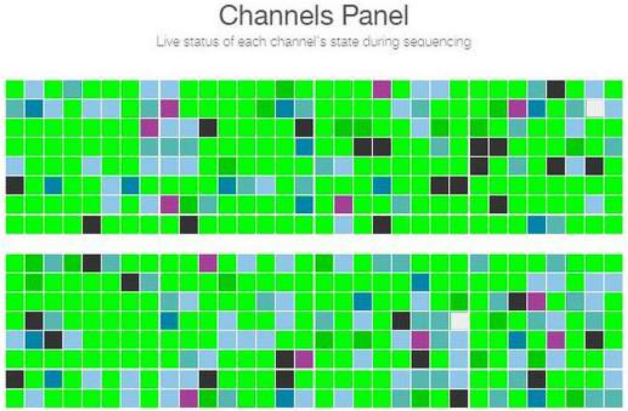
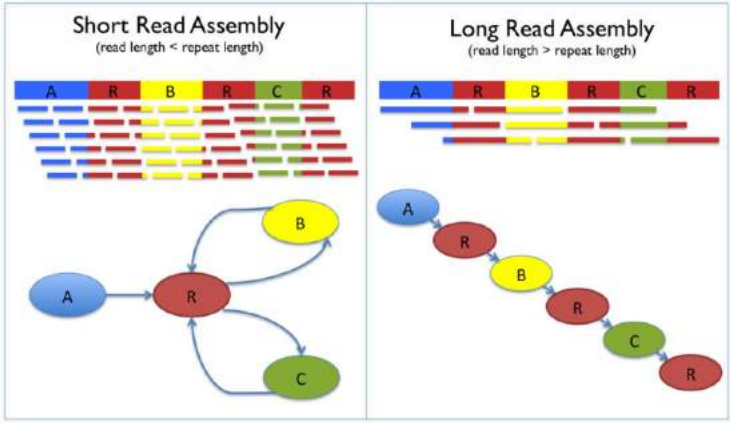


PACBIO®

Experimental design and obtaining DNA for long-read sequencing

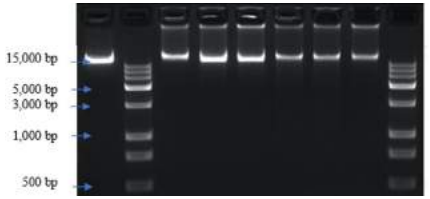


Outline

- Select your species and/or individual

- Choose sequencing platform

- Generate high quality DNA



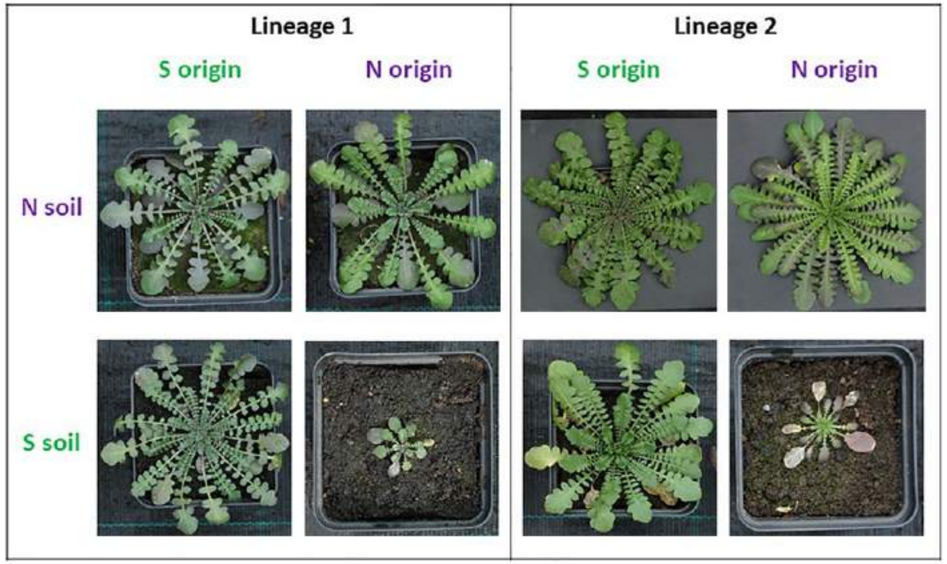
- Some examples

What do all these things have in common?

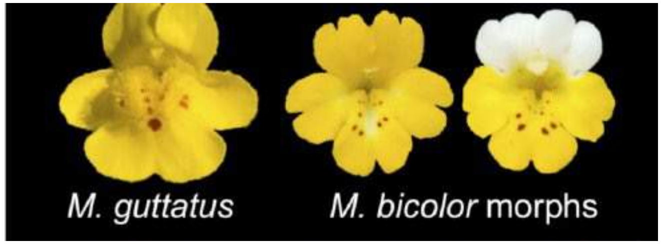
Shifts in pollinators

Evolution of carnivorous plants

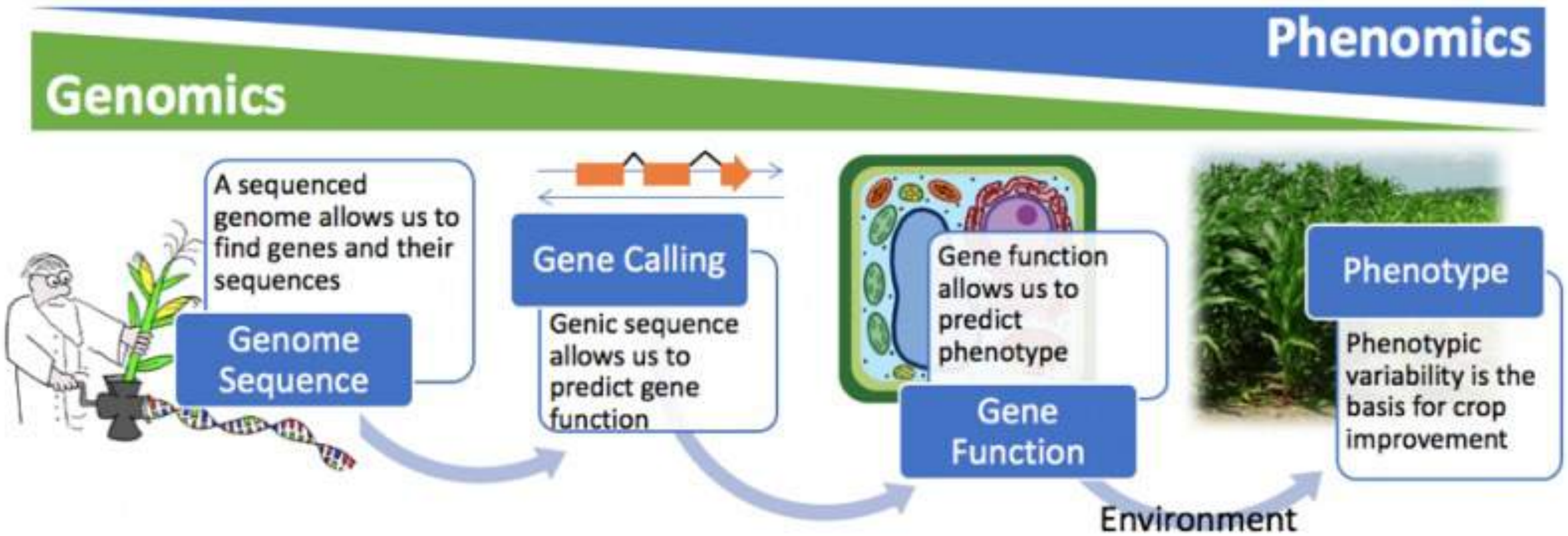
Response to nutrients



Flower color morphs



Genome as a puzzle

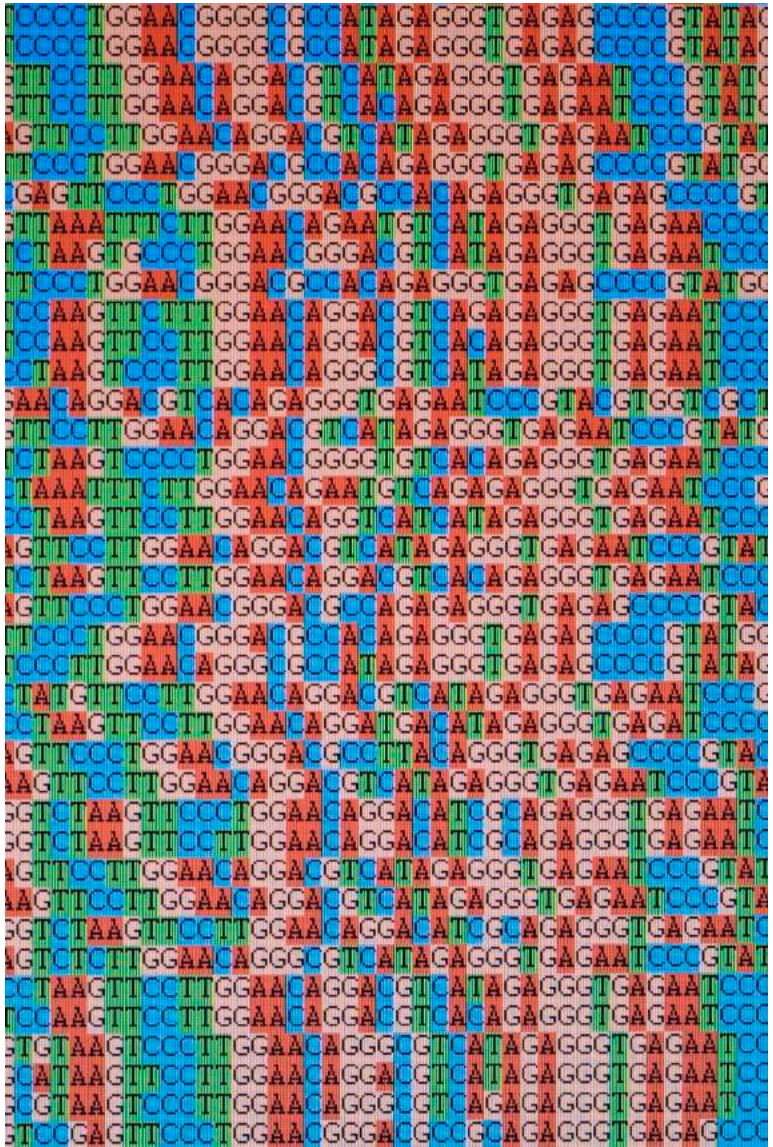


A sequenced genome allows us to find genes and their sequences

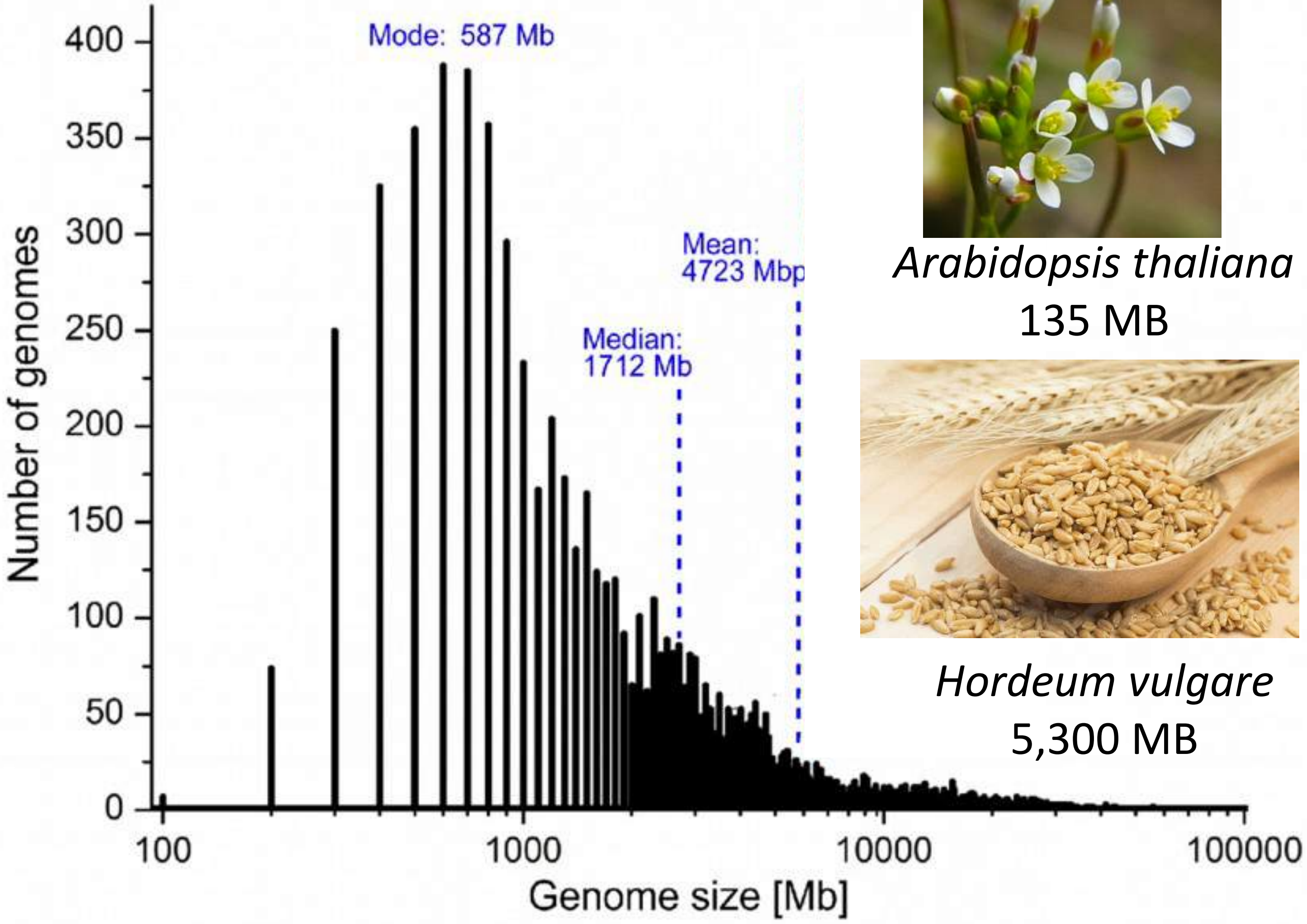
Genic sequence allows us to predict gene function

Gene function allows us to predict phenotype

Phenotypic variability is the basis for crop improvement



Genome size of angiosperms



Arabidopsis thaliana
135 MB

Oryza sativa
430 MB

Zingiber officinale
1,582 MB

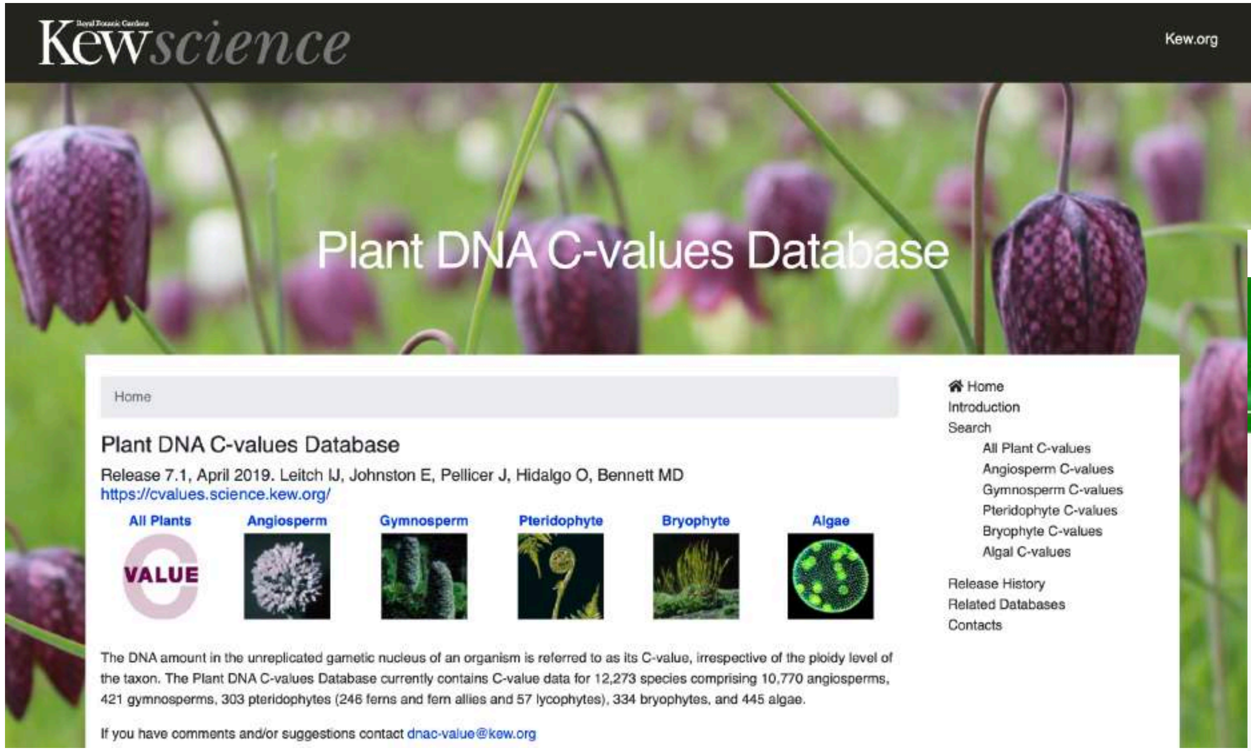
Hordeum vulgare
5,300 MB

Allium cepa
16,000 MB

Tulipa sylvestris
59,241 MB

Adapted from Wicker et al. 2017 *Mobile DNA*

Chromosome number and ploidy



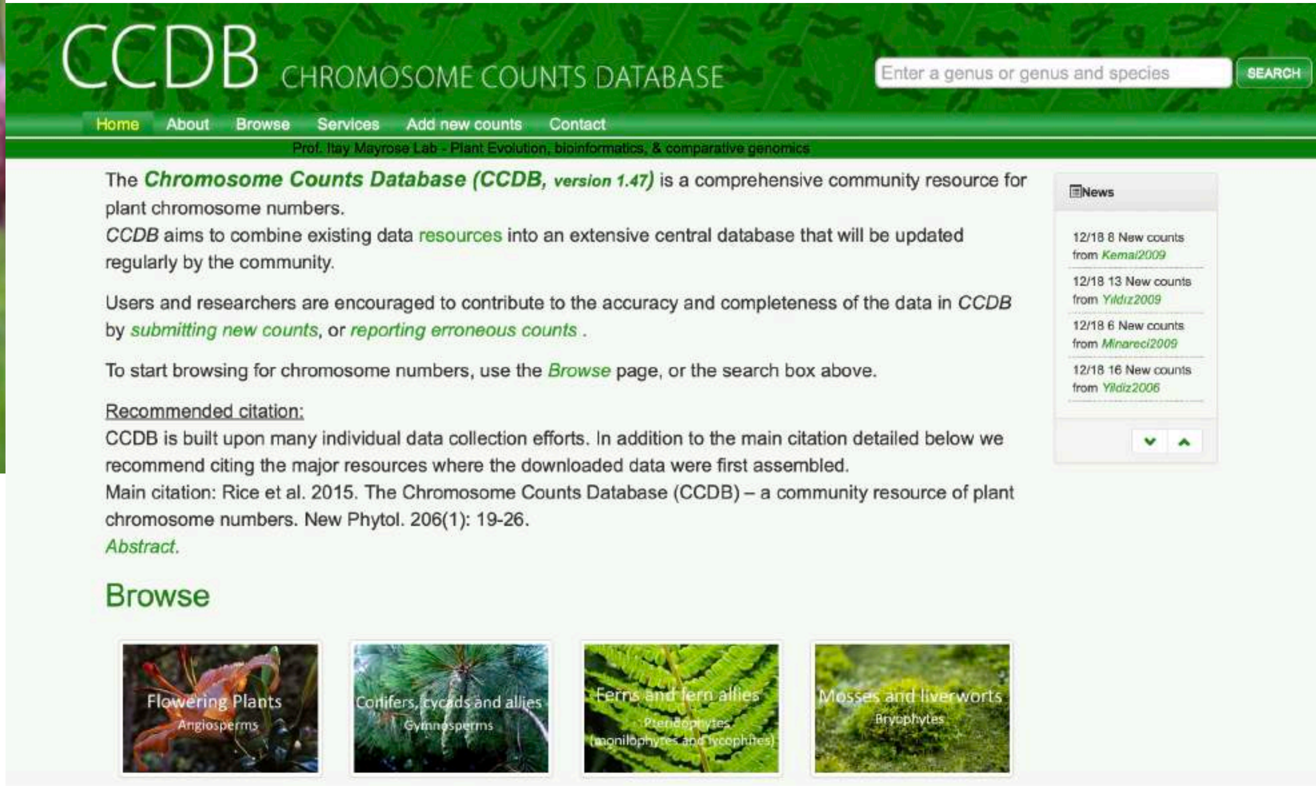
Plant DNA C-values Database

Release 7.1, April 2019. Leitch IJ, Johnston E, Pellicer J, Hidalgo O, Bennett MD
https://cvalues.science.kew.org/

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

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
News
Flowering Plants
Angiosperms
Conifers, cycads and allies
Gymnosperms
Ferns and fern allies
Pteridophytes
(monilophytes and lycophytes)
Mosses and liverworts
Bryophytes

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.

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*

Browse

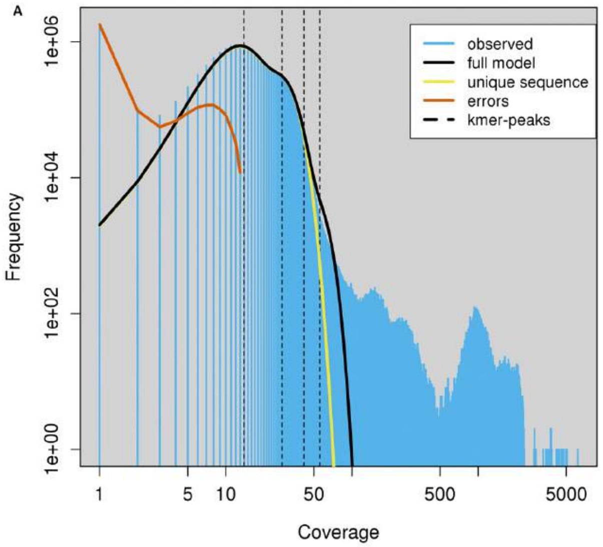
Genome size estimate

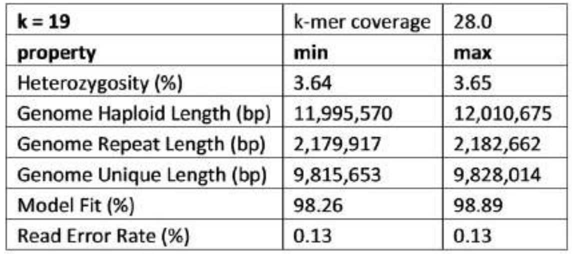
- For many nonmodel systems, there are no entries in the Kew database
    - Even if the genus is there, genome size can vary between species

Kmer (estimates)

- Cleaned Illumina data

- Jellyfish paired with GenomeScope or RESPECT



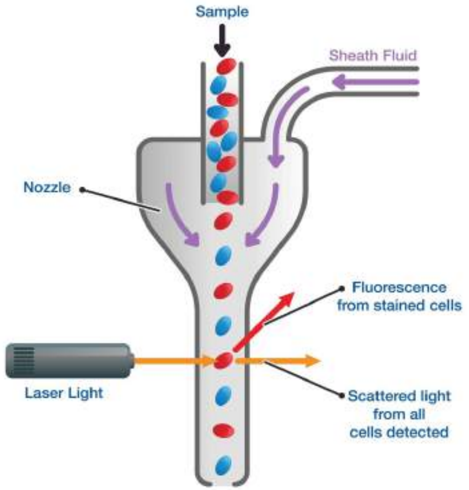


k = 19	k-mer coverage	28.0
property	**min**	**max**
Heterozygosity (%)	3.64	3.65
Genome Haploid Length (bp)	11,995,570	12,010,675
Genome Repeat Length (bp)	2,179,917	2,182,662
Genome Unique Length (bp)	9,815,653	9,828,014
Model Fit (%)	98.26	98.89
Read Error Rate (%)	0.13	0.13

Flow cytometry (more reliable)

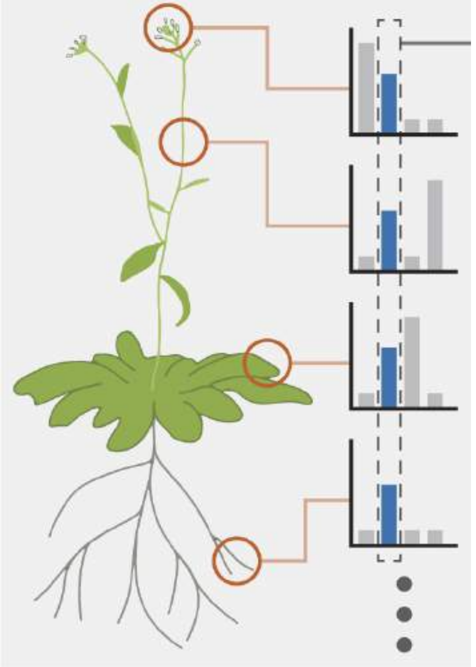
- Fresh or silica dried matieral; protocols can vary a lot between species

- Need accurate references to compare


Sample
Sheath Fluid
Nozzle
Fluorescence from stained cells
Laser Light
Scattered light from all cells detected

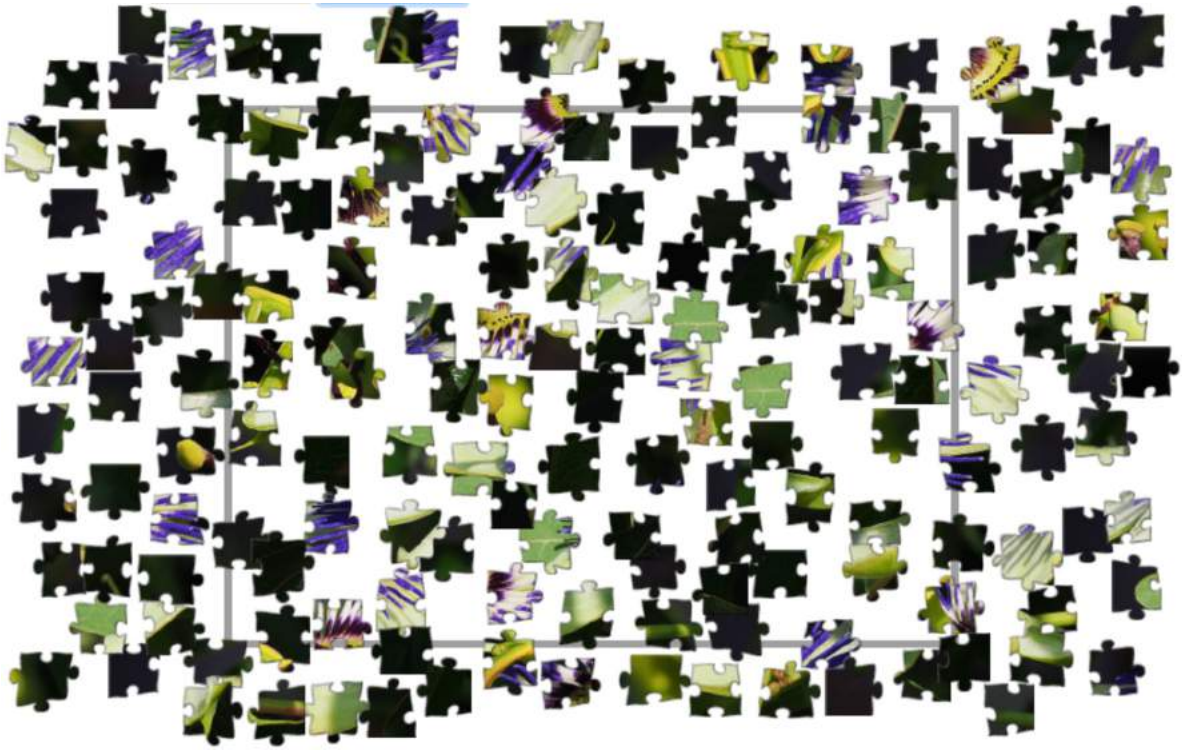
Ideal scenario for sample selection

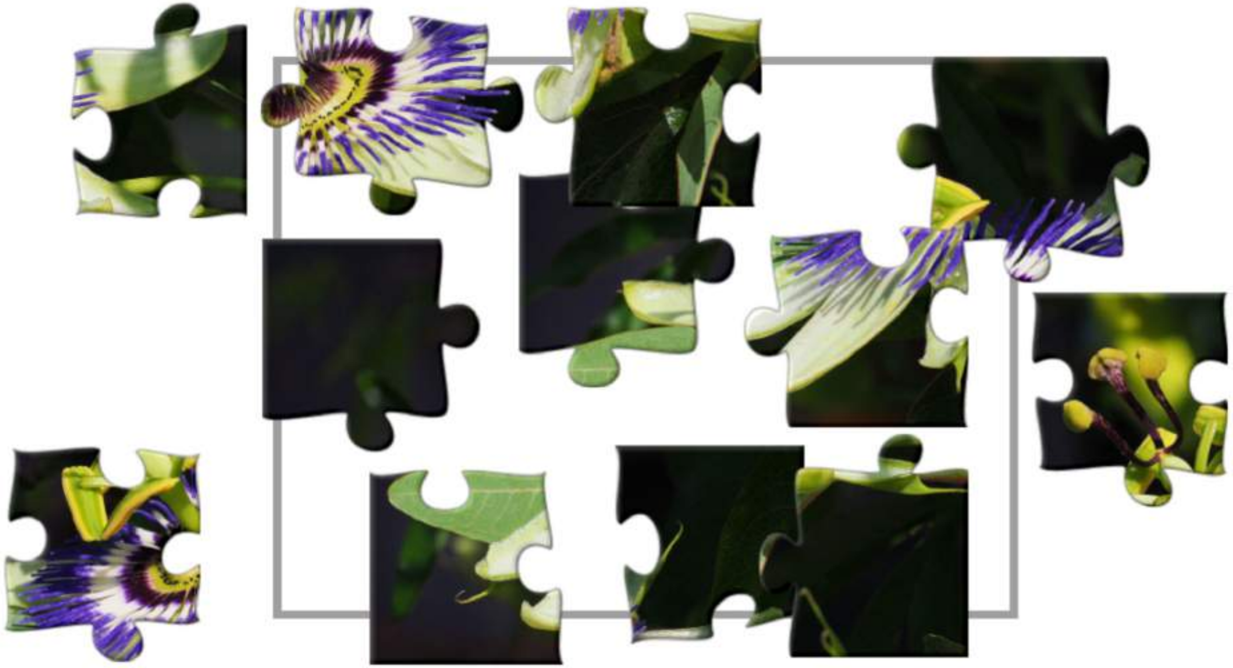
- Individual with lots of fresh material available
  - Single induvial for sequencing and assembly
  - Scaffolding and/or annotation material can come from different individuals/species
  - Generate large amounts of high molecular weight DNA (often multiple micrograms)
  - RNA from multiple tissues and/or developmental stages
  - Fresh tissue for Hi-C sequencing
- ”Clean” – less exposure to microorganisms or other organisms
  - If others are sequenced (including yourself), won’t scaffold and be annotated



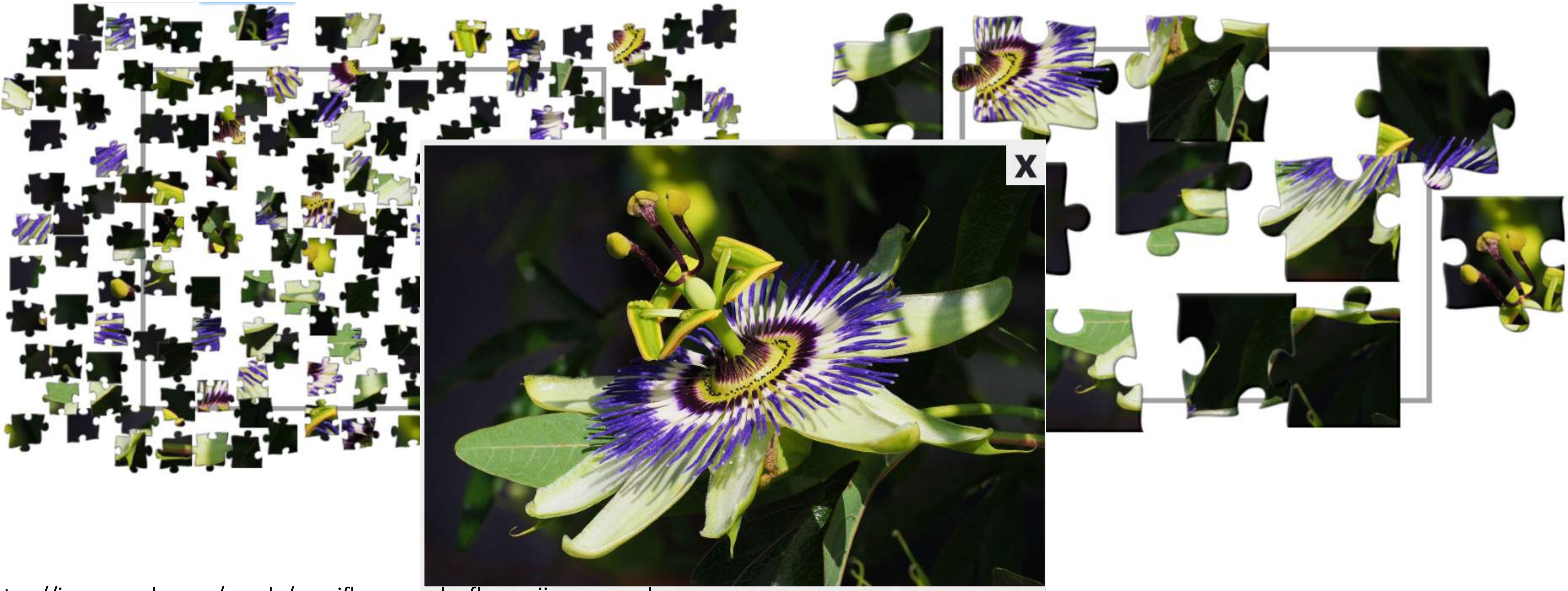
Obtaining Sequencing data

Which puzzle is easier to put together?



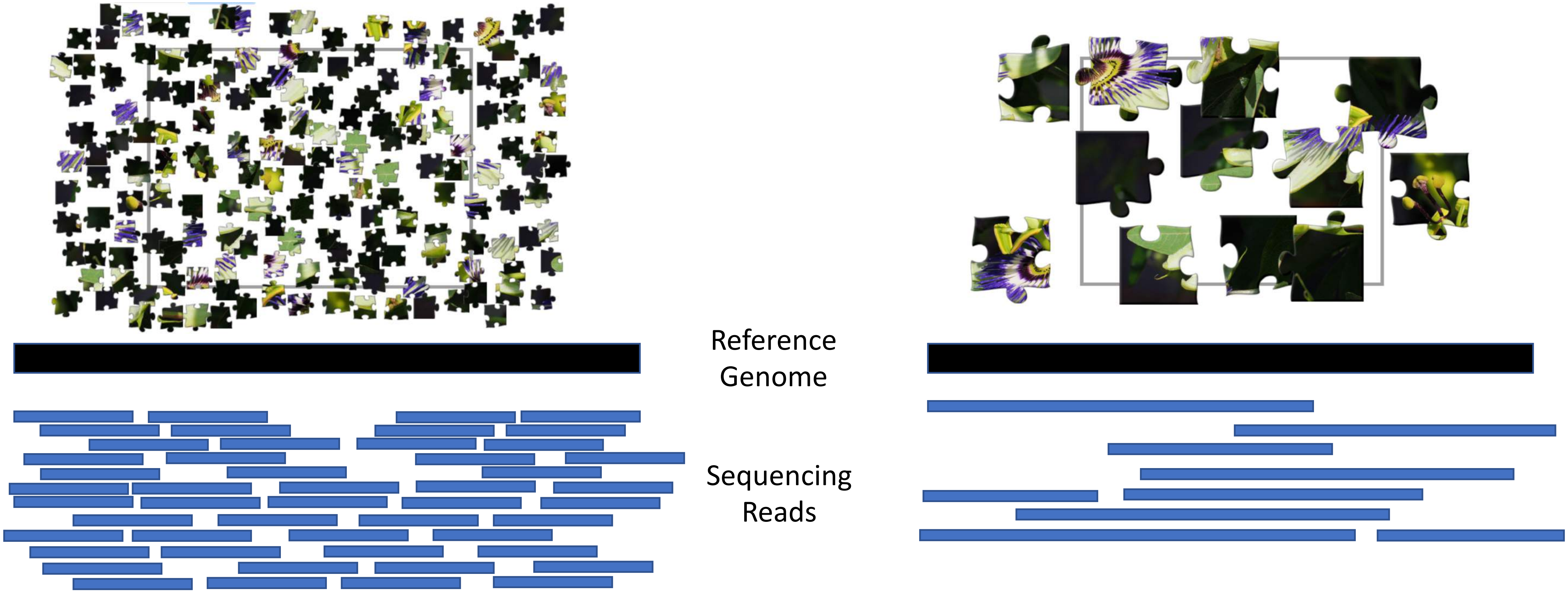


Which puzzle is easier to put together?



https://im-a-puzzle.com/puzzle/passifloracaeruleaflower-jigsaw-puzzle

In the world of genomes

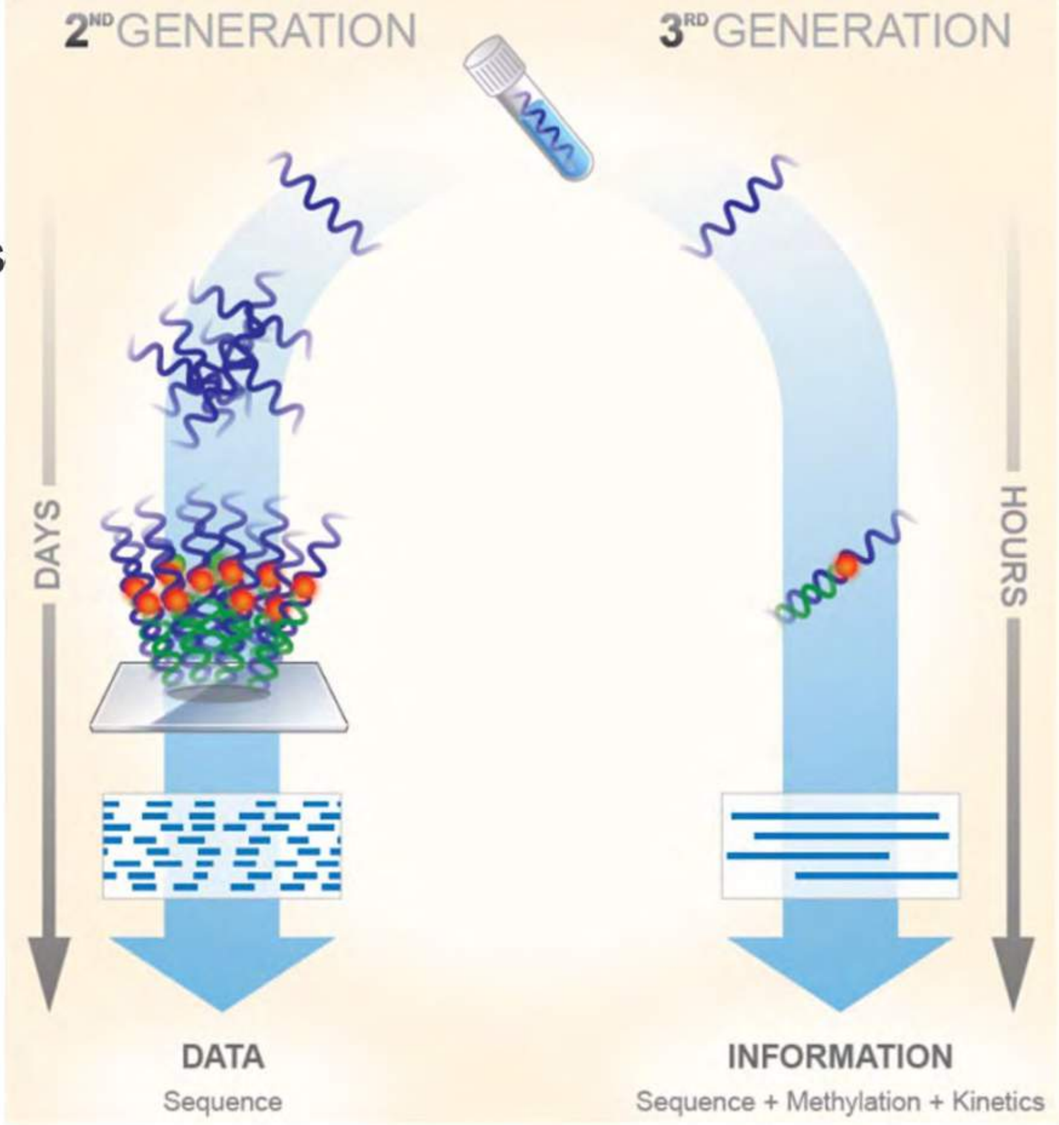


Reference Genome

Sequencing Reads

Short vs Long-reads

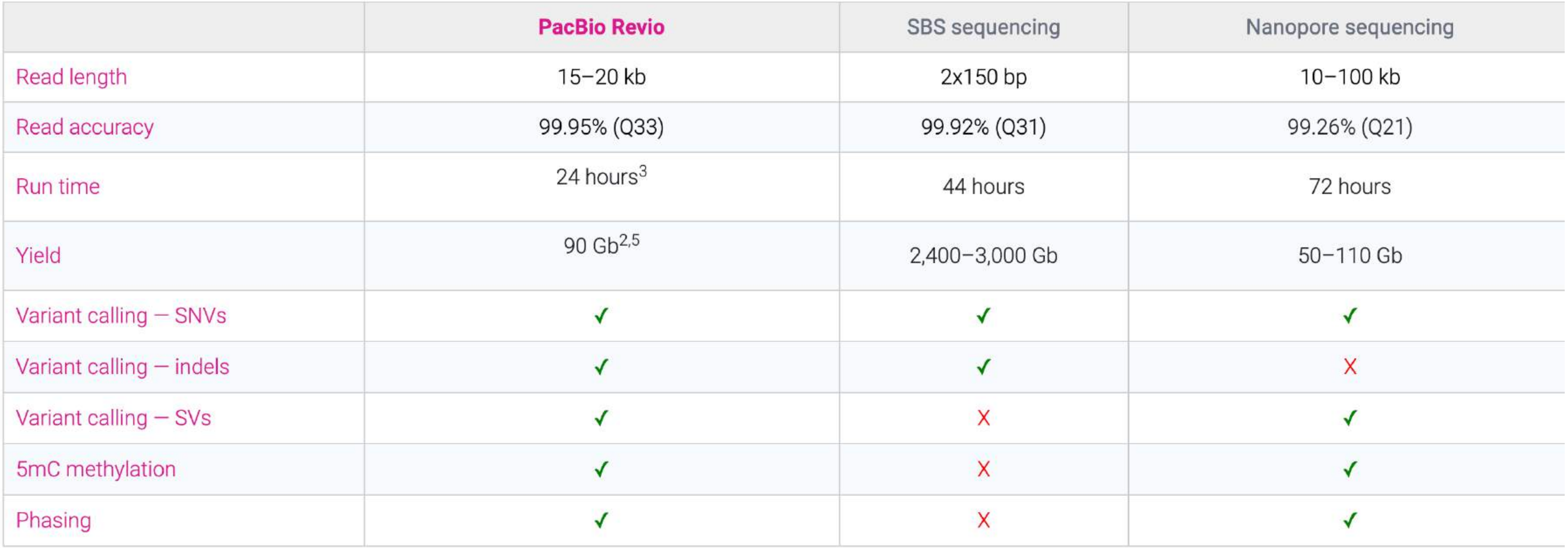
- Short reads
- Amplification errors and bias
- Several enzymatic steps
- Multi-molecule raw accuracy
- Errors tend to be systematic
- More coverage required



- Long reads
- No required amplification
- Simple sample prep
- Single molecule raw accuracy
- Errors tend to be random (vs. systematic)
- Less coverage required

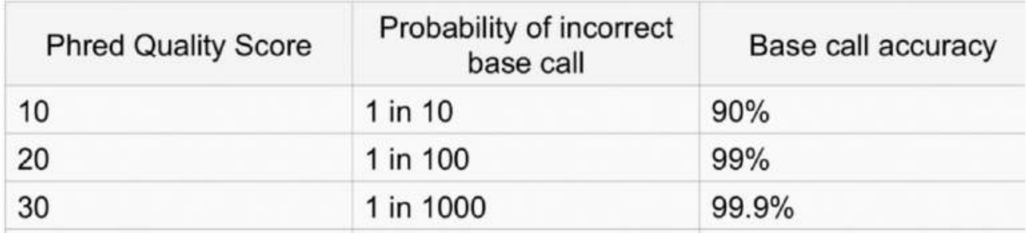
https://get.genotoul.fr/wp-content/uploads/2017/06/150519_Technology_and_Applications_PACBIO_GerritKuhn_low.pdf

Long-read options



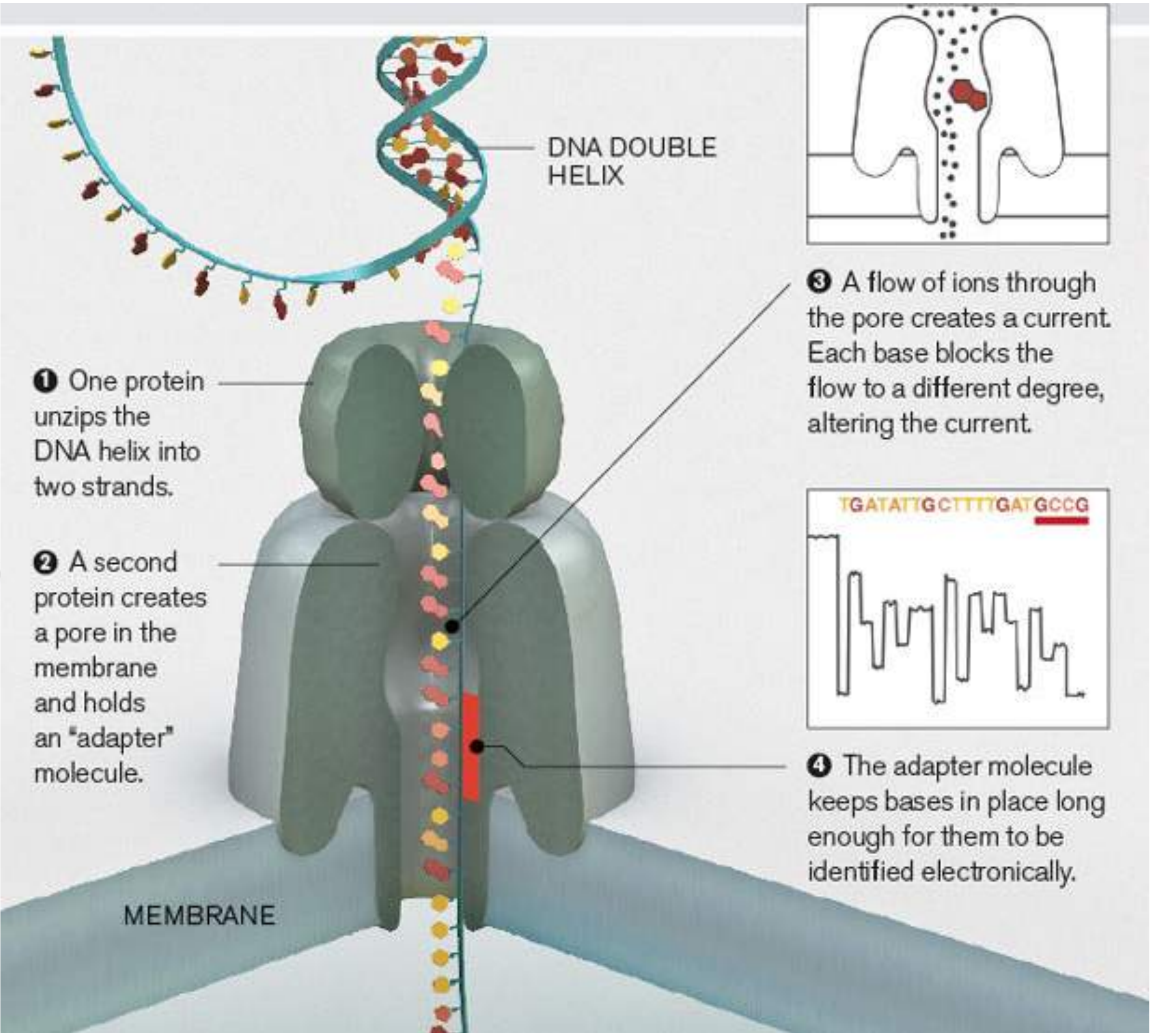
	PacBio Revio	SBS sequencing	Nanopore sequencing
Read length	15–20 kb	2x150 bp	10–100 kb
Read accuracy	99.95% (Q33)	99.92% (Q31)	99.26% (Q21)
Run time	24 hours[3]	44 hours	72 hours
Yield	90 Gb[2,5]	2,400–3,000 Gb	50–110 Gb
Variant calling – SNVs	✓	✓	✓
Variant calling – indels	✓	✓	X
Variant calling – SVs	✓	X	✓
5mC methylation	✓	X	✓
Phasing	✓	X	✓

HiFi targets 10 kbp, while Nanopore works “best” around 10-20 kbp (“best” can vary if fragment size or output is most desired)



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Nanopore



❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

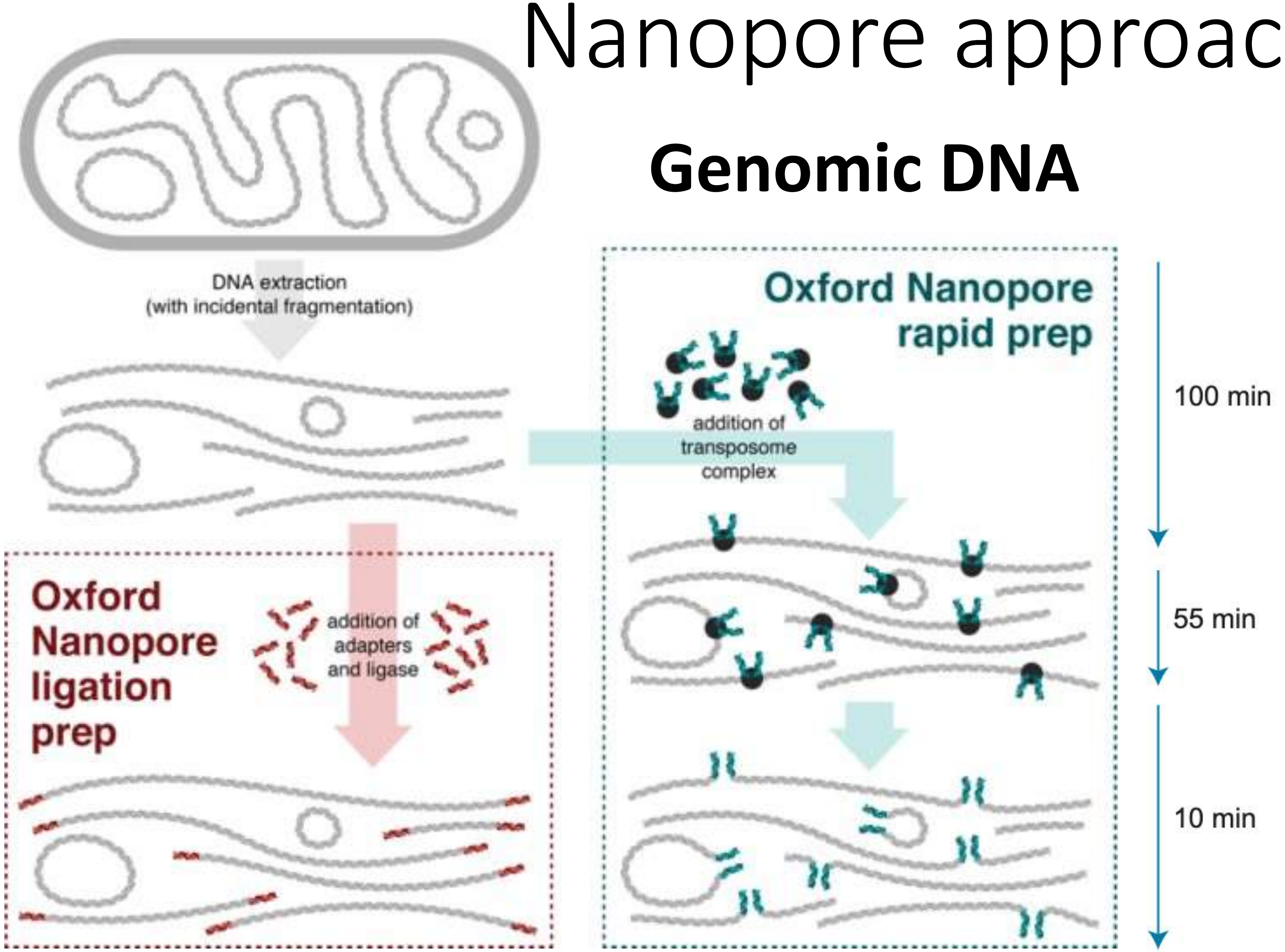
❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

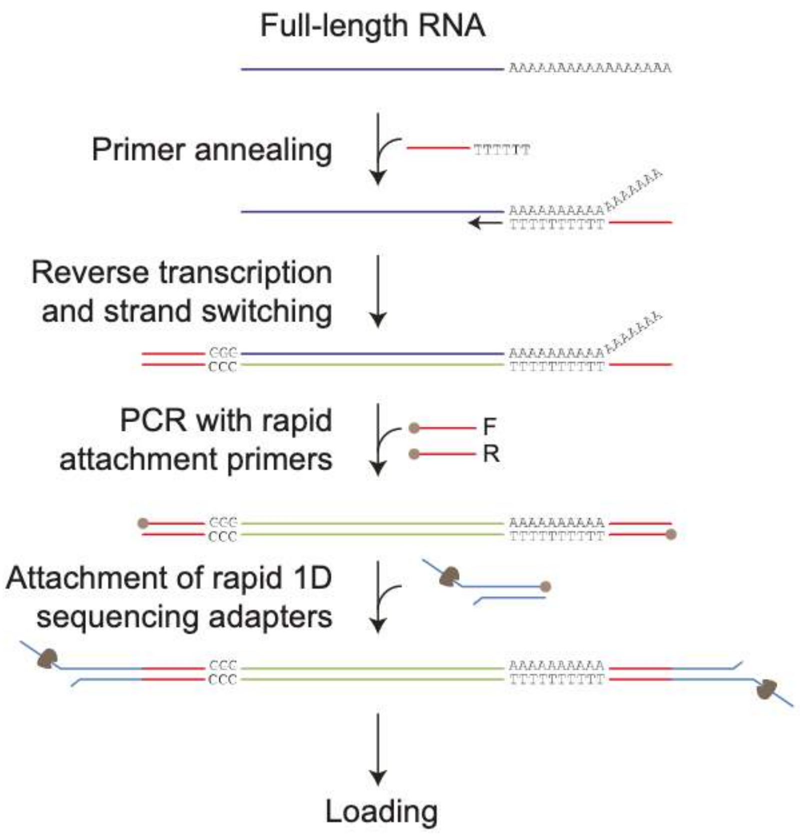
https://finlaymagui.re/presentations/robot_revolution.html

Nanopore approaches

Genomic DNA


DNA extraction
(with incidental fragmentation)
Oxford Nanopore ligation prep
addition of adapters and ligase
Oxford Nanopore rapid prep
addition of transposome complex
100 min
55 min
10 min

RNA

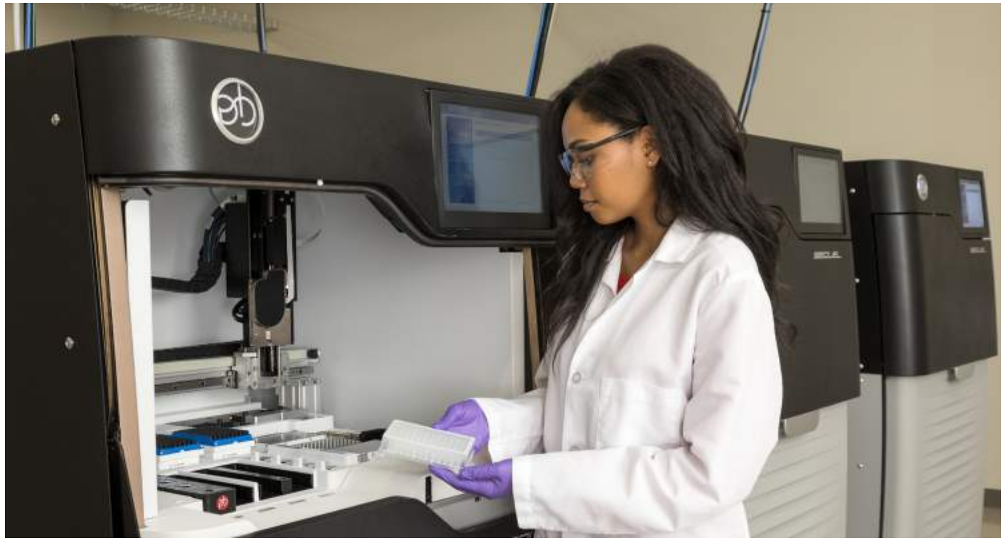


Library Prep	~2.5 hours	10 minutes	165 minutes
Output	10-20 GB in 96 hours	8-10 GB in 96 hours	5 – 7 Million transcripts in 48 hours

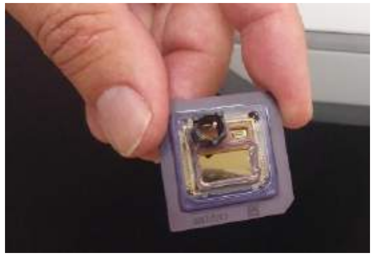
Getting started with Nanopore MinION

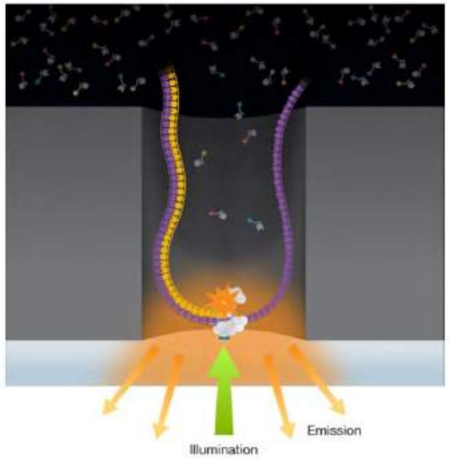
- Starter pack - $1,999
  - Two 10.4.1 flow cells and 1 6-reaction library kit
- Flow cells – typically guaranteed to last 3 months or less in the fridge
  - Individually $700 each
  - 24 bundle $500 each
  - 48 bundle $475
- Library prep is ~$150 per sample (kit is $100/sample + extra reagents)
  - 1.5-2 hour prep time, need 1 ug starting DNA
- PromethION is $900 per flow cell, similar prep method, with increased sequencing output

PacBio

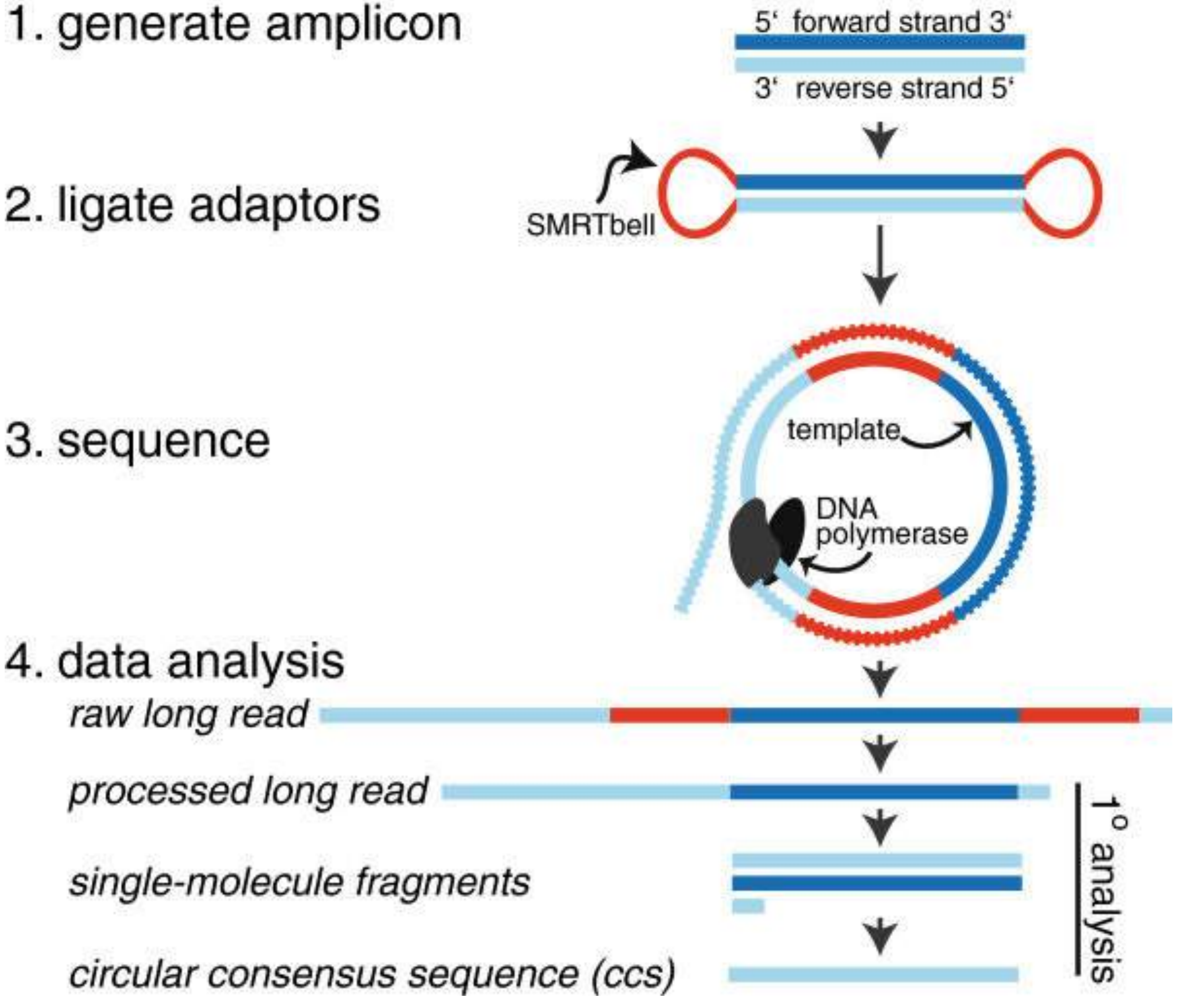


SMRT Cell





Science, Vol 299, Jan 31 2003, pp682-686
J. Appl. Phys. 103, 034301 (2008)



1. generate amplicon
2. ligate adaptors
3. sequence
4. data analysis

raw long read

processed long read

single-molecule fragments

circular consensus sequence (ccs)

Fichot and Norman 2013; *Microbiome*

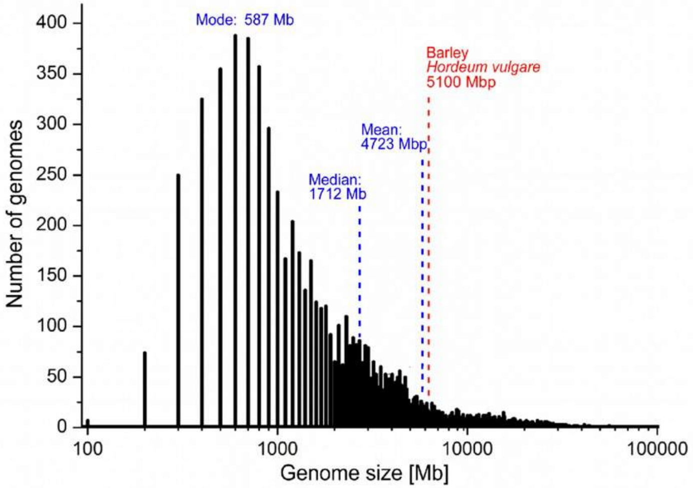
Revio

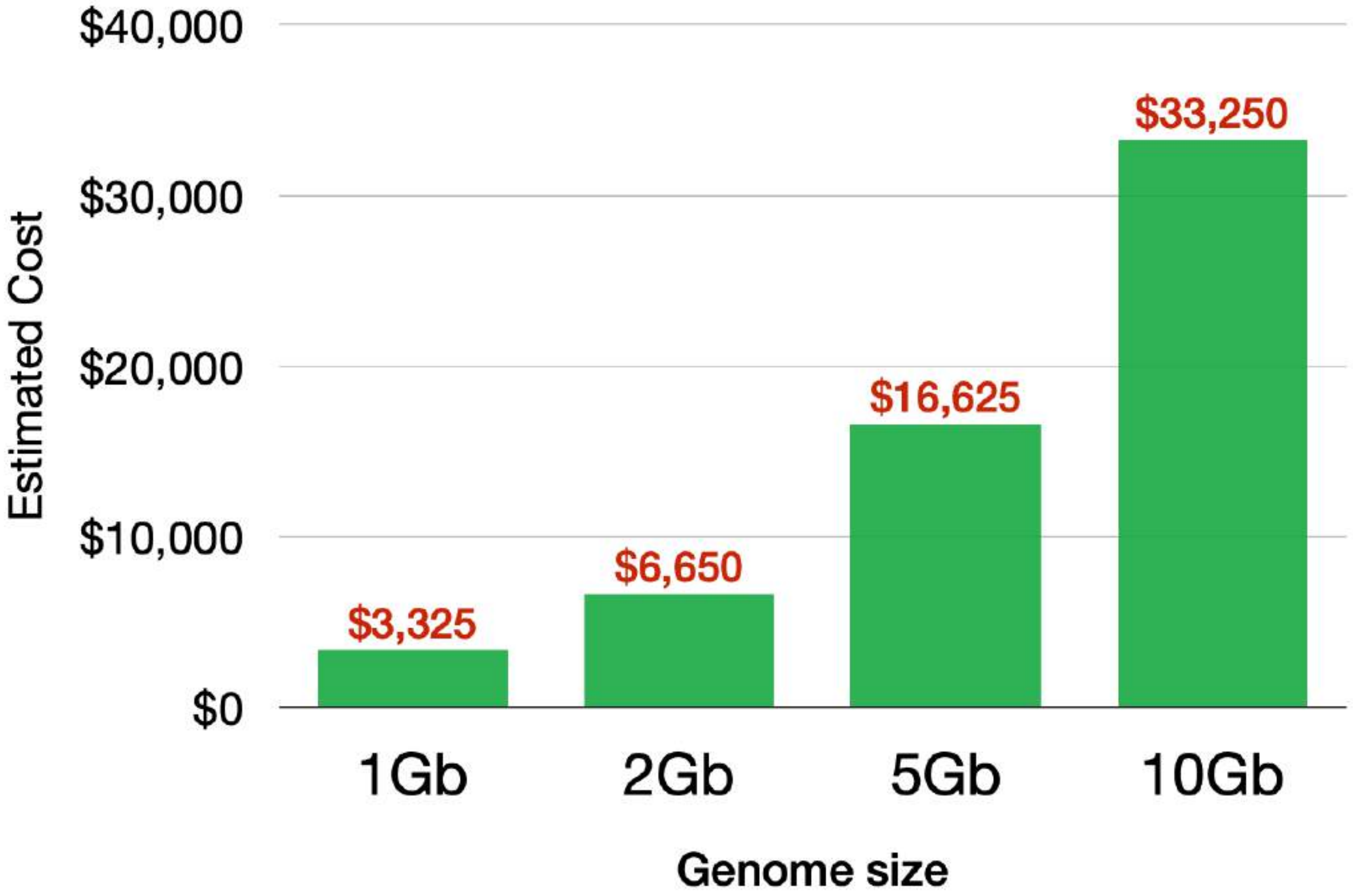
- 90% of bases ≥Q30 and median read accuracy ≥Q30
- 15x increase in throughput over the Sequel II system
- Little less than 100 GB per SMRT cell
    - If DNA fragments are less than 10 KB, total output drops
- HiFi sequencing provides structural variants, repeat expansions, methylation, and haplotype phasing from a single library
    - **The $1000 complete, phased genome**
- Typically outsourced as opposed to what can be done inhouse with minION

Cost of sequencing genomes

- 50X Illumina:
  - 50Gb x $26.5/Gb = **$1,325**
- 50X nanopore:
  - 50Gb x $40/Gb = **$2,000**

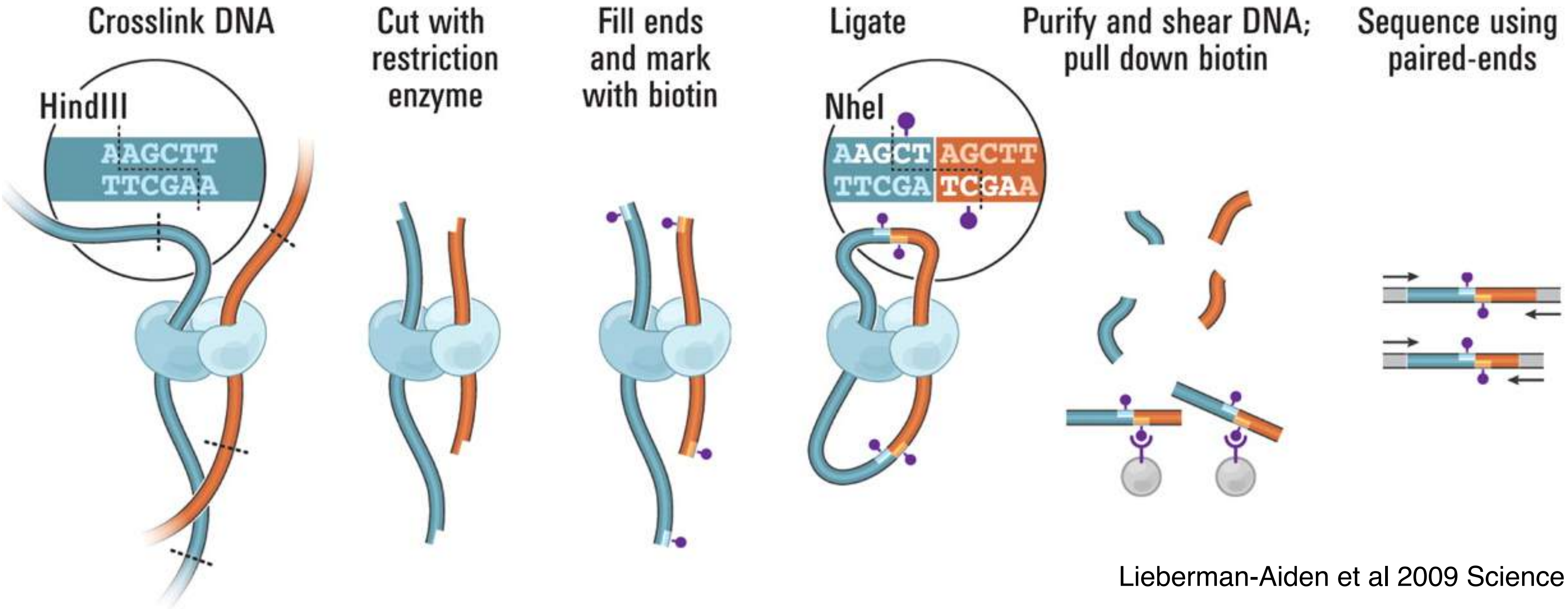
$3,325





Hi-C

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


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Lieberman-Aiden et al 2009 Science

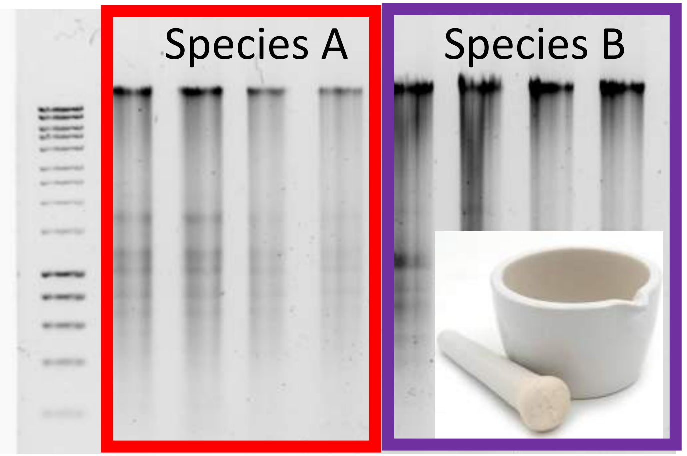
Scaffolding (1 GB genome)

- Hi-C Library kit $500 + 50x Illumina $550 = **$1,050**
  - Library prep is not trivial, two-day protocol
  - Comes in sets of two
- Outsource to Phase Genomics
  - Send frozen samples on dry ice
  - Library prep $1,500 + 150 M Read-pair Illumina $750 = **$2,250**
  - Guaranteed to get usable data
- Arima Genomics has a 6 hour rapid protocol
- Optical mapping by Bionano
  - Outsource (HWM extraction + Saphyr chip + analysis) = **~$3,000**

Extracting good DNA

Tips for nonmodel systems

Tissue Grinding

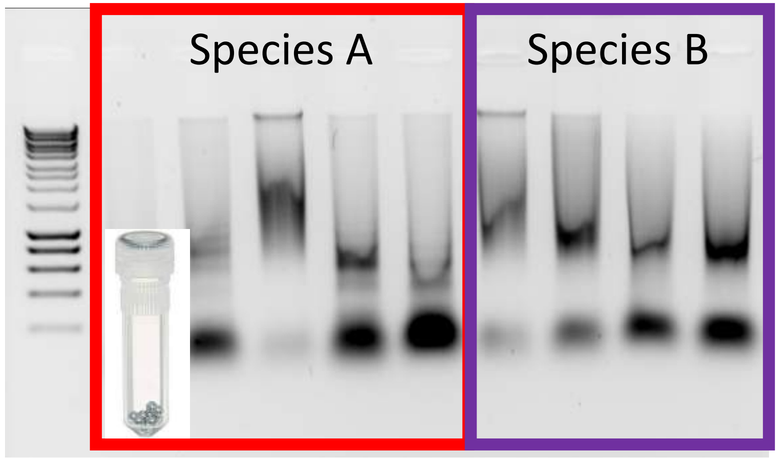


Mortar and pestle

- Better yield, larger fragments
- Nanopore flow cell generated 18.5 GB with an N50 of 6.5 kb

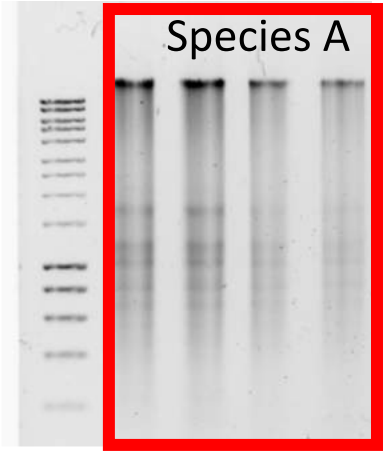
Grinding beads

- Much lower yield; highly fragmented DNA
- Nanopore flow cell generated 12 GB with an N50 of 4.2 kb

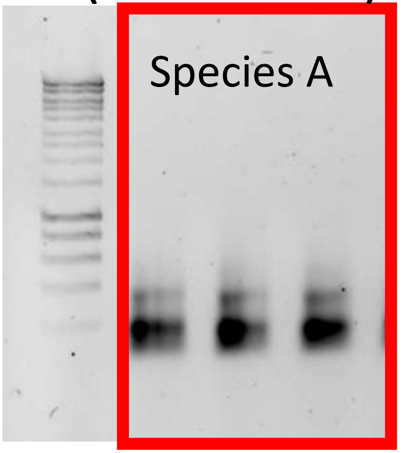


Extraction method

SDS
(Monocots)



CTAB
(Monocots)

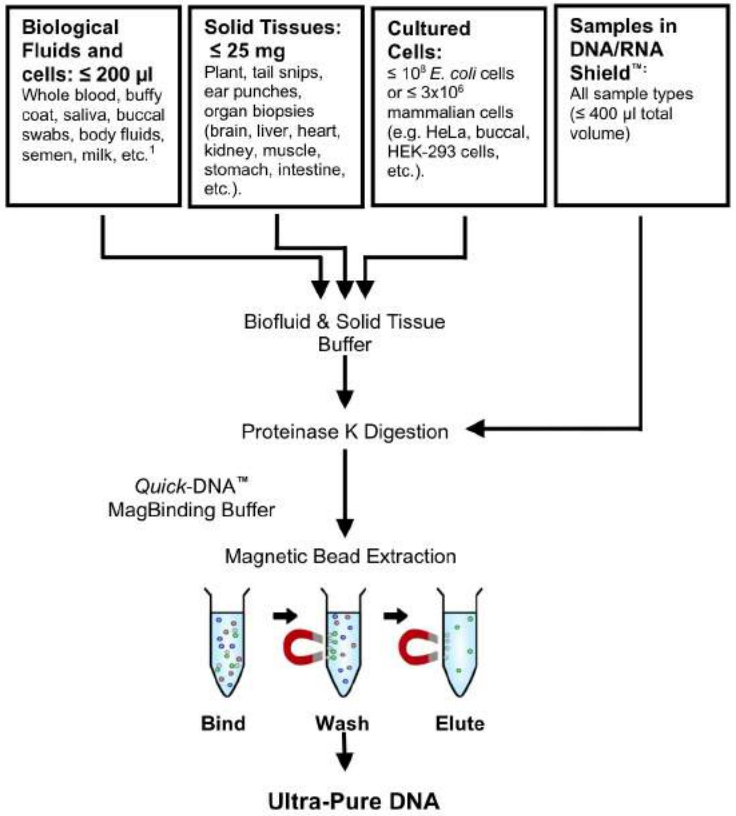


Modified SDS for Monocots
Modified CTAB for Eudicots

De La Cerda et al 2023; *APPS*

Kit based approaches

- Several options available
- Zymo's is supposed to work in 45 minutes



Biological Fluids and cells: ≤ 200 µl Whole blood, buffy coat, saliva, buccal swabs, body fluids, semen, milk, etc.[1]

Solid Tissues: ≤ 25 mg Plant, tail snips, ear punches, organ biopsies (brain, liver, heart, kidney, muscle, stomach, intestine, etc.).

Cultured Cells: ≤ 10^8 *E. coli* cells or ≤ $3x10^6$ mammalian cells (e.g. HeLa, buccal, HEK-293 cells, etc.).

Samples in DNA/RNA Shield™: All sample types (≤ 400 µl total volume)

Biofluid & Solid Tissue Buffer

Proteinase K Digestion

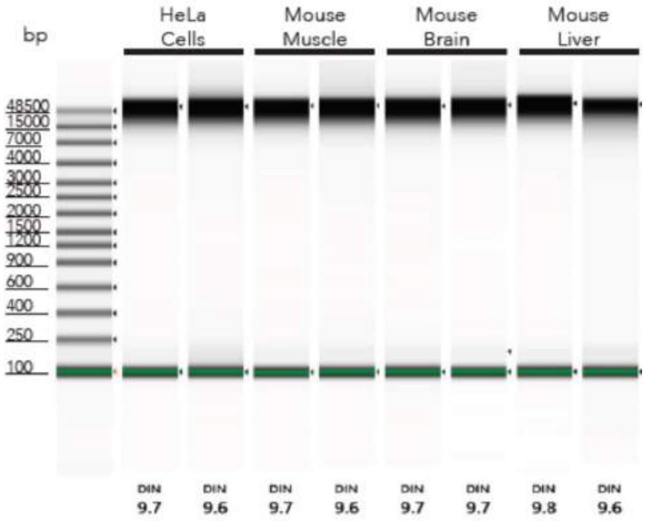
Quick-DNA™ MagBinding Buffer

Magnetic Bead Extraction

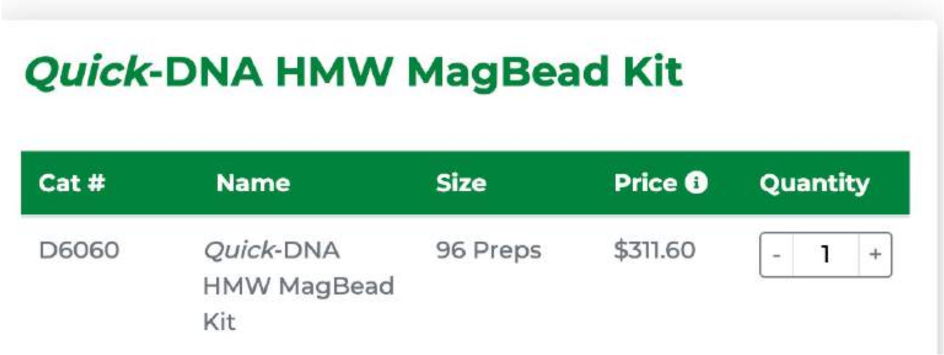
Bind **Wash** **Elute**

Ultra-Pure DNA

DNA up to 150+ kb



Quick-DNA HMW Magbead Kit
ZYMO RESEARCH
The Beauty of Science is to Make Things Simple

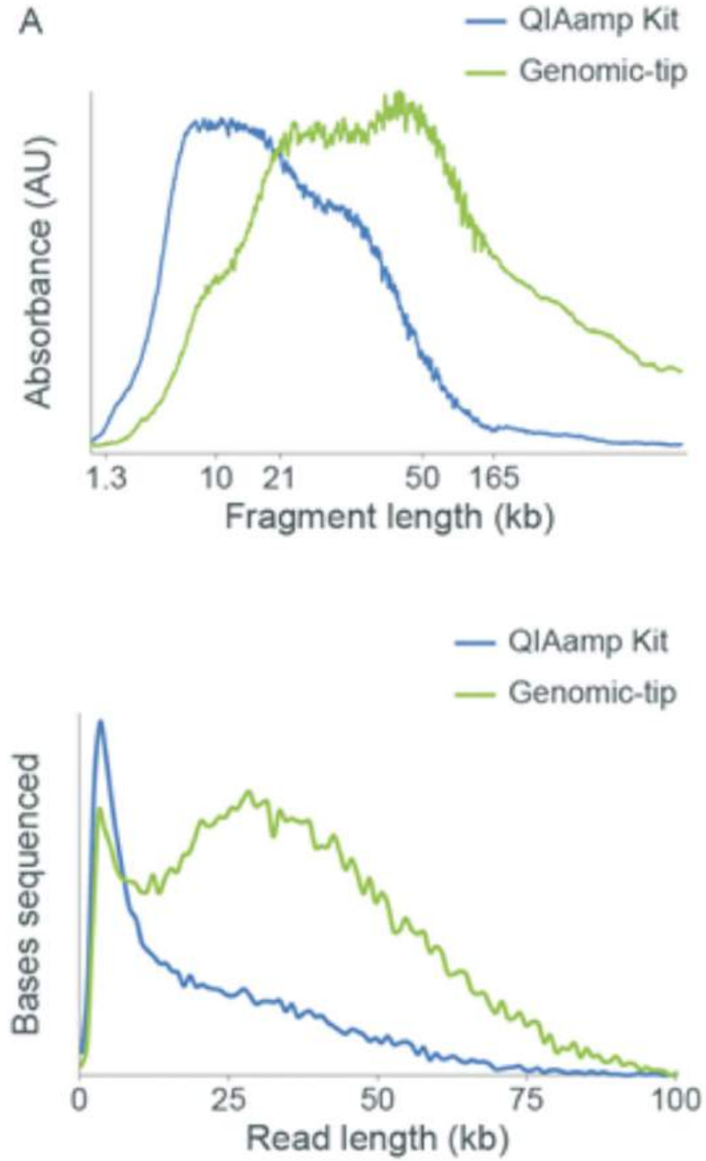


Quick-DNA HMW MagBead Kit

Cat #	Name	Size	Price	Quantity
D6060	*Quick*-DNA HMW MagBead Kit	96 Preps	$311.60	- 1 +

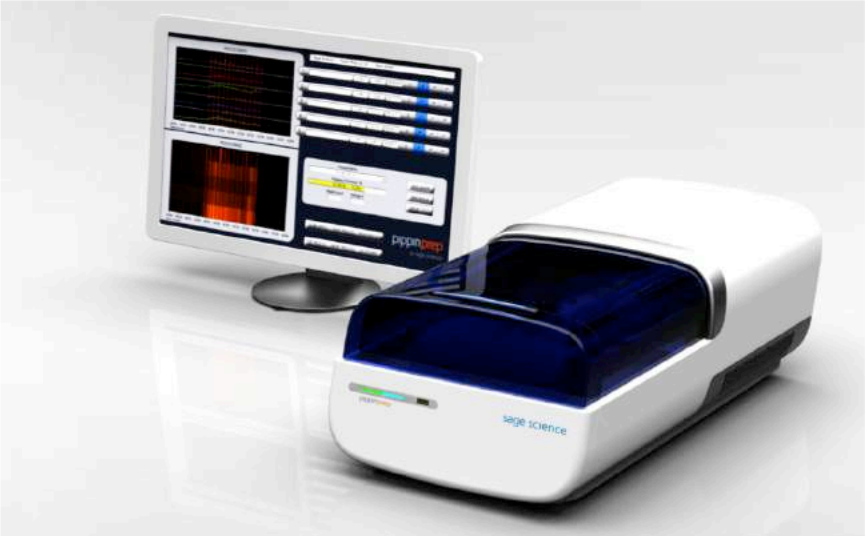
QC

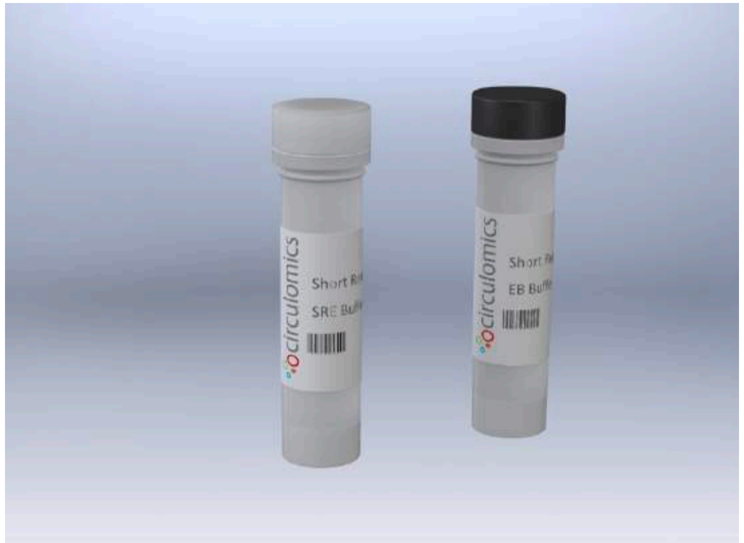
- Quantification – Need to rely on Qubit
  - Nanodrop drastically overestimates concentrations
  - 1 ug for each sequencing run; if size selection need around 5 ug starting out
- Purity/Cleanliness – Nanodrop
  - 260/280 values should be 1.8-2, while 260/230 values 2.0-2.2
  - If pure DNA, concentrations should be close to 1:1 (Nanodrop:Qubit)
- Integrity
  - Bioanalyzer or Femto Pulse
  - Low percentage agarose gel (0.5-1%) with low voltage
  - NEB 1 KB Extend Ladder (top band 48.5 kb)



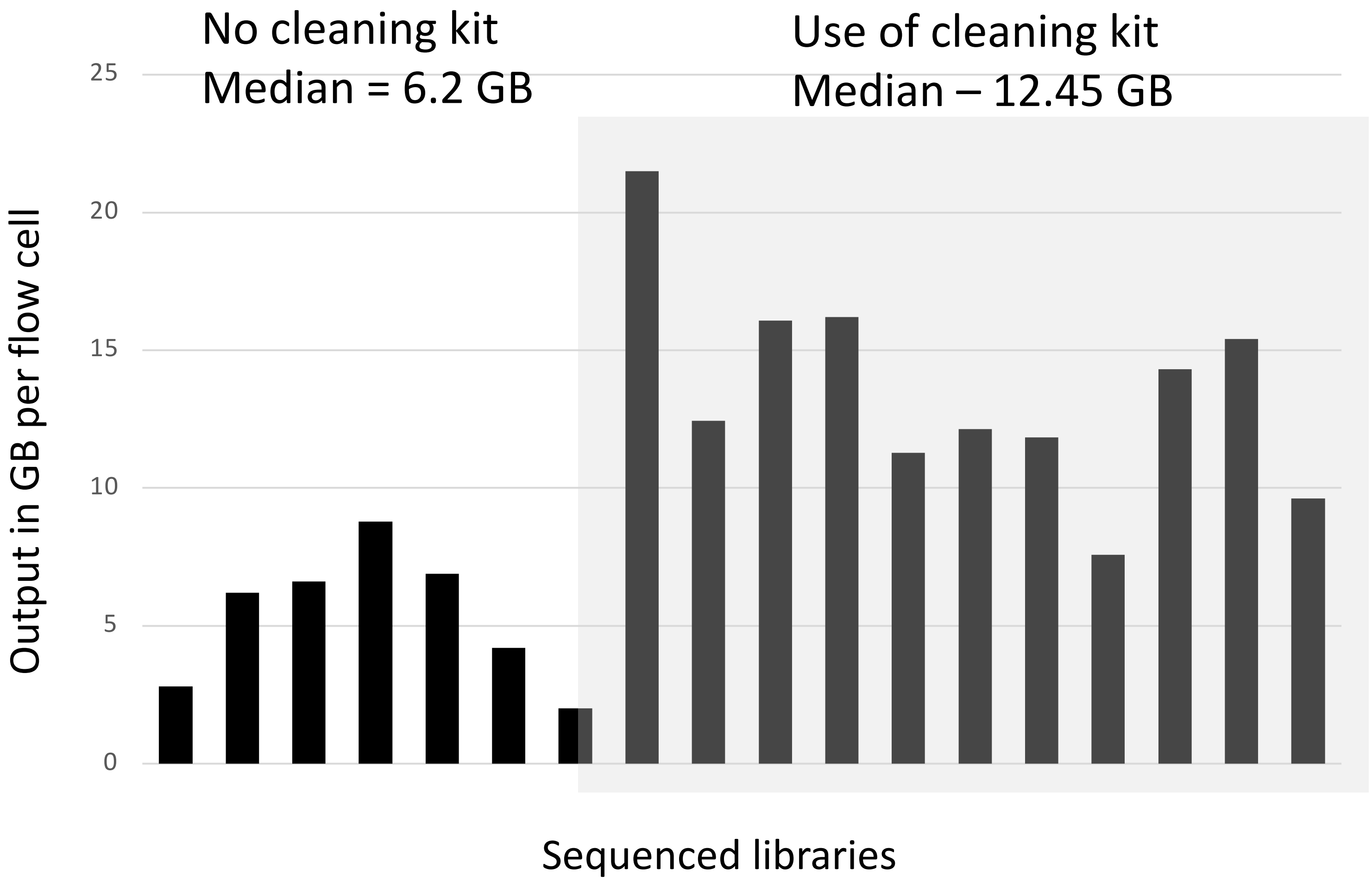
Size selection

- Blue pippin
- Ciculomics Short Read Eliminator Kit
  - Target cutoff size: XS (10 kb), Regular (25 kb), XL (40 kb)
- Some/most DNA will be lost, but what remains is highly valuable
  - Be prepared for around a 40% quantity reduction with each cleaning step





Cleaning the DNA


No cleaning kit
Median = 6.2 GB

Use of cleaning kit
Median – 12.45 GB

- Some species can be very difficult to get pure DNA
  - 1:1 Nanodrop:Qubit
- DNAeasy ProClean kit increases sequencing yield
  - DNA is sheared somewhat

De La Cerda et al 2023; *APPS*

A couple of examples and costs

Small(ish) genome

- Estimated genome size 750 MB – 1 GB
- Two Nanopore flow cells ($1,200)
- 50x Illumina ($570)
- 1 SMRT cell Revio ($2,760)
- 30x Hi-C ($1,880)
- Total: **$6,410**
- Chromosome scale with 90% of estimated size in appropriate number of scaffolds

Mid sized genome

- Estimated genome size 1.8 GB
- Four Nanopore flow cells ($2,400)
- 50x Illumina ($570)
- 3 RSII SMRT cells ($7,235)
- 30x Hi-C ($2,250)
- Total: **$12,545**
- Chromosome scale with 90% of estimated size in appropriate number of scaffolds

Larger genome

- Estimated genome size 3 GB
- 4 RS II SMRT cells ($9,090)
- 30x Hi-C ($3,500)
- Total: **$12,590**
- Chromosome scale with 90% of estimated size in appropriate number of scaffolds

Questions

@JLandisBotany jbl256@cornell.edu