Designing a genome sequencing project

Fay-Wei Li
Boyce Thompson Institute &
Cornell University

@fern_way

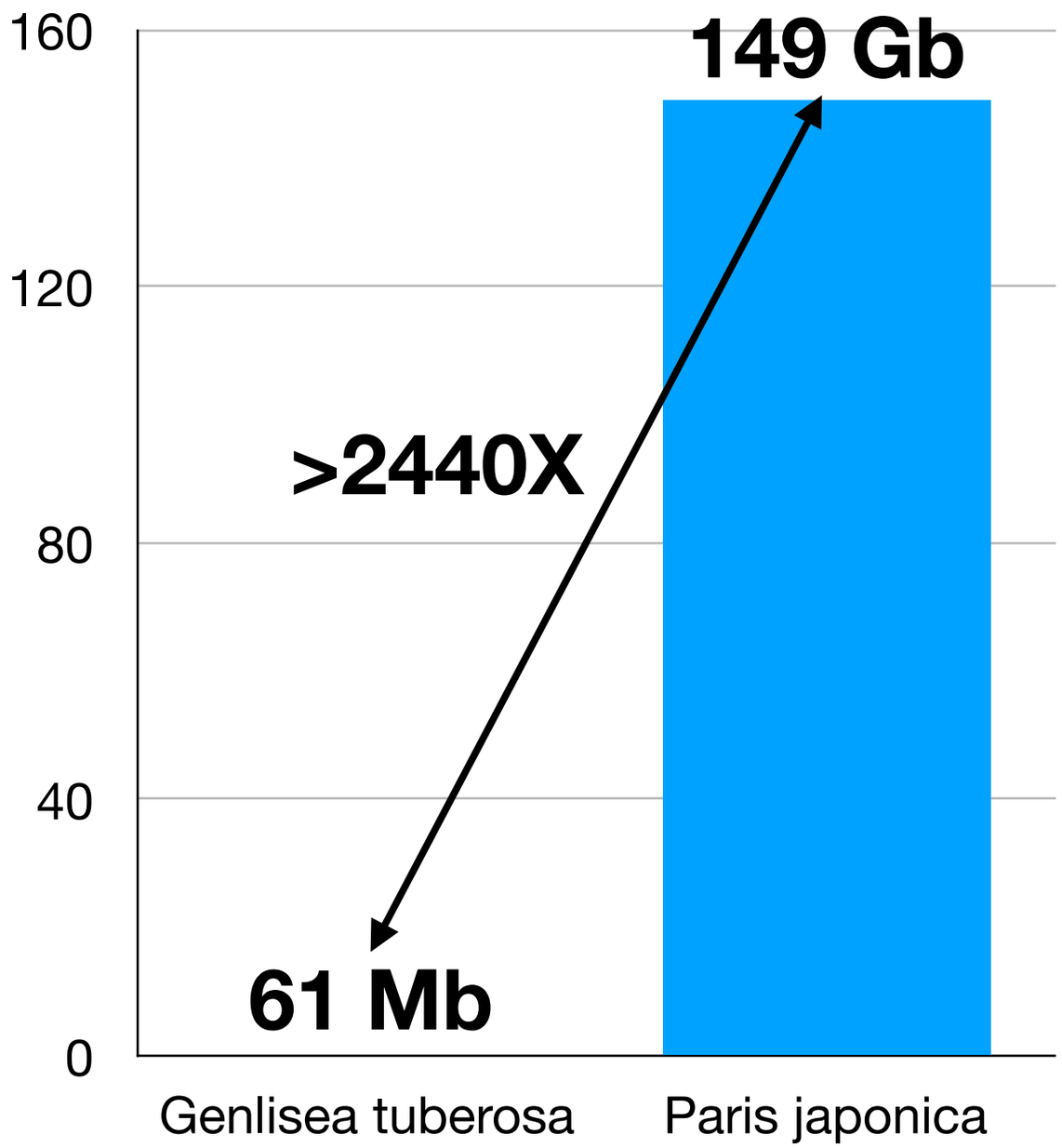
Outline

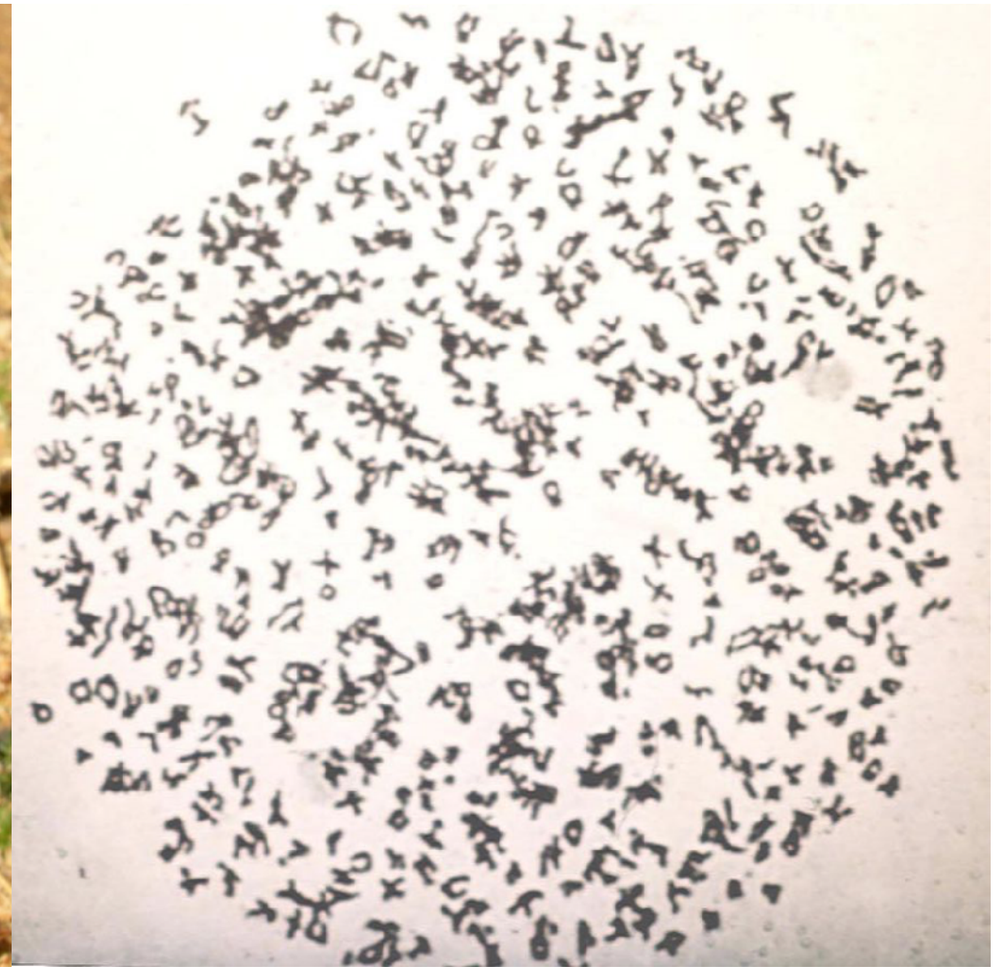
- Select your species and individual
- Choose the sequencing platforms
- Get DNA and sequence it
- Example

Outline

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

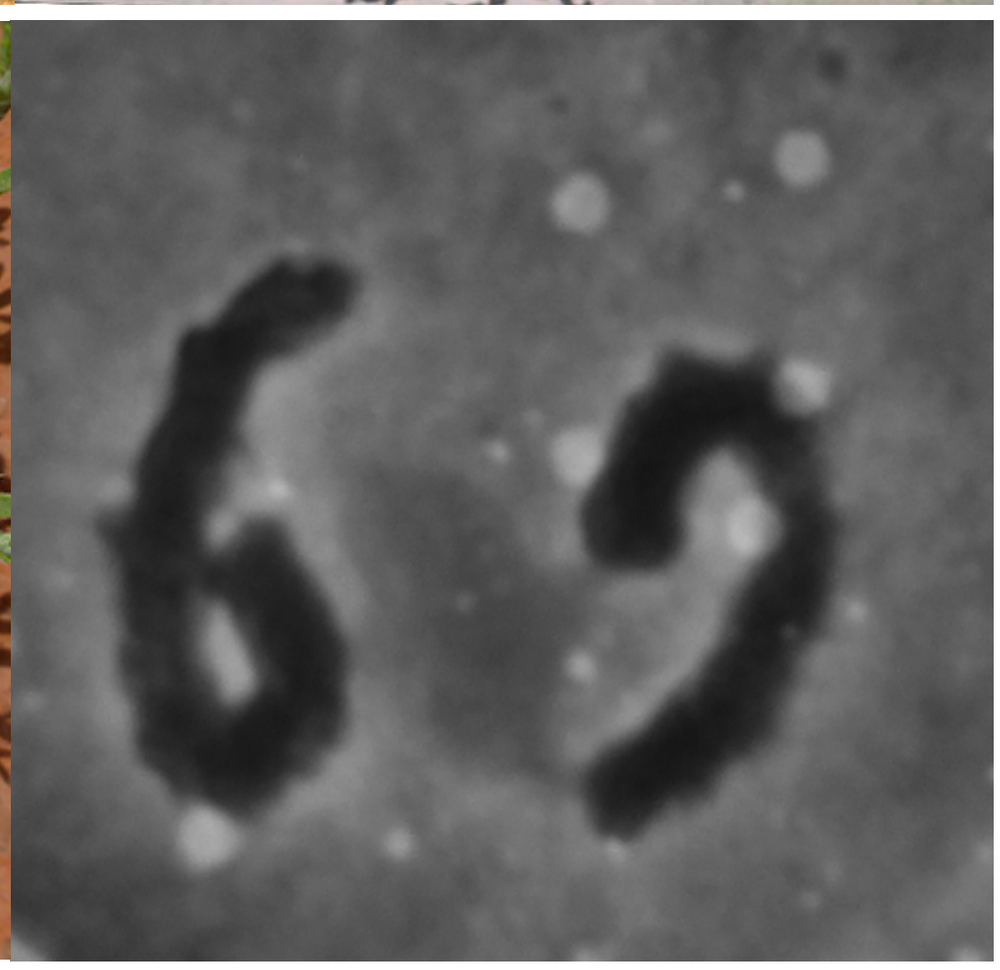
Not all plants are equally sequenceable (yet)





Ophioglossum reticulatum

2n = 1440



Xanthisma gracile

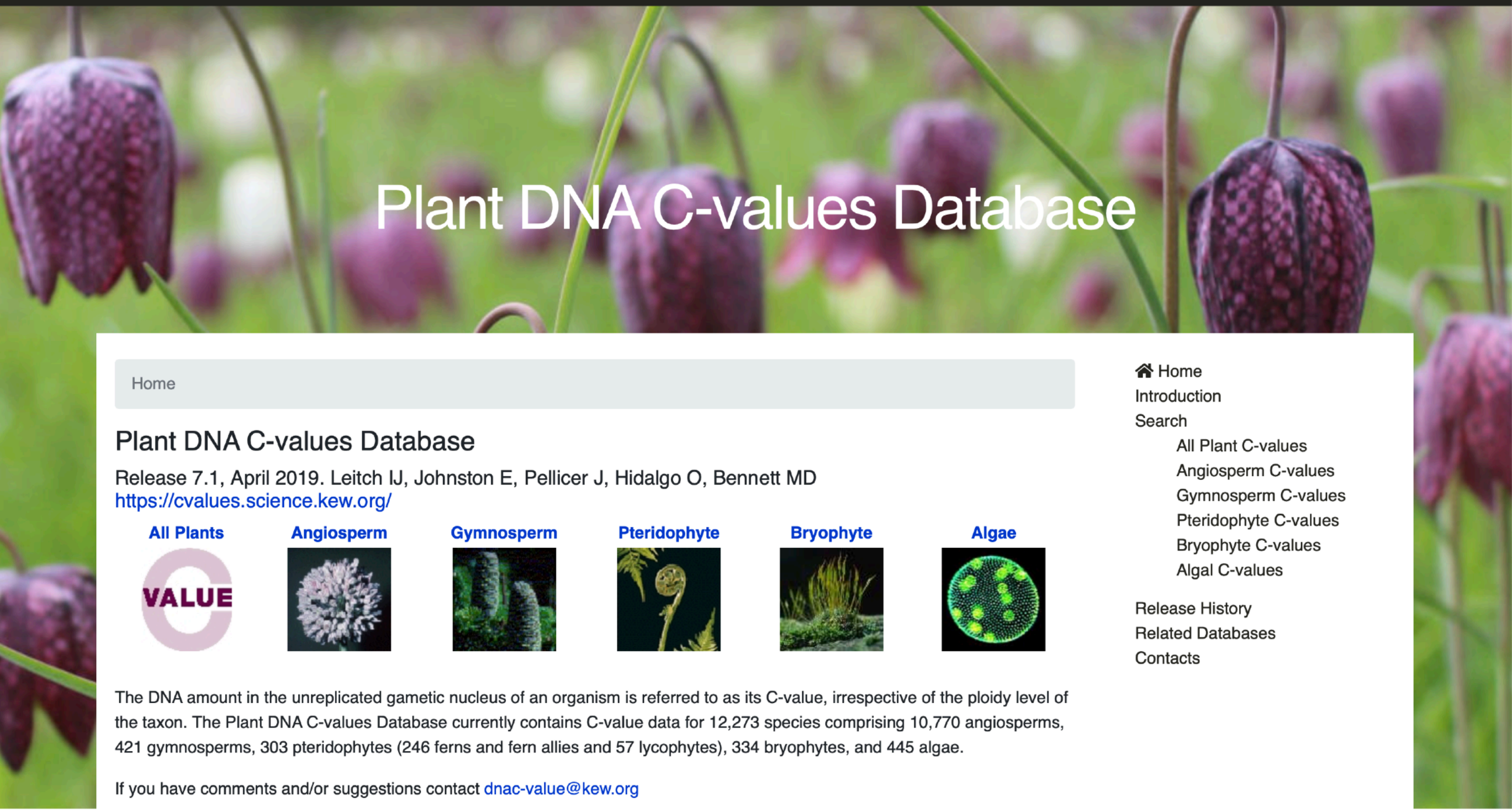
2n = 4

KNOW THE GENOME SIZE AND PLODY!!

https://cvalues.science.kew.org

Royal Botanic Gardens Kew science

Kew.org



Plant DNA C-values Database

Home

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

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

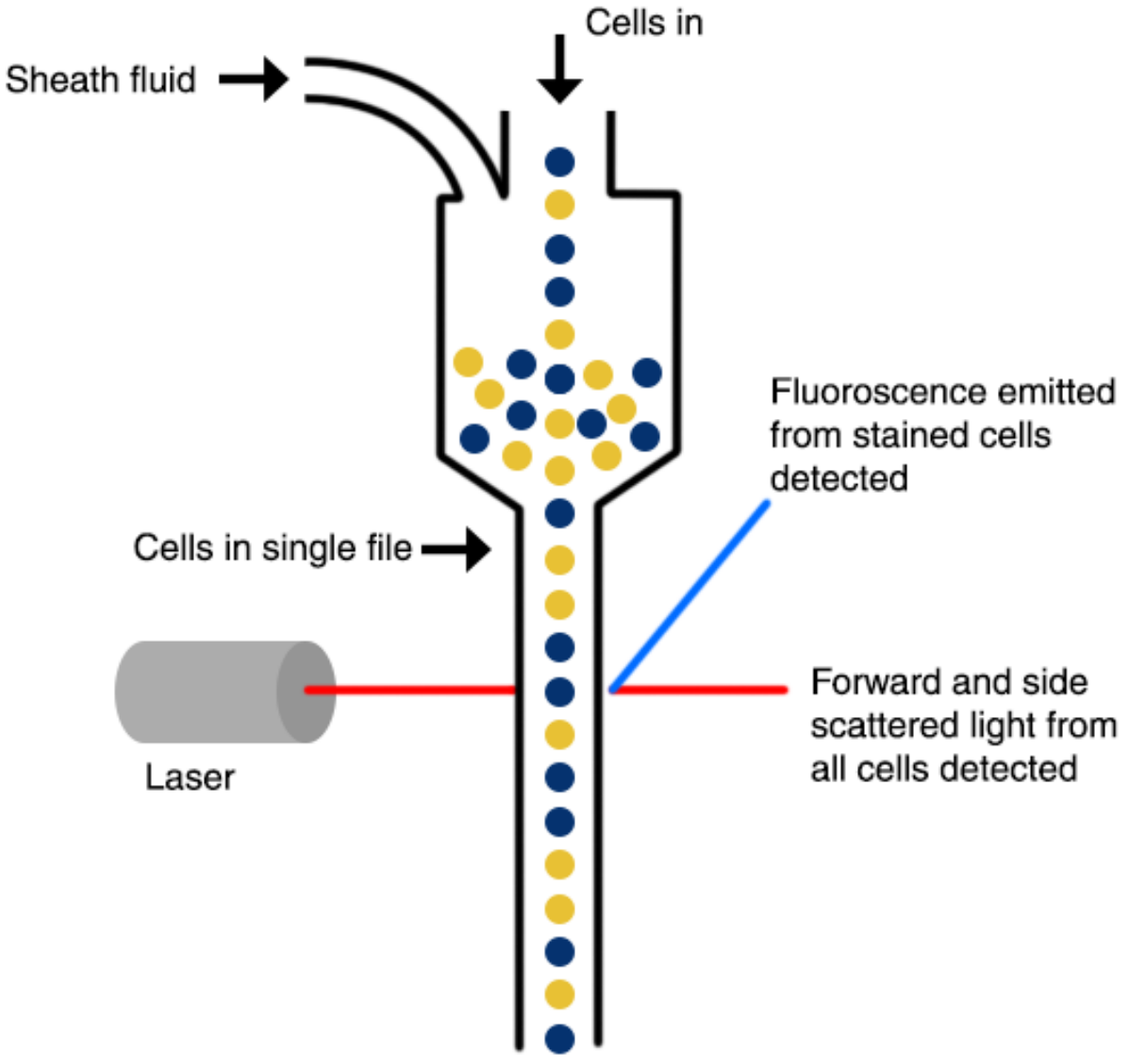
Genus	Species	Subspecies	DNA Amount 1C (pg)	Original Reference
Paris	verticillata		31.21	Pellicer et al.,2014
Paris	tetraphylla		40.75	Pellicer et al.,2014
Paris	incompleta		42.25	Pellicer et al.,2014
Paris	quadrifolia		50.52	Pellicer et al.,2014
Paris	thibetica	var. thibetica	52.52	Pellicer et al.,2014
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

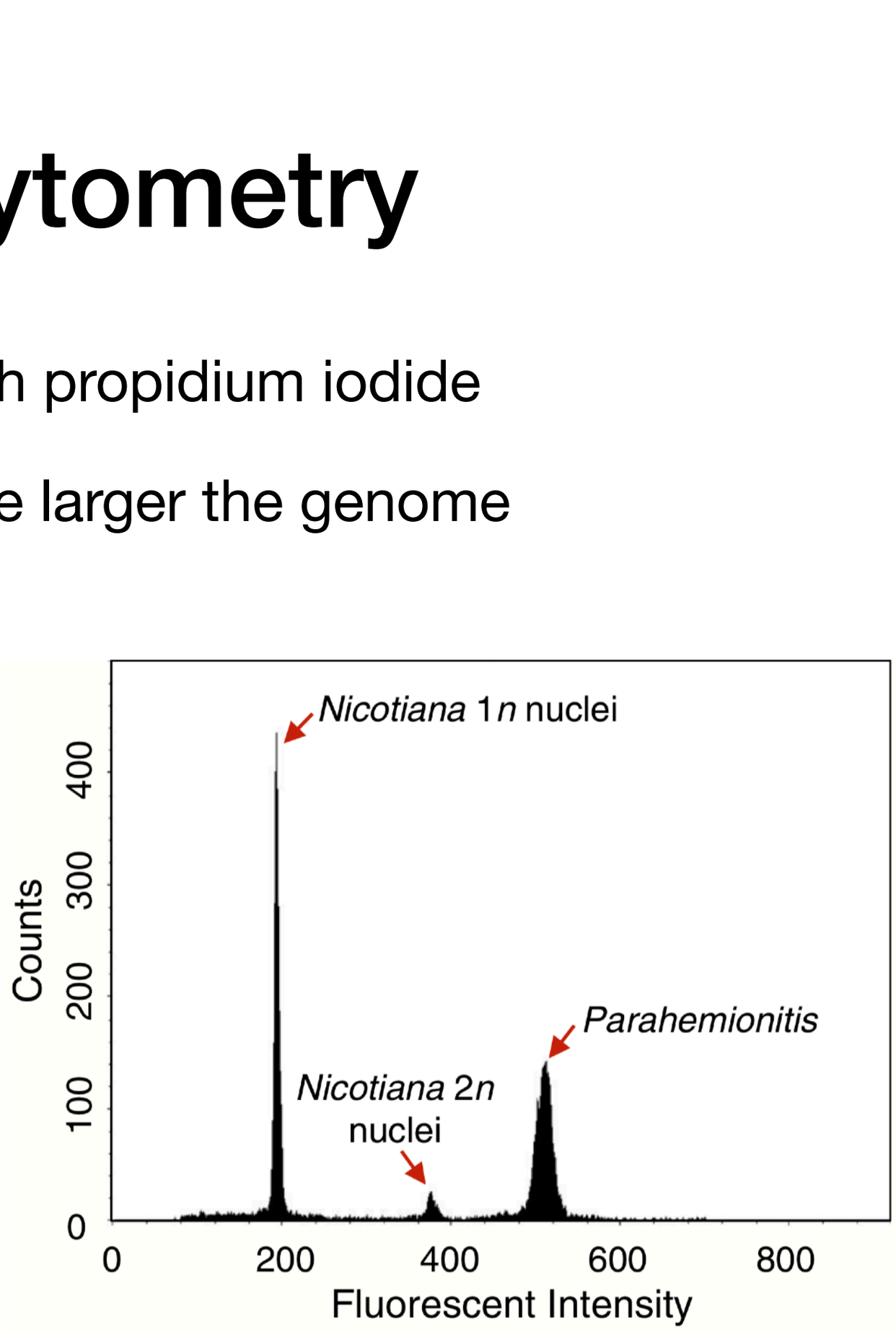
1C = haploid genome size

1pg = 978 Mbp

Flow cytometry

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

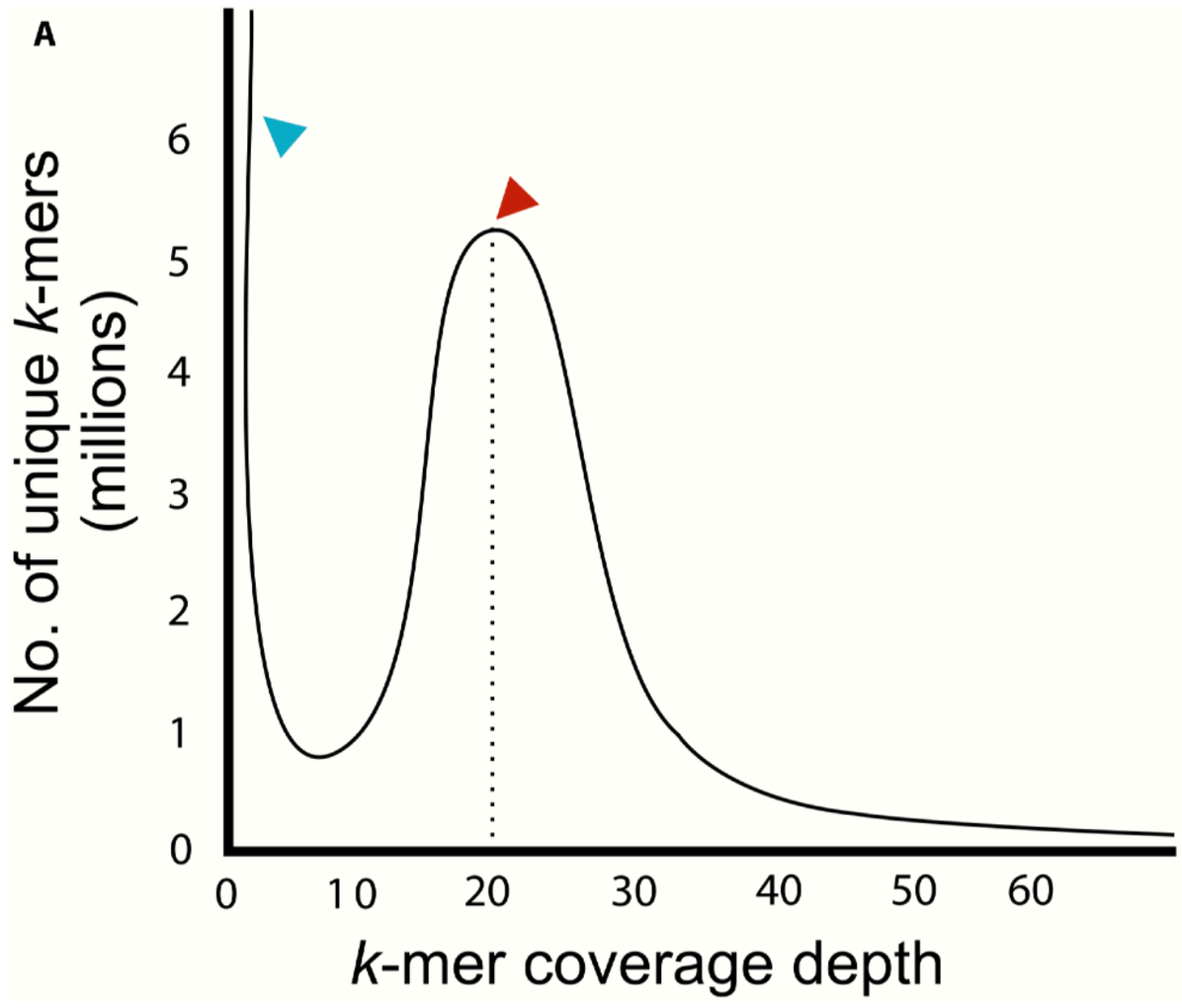




K-mer frequency plot

```
ATGCTAGCTAACTAGACTACTAAGCTAGCAT
ATGCTAGCTAACTAGACTAC
 TGCTAGCTAACTAGACTACT
  GCTAGCTAACTAGACTACTA
   CTAGCTAACTAGACTACTAA
    TAGCTAACTAGACTACTAAG
     AGCTAACTAGACTACTAAGC
      GCTAACTAGACTACTAAGCT
       CTAACTAGACTACTAAGCTA
        TAACTAGACTACTAAGCTAG
         AACTAGACTACTAAGCTAGC
          ACTAGACTACTAAGCTAGCA
           CTAGACTACTAAGCTAGCAT
```

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



http://qb.cshl.edu/genomescope/

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.

http://ccdb.tau.ac.il

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.

News

12/18 8 New counts from *Kemal2009*

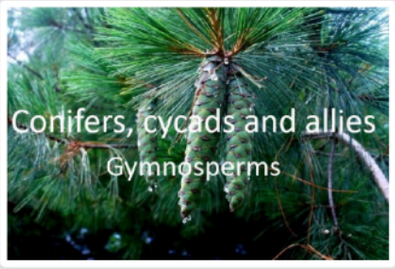
12/18 13 New counts from *Yıldız2009*

12/18 6 New counts from *Minareci2009*

12/18 16 New counts from *Yildiz2006*

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



46 species in *"Paris"* :

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

Applications in Plant Sciences

BOTANICAL SOCIETY OF AMERICA
leading scientists and educators since 1893

PROTOCOL NOTE

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

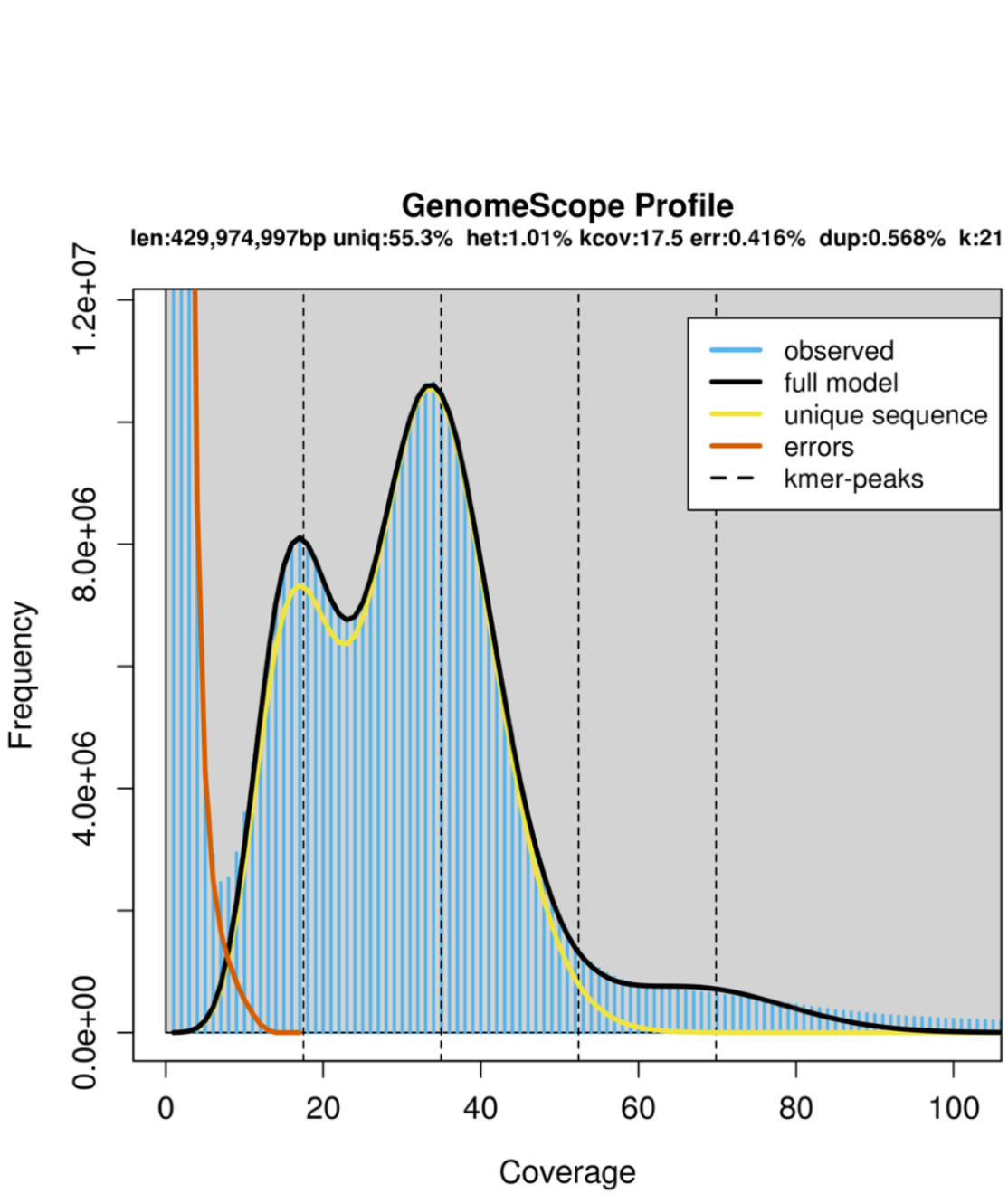
Michael D. Windham[1], Kathleen M. Pryer[1], Derick B. Poindexter[2], Fay-Wei Li[3], Carl J. Rothfels[4], and James B. Beck[5,6,7]

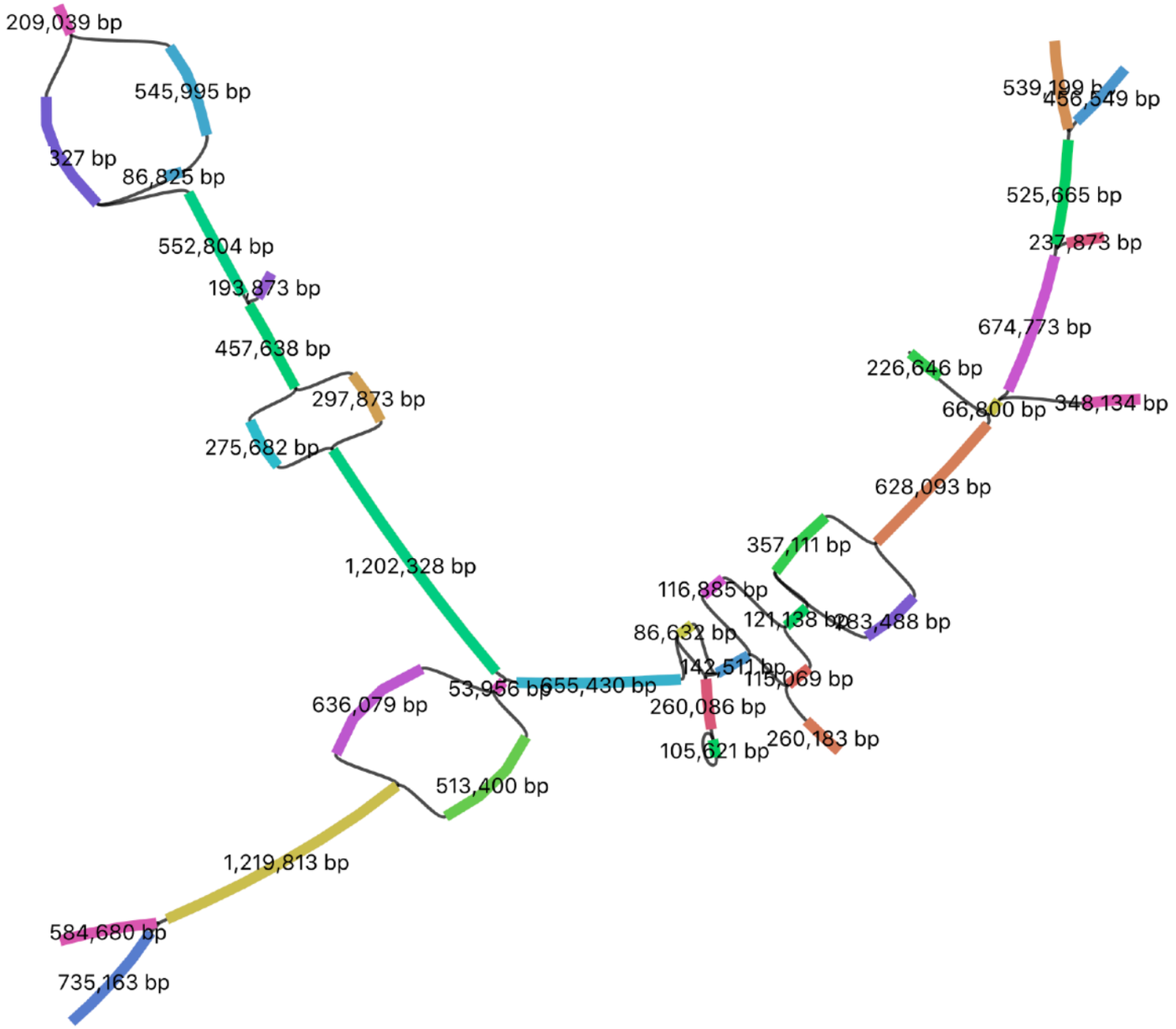
KNOW THE GENOME SIZE AND PLODY!!

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





Genome survey

- >20X Illumina coverage
- Genome size
- Heterozygosity
- Repeat content

GenomeScope
http://qb.cshl.edu/genomescope/

- Contamination

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:

ATGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGACTTT

GAGAGAGAGAGA

ATGAGAGAGAGAGAGA

GAGAGAGAGAGAGA

GAGAGAGAGAGA

GAGAGAGAGAGACTT

GAGAGAGAGAGA

GAGAGAGAGAGA

Solving the assembly puzzle

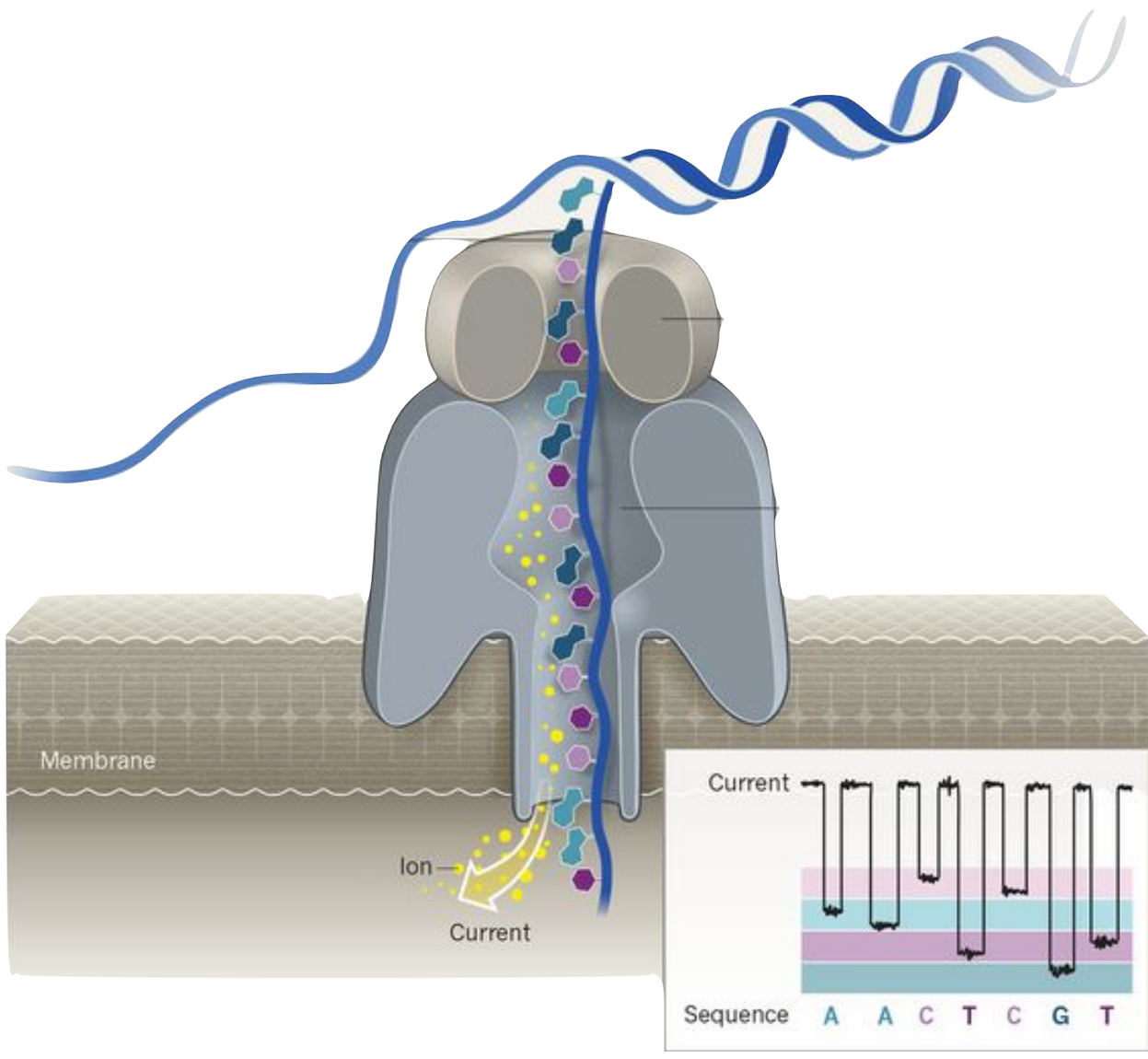
Short reads = small pieces

Long reads = large pieces

https://vertebrategenomesproject.org/technology

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

PromethION

SmidgION

Flowcell

MinION device

MinION

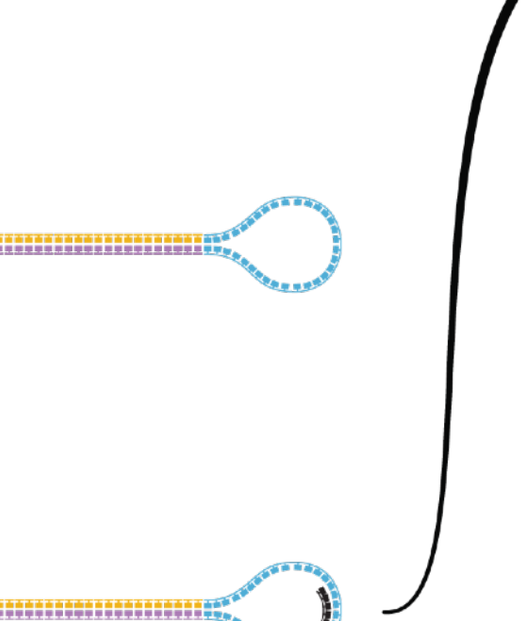
- R9 flowcell
  - $500
  - 10-20 Gb
- Flongle flowcell
  - $90
  - 1-2 Gb

- Library prep
  - $150
  - 1hr
- 1ug DNA
- (RNA, cDNA, amplicon, …)

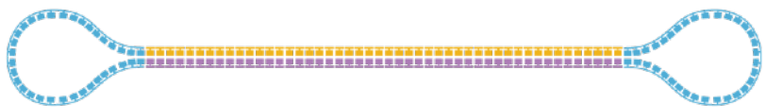
- Starter pack
  - $1000
  - 2 R9 flowcells
- 1 library prep kit

PACBIO®

Start with high-quality double stranded DNA



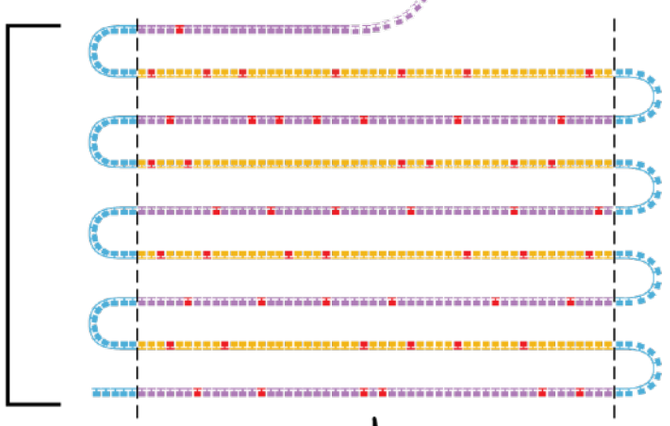
Ligate SMRTbell adapters and size select



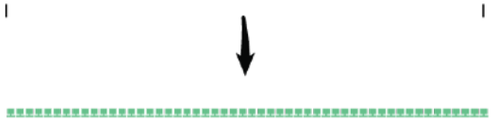
Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads



Consensus is called from subreads



HiFi READ

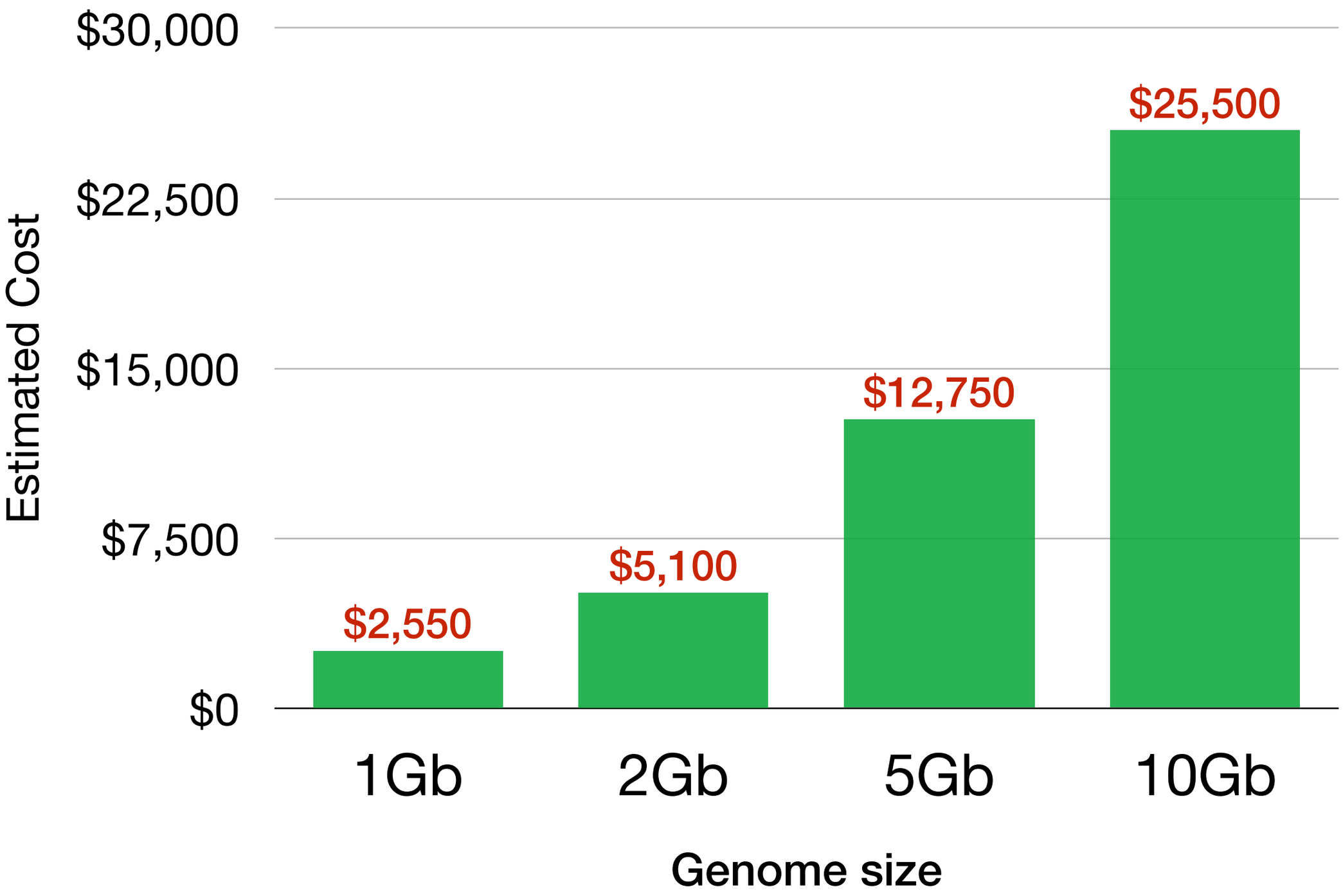
(>99% accuracy)

How much $$ for an 1Gb genome?

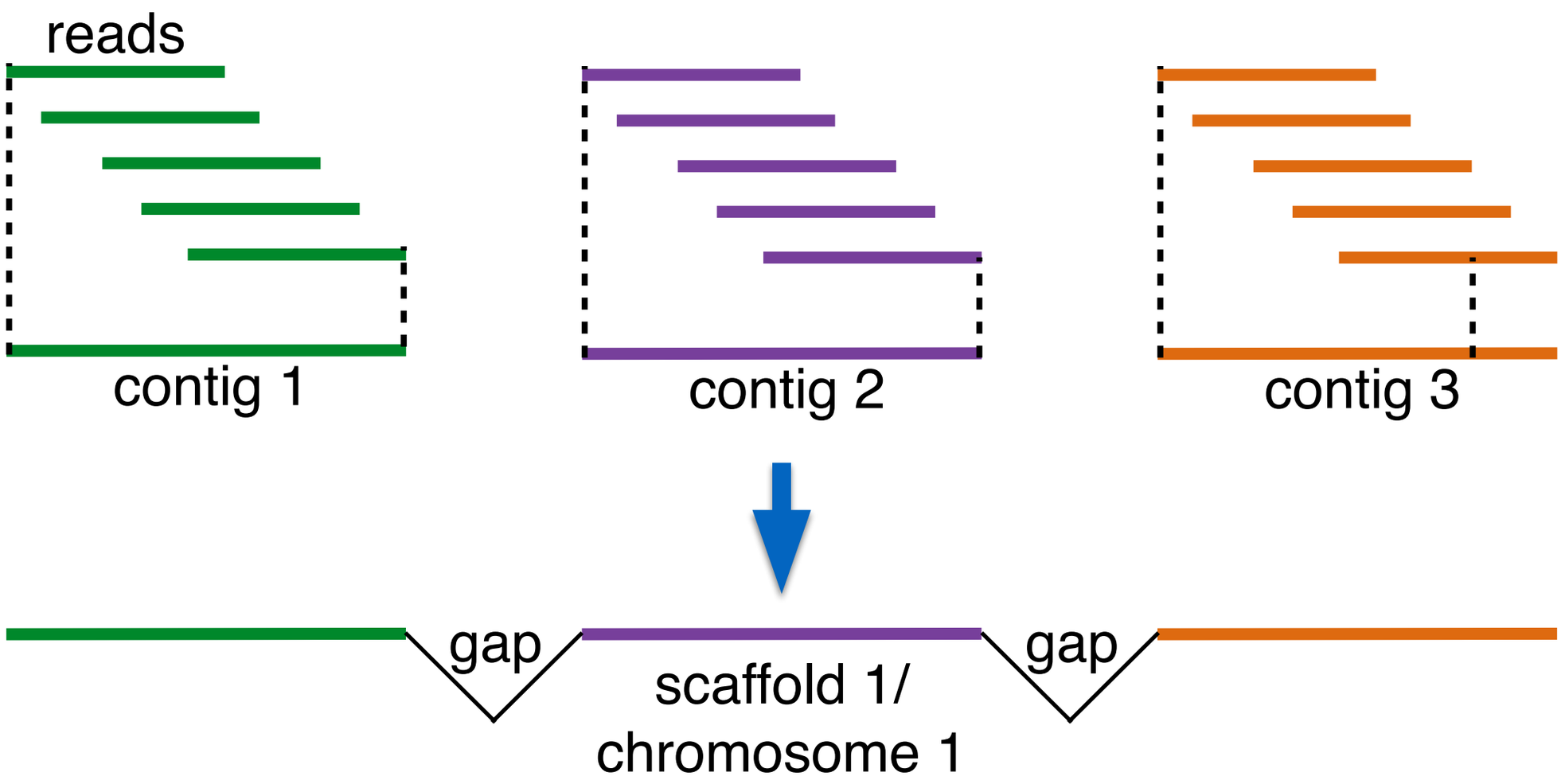
- 50X Illumina:
  - 50Gb x $11/Gb = **$550**
- 50X nanopore:
  - 50Gb x $40/Gb = **$2,000**

$2,550

How much $$?

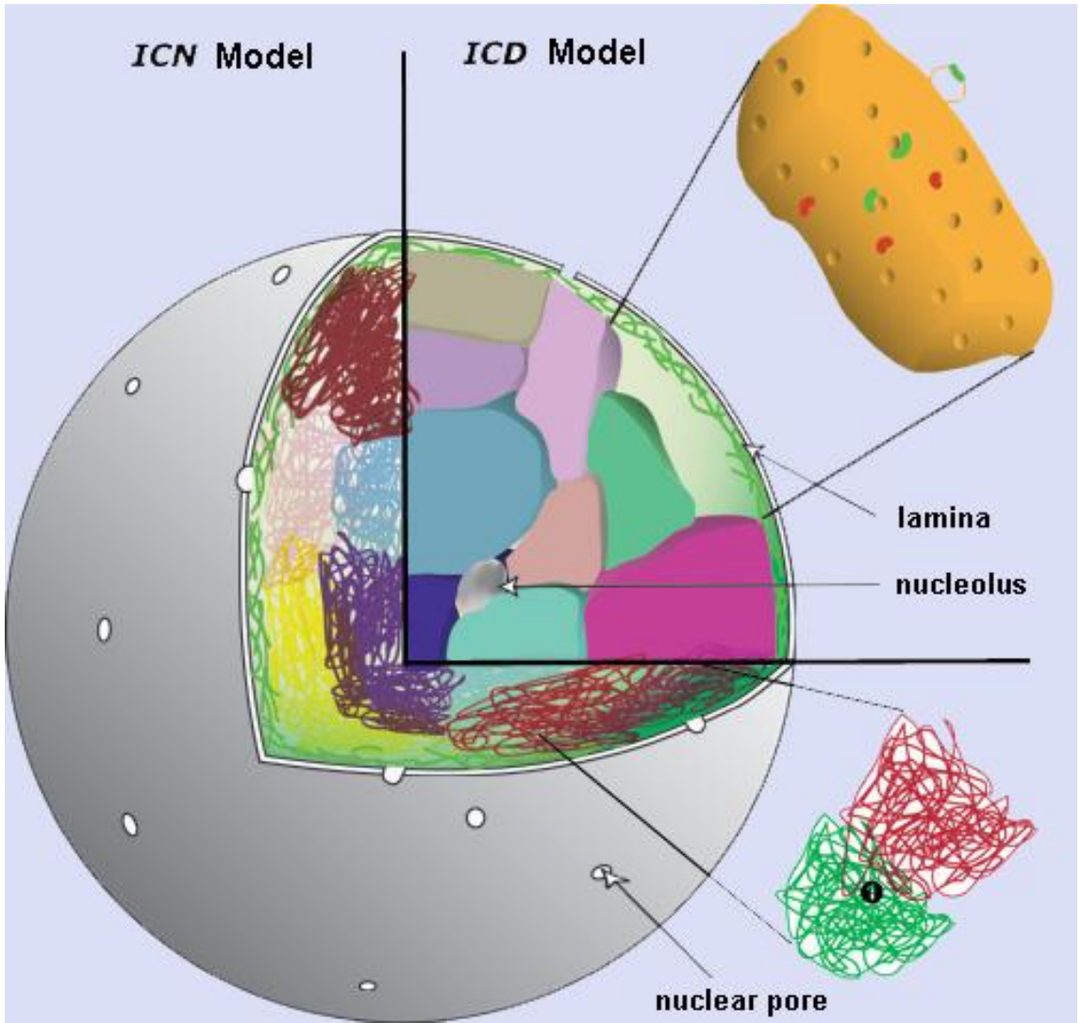

$30,000
$22,500
$15,000
$7,500
$0
Estimated Cost
$2,550
$5,100
$12,750
$25,500
1Gb
2Gb
5Gb
10Gb
Genome size

Reads -> Contigs -> Scaffolds



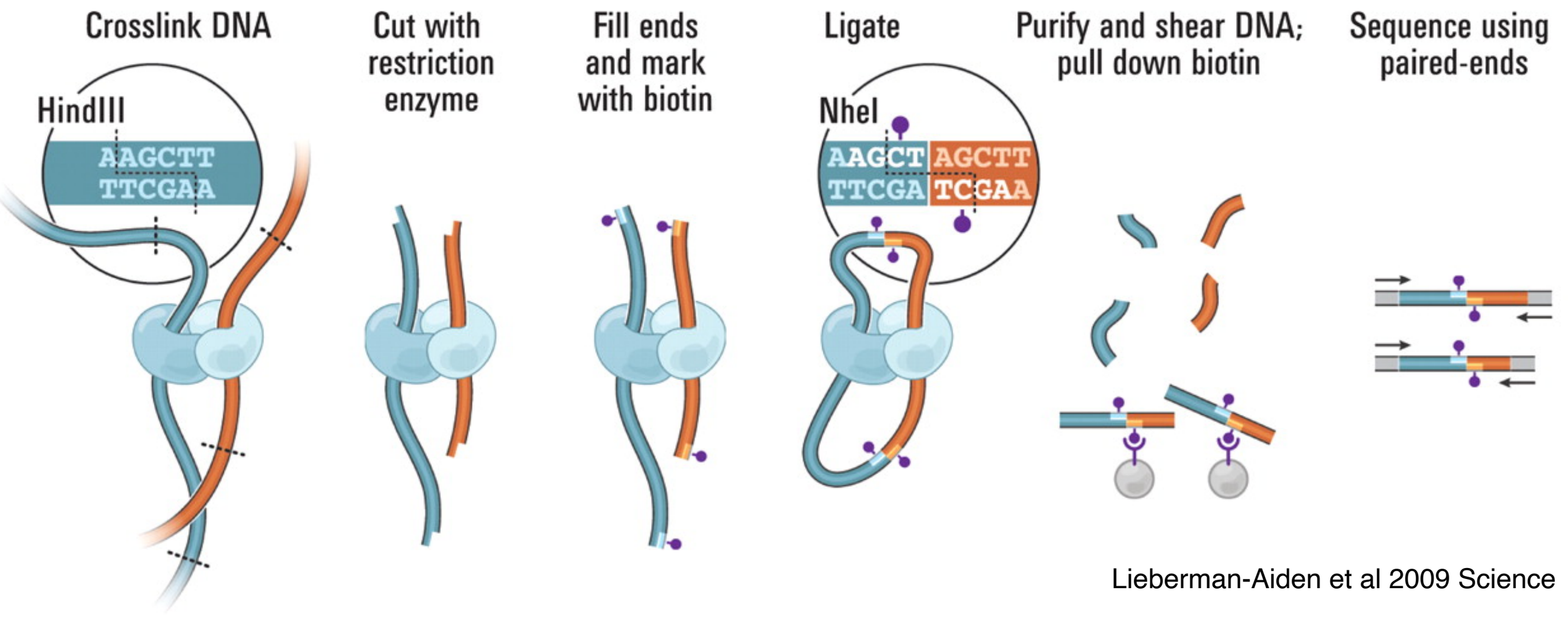
Hi-C

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



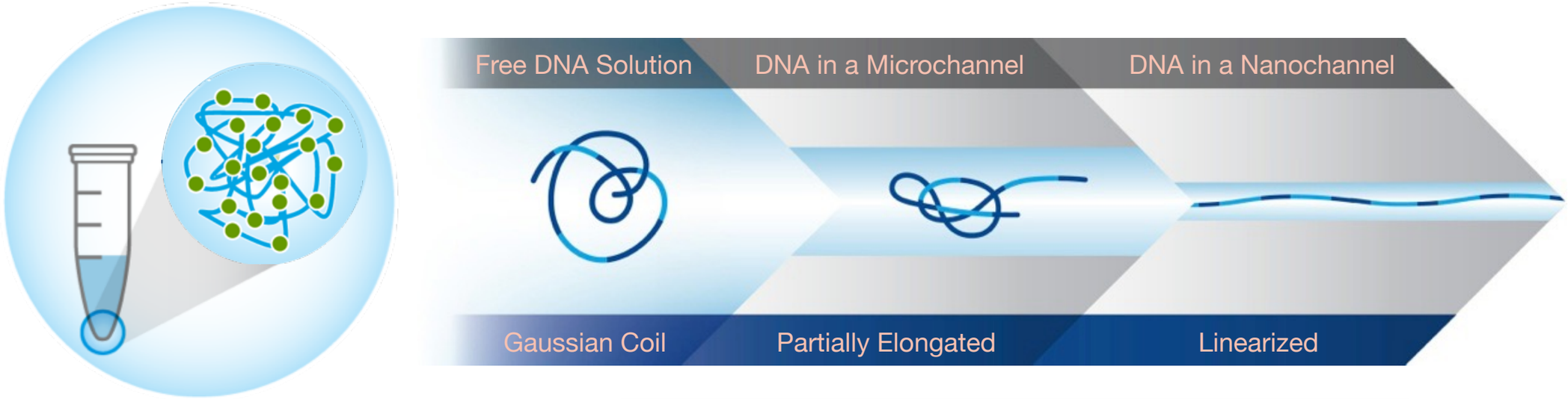
Hi-C

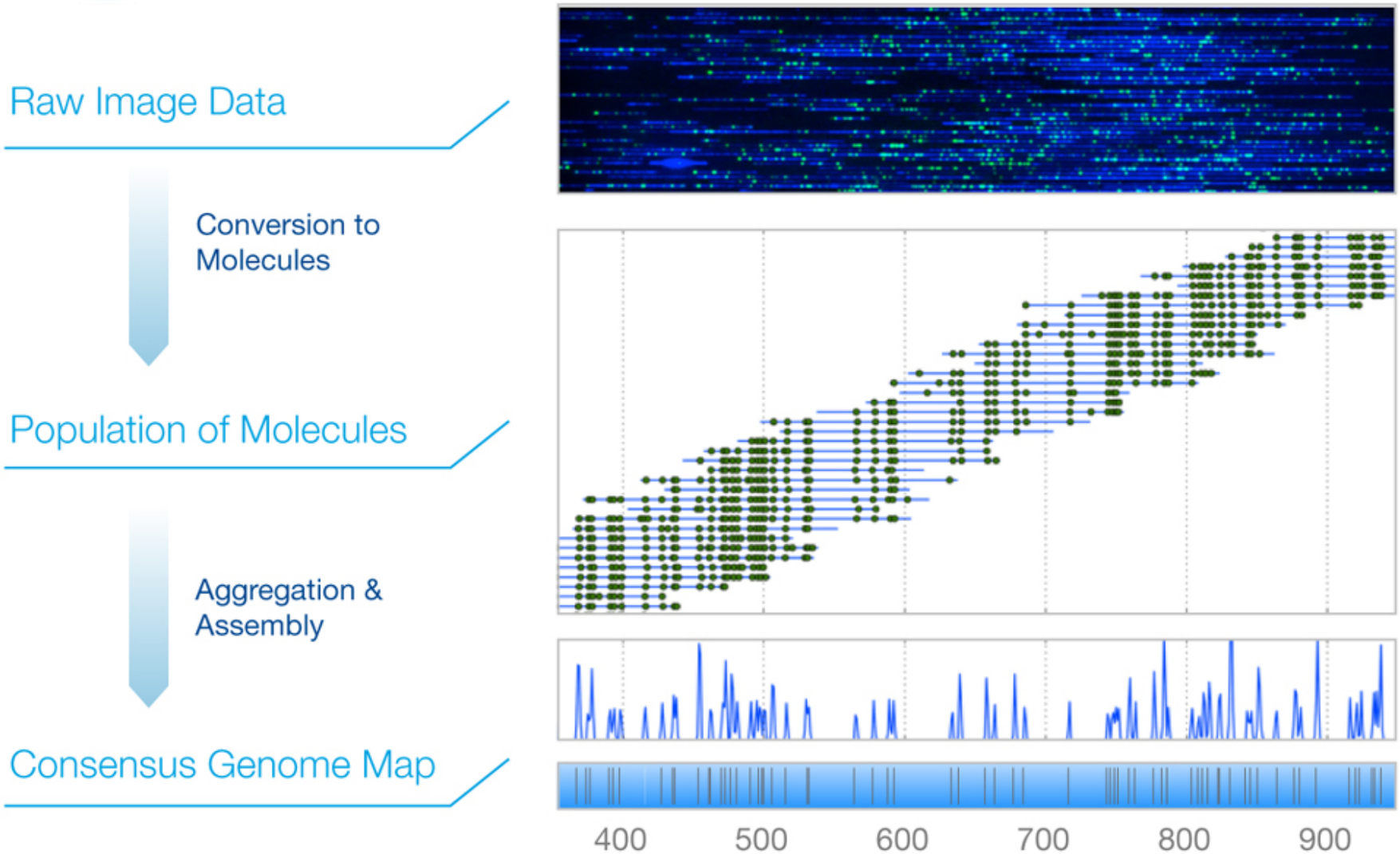
- Hi-C = high throughput chromatin conformation capture
- DNA of the same chromosome will be ***spatially*** close



Lieberman-Aiden et al 2009 Science

Optical mapping



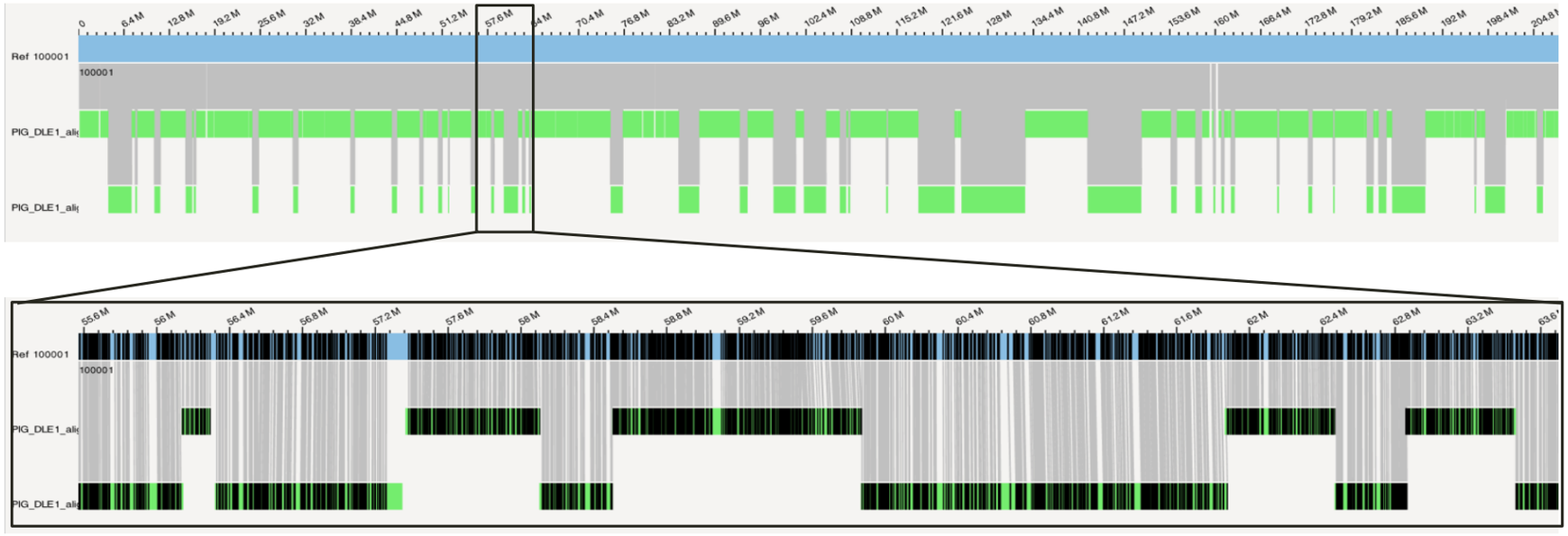


Optical mapping

Use optical map to scaffold contigs

Genome map

contigs



How much $$ for scaffolding an 1Gb genome?

- Hi-C
  - Lib prep kit **$500** + 50X Illumina **$550** = **$1,050**
- 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

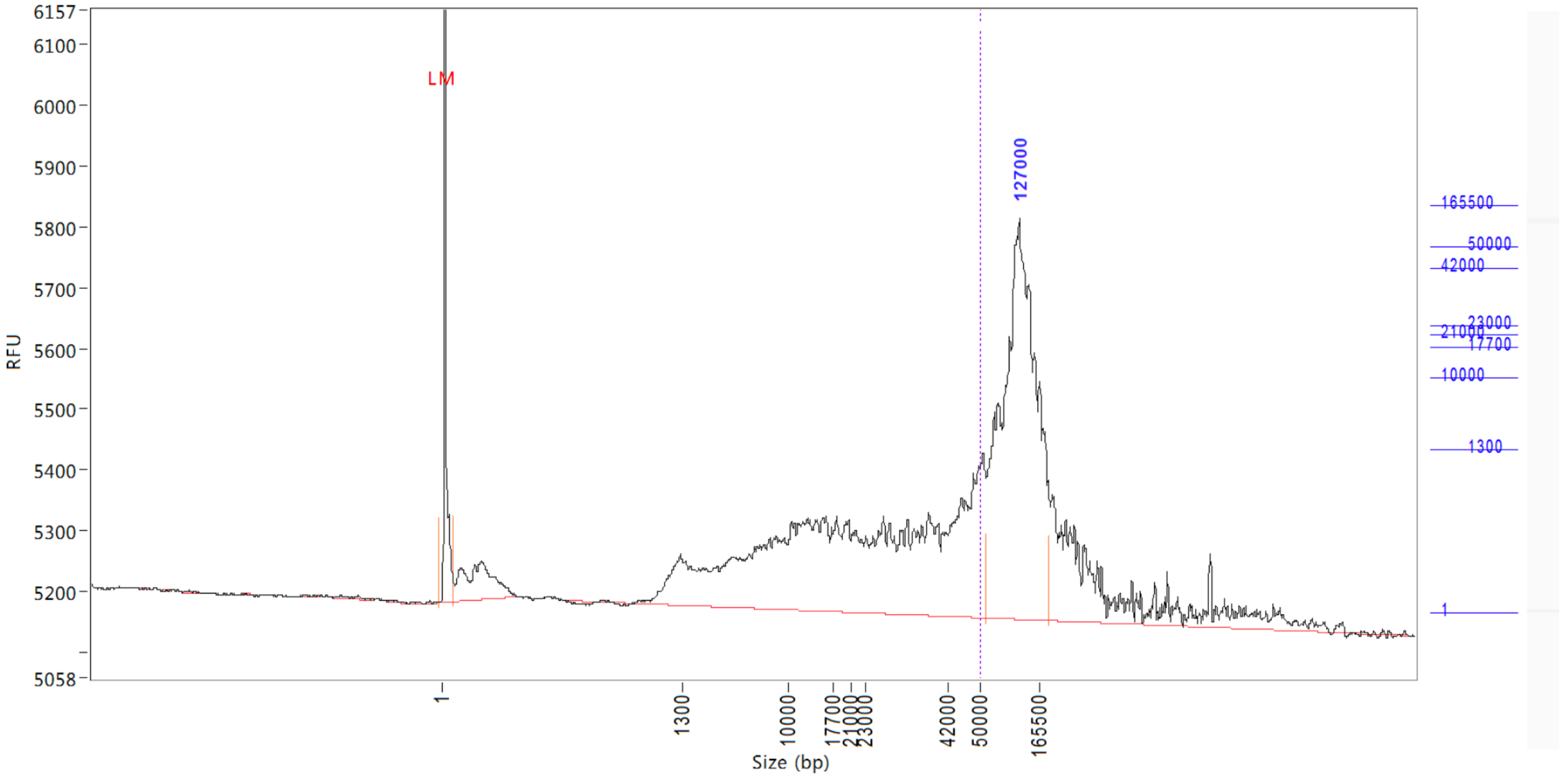
High molecular weight DNA

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

QC HMW DNA

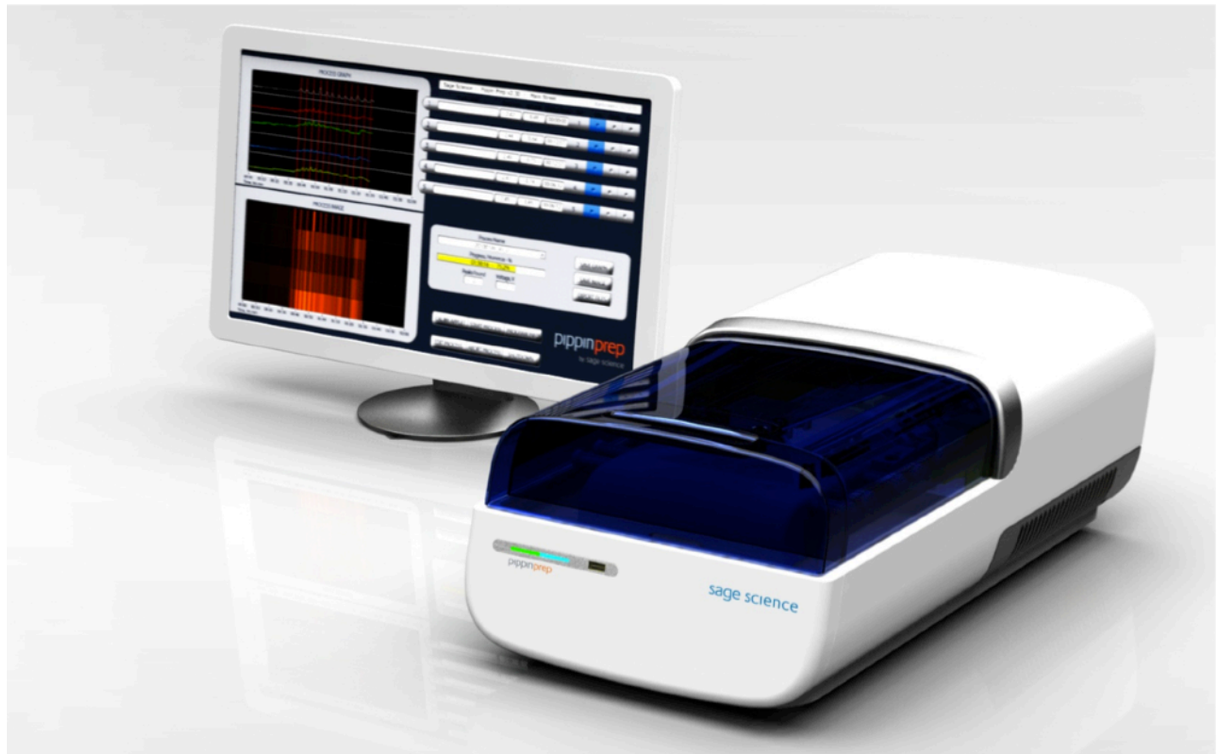
- **Quantification** - Qubit (never trust nanodrop conc!)
    - 1ug min. each nanopore run, ~5ug if size selection
- **Purity** - nanodrop
    - 260/280, 260/230 close to 2
- **Integrity**
    - Pulse-field gel electrophoresis
    - Bioanalyzer, Femto Pulse
    - Low percentage (0.5%) agarose gel with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

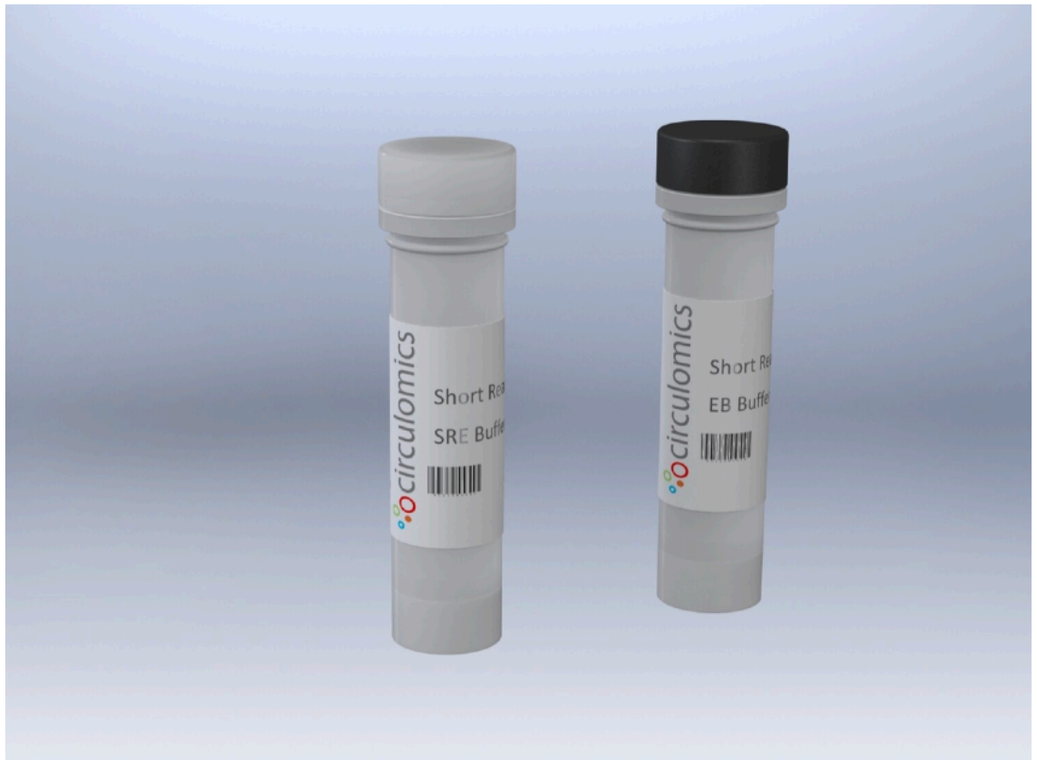
“Slow” CTAB on *Aconitum noveboracense* -> Femto Pulse QC



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

Outline

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

Assembly size
Scaffold N50
Contig N50

Mosses
Bryophytes
Liverworts
Hornworts
Lycophytes
Vascular plants
Ferns
Seed plants

Hookeriales
Hypnales — Pleurozium schreberi; Fontinalis antipyretica
Hypoterygiales
Ptychomiales
Hypnodendrales
Aulacomniales
Orthodontiales
Orthotrichales
Rhizogoniales
Bryales
Bartramiales
Hedwigiales
Splachnales
Dicranales — **Ceratodon purpureus** (GG1, R40)
Pottiales
Grimiales
Timmiales
Funariales — **Physcomitrium patens**; Funaria hygrometrica
Gigaspermales
Diphysicales
Buxbaumiales
Tetraphidales
Polytrichales
Oedipodiales
Andreaeales
Andreaeobryales
Sphagnales — **Sphagnum fallax**; **Sphagnum magellanicum**
Takakiales

Ptilidales
Porellales
Jungermanniales
Pleurzoziales
Metzgeriales
Fossombroniales
Pallaviniales
Pelliales
Marchantiales — **Marchantia polymorpha** (ruderalis 'Tak1', polymorpha, montivagans); Marchantia paleacea; Marchantia inflexa
Shpaeocarpales
Neohodgsoniales
Blasiales
Treubiales
Haplomitriales

Dendrocerotales
Phymatocerotales
Notothyladales
Anthocerotales — **Anthoceros agrestis** (Bonn, Oxford); **Anthoceros punctatus**; **Anthoceros augustus**
Leiosporocerotales

Isoetales
Selaginellales — **Selaginella moellendorffii**; **Selaginella lepidophylla**; **Selaginella tamariscina**
Lycopodiales

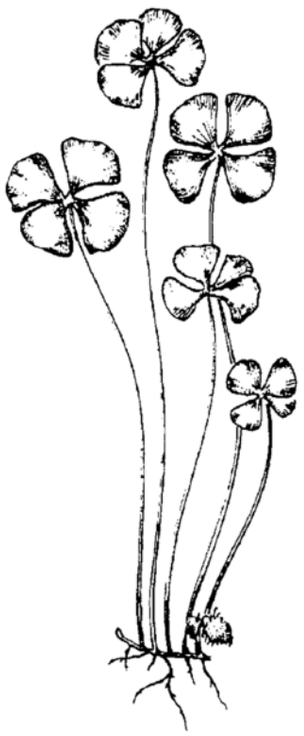
Polypodiales
Cyatheales
Salviniales — **Azolla filiculoides**; **Salvinia cucullata**
Schizaeales
Gleicheniales
Hymenophyales
Osmundales
Marattiales
Psilotales
Ophioglossales
Equisetales

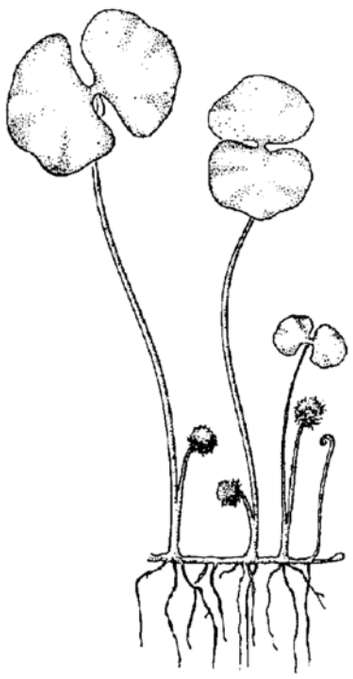
Seed plants — **Penium margaritaceum**

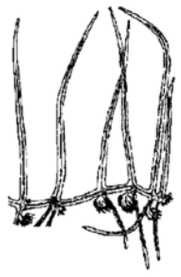
Zygnematophyceae — **Mesotaenium endlicherianum**; **Spirogloea muscicola**
Coleochaetophyceae
Charophyceae — **Chara braunii**
Klebsormidiophyceae — **Klebsormidium nitens**
Mesostigmatophyceae — **Mesostigma viride**
Chlorokybophyceae — **Chlorokybus atmophyticus**

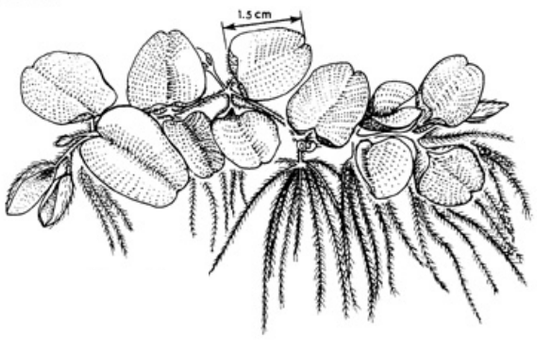
Szovenyi, Gunadi, Li (2021) Nature Plants

Mosses
Andreaeobryales
Sphagnales
Takakiales
Sphagnum fallax
Sphagnum magellanicum
Ptilidales
Porellales
Jungermanniales
Pleurzoziales
Metzgeriales
Fossombroniales
Pallaviniales
Pelliales
Marchantiales
Shpaeocarpales
Neohodgsoniales
Blasiales
Treubiales
Haplomitriales
Marchantia polymorpha
ruderalis ‘Tak1’
polymorpha
montivagans
Marchantia paleacea
Marchantia inflexa
Bryophytes
Liverworts
Dendrocerotales
Phymatocerotales
Notothyladales
Anthocerotales
Leiosporocerotales
Anthoceros agrestis
Bonn
Oxford
Anthoceros punctatus
Anthoceros augustus
Hornworts
Isoetales
Selaginellales
Lycopodiales
Selaginella moellendorffii
Selaginella lepidophylla
Selaginella tamariscina
Lycophytes
Polypodiales
Cyatheales
Salviniales
Schizaeales
Gleicheniales
Hymenophyales
Osmundales
Marattiales
Psilotales
Ophioglossales
Equisetales
Azolla filiculoides
Salvinia cucullata
Vascular plants
Ferns
Seed plants
Zygnematophyceae
Coleochaetophyceae
Charophyceae
Klebsormidiophyceae
Mesostigmatophyceae
Chlorokybophyceae
Penium margaritaceum
Mesotaenium endlicherianum
Spirogloea muscicola
Chara braunii
Klebsormidium nitens
Mesostigma viride
Chlorokybus atmophyticus









Marsilea Regnellidium Pilularia Salvinia Azolla

Marsileaceae

187MYA

Salviniaceae

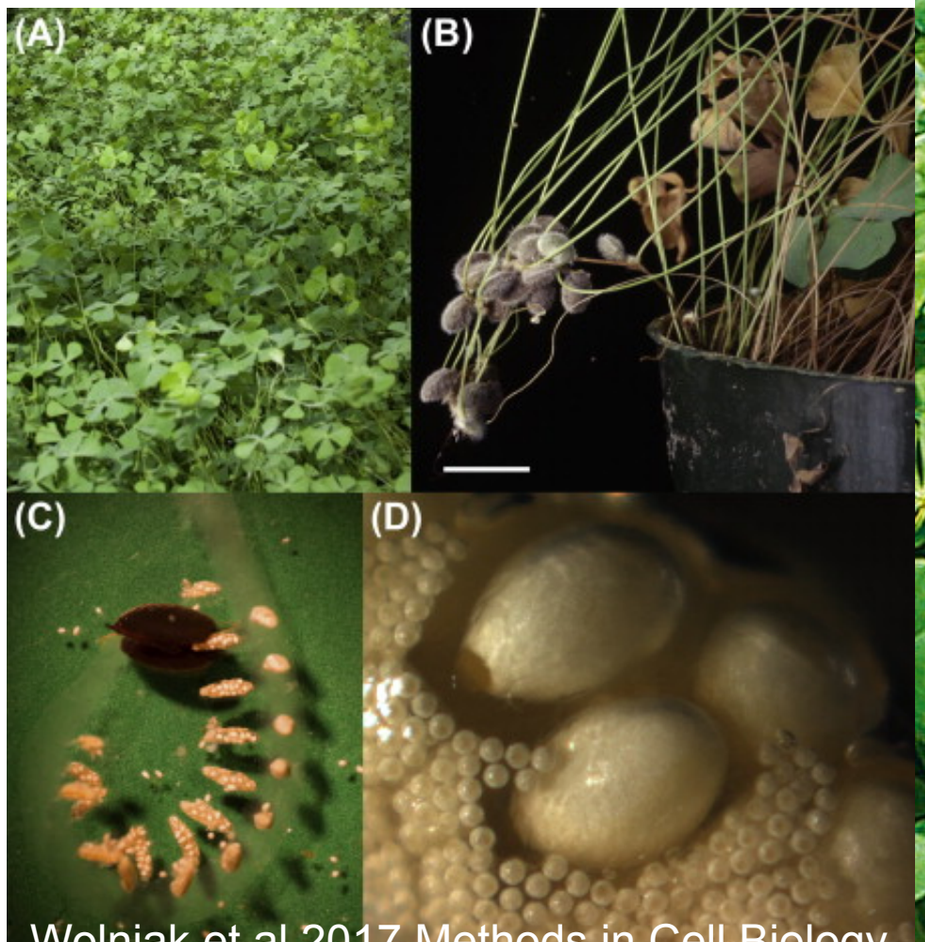
Salviniales

Li et al 2018 Nature Plants

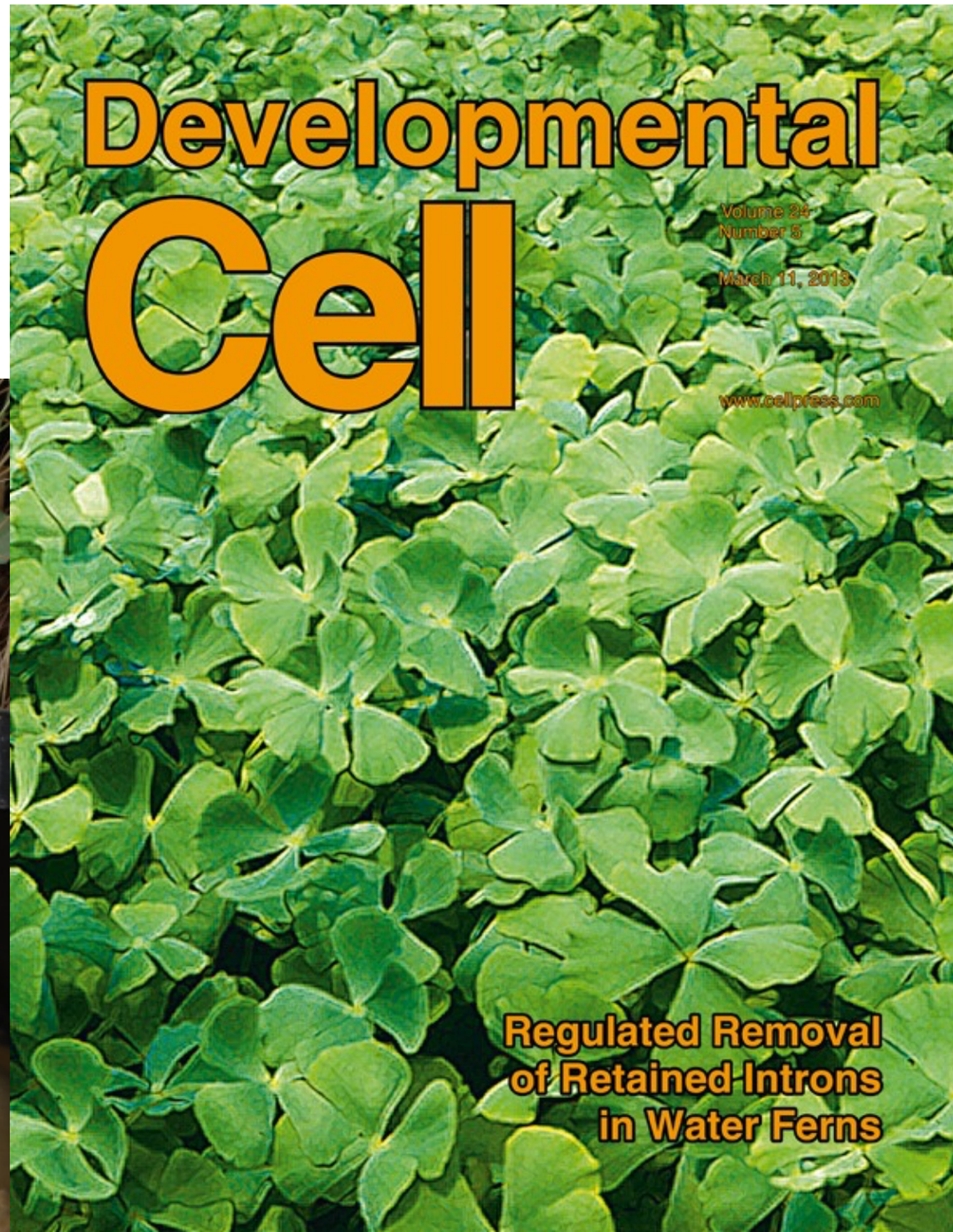
Marsilea vestita

- A model system to study rapid spermatogenesis by S. Wolniak

- First time RNAi was used in ferns



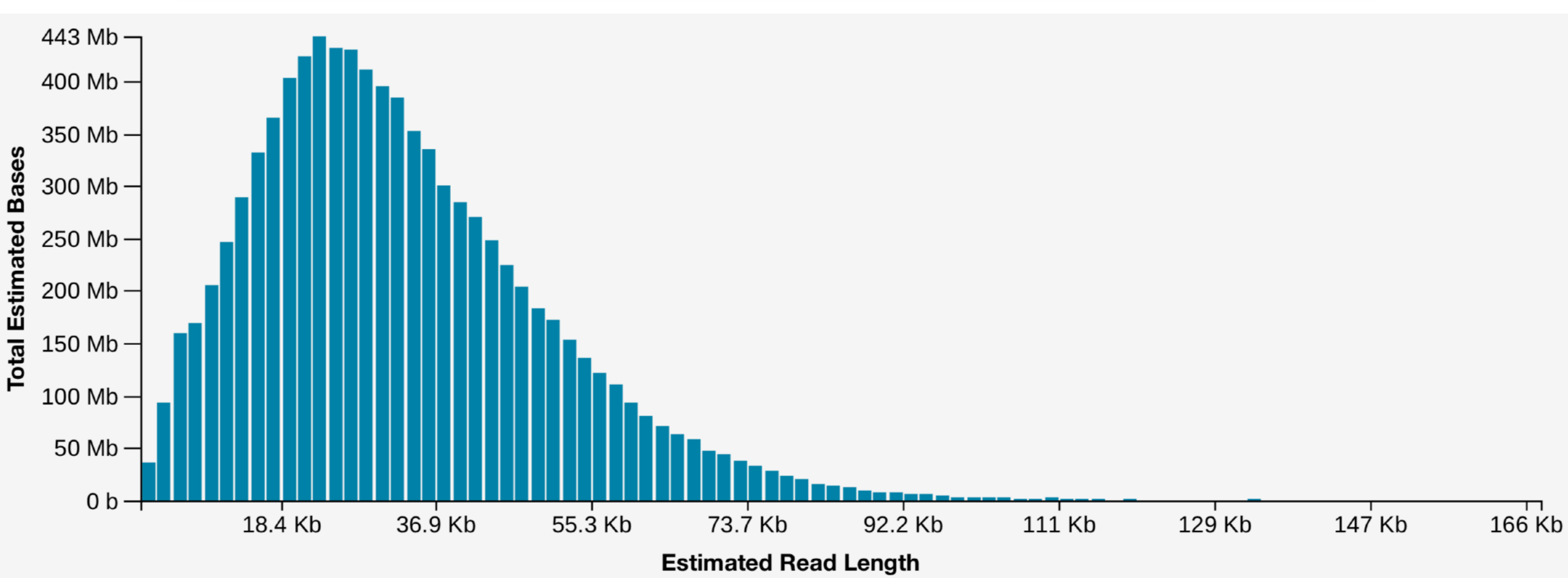
Wolniak et al 2017 Methods in Cell Biology



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	**25.8 Kb**	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	**3.0 Mb (n=97)**	276 Kb (n=485)	**21.2 Mb**	**97.1%**

Hi-C scaffolding

Comparative Genomics/Transcriptomics

Rahmatpour, Nasim [1], Kuo, Li-Yaung [2], Kang, Jessica [3], Herman, Elaina [3], Zhou, Junhui [3], Wolniak, Stephen [3], Zipper, Richard [4], Delwiche, Charles [5], Mount, Stephan M [6], Li, Fay-Wei [7].

A chromosome-scale genome assembly of the fern *Marsilea vestita*.

To date, only two fern genomes have been published,neither of which was assembled at the chromosomal level. With such a limited genomic resource, it has been difficult to gain a holistic view of the fern genome space. Here, we report the genome assembly and annotation of *Marsilea vestita* (Marsileaceae), which has been a model for male gametophyte development, including posttranscriptional regulation and intron retention. Using a combination of nanopore long reads, Illumina short reads, and high-throughput chromosome conformation capture (Hi-C), we assembled the genome into 20 chromosomes. We annotated the genome based on protein evidence and 18 RNA-seq libraries from leaf tissues, as well as different time intervals of spermatogenesis and treatments. Our data provides a much-needed basis for future comparative genomic studies to understand the history of whole genome duplication, gene family evolution, RNA processing and the dynamics of transposable elements.

Log in to add this item to your schedule

1 - Boyce Thompson Institute, 533 Tower Road, Ithaca, NY, 14850, USA
2 - National Tsing Hua University, Institute of Molecular & Cellular Biology, Hsinchu, Taiwan
3 - Department of Cell Biology and Molecular Genetics, University of Maryland, Maryland, USA
4 - Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, USA
5 - University Of Maryland, CELL BIOLOGY AND MOLECULAR GENETICS, Biosciences Research Building, 4066 Campus Drive, College Park, MD, 20742, United States
6 - Institute of Molecular & Cellular Biology, National Tsing Hua University, Hsinchu, , Taiwan
7 - Boyce Thompson Institute, 533 Tower Rd, Ithaca, NY, 14853, United States

Keywords:
Marsilea vestita
chromosome number
genome evolution.

Presentation Type: Oral Paper
Session: CGT3, Comparative Genomics/Transcriptomics III
Location: /
Date: Tuesday, July 20th, 2021
Time: 4:00 PM(EDT)
Number: CGT3005
Abstract ID:*864*

Nasim Rahmatpour

$4,875

$1,950 Nanopore seq
$1,325 Illumina seq
$1,600 Hi-C lib + Illumina seq