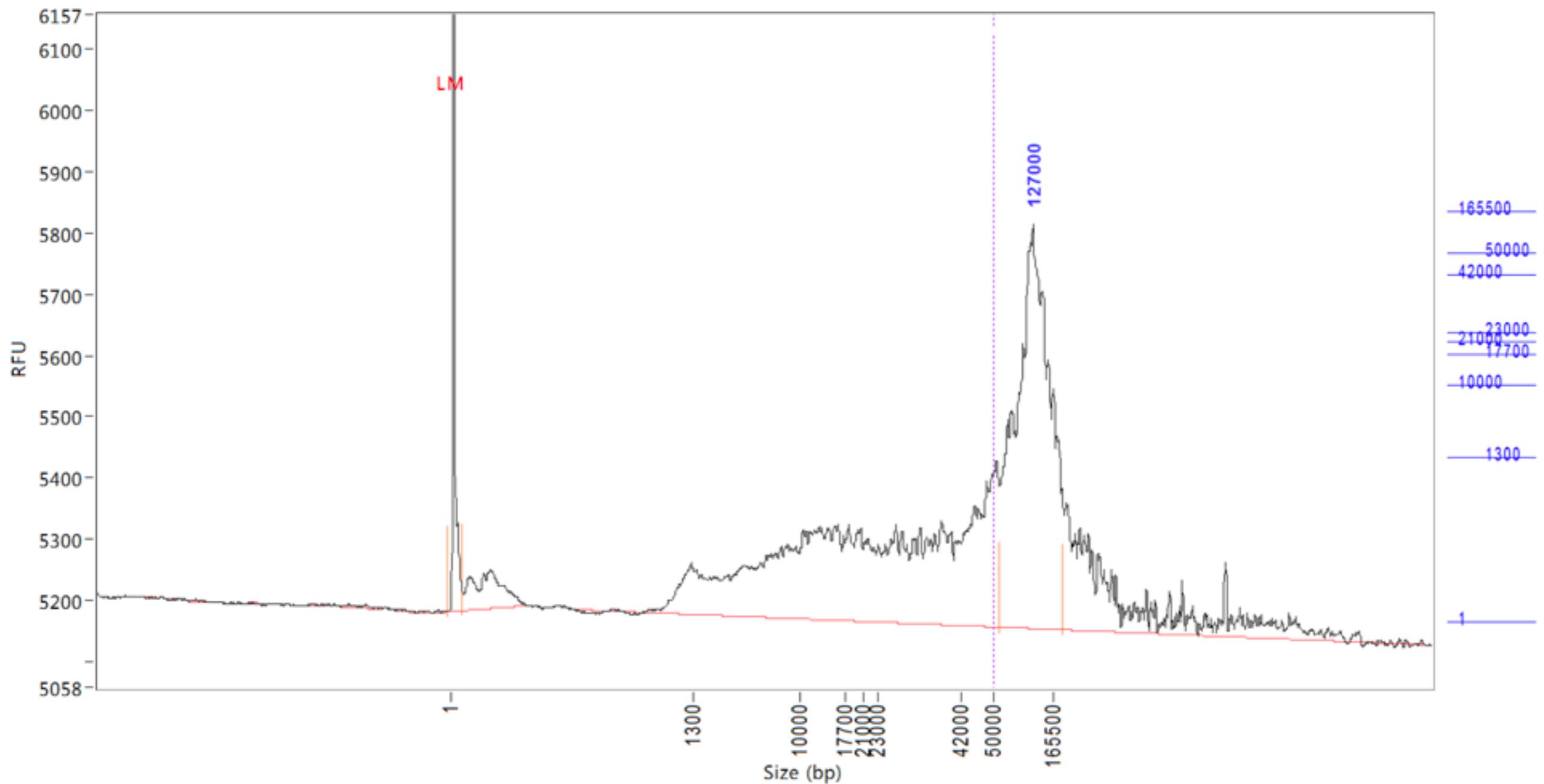
# High molecular weight DNA

- Nuclei prep
- Kits (MagAttract, Nanobind, ...)
- “Slow” CTAB
  - Young leaves
  - Gentle in EVERY step
  - No vortexing
  - Wide-bore tips (cut tips)

# QC HMW DNA

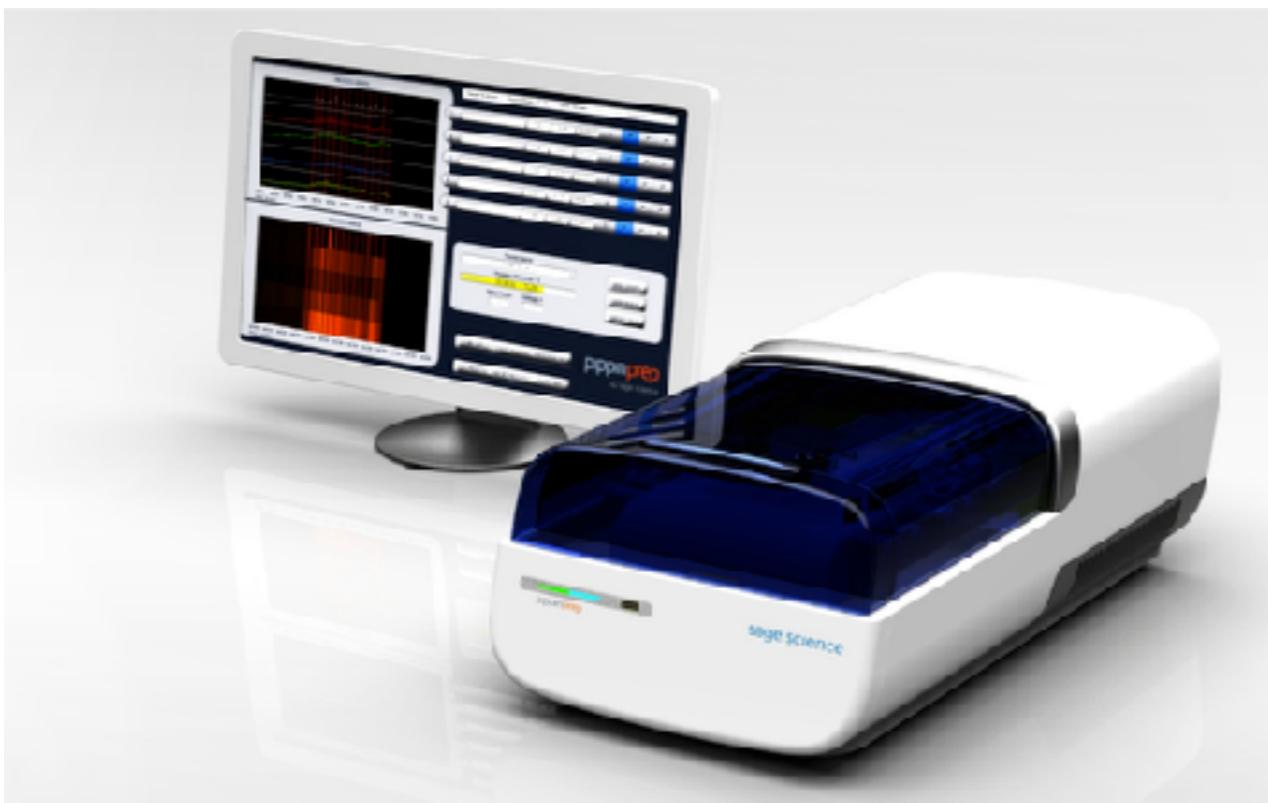
- **Quantification** - Qubit (never trust nanodrop conc!)
  - 1ug min. each nanopore run, ~5ug if size selection
- **Purity** - nanodrop
  - 260/280, 260/230 close to 2
- **Integrity**
  - Pulse-field gel electrophoresis
  - Bioanalyzer, Femto Pulse
  - Low percentage (0.5%) agarose gel with NEB 1kb Extend DNA Ladder (highest band @48.5 kb)

## “Slow” CTAB on *Aconitum noveboracense* -> Femto Pulse QC



# Size selection

- Blue pippin
- Circulomics Short Read Eliminator kit
  - XS (10kb), Regular (25kb), XL (40kb)
- Some DNA will be lost



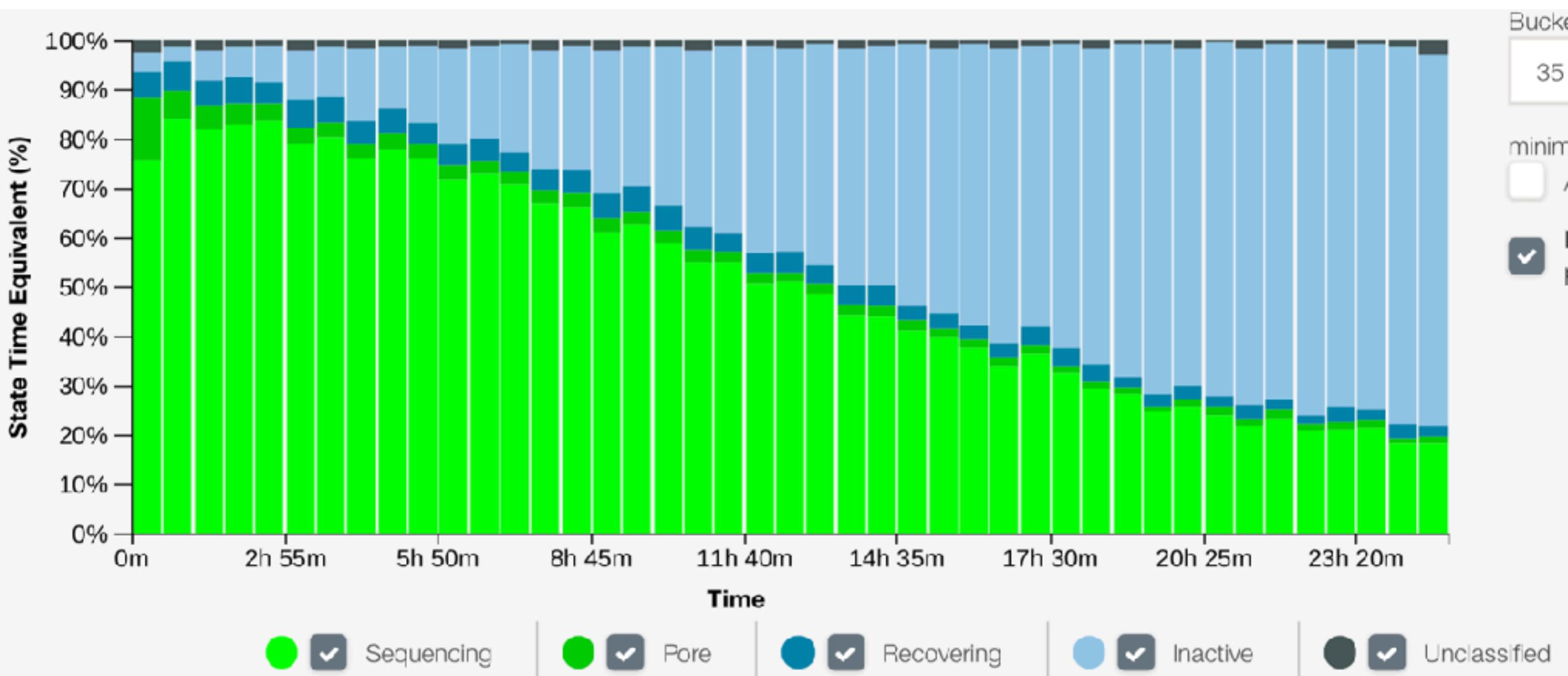
# Nanopore library prep kits

For genomic DNA:

- Ligation Sequencing Kit
  - 60 min prep time
  - Highest yield
- Rapid Sequencing Kit
  - 10 min prep time
  - Lower yield
- Field Sequencing Kit

# Nuclease flush

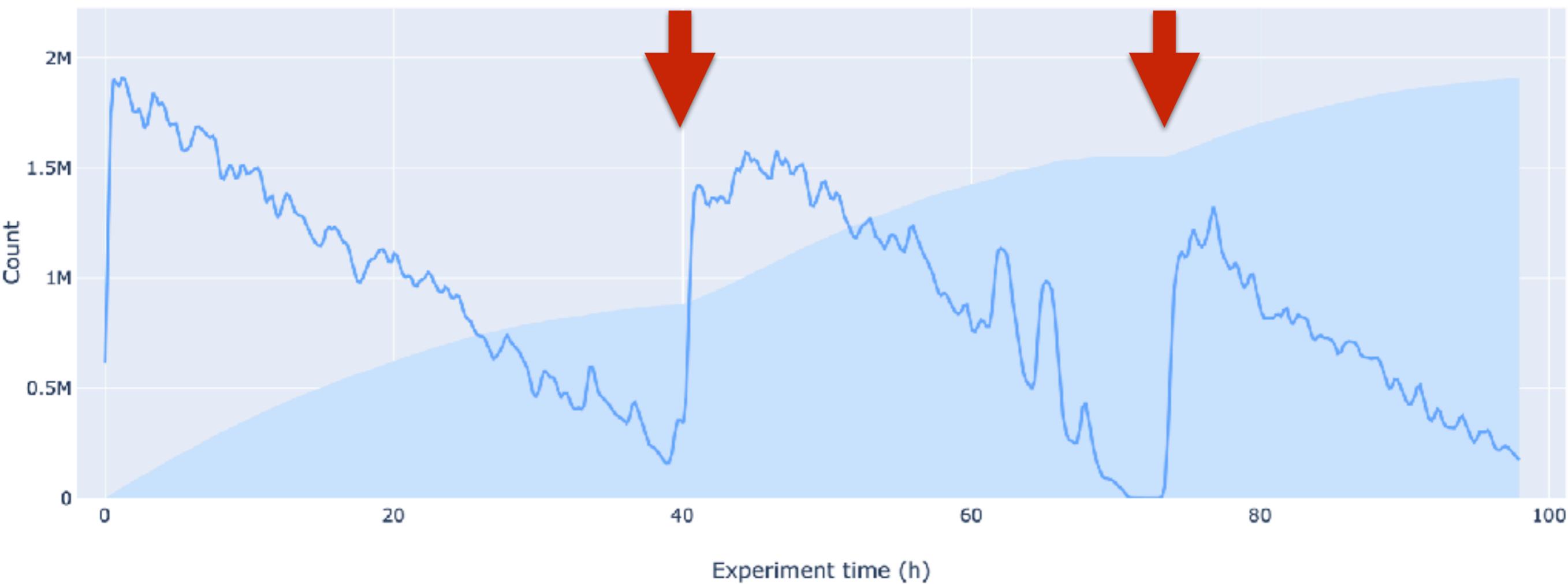
- Pore blockage through time
  - Nuclease flush revives dead pores



# Nuclease flush

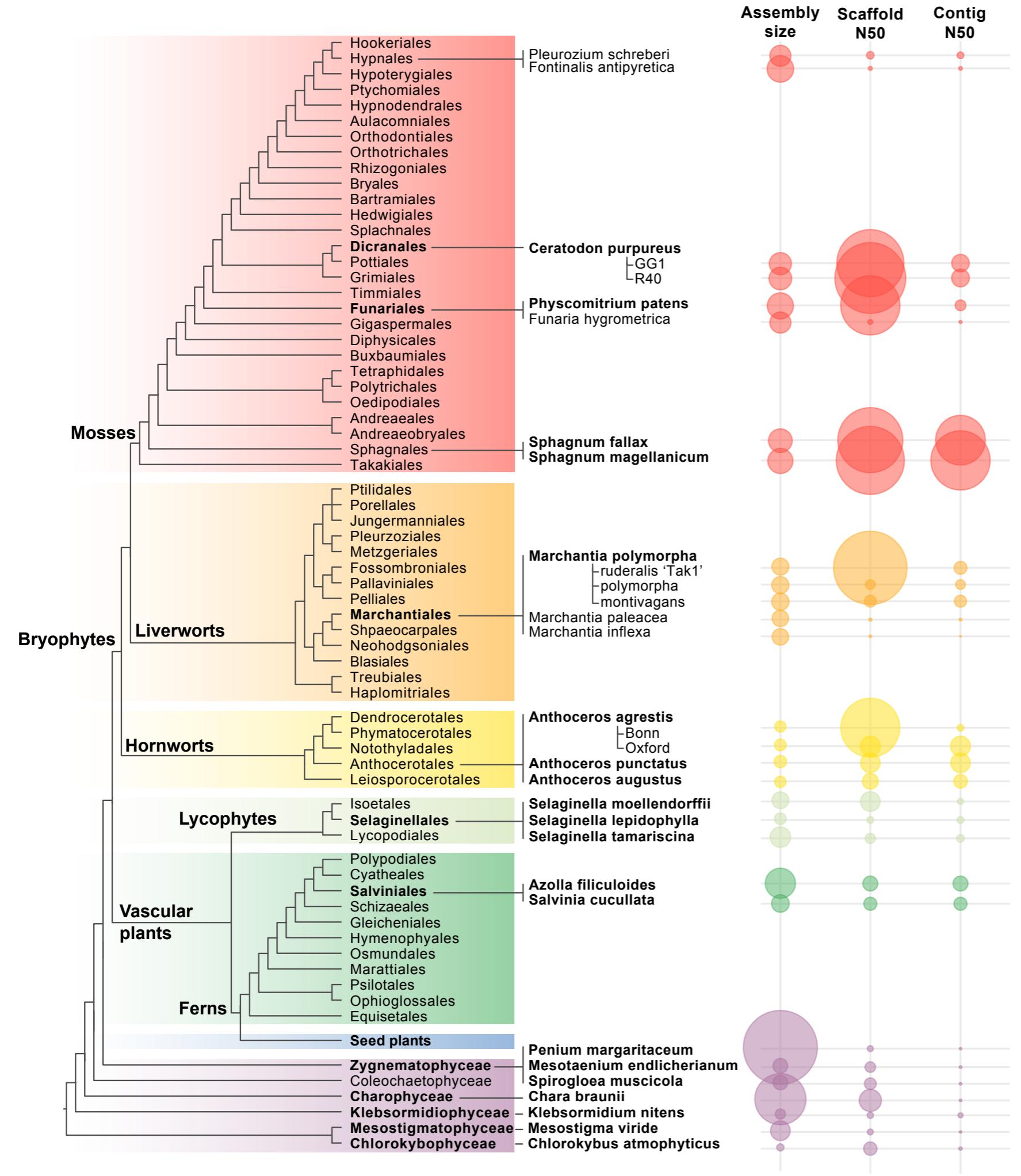
- Pore blockage through time
- Nuclease flush *revitalize* dead pores

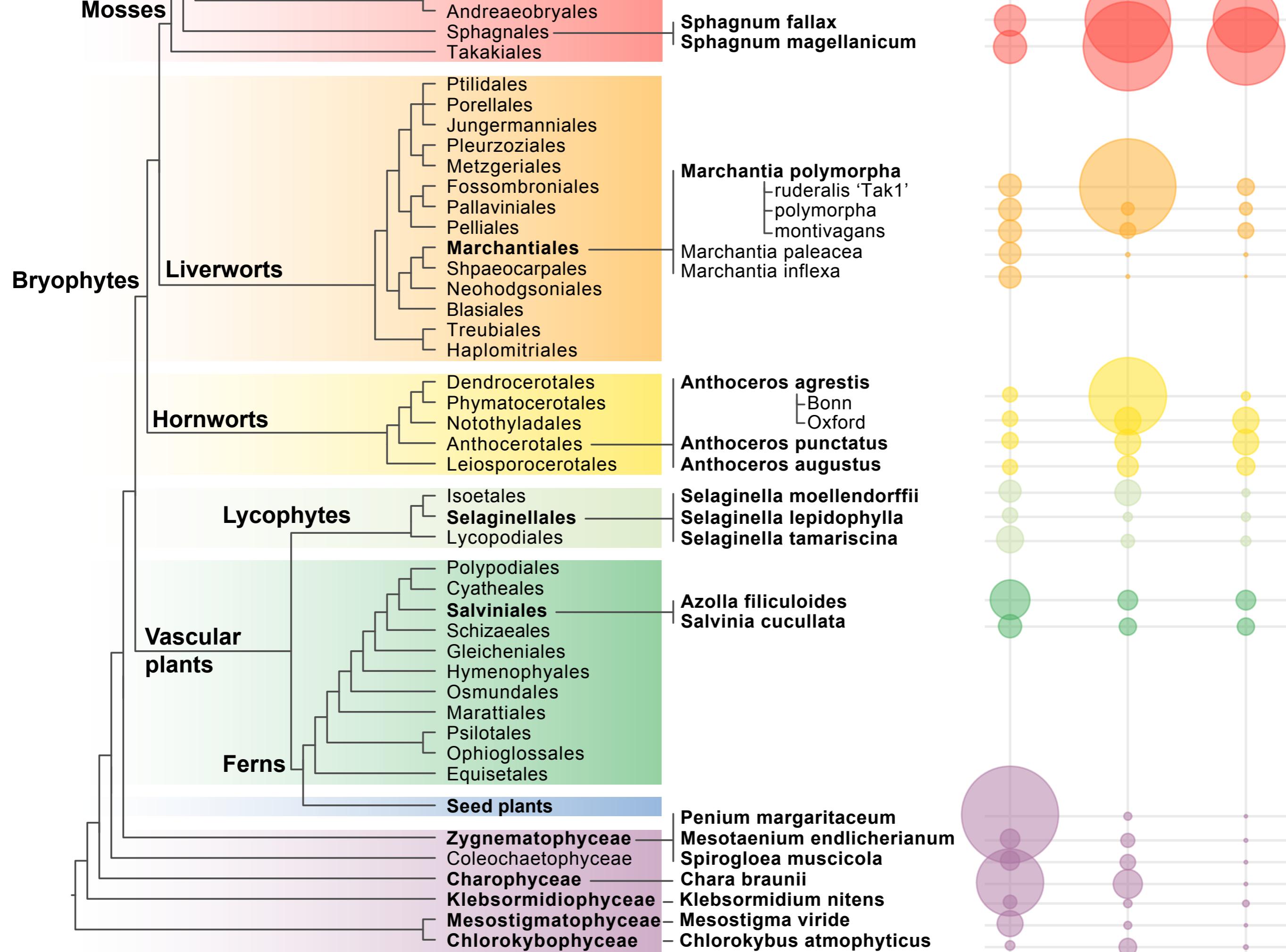
#reads sequenced through time



# Outline

- Select your species and individual
- Choose the sequencing platforms
- Get DNA and sequence it
- **Example**







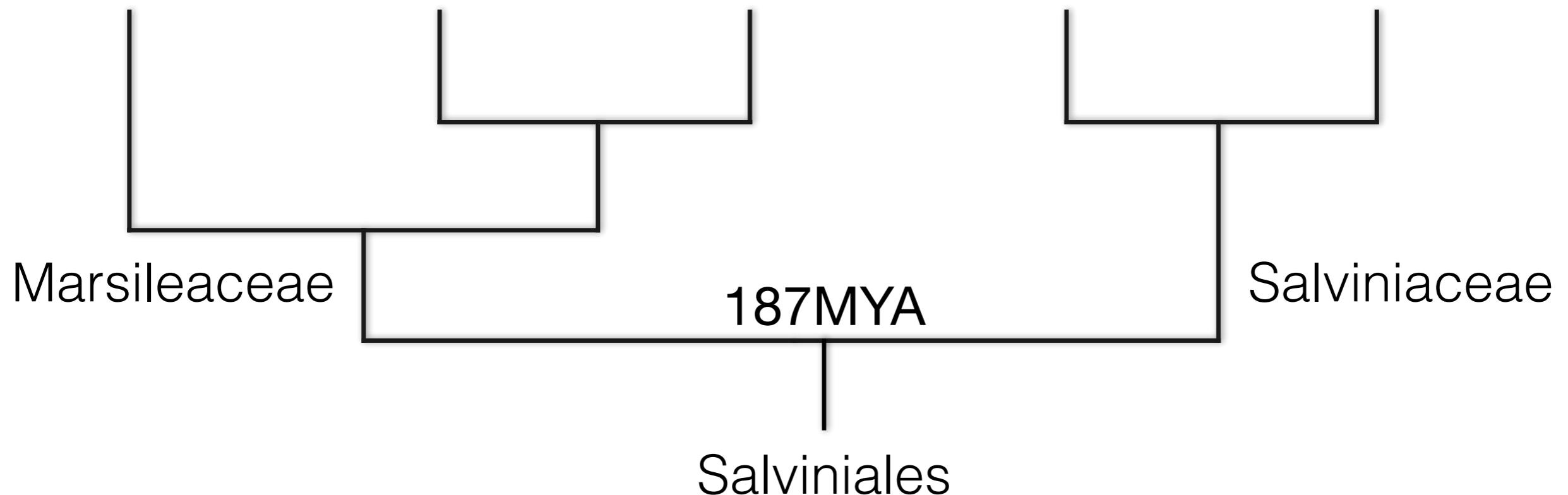
**Marsilea**

Regnellidium

Pilularia

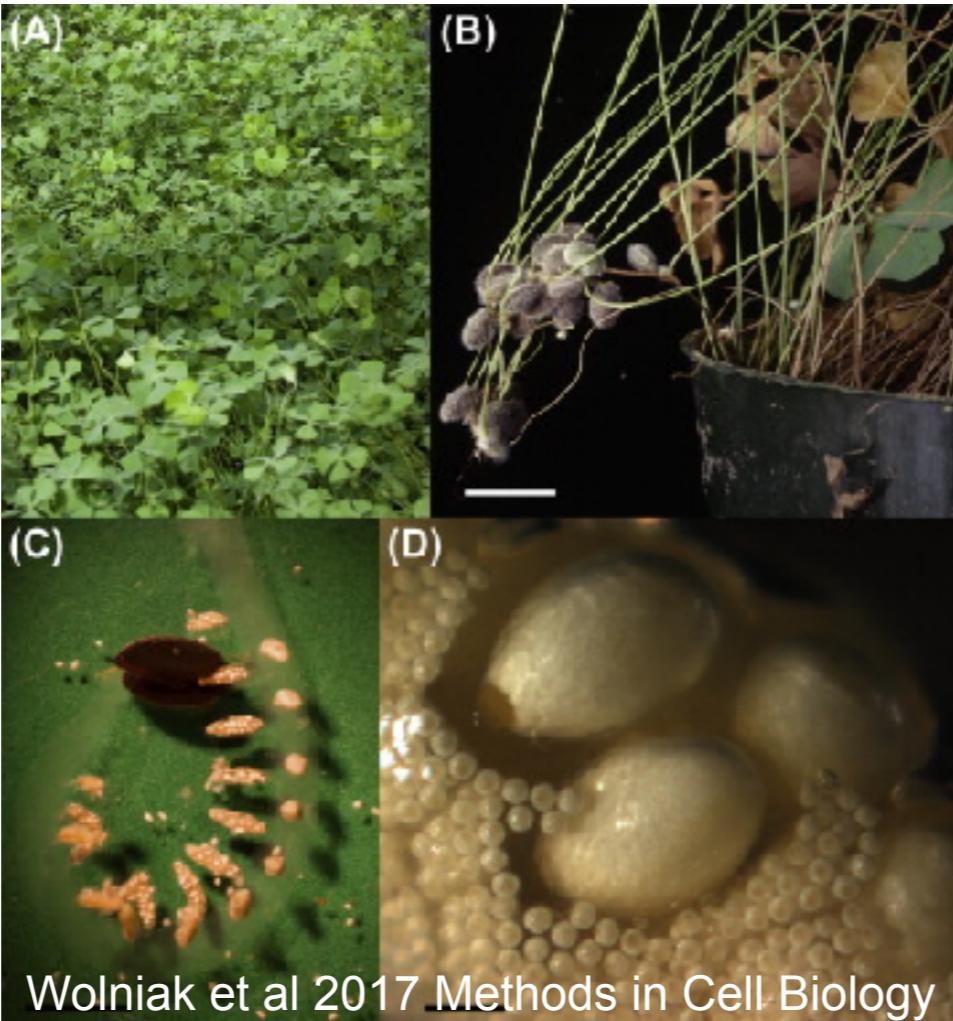
**Salvinia**

**Azolla**

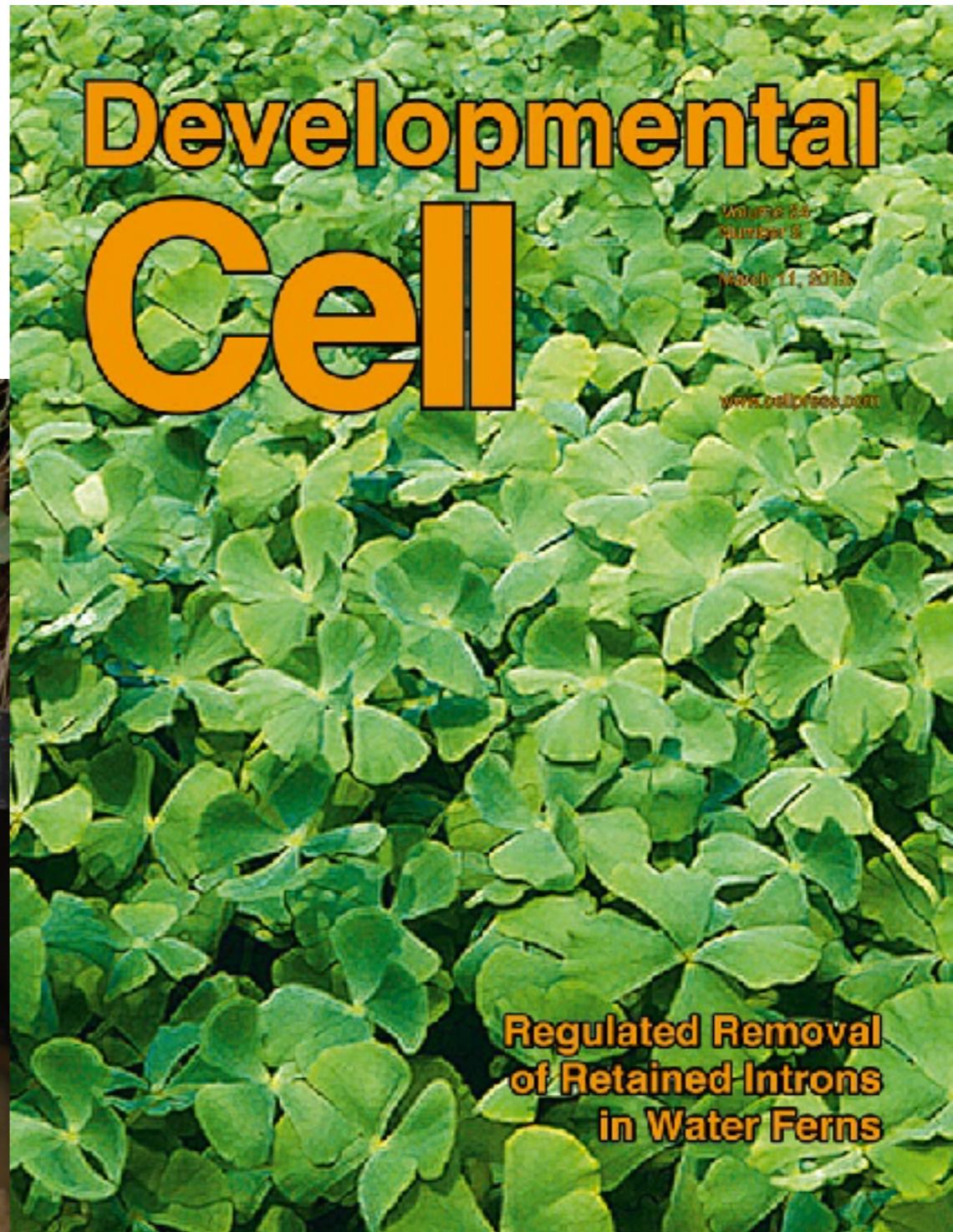


# *Marsilea vestita*

- A model system to study rapid spermatogenesis by S. Wolniak
- First time RNAi was used in ferns



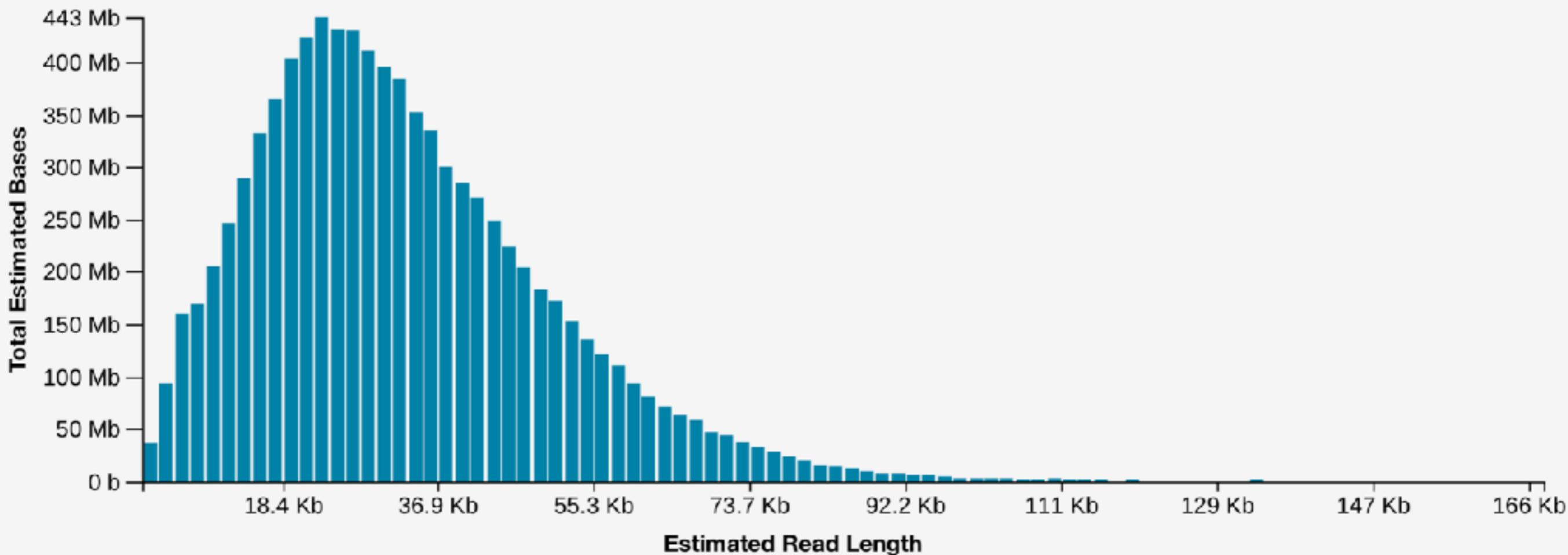
Wolniak et al 2017 Methods in Cell Biology



# Nanopore sequencing

“slow” CTAB extraction -> short-read eliminator kit ->  
ligation lib prep -> 3 nanopore flowcells

Total bases	Number of reads	Read length N50	Read length mean	Longest read
34.1 Gb (~33X)	1.9 M	<b>25.8 Kb</b>	18.3 Kb	170 Kb



# Hybrid assembly

33X nanopore + 104X Illumina

	Genome assembled	Contig No.	Contig N50	Contig N90	Longest contig	BUSCO
masurca-CA	1035 Mb	3004	<b>3.0 Mb (n=97)</b>	276 Kb (n=485)	<b>21.2 Mb</b>	<b>97.1%</b>

# Hi-C scaffolding

PhaseGenomics plant kit -> Illumina seq -> AllHiC -> TGS-GapCloser

- 93.6% of the contigs scaffolds in 20 chromosomes
- 97.5% of the total bases in 20 chromosomes
- After gap closing, contig N50 increased from 3 to 12.5 Mb

**\$4,875**

\$1,950 Nanopore seq

\$1,325 Illumina seq

\$1,600 Hi-C lib + Illumina seq