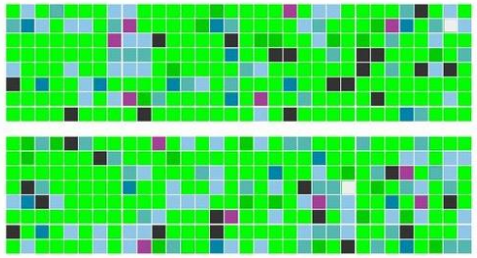


de novo Genome Assembly Workshop

September 15th, 2026

BRC Bioinformatics

Qi Sun

Co-Director
Bioinformatics Facility

Genomics

Jacob B. Landis

Research Associate

Plant Biology

and

Genomics Innovation Hub

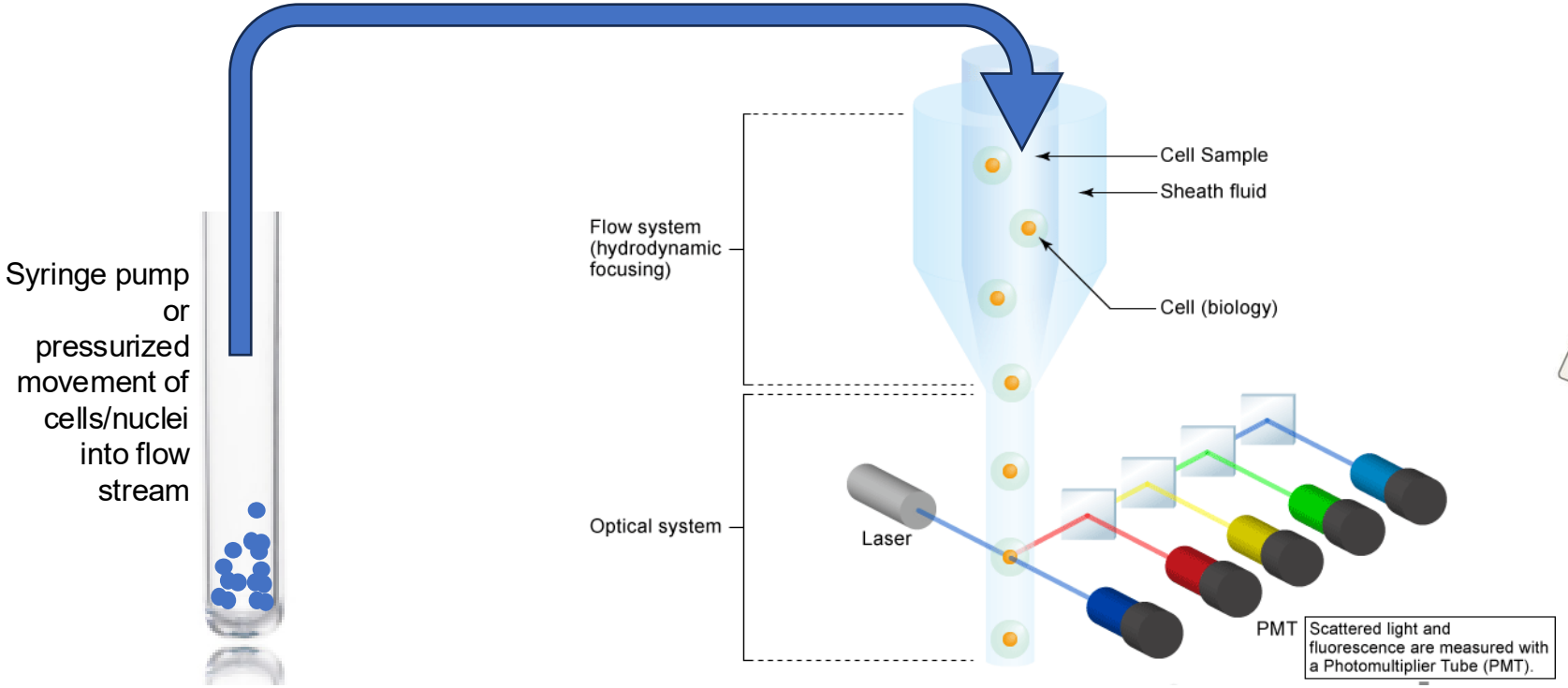
Flow Cytometry

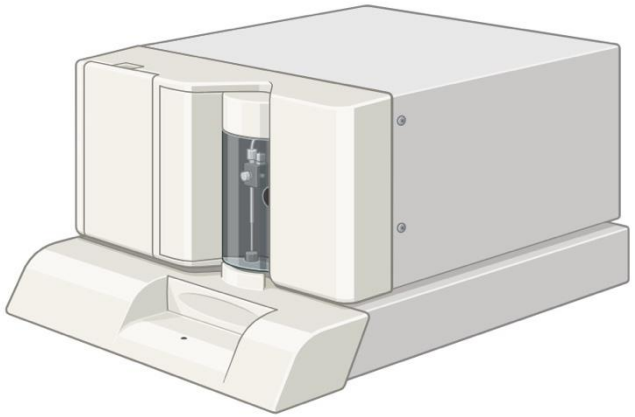
Julie Sahler

Director Flow Cytometry
Facility

Estimating Genome Size and Ploidy by Flow Cytometry

- Flow cytometry measures relative fluorescence of an individual cell/nuclei at speed of 1000's events/sec
- Common DNA dyes
  - Propidium iodide (PI) is more commonly used for genome size and ploidy analyses; requires RNAse treatment
  - Dapi can also be used for ploidy analyses
- Need: diploid/tetraploid/etc. and/or known genome sized **reference controls**

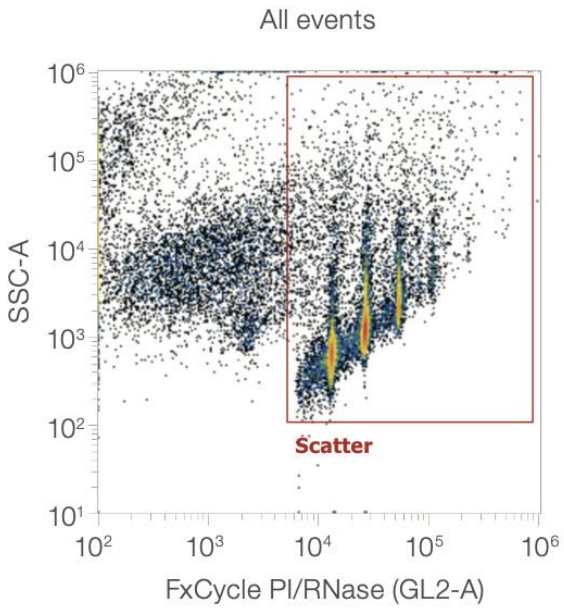


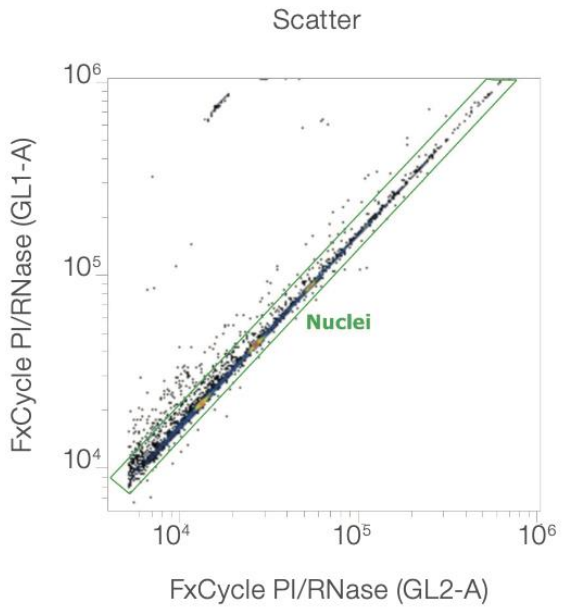


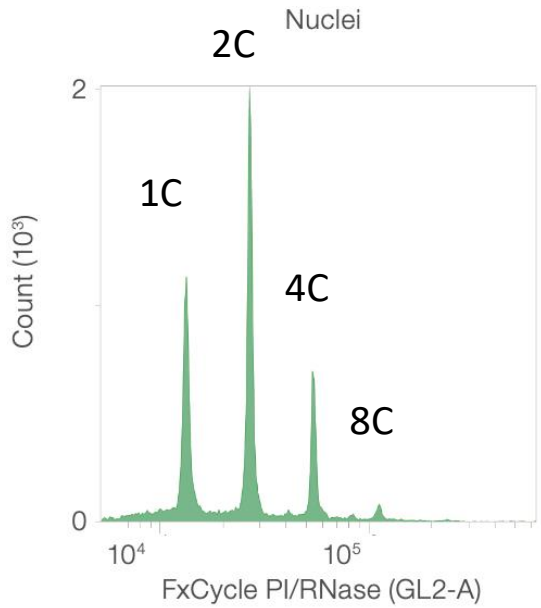
Institute of Biotechnology
Cornell Research&Innovation

Estimating Genome Size and Ploidy by Flow Cytometry

Example Data:







Genome size estimation: 2Ctest(pg) = FItest/FIstandardx 2Cstandard(pg)

BRC Flow Cytometry Core: services we offer:

- Full service sample acquisition/sorting and analysis
- User training to independently use cytometers
- https://blogs.cornell.edu/brcflowdev/
- Brc_facs@cornell.edu

ThermoFisher
DOI: 10.31383/ga.vol2iss2pp1-12

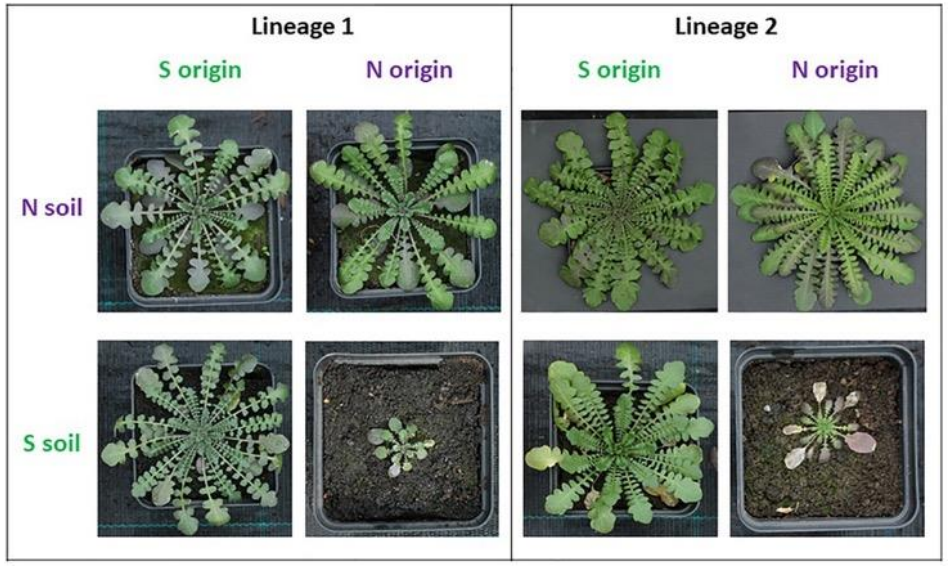
Institute of Biotechnology
Cornell Research&Innovation

What do all these things have in common?

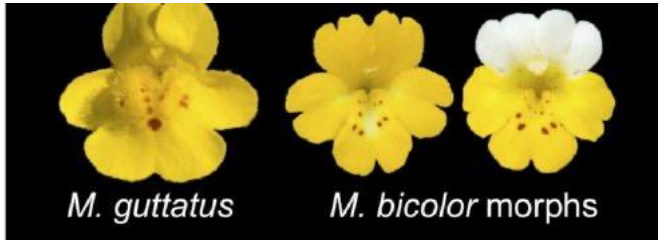
Shifts in pollinators

Evolution of carnivorous plants

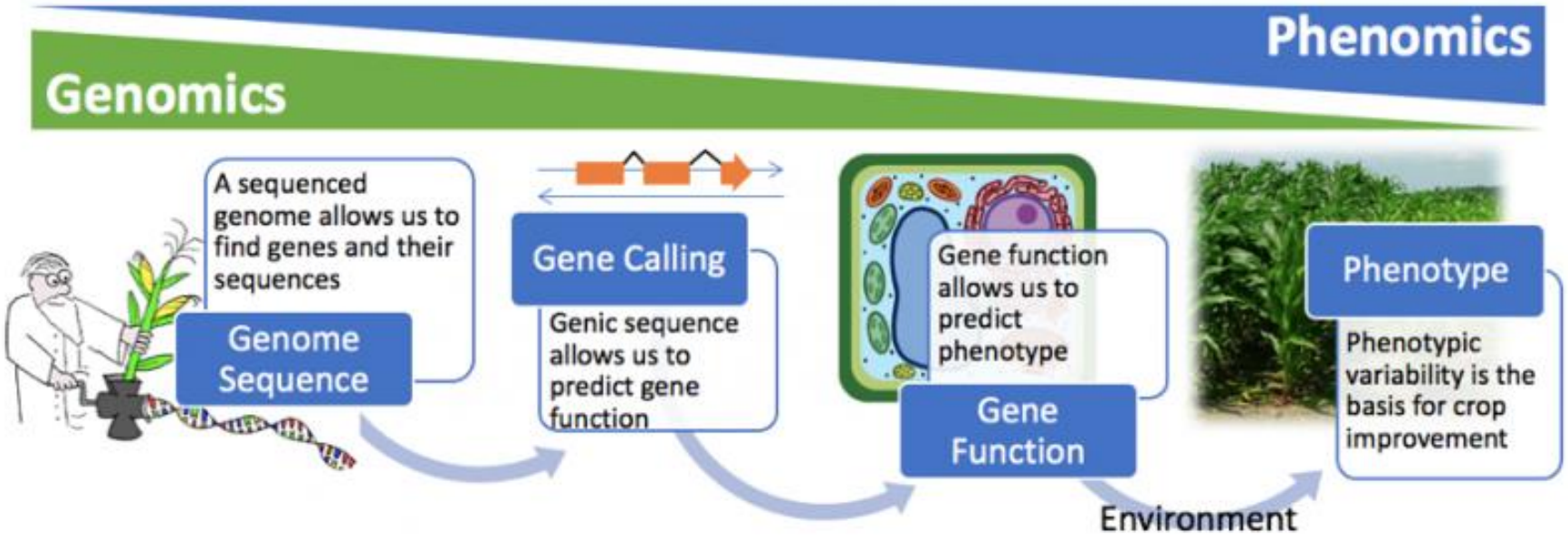
Response to nutrients

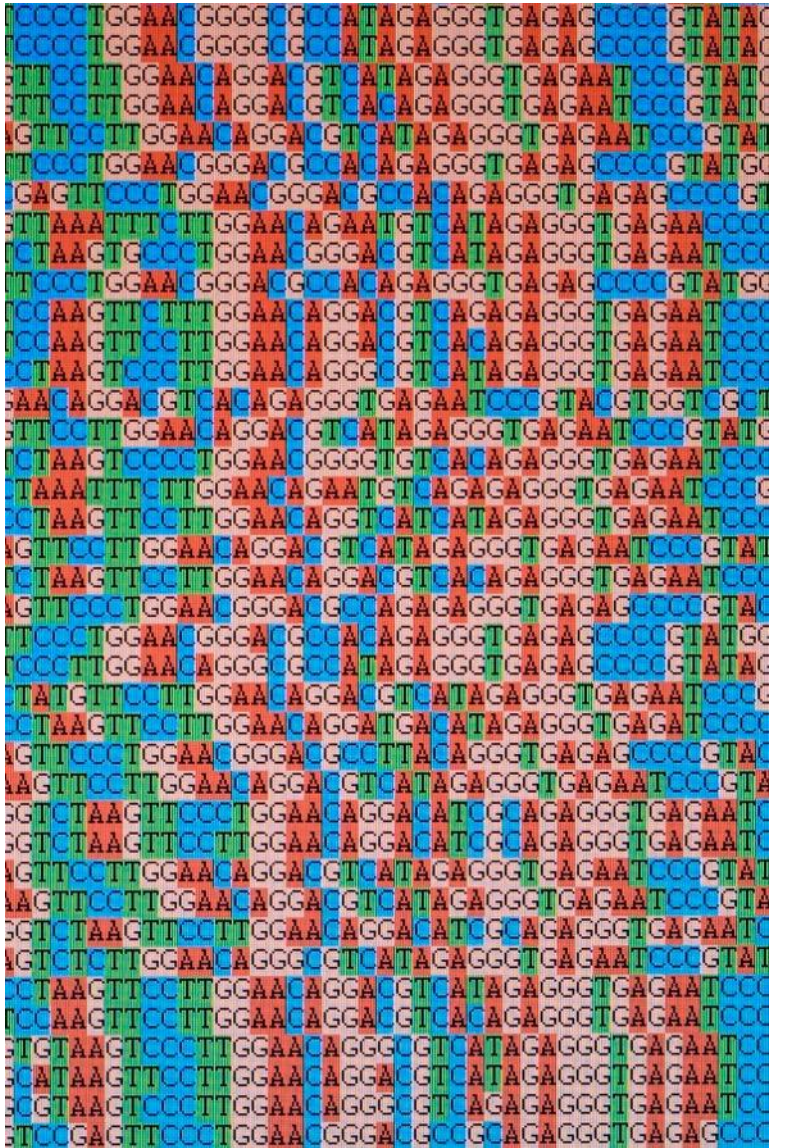


Flower color morphs

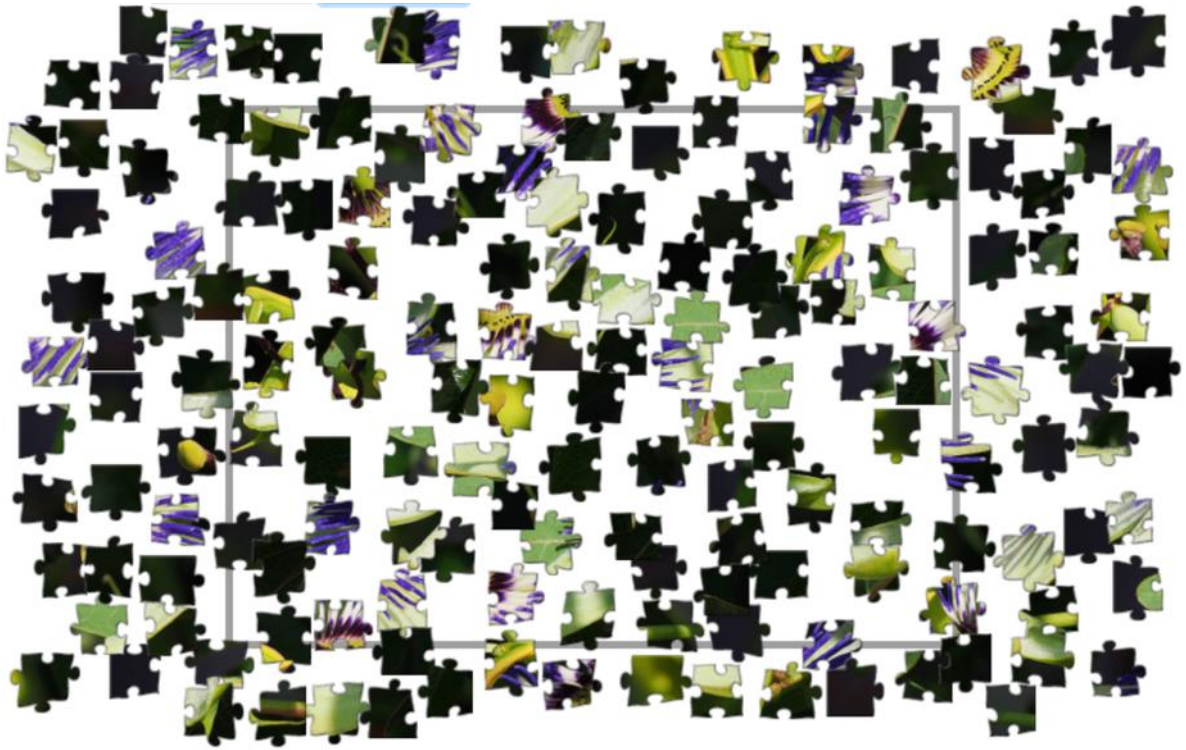


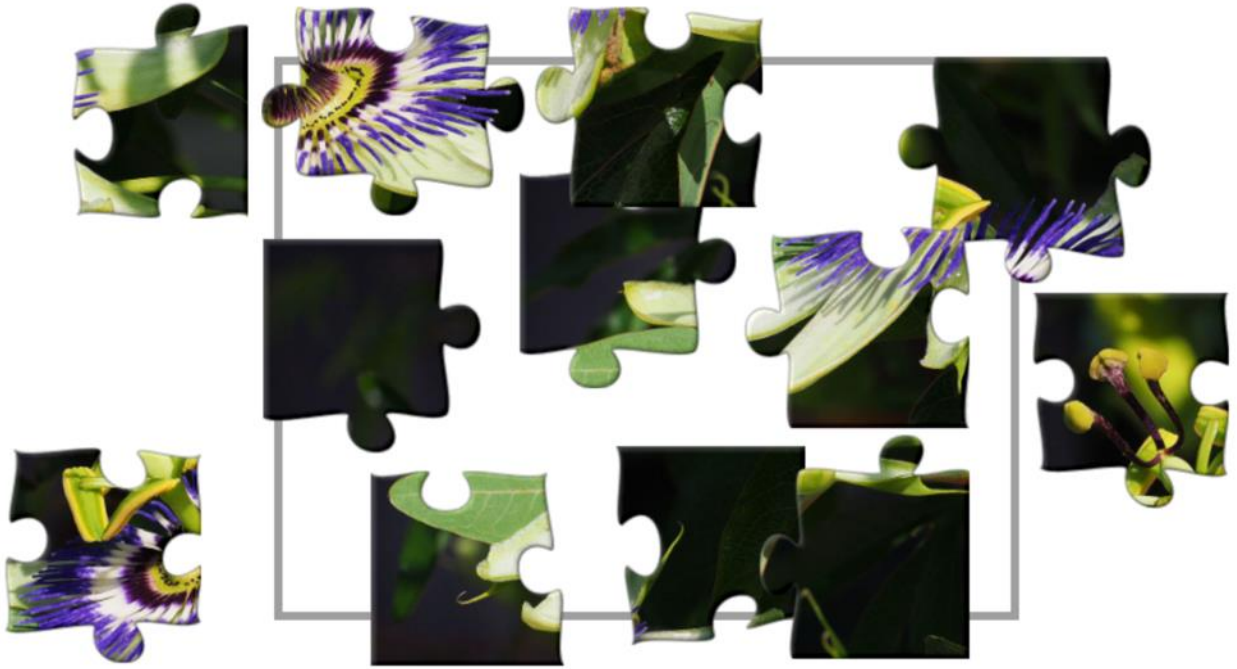
Genome as a puzzle





Which puzzle is easier to put together?

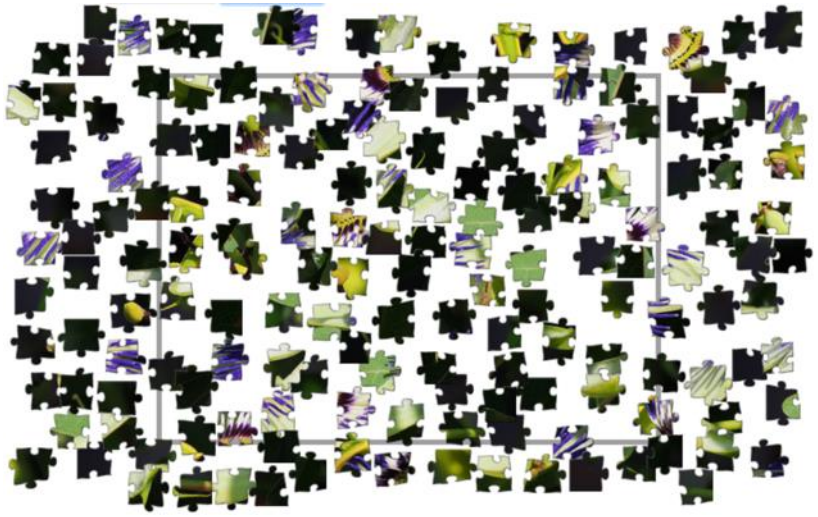




Which puzzle is easier to put together?

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

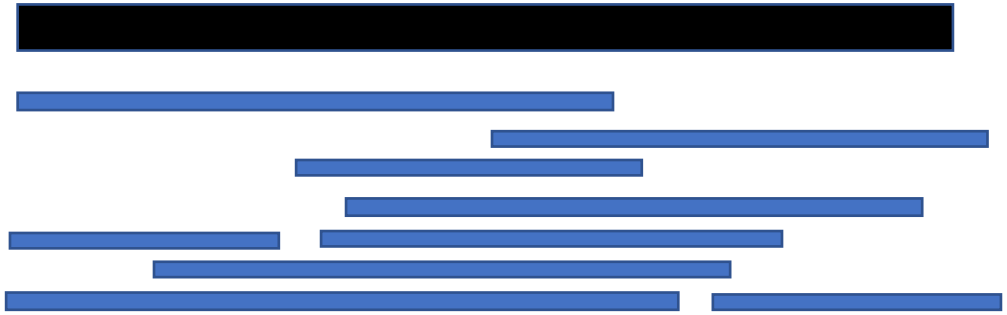
In the world of genomes



Reference Genome

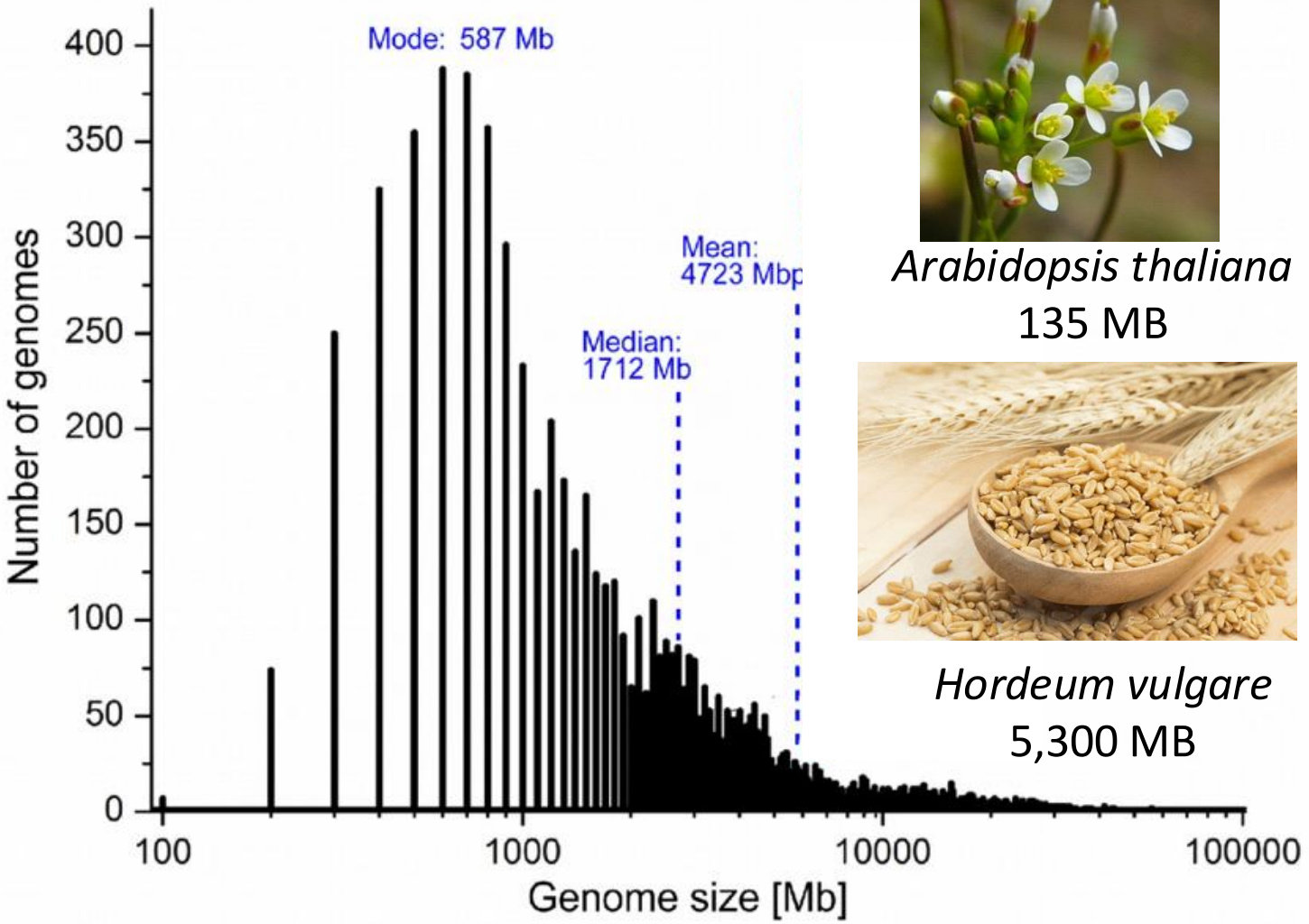
Sequencing Reads

Illumina reads



PacBio reads

Genome size of angiosperms



Arabidopsis thaliana
135 MB

Oryza sativa
430 MB

Zingiber officinale
1,582 MB

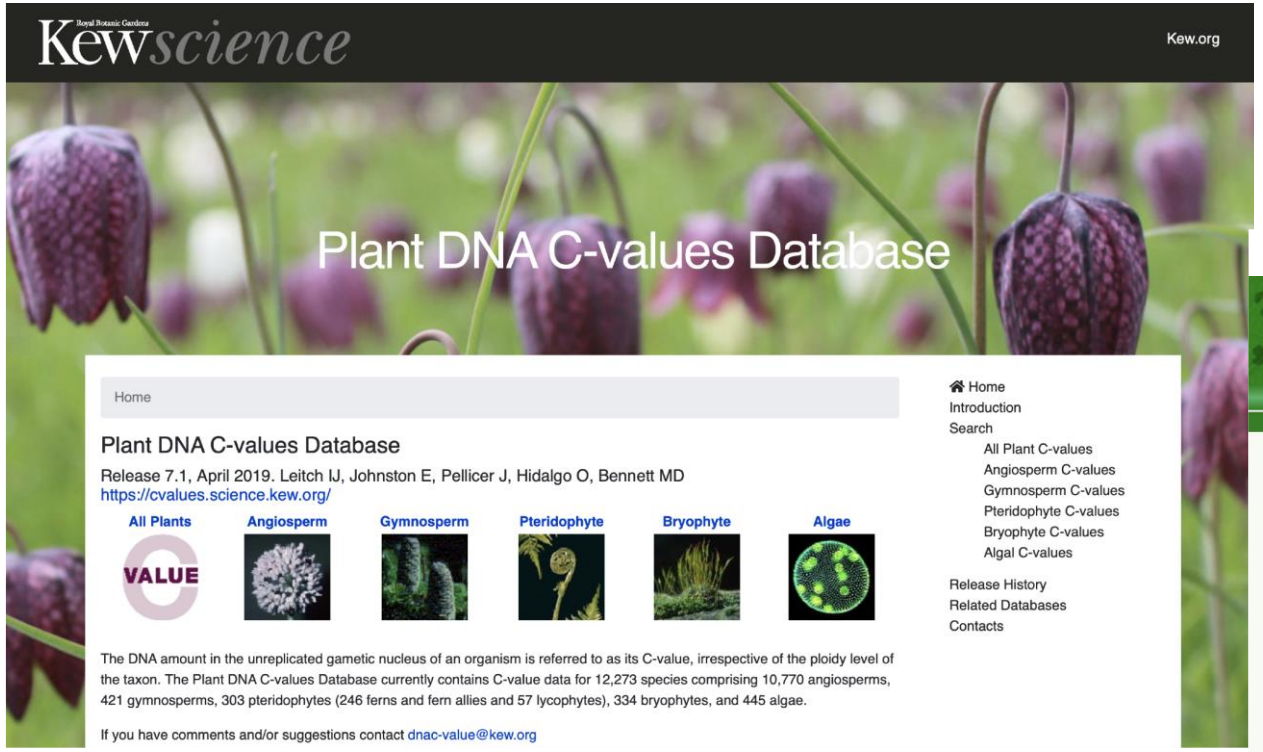
Hordeum vulgare
5,300 MB

Allium cepa
16,000 MB

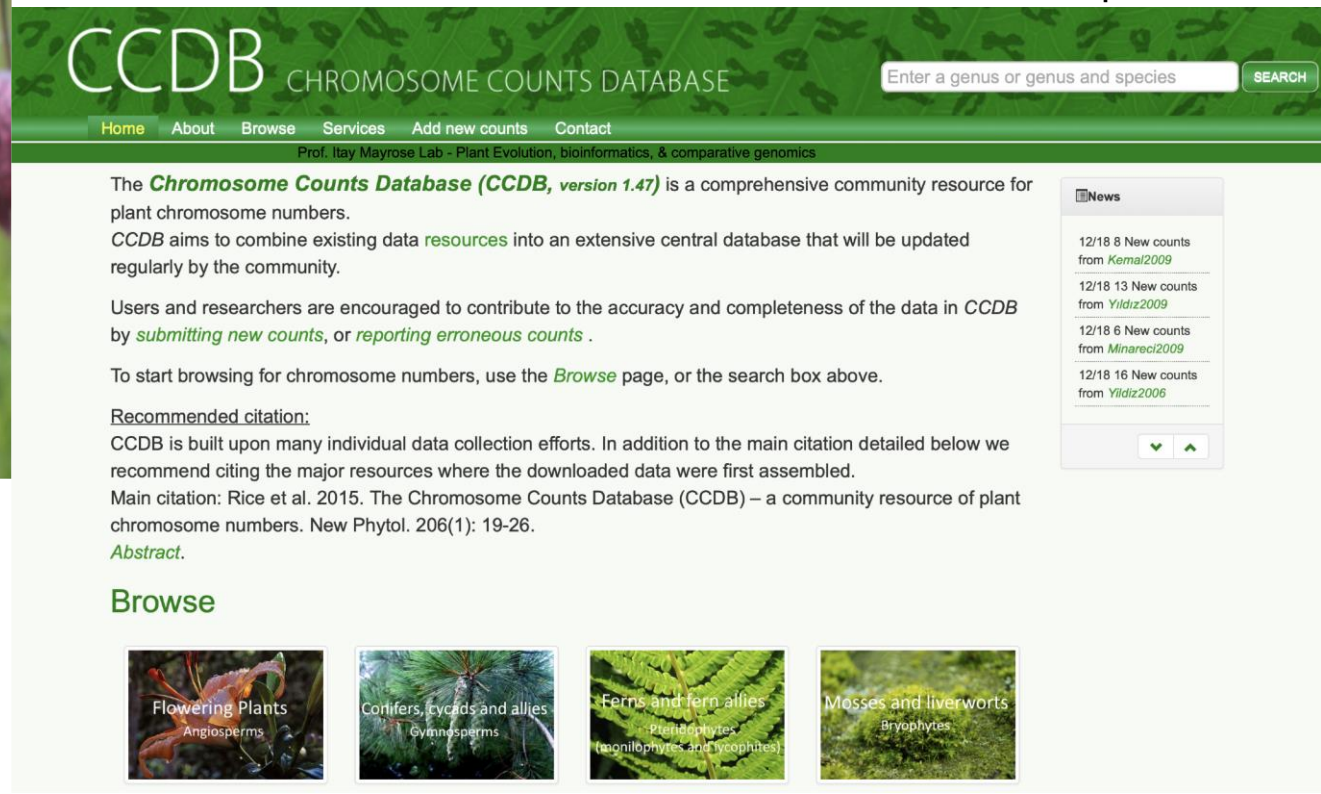
Tulipa sylvestris
59,241 MB

Adapted from Wicker et al. 2017 *Mobile DNA*

Chromosome number and genome size



http://ccdb.tau.ac.il

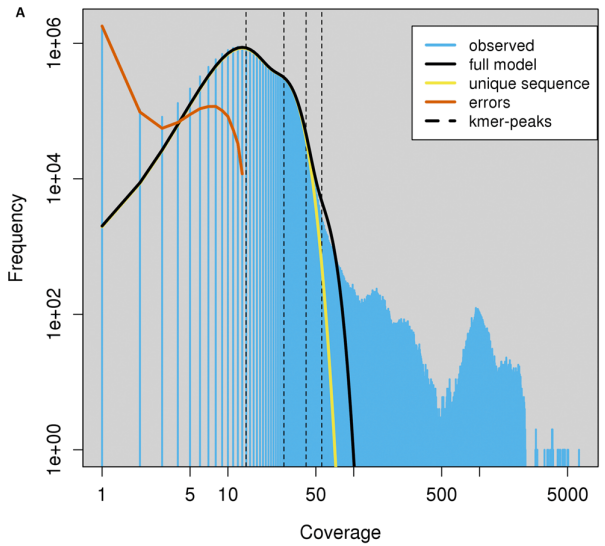


Genome size estimate

- 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)

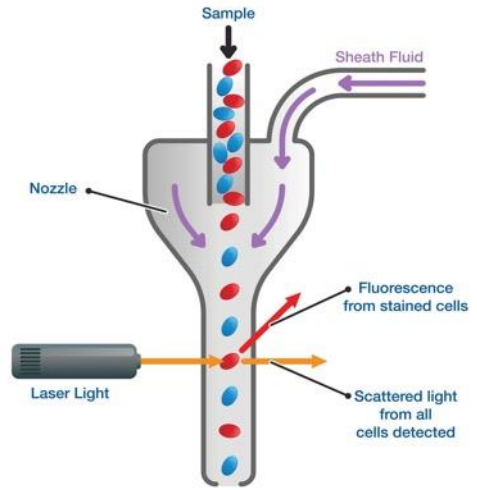
- Cleaned Illumina data
- Jellyfish paired with GenomeScope or RESPECT



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

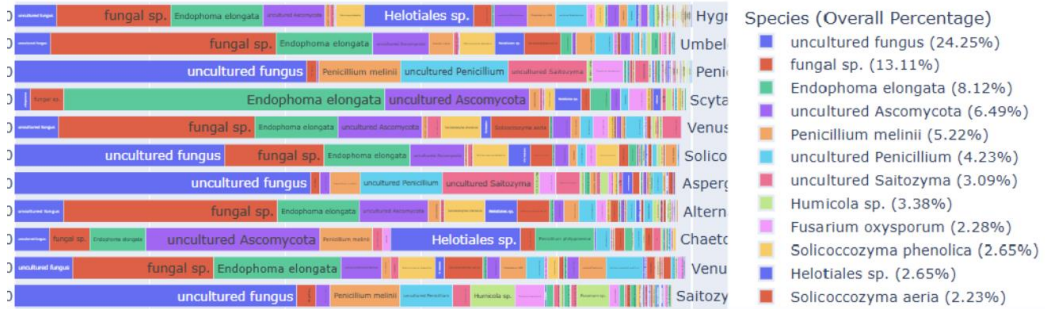
Flow cytometry (more reliable)

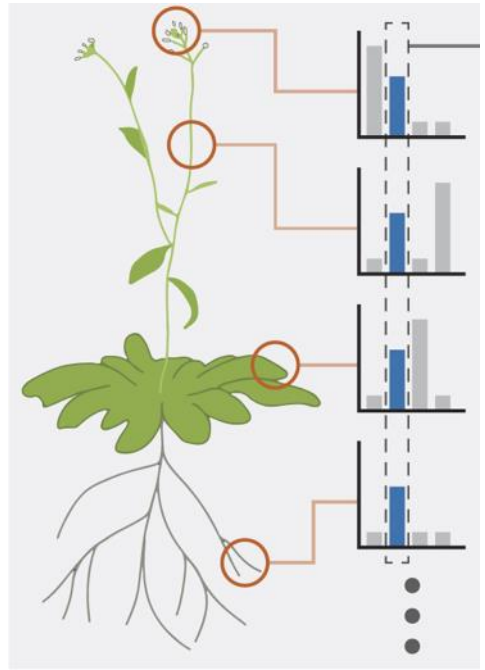
- Fresh or silica dried material; protocols can vary a lot between species
- Need accurate references to compare



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/frozen tissue for Hi-C sequencing
- “Clean” – less exposure to other organisms
  - SpeciesDetector can help detect other species



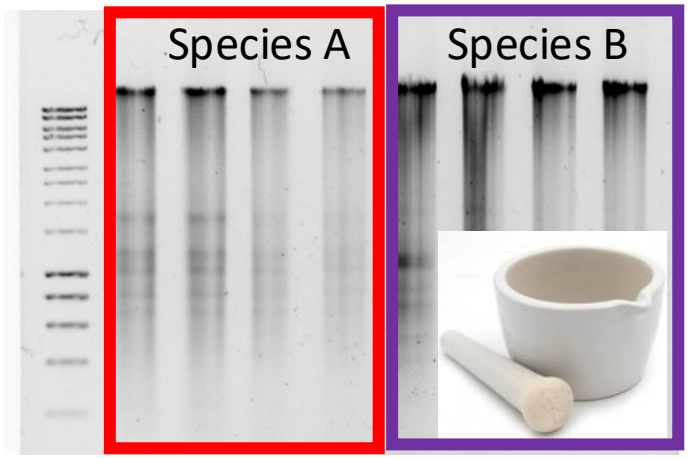


Getting good DNA

- Obtaining enough high molecular weight DNA is the hardest part
- Plants are more difficult than other species; have to deal with the cell wall
- May be very species specific and require troubleshooting

Tips for nonmodel systems

Tissue Grinding

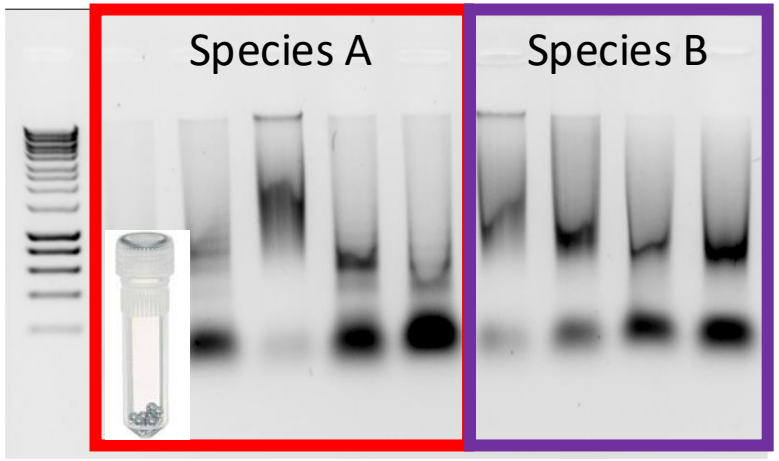


Mortar and pestle

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

Grinding beads

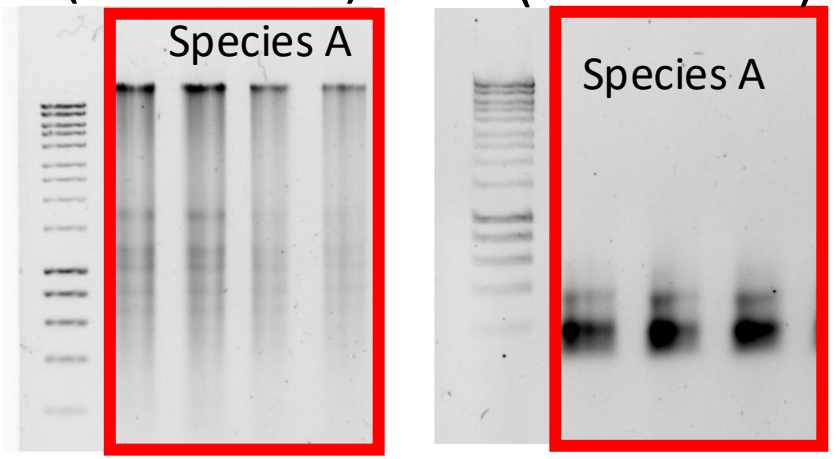
- Much lower yield; highly fragmented DNA
- Nanopore flow cell generated 12 GB with an N50 of 4.2 kb



Extraction method

SDS (Monocots)

CTAB (Monocots)



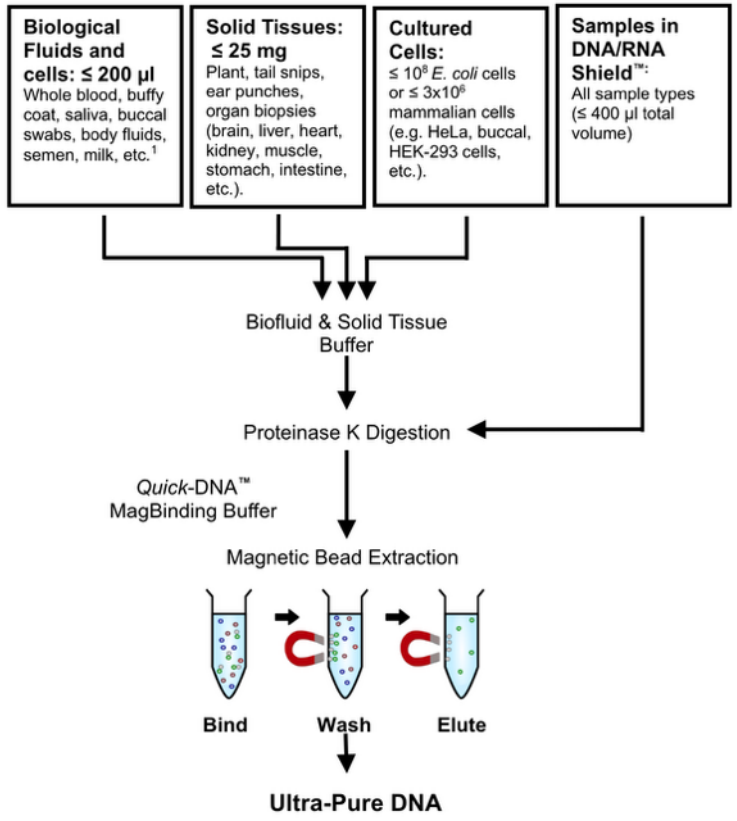
Modified SDS for Monocots

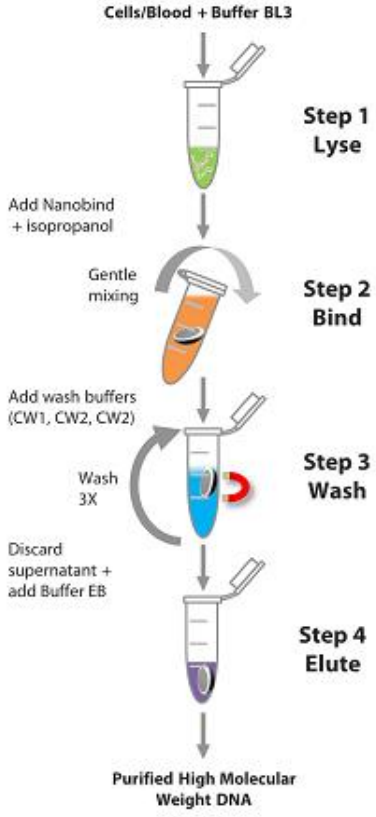
Modified CTAB for Eudicots

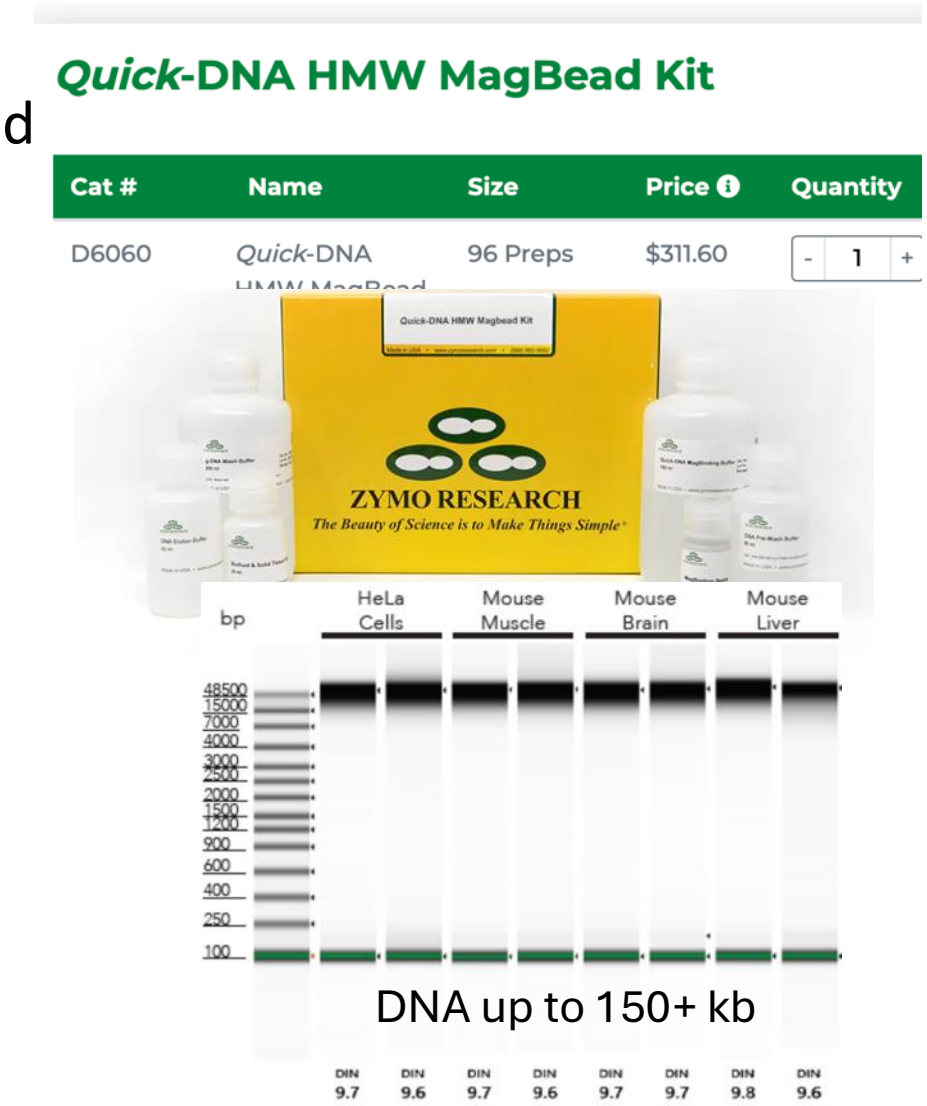
De La Cerda et al 2023; *APPS*

Kit based approaches

- Several commercial options
  - Nanobind PacBio/Cicrulomics is highly recommended
- Zymo's is supposed to work in 45 minutes

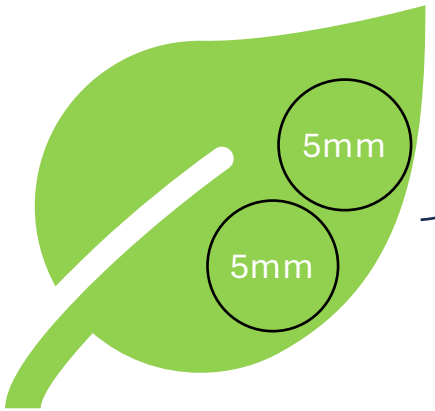






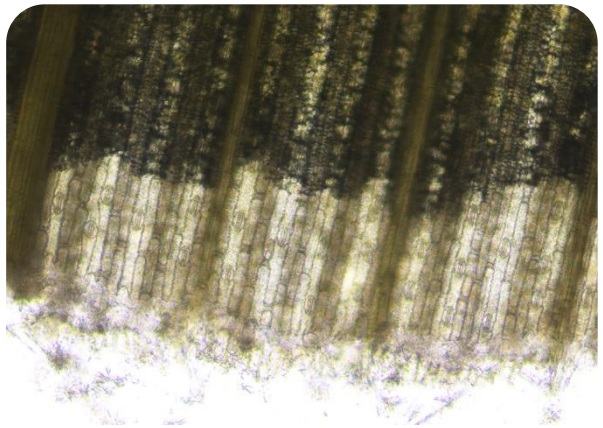
Enyzematic approach GIH

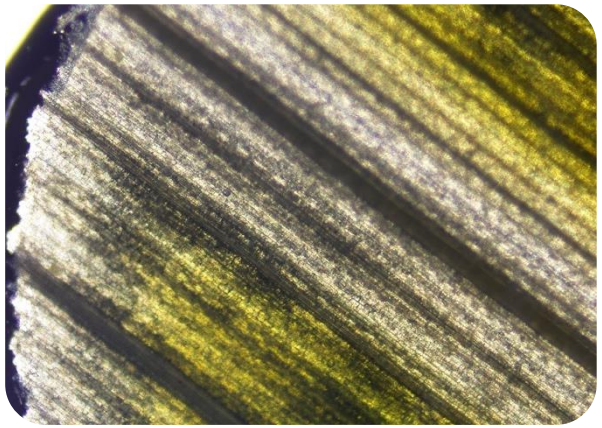
x 10



chloroform
30s

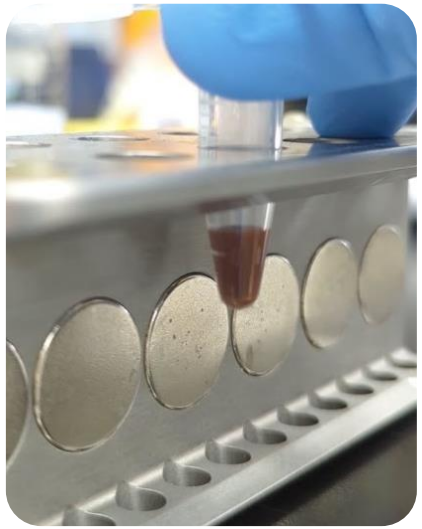
2hr Enzyme cocktail + Digestion buffer
Overnight proteinase K



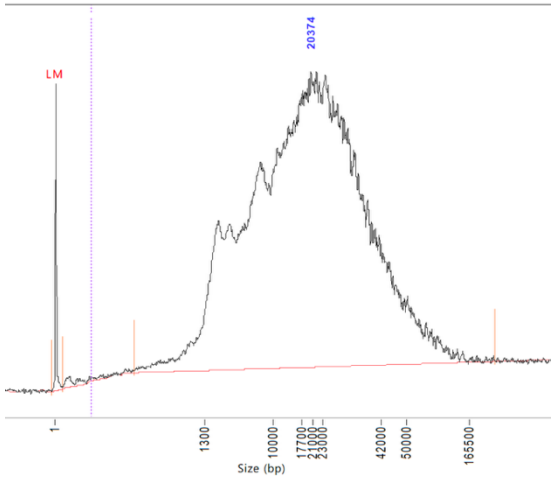


1hr Lysis Buffer + RNase

Bead clean



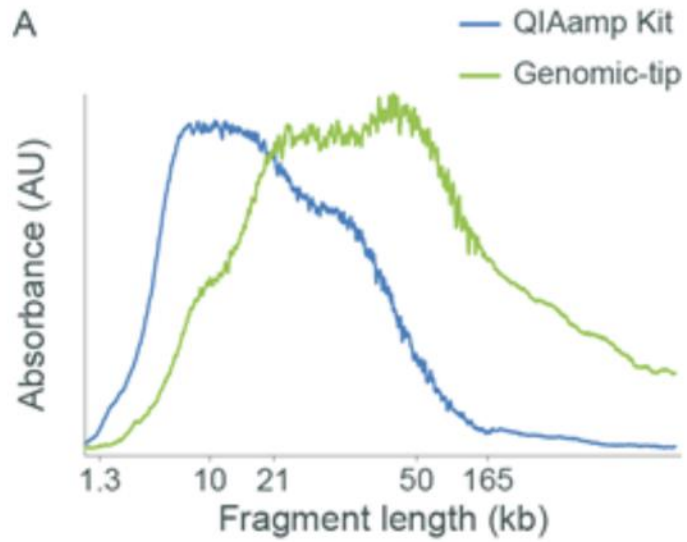
Femto

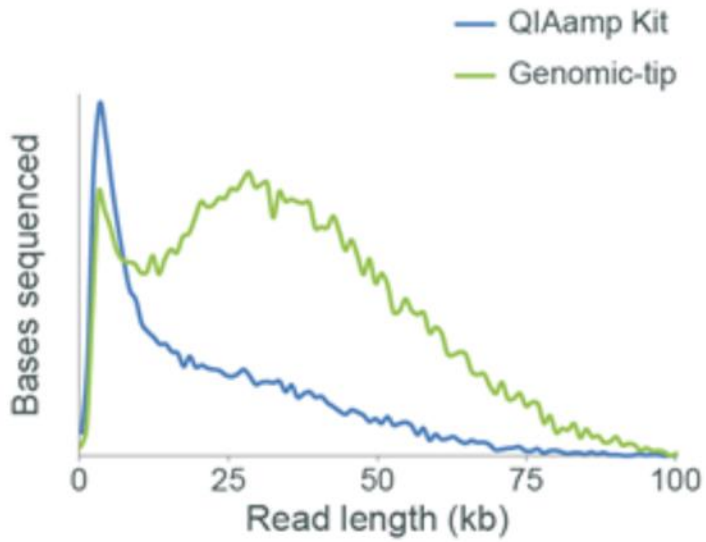


DNA peak = 20 kb
Amount = 1-3 ug

DNA QC

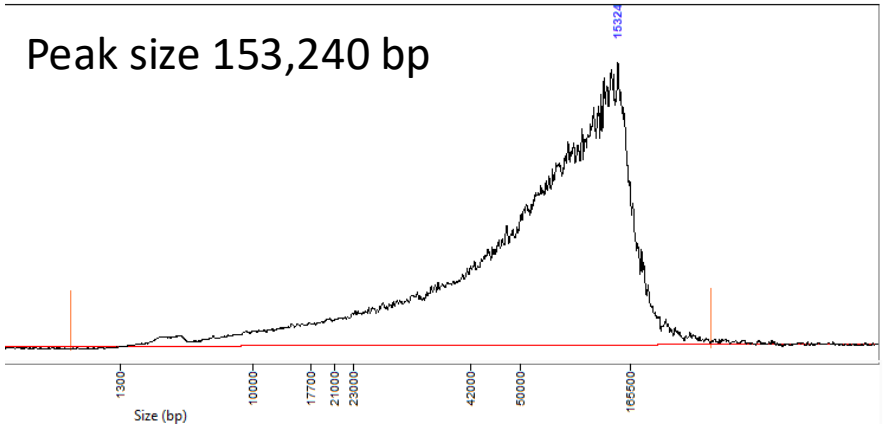
- Quantification – Need to rely on Qubit
  - Nanodrop drastically overestimates concentrations
  - 1 ug for each sequencing run; if size selection need around 5 ug starting out
- Purity/Cleanliness – Nanodrop
  - 260/280 values should be 1.8-2, while 260/230 values 2.0-2.2
  - If pure DNA, concentrations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

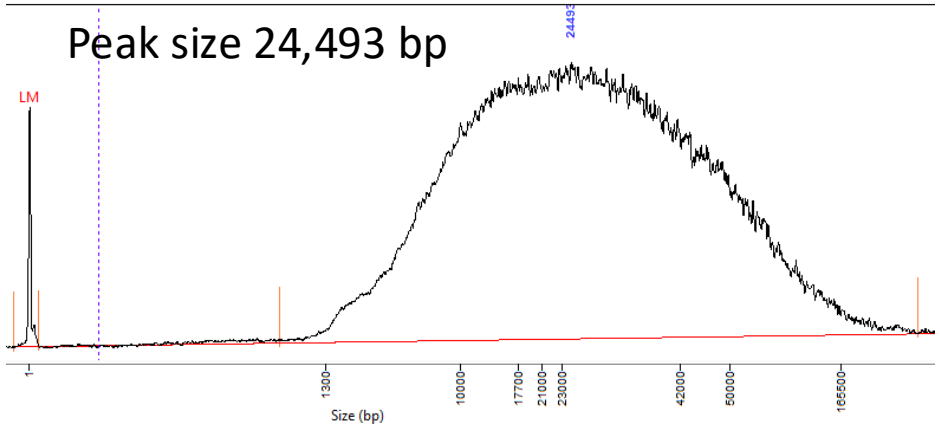


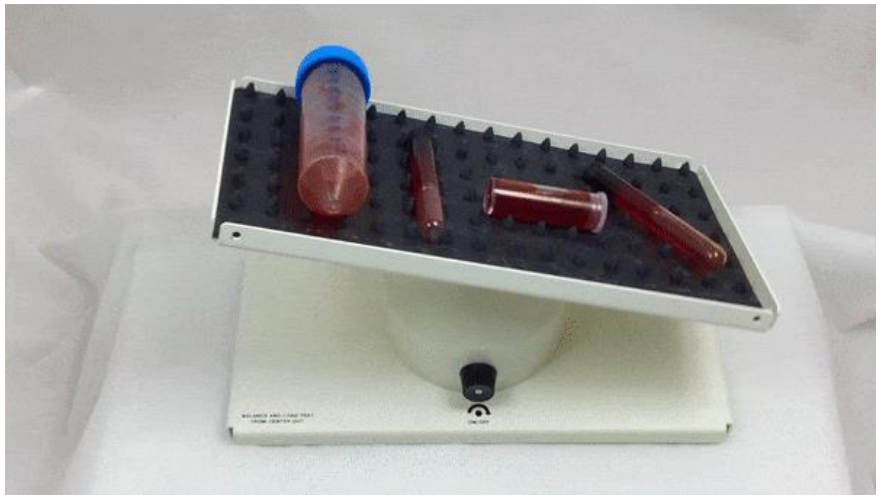


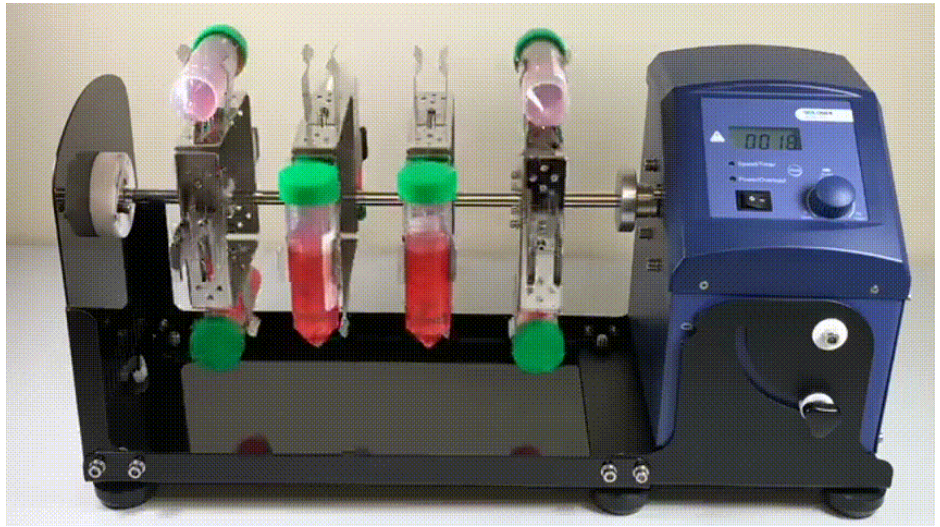
How fragile is DNA?

What can fragment DNA during an extraction?



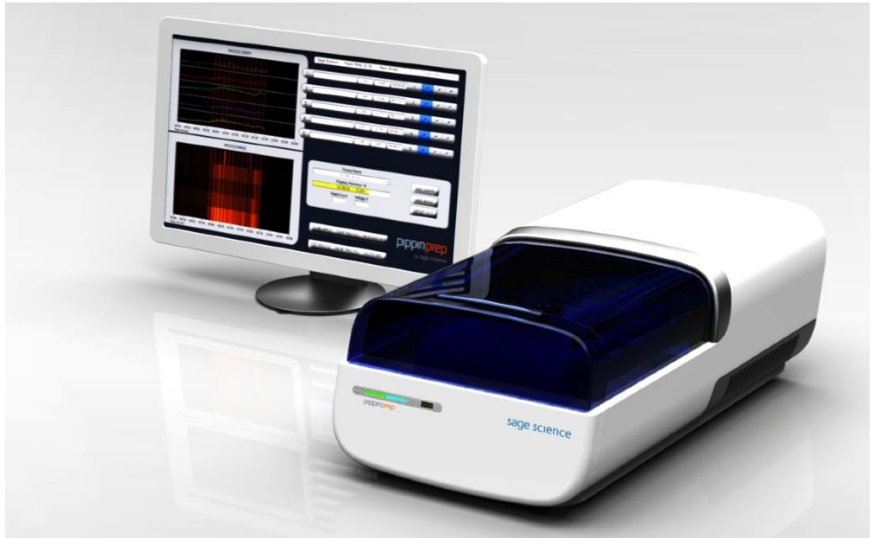






Size selection

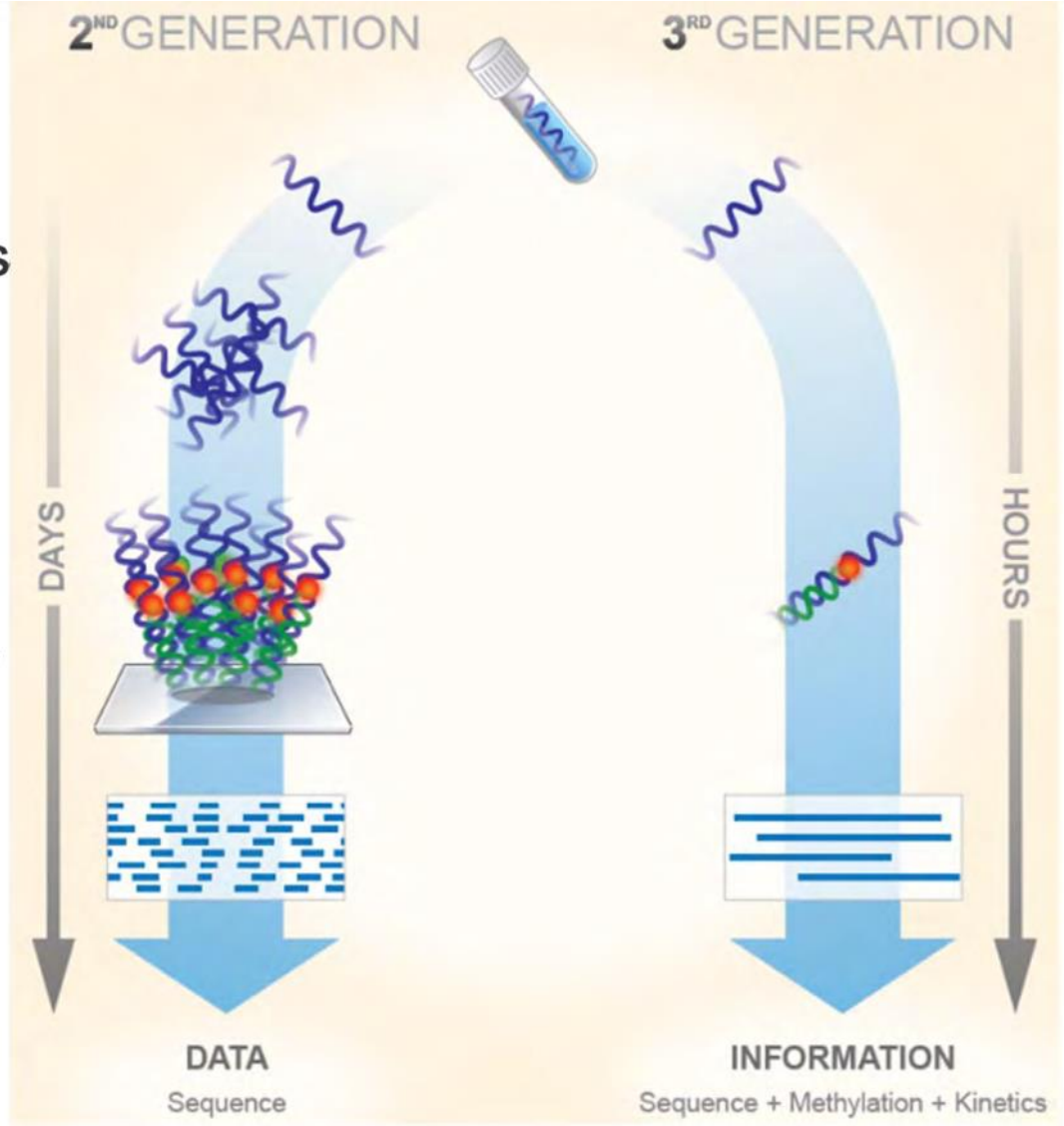
- Blue pippin
- Ciculomics Short Read Eliminator Kit
  - Target cutoff size: XS (10 kb), Regular (25 kb), XL (40 kb)
- Modified SPRI beads to retain large DNA fragments
  - MagBio or homemade with low-salt buffer
- Some/most DNA will be lost, but what remains is highly valuable
  - Be prepared for around a 40% quantity reduction with each cleaning step



Sequencing platforms

Short vs Long-reads

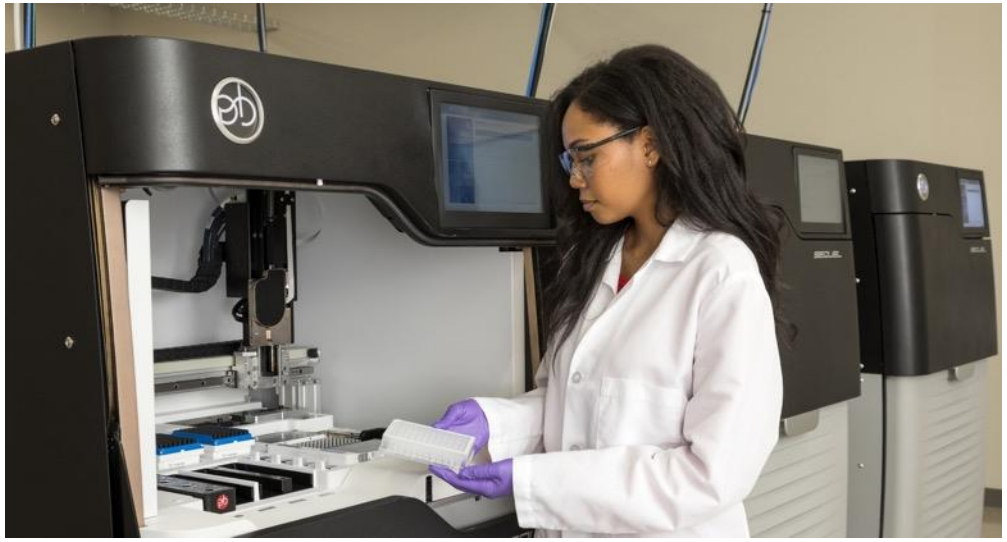
- Short reads
- Amplification errors and bias
- Several enzymatic steps
- Multi-molecule raw accuracy
- Errors tend to be systematic
- More coverage required



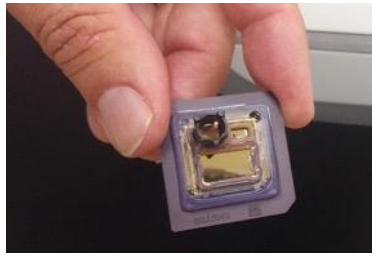
- Long reads
- No required amplification
- Simple sample prep
- Single molecule raw accuracy
- Errors tend to be random (vs. systematic)
- Less coverage required

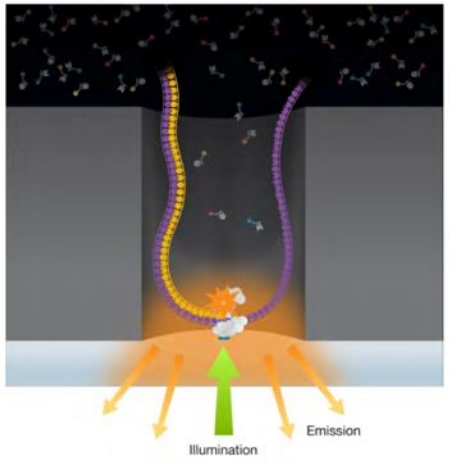
https://get.genotoul.fr/wp-content/uploads/2017/06/150519_Technology_and_Applications_PACBIO_GerritKuhn_low.pdf

Long-read - PacBio

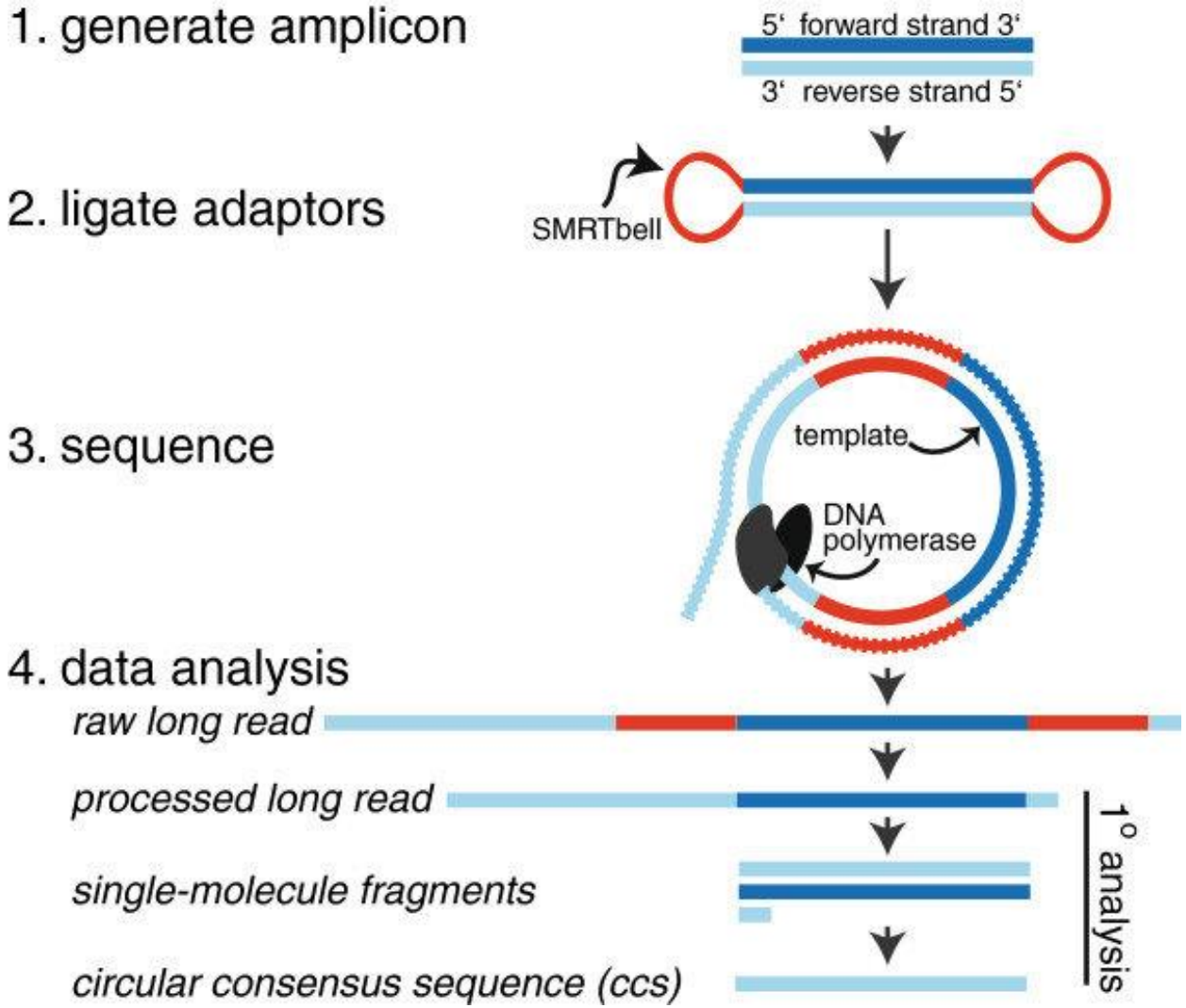


SMRT Cell





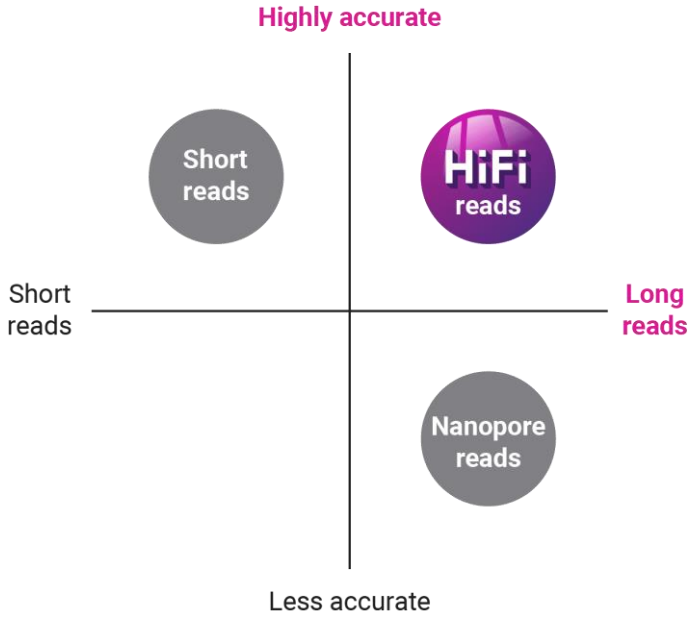
Science, Vol 299, Jan 31 2003, pp682-686
J. Appl. Phys. 103, 034301 (2008)

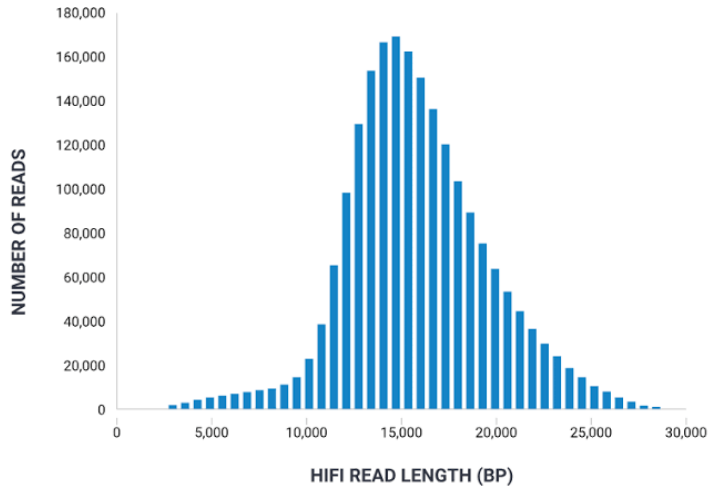


Fichot and Norman 2013; *Microbiome*

PacBio Revio

- 90% of bases ≥Q30 and median read accuracy ≥Q30
- 15x increase in throughput over the Sequel II system
- Supposedly a little less than 100 Gbp per SMRT cell
  - If DNA fragments are less than 10 KB, total output drops
  - Many plant runs may get lucky to generate 70 Gbp per SMRT cell
- Typically outsourced to the Genomics Core or other locatoins as opposed to what can be done inhouse with minION



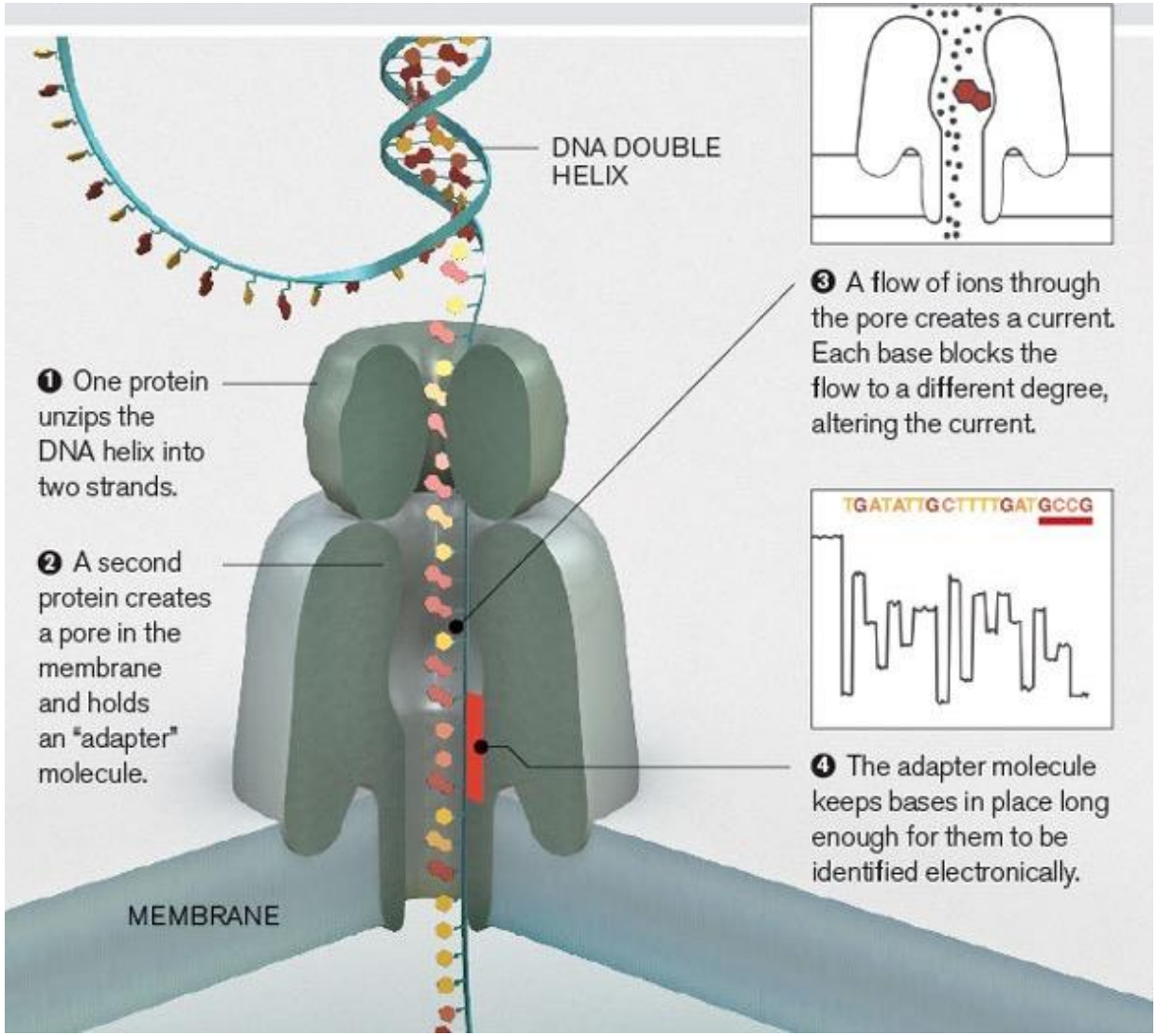


Long-read - Nanopore

Flowcell

MinION device

PromethION



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

HiFi Sequencing at Cornell

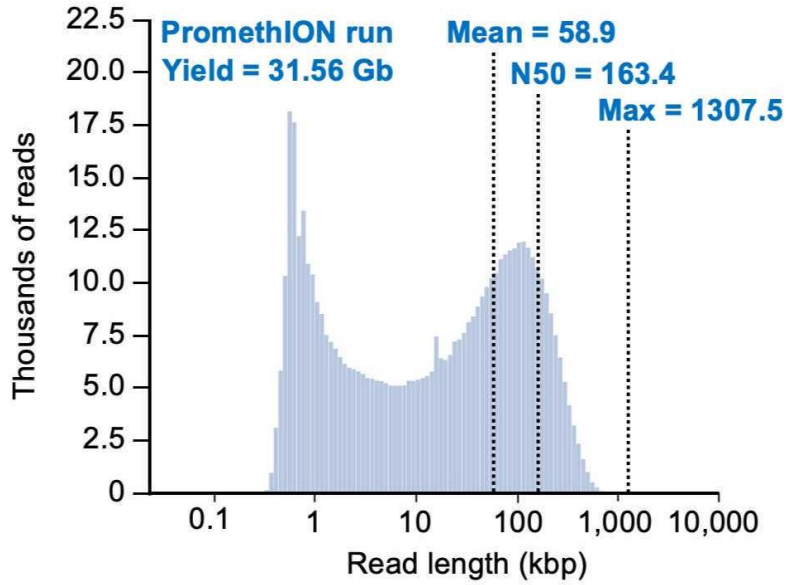
- Genomics Core now offering PacBio HiFi sequencing
- PacBio Revio: SMRTbell library prep (1 sample, pricing includes femto QC, SMRTbell library prep, 25M Revio cell): **$2,269**
- PacBio Revio: SMRTbell library prep (multiplexed, pricing includes femto QC, SMRTbell library prep, 25M Revio cell, **$270 each** additional library): **$2,319**
- Please submit at least **3ug** (per Gbp genome size) of high molecular weight DNA in 50uL (minimum 60ng/ul)
  - To include the gDNA QC check with the sequencing order, please prepare a separate aliquot with concentration ≤ 0.5 ng/ul in minimum 5ul volume

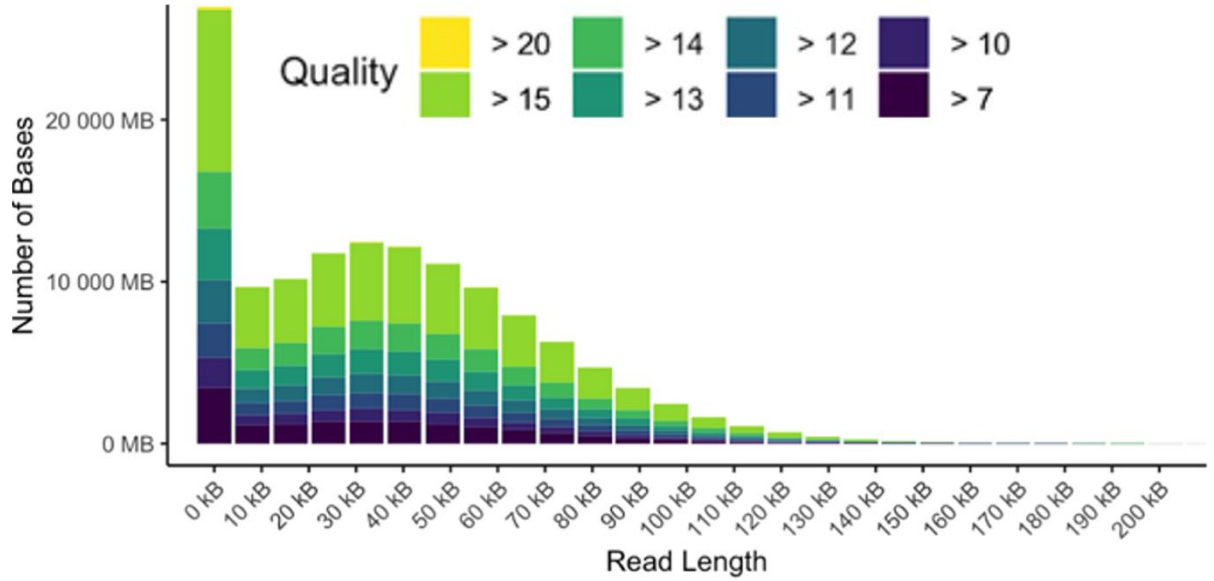
Why so much DNA?
- Size selection
- Redo if prep fails
- Large genome

Library type	Library size	Minimum Amount required per SMRT cell (ng)* Volume requirement: 25µl
WGS	10Kb	120ng
	15Kb	180ng
	18Kb	210ng
	21Kb	250ng

Sequencing at Cornell - Nanopore

- If you want/need ultra-long reads (>100 kb), Nanopore is the best option
  - Can combine HiFi and Nanopore reads in an assembly; best of both worlds
- **PromethION:** Flowcell contains 10,700 pores and typical outputs range from 25-100 Gbp depending on sample type and quality - **$2,900**
- **MinION:** Flowcell contains 2,048 pores and typical outputs range from 5-15 Gbp of data depending on sample type and quality **- $2,185**





How does this translate to costs?

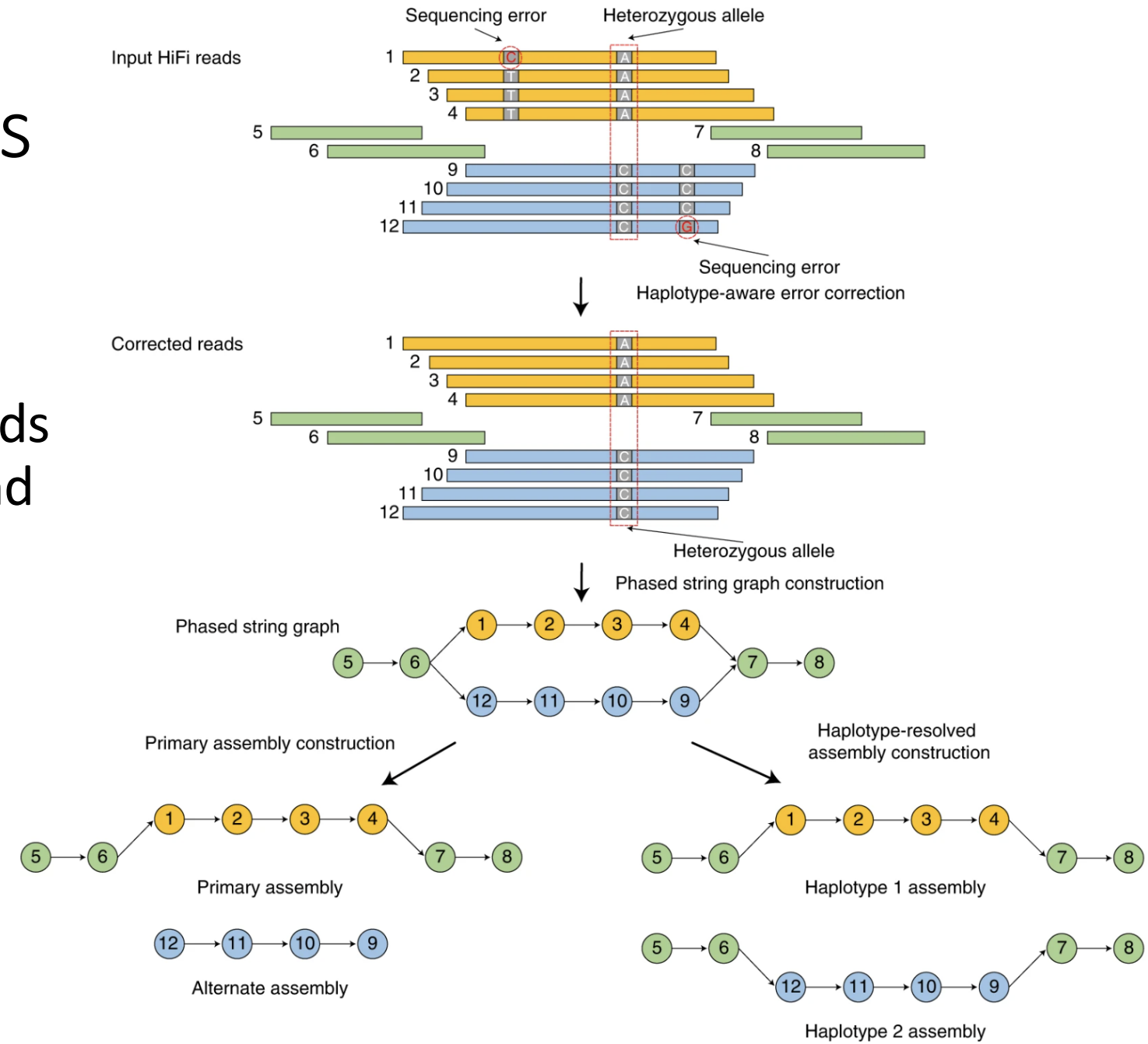
- PacBio says that 15x coverage is enough to assemble a high-quality genome
  - In my experience closer to 30x coverage is needed (15x per haploid copy)

- One sample 1 Gbp genome
  - Minimum of 30 Gbp - **$2,269**
- Three samples 1 Gbp genome each - **$2,859 total**
  - **$953 each**

- One sample 5 Gbp genome
  - Minimum of 150 Gbp, two SMRT cells - **$4,538**
- One sample 10 Gbp genome
  - Minimum of 300 Gbp, three (or four SMRT cells) - **$6,807-9,076**

Long-read assemblers

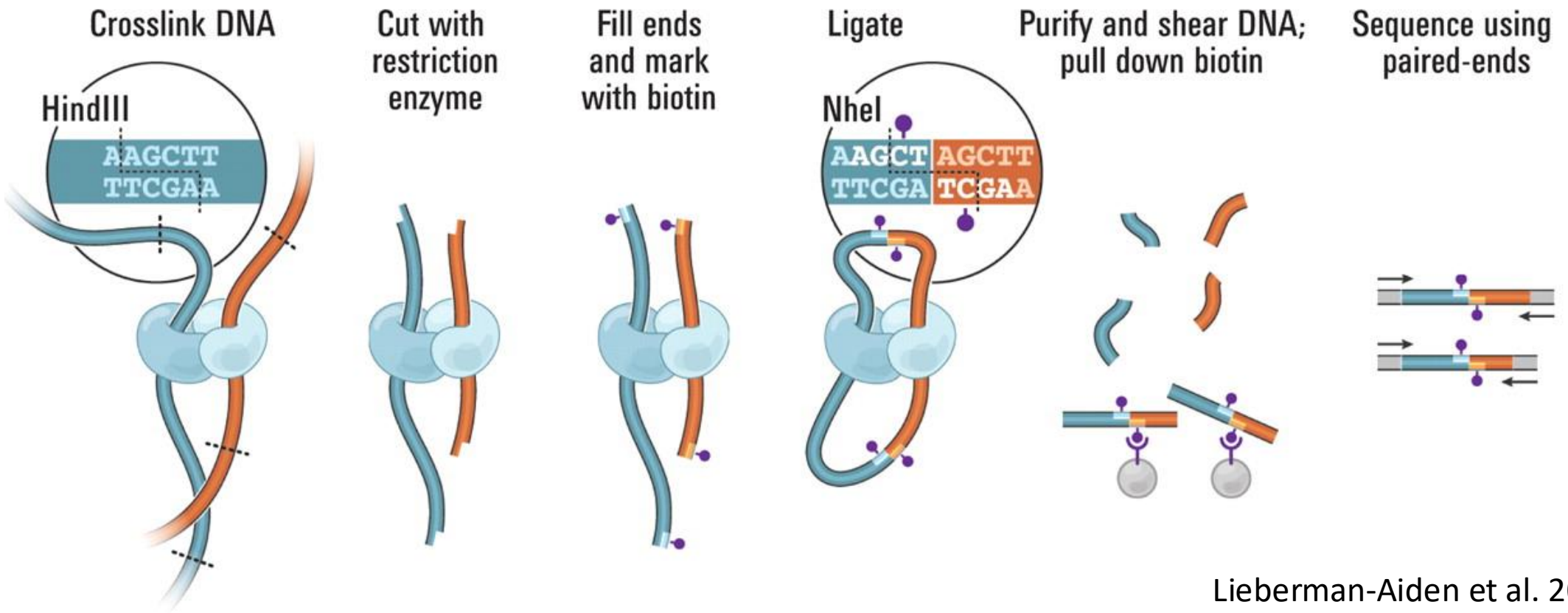
- Most favored long-read assembler by far is **HiFiasm**
- Input HiFi reads, Nanopore reads (--ul), and/or Hi-C data (--h1 and --h2)
- Corrects sequencing errors, handles heterozygous alleles
- Outputs primary assembly or haplotypes



Cheng et al. 2021 – *Nature Methods*

Hi-C for scaffolding

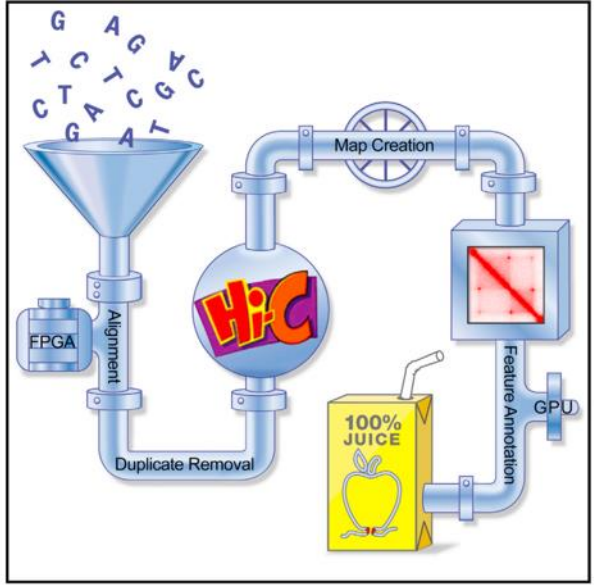
- Hi-C = high throughput chromatin conformation capture
- DNA of the same chromosome will be ***spatially*** close



Lieberman-Aiden et al. 2009; *Science*

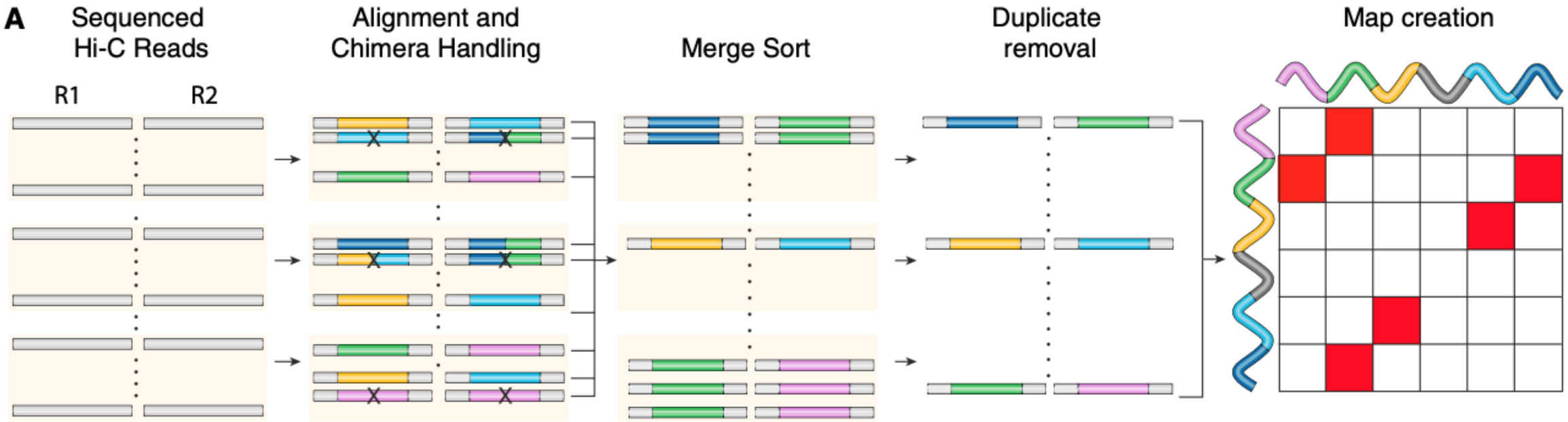
Scaffolding (1 Gbp genome)

- 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; though for some difficult species they struggled to get the necessary data
- Arima Genomics has a 6-hour rapid protocol
- Optical mapping by Bionano
  - Outsource (HWM extraction + Saphyr chip + analysis) = **~$3,000**



Scaffolding with Juicer

- Hi-C reads mapped to genome
- Read pairs aligning to more than two positions are excluded
- Remaining are sorted and merged into a single list
- Numerous tutorials online



Durand et al. 2016; *Cell Systems*

Scaffolding with YAHS (Yet another Hi-C scaffolding tool)

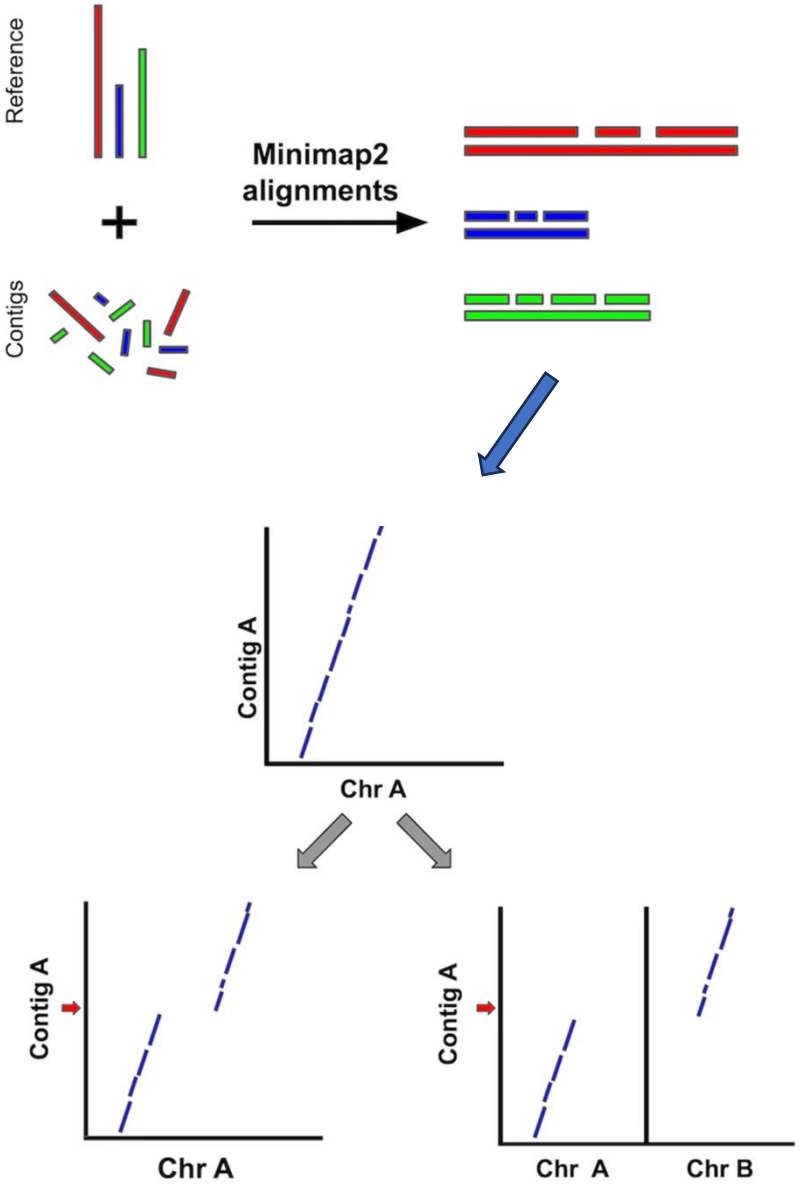
- Super easy: bwa-mem2 (mapping), samtools (BAM sorting), piccard (mark duplicates), YAHS (scaffolding)
- Major drawback: YAHS can only handle up 45,000 contigs; no input for expected number of chromosomes
- Built in scripts to make Hi-C map for Juicebox
- Changing resolution values can drastically change output

	pass 1	pass 2	pass 3	pass 4	pass 5	pass 6	pass 7	pass 8
scaffold_1	487,845,246	201,791,392	201,791,392	536,081,994	443,652,476	738,930,895	908,836,805	908,836,805
scaffold_2	391,778,060	166,829,253	166,829,253	296,681,732	373,050,187	369,342,481	442,872,607	442,872,607
scaffold_3	363,252,375	158,873,270	158,873,270	296,464,256	360,592,181	296,132,661	378,717,349	378,717,349
scaffold_4	297,174,338	146,650,840	146,650,840	264,631,872	294,475,627	253,808,656	318,505,497	297,763,748
scaffold_5	237,614,215	143,171,990	143,171,990	256,939,548	185,231,450	246,475,533	294,922,786	294,922,786

201 MB – 908 MB
166 MB – 442 MB

RagTag (supersedes RaGOO)

- Use a high-quality reference to scaffold query sequences
- Does not alter input sequences in any way; only orders and orients sequences, joining them with gaps
  - Any local inversions or indels in a query contig are carried through unchanged
- Appends unplaced query sequences as-is to the end of the output
  - Option to concatenate all unplaced into a “Chr0”



Alonge et al. 2019; *Genome Biology*

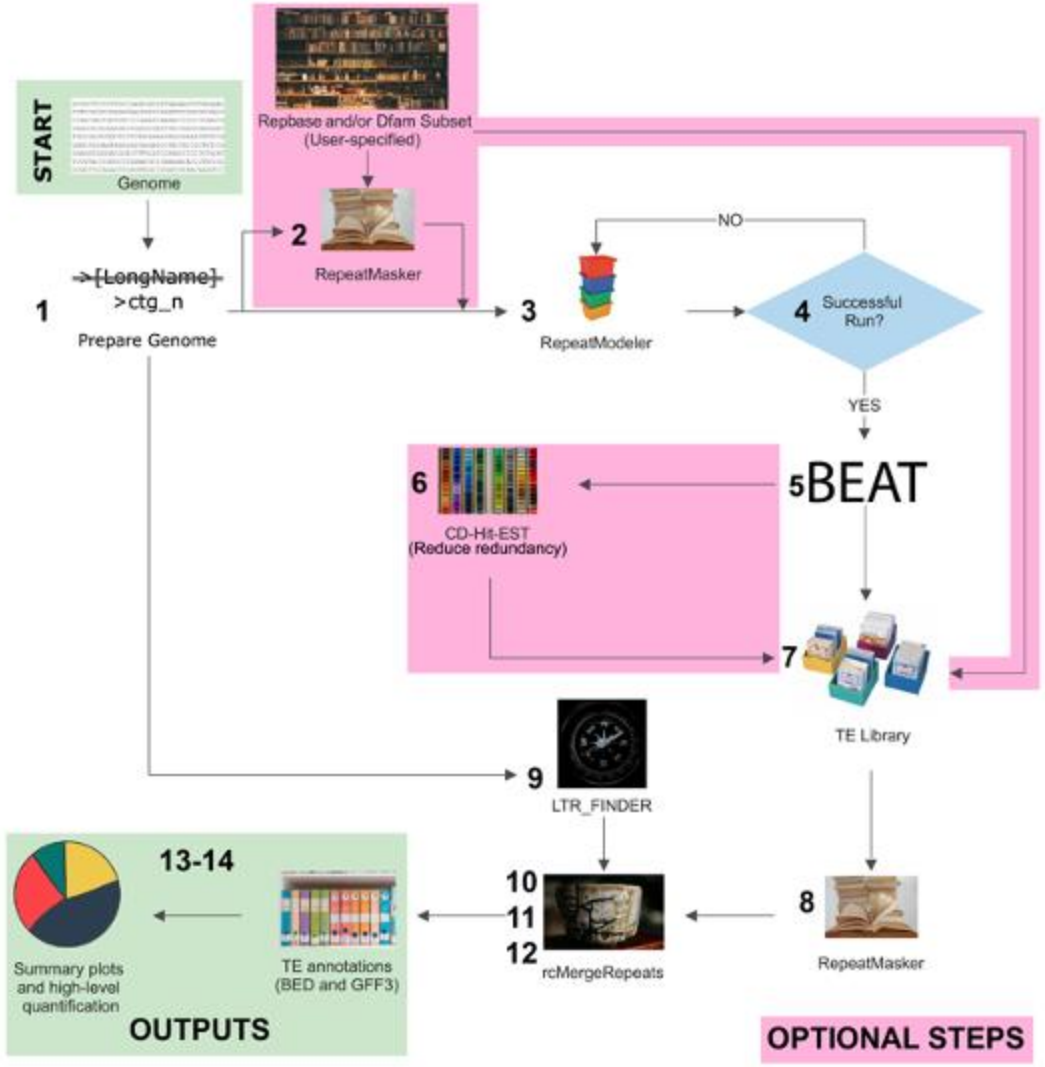
Annotations

- For some analyses you just need/want the genome sequence
  - Population genomic or phylogenomic analyses
- Other times you need to know where coding gene sequences are in the genome and what those genes are doing
- For publishing high-quality genomes, genome annotations are usually required

Prepare genome - Mask repetitive regions

Earl Grey

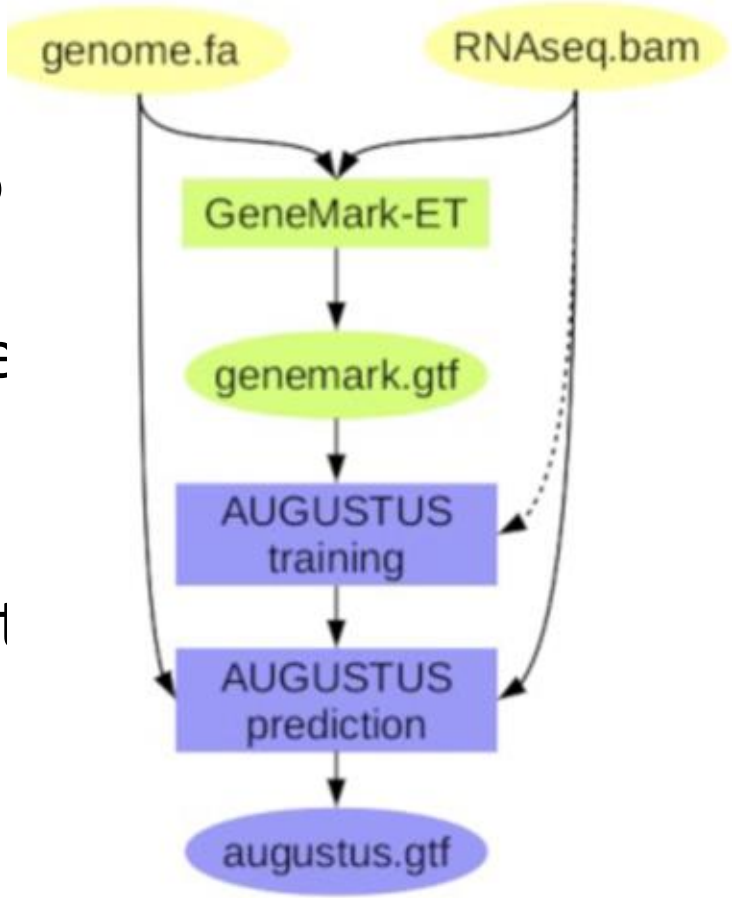
- Initial masking step against pre-curated library
- Model TEs from genome to add to library
- Mask all repeat elements
- Generate repeat annotations and repeat statistics
- Soft masked genome



Baril et al. 2024; *Molecular Biology and Evolution*

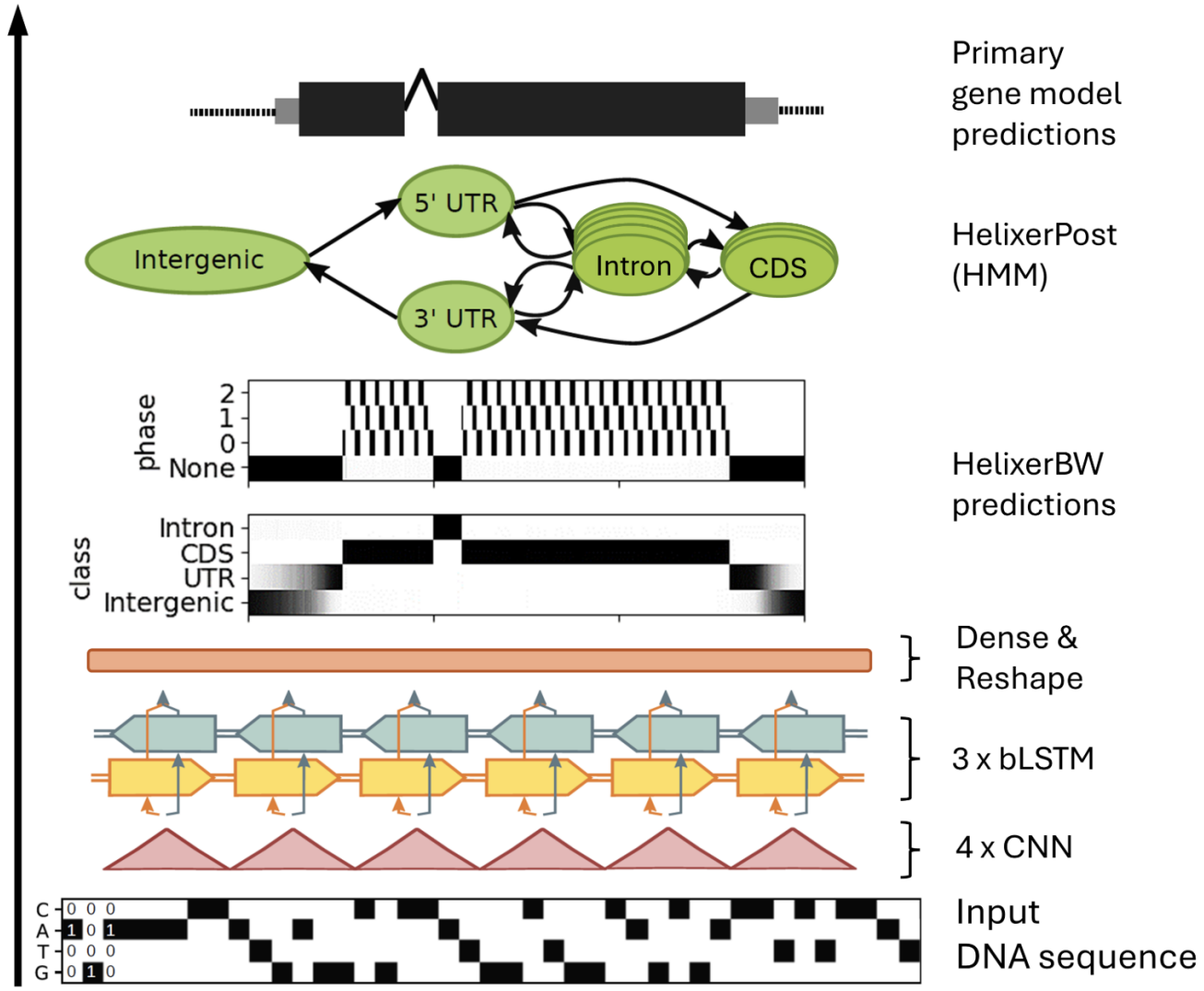
Braker4 for Gene Annotations

- Complete rewrite of BRAKER in Snakemake.
- Gene prediction logic is unchanged
  - GeneMark trains on extrinsic evidence, AUGUSTUS trained o GeneMark predictions, and TSEBRA merges results
- **Automatic repeat masking - I**f you provide an unmaske genome, BRAKER4 runs RepeatModeler2 and RepeatMasker before gene prediction
- **Automated RNA-Seq sampling -** If missing RNA-Seq dat BRAKER4 can use VARUS to automatically download RNA-Seq libraries from NCBI



Helixer – AI option for annotations

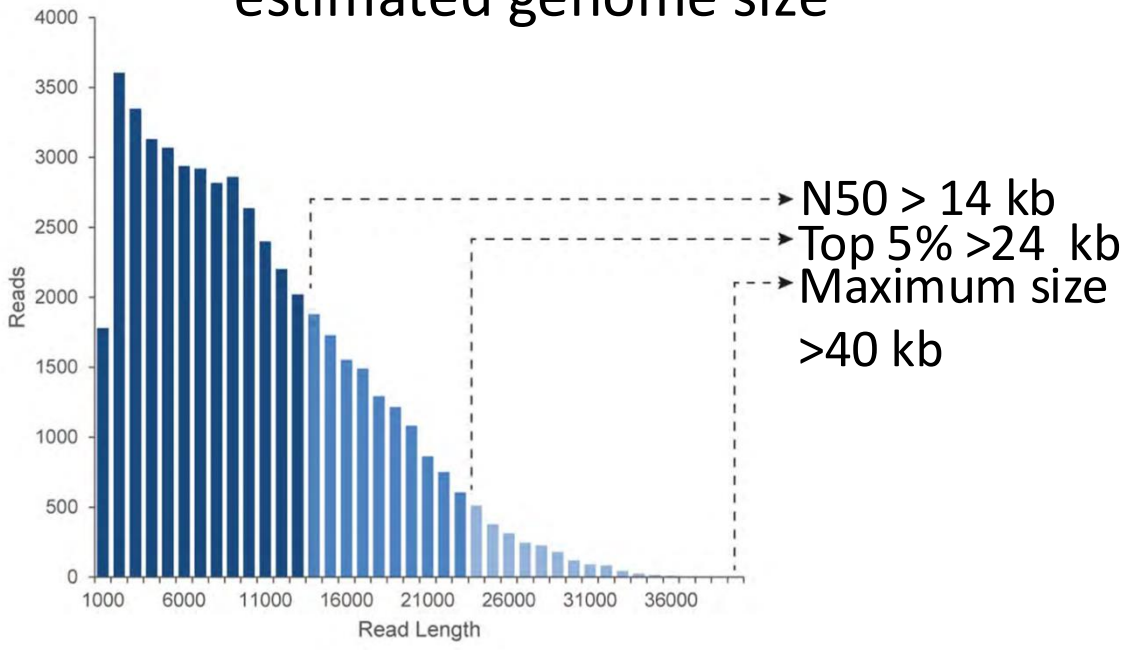
- Utilizes Deep Neural Networks and a Hidden Markov Model to directly provide primary gene models in a gff3 format
- Performs gene calling and UTR/CDS/Intron regions of genes
- Four trained models available: fungi, land plant, vertebrate and invertebrate
  - No need for species specific RNA-Seq data



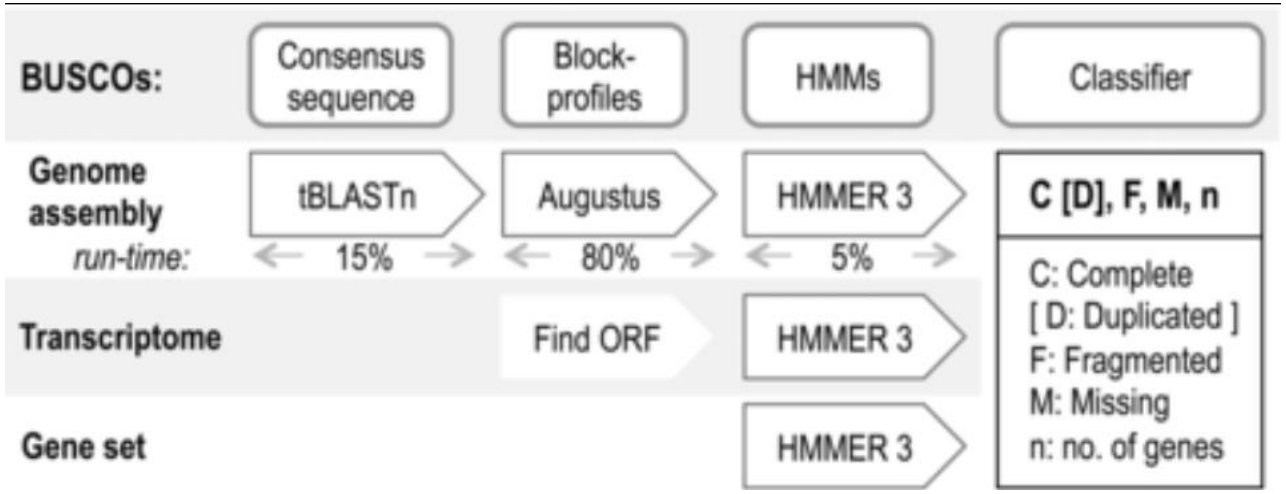
Holst et al. 2025; *Nature Methods*

Size and completeness

- Size metrics of the genome
  - Number of contigs
  - N50
  - Pseudomolecules/chromosomes
  - Assembled bases; percentage of estimated genome size



- Completeness of the genome
- Lineage libraries for single copy genes
- Assembly statistics using BBTools



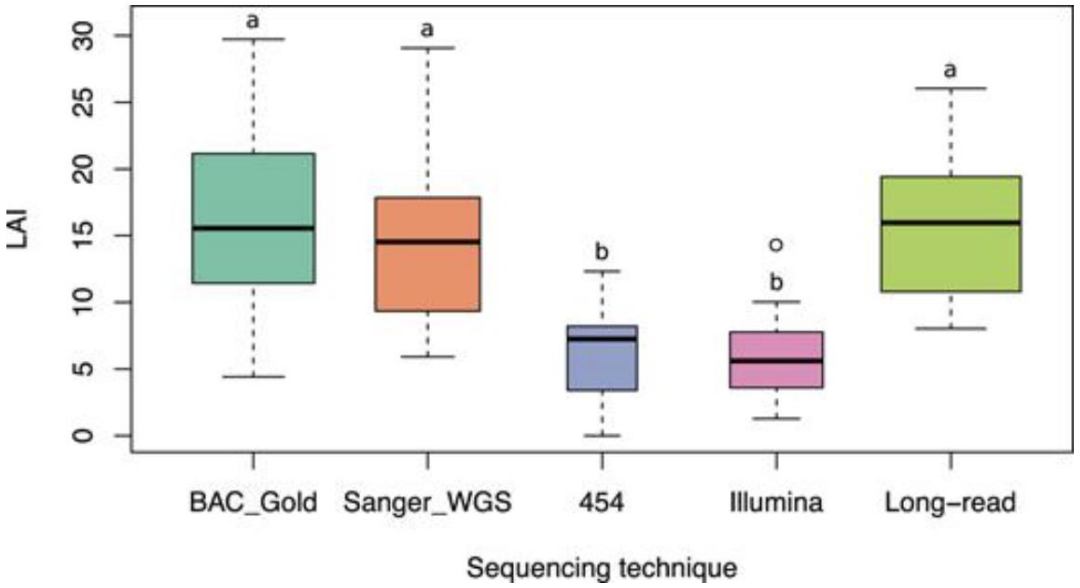
Simao et al. 2015; *Bioinformatics*

LAI – LTR Assembly Index

- Quantitative value indicating how contiguous your genome is
- Accurate identification of LTR retrotransposons (LTR-RTs)
  - Screens output from LTRharvest and LTR_FINDER
- When contigs are short, long LTR elements are less likely to be assembled, resulting in low LAI

$$\text{Raw LAI} = \frac{\text{Intact LTR retrotransposon length}}{\text{Total LTR sequence length}} \times 100$$

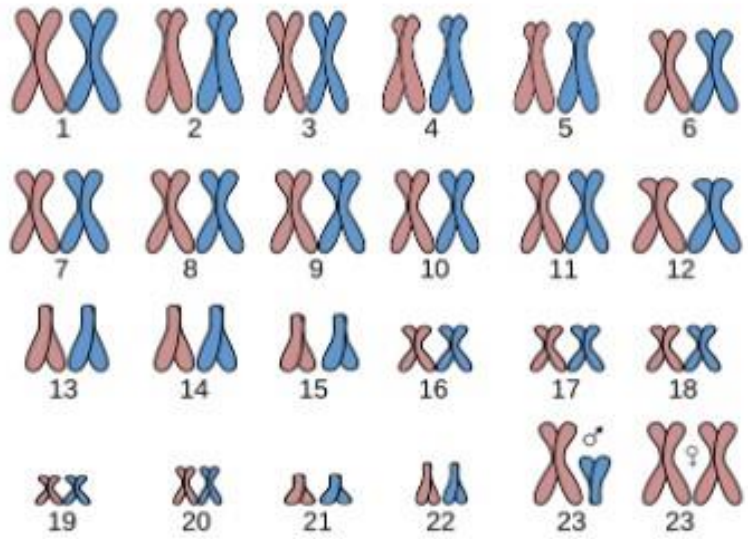
Category	LAI	Examples
Draft	0 ≤ LAI < 10	Apple (v1.0), Cacao (v1.0)
Reference	10 ≤ LAI < 20	Arabidopsis (TAIR10), Grape (12X)
Gold	20 ≤ LAI	Rice (MSUv7), Maize (B73 v4)



Ou et al. 2018; *Nuclei Acids Research*

Levels of Genome Assembly Completeness

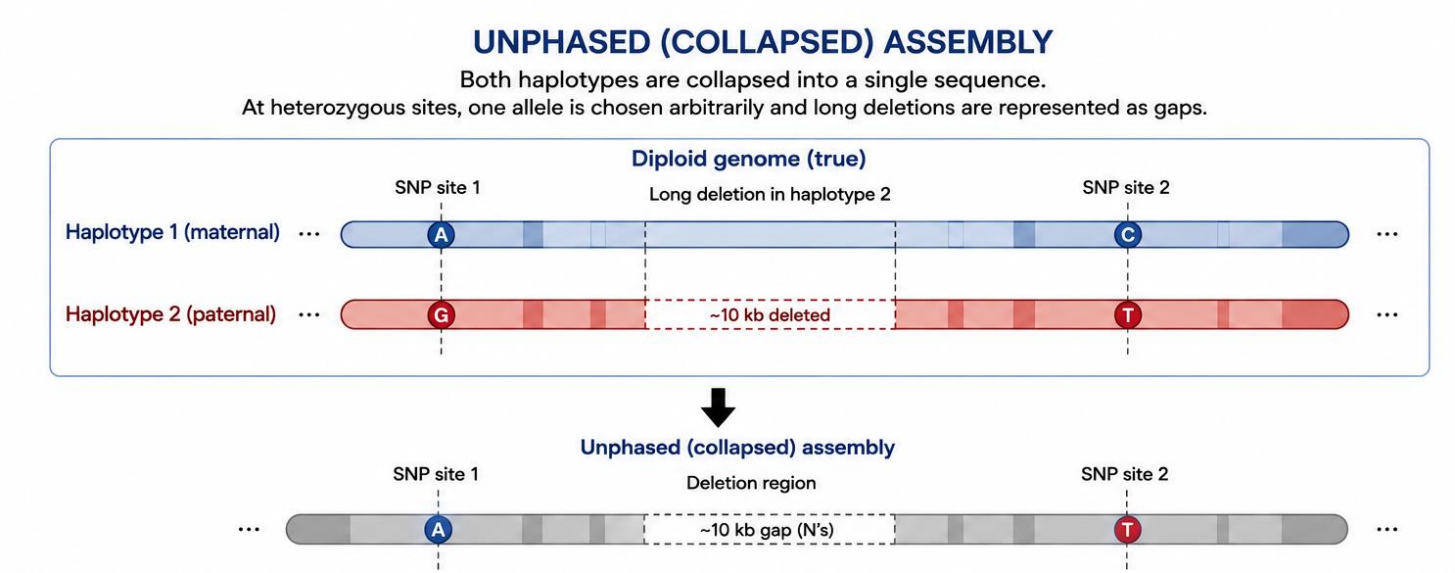
The "holy grail" of genome assembly
- telomere-to-telomere (T2T)



Human genome assembly:

46 sequences (23 pairs) with no gaps

-> One step lower: unphased genome assembly



Unphased Assemblies

- Chromosome-scale (pseudo-molecules)
- Scaffold-level
- Contig-level

Haplotype-resolved

- T2T
- Chromosome-scale haplotype-phased assembly
- Haplotype-phased assembly

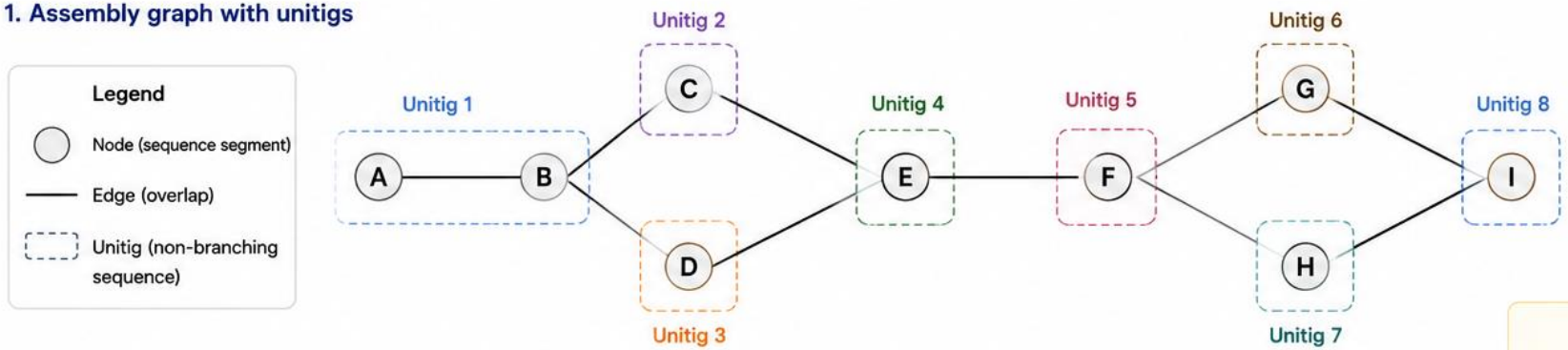
Assembly graph, unitigs, and two haplotype paths

Assembly graph
Nodes are sequence segments.
Edges connect segments based on overlaps.

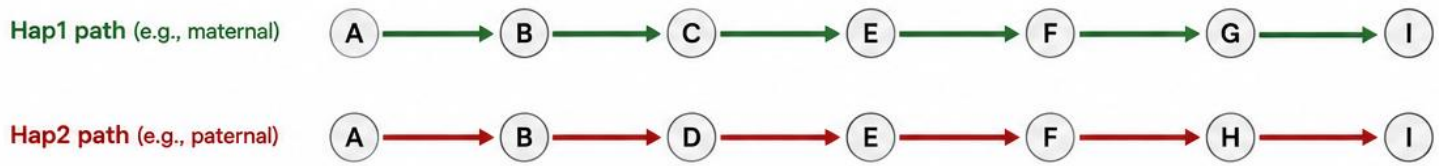
Unitig
A maximal non-branching path in the graph.
Each unitig is a contiguous sequence.

Hap1 and Hap2 paths
Two haplotype-resolved paths through the graph.
They represent the two chromosome copies.

1. Assembly graph with unitigs



2. Two haplotype paths through the graph



Both haplotypes **share** unitigs 1, 4, 5, and 8 (A–B, E, F, and I).
They **differ** in unitigs 2 vs 3 (C vs D) and 6 vs 7 (G vs H).

3. Unitigs (sequence of each non-branching path)

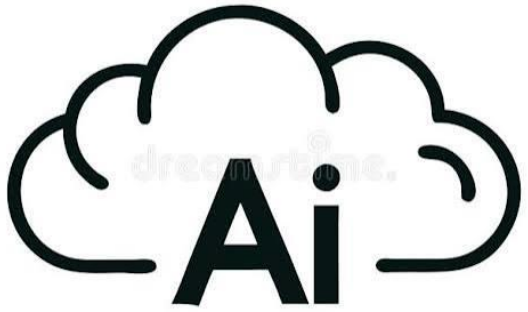
Unitig 1 (A–B)	Unitig 2 (C)	Unitig 3 (D)	Unitig 4 (E)	Unitig 5 (F)	Unitig 6 (G)	Unitig 7 (H)	Unitig 8 (I)
ATGCGT... ...GCTTAA	GGAACC... ...TTGGCC	GGTGTA... ...TACACC	CCATGA... ...TTCGGA	TGGCAA... ...AGTCCG	AACCTT... ...GGTCAA	AAAGGT... ...ACCTTT	TTGACC... ...CCGTAA

Key points

- **Unitigs** are the building blocks (non-branching sequences) of the graph.
- **Hap1** and **Hap2** are two different paths through the same graph.
- Shared unitigs appear in both paths.
- Alternative unitigs capture differences between the two haplotypes.

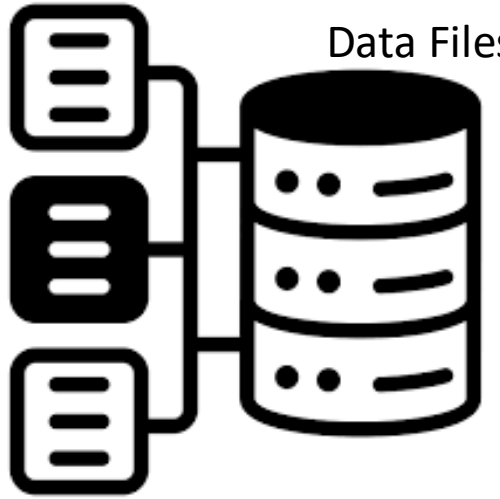
Hap1 path
Hap2 path
Unitig (non-branching)

What Do You Need for AI Agent-Assisted Data Analysis?



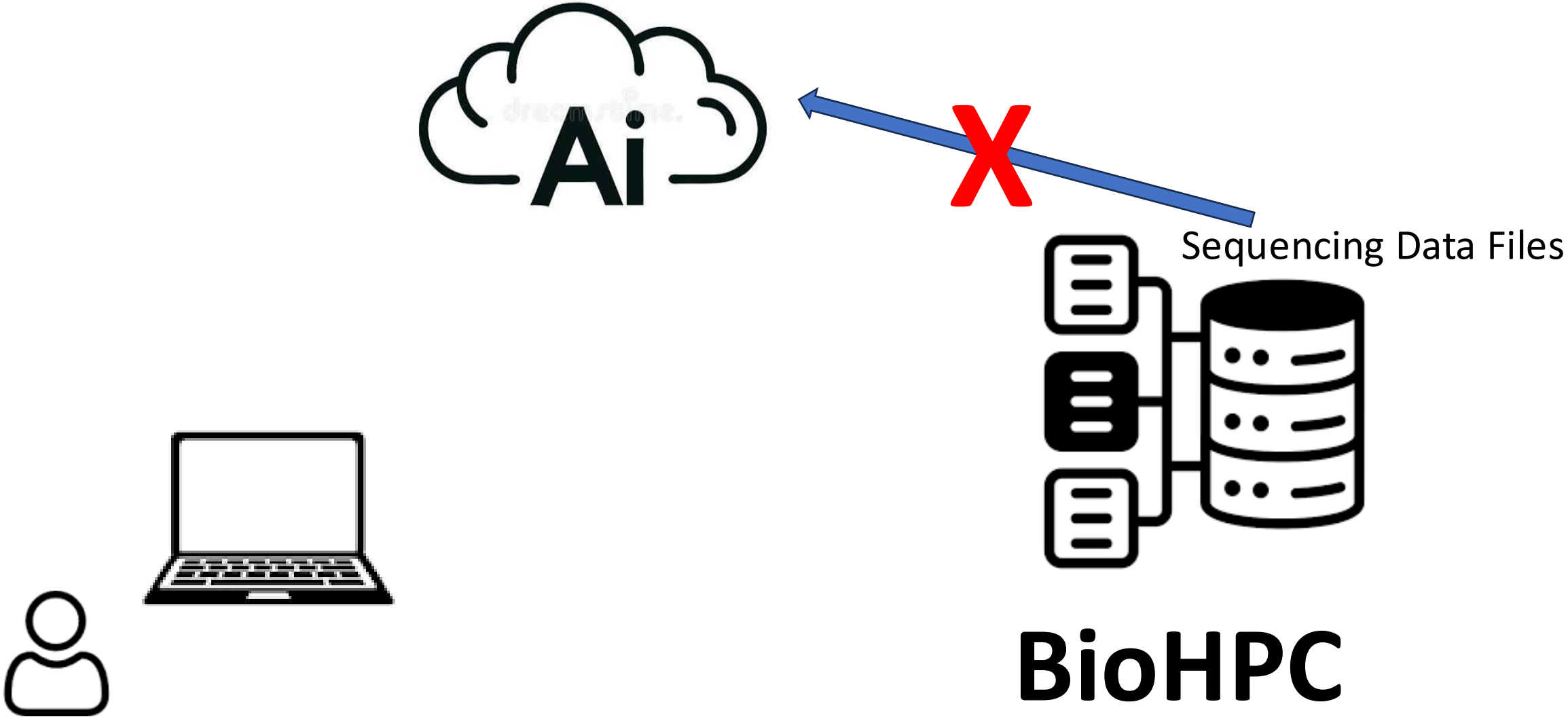
AI Subscription
ChatGPT or Claude

Your Laptop

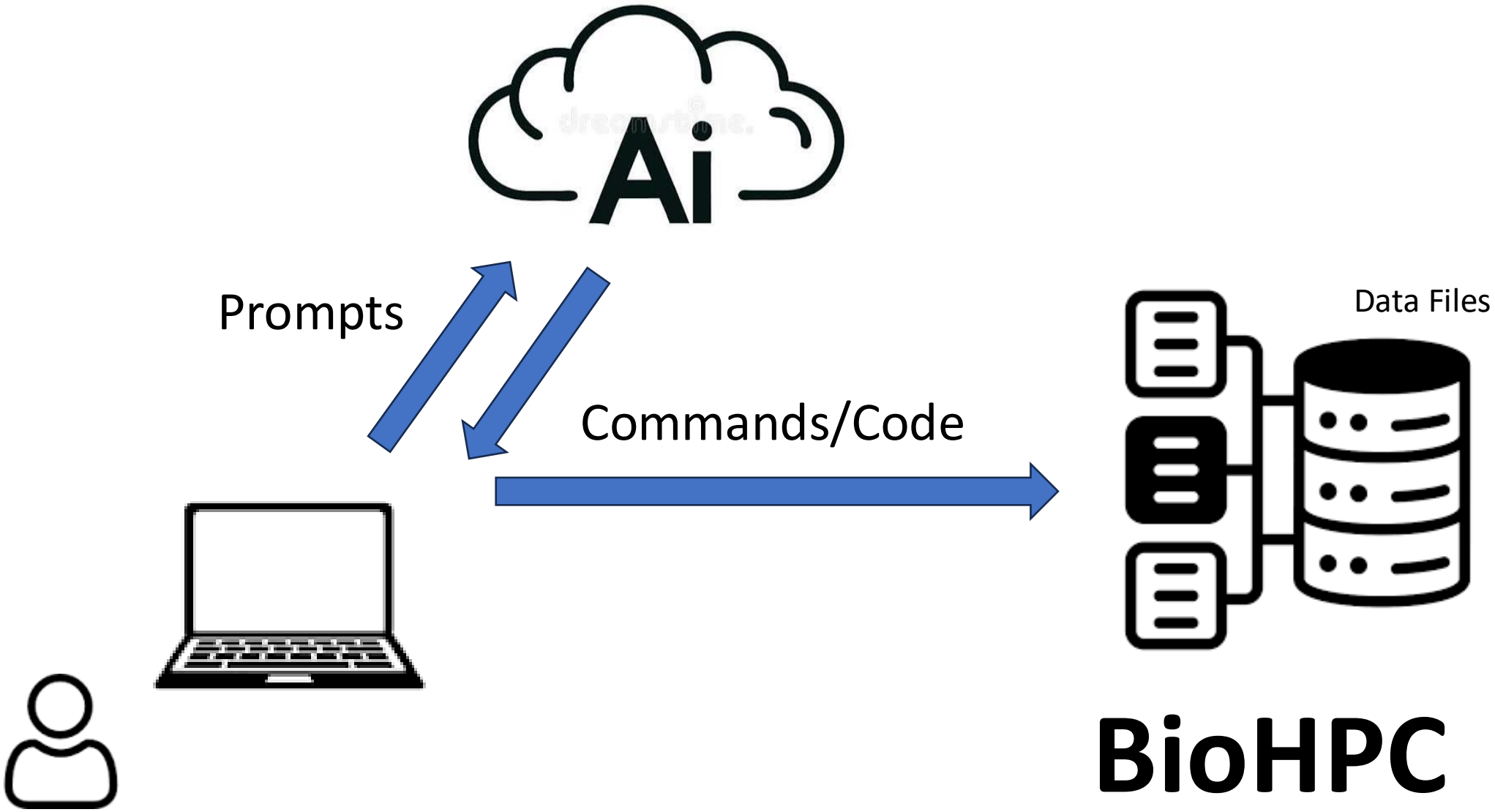


HPC / Linux Server
BioHPC / BioSlurm

Your sequencing data are NOT sent to the LLM for analysis.

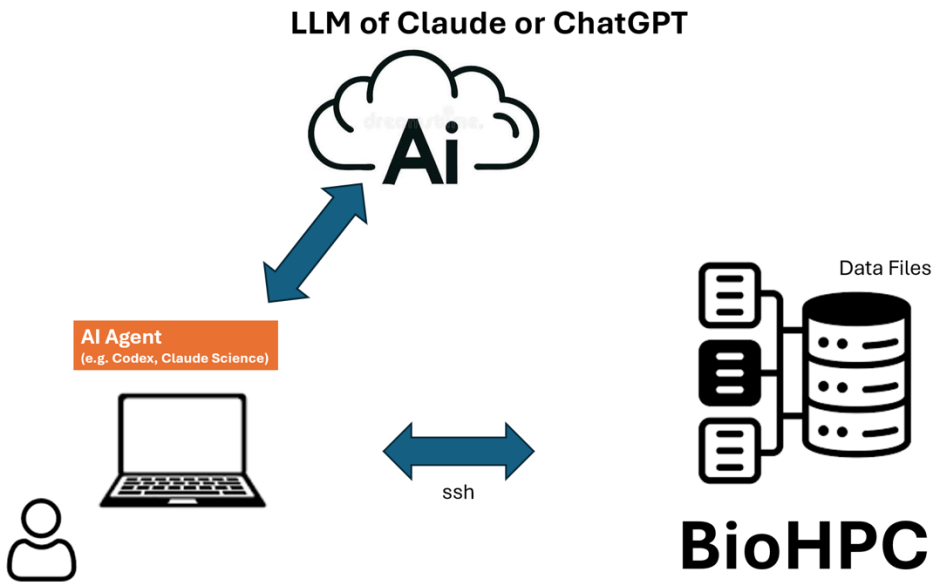


How Does the AI Agent Work?

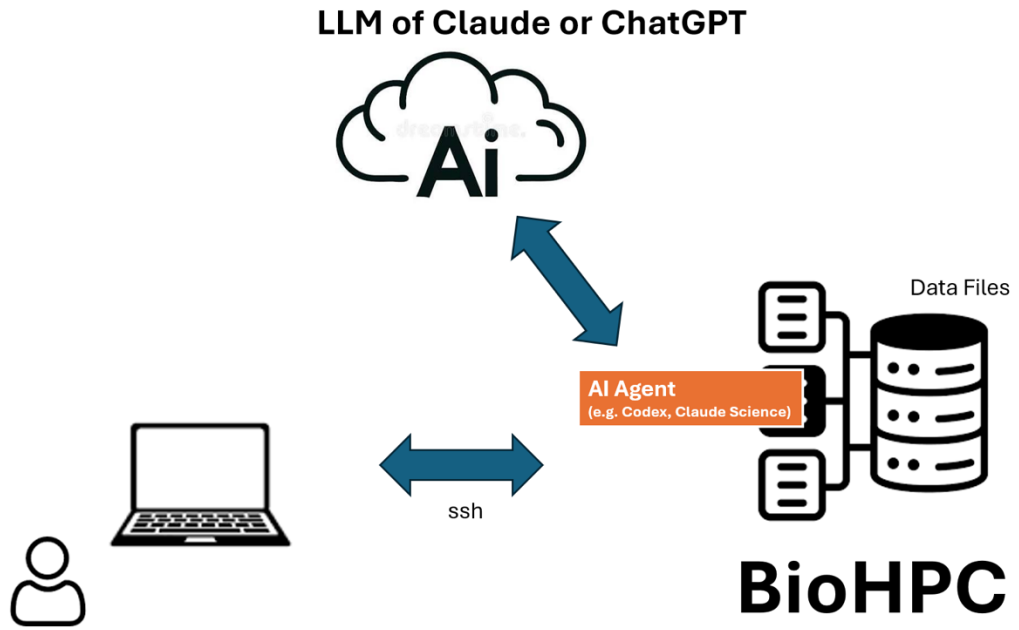


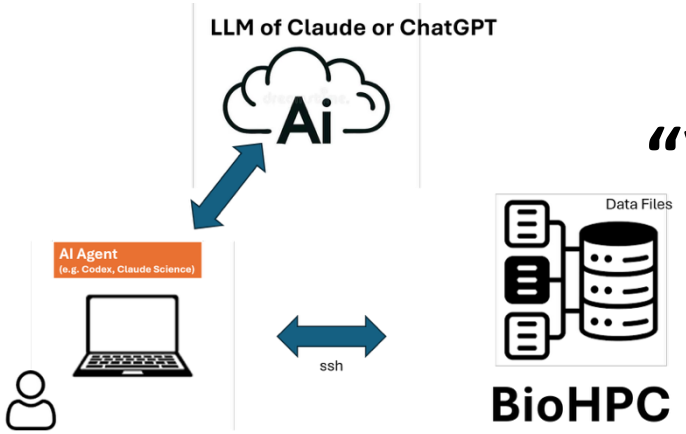
Where Does the AI Agent Run?

Agent runs on your laptop



Agent runs on BioHPC server





“Vibe Coding”: Prompts Instead of Commands

Traditional Bash command:

```
hifiasm -o sample.asm -t 32 reads.fastq.gz
```

Vibe coding:

> Assemble the HiFi reads in file reads.fastq.gz, using hifiasm with 32 CPU cores

You still need to understand the analysis — you don't necessarily need to write every command.

Agent Instructions and Analysis Skills

Agent Instruction — AGENTS.md

How to work on this computing system

- Available software and where it is installed
- CPU, memory, GPU limits
- BioSlurm usage
- File and temporary-directory locations
- Containers
- Rules and constraints

Analysis Skill — GENOME_ASSEMBLY_SKILL.md

How to perform a specific analysis

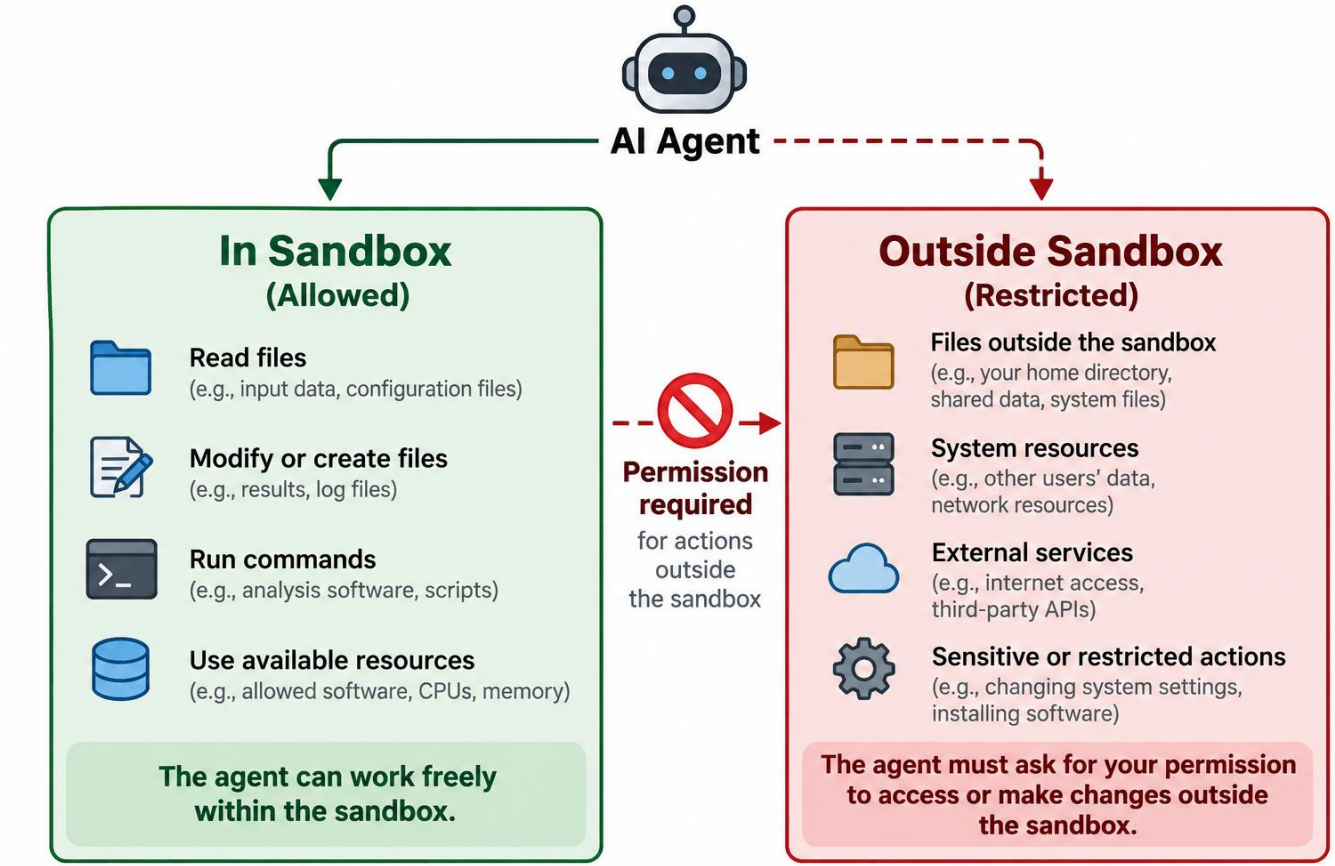
- Select appropriate analysis software
- Recommended workflow
- Important parameters
- QC and evaluation
- Interpretation of results

Data Security — What is sent to the LLM?

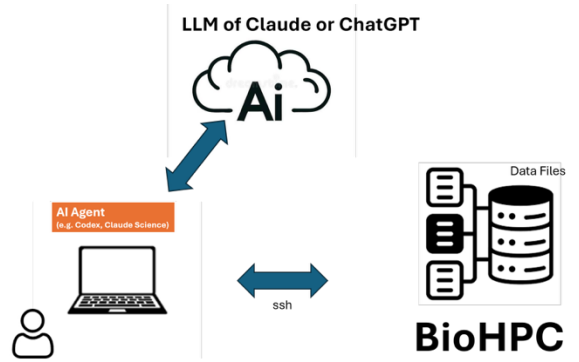
Prompt:	Inspect assembly.log
Agent:	Read -> Send selected content to LLM for interpretation

Cornell-managed enterprise AI
→ Additional institutional privacy/data-use protections.

Sandbox — What can the agent access or change?

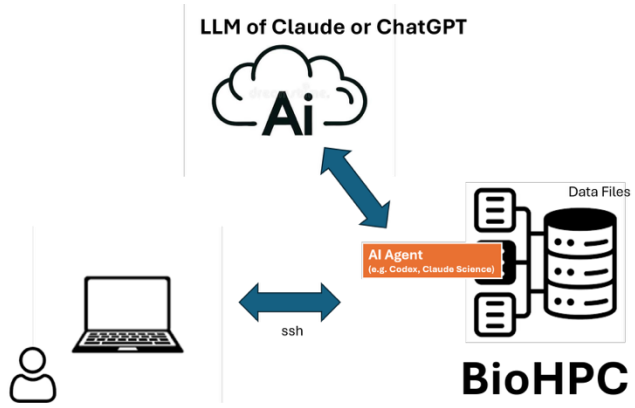


Agent on laptop



-- Agent is sandboxed locally, but BioHPC files may not be.

Agent on BioHPC



-- The BioHPC project workspace is sandboxed.

Context and Token usage

Your relationship with the AI Agent

Too little context	Good context	Too much / stale context
😕	🙂	😨
New friend	Best friend	Complicated

Some techniques for context management

- **Be explicit** Clearly state the goal, inputs, constraints, and expected output
- **Keep** Save important information in files, not only in conversation history
- **Compact** long conversations to reduce unnecessary context
- **Split** complex analyses into separate contexts/stages
- **Restart** with a fresh context when the conversation becomes confused or outdated

(more in hands-on)

Accuracy and Reproducibility

Code is precise; natural language can be ambiguous

- **Be explicit** about inputs, methods, parameters, and expected outputs
- **Review** commands/code generated by the agent
- **Save** scripts, commands, results, and an AI-generated summary/checkpoint file
- **Validate** results using appropriate QC and scientific knowledge

AI-generated summary/checkpoint file

Long AI session
↓
Ask agent to summarize
↓
assembly_summary.md
↓
New AI session
↓
Read assembly_summary.md
↓
Continue analysis

Reproducibility

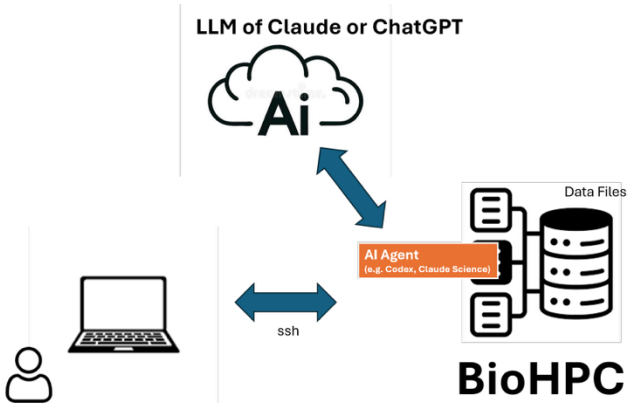
- The resulting analysis must be precise and reproducible.
- Reproducible does not mean byte-for-byte identical output.

Beyond Running Pipelines: AI-Assisted Research

- **Generate data visualization**
- **Generate hypotheses**
- **Challenge interpretations**
- **Identify alternative explanations and artifacts**
- **Make testable predictions**
- **Design analyses or experiments to test them**
- **Refine hypotheses based on evidence**

AI can assist throughout the research cycle — not just process data.

Agent on BioHPC



Hands-on project

Tasks:

- Assemble a genome
- QC and genome annotation

Workshop web site:

https://biohpc.cornell.edu/ww/1/Default.aspx?wid=172

- Self-guided
- Instructions provided
- Office hours available