HARVARD
Faculty of Arts and Sciences

Long Read Workshop

Kelly Cribari, Danielle Khost, Lei Ma

Bauer Core

- Offer instrumentation and expertise to serve our scientists
- Advance research goals and efforts
- Offer services and training to internal and external users
  - QC
  - Library Preparation
  - Sequencing

Bauer Core

- Kelly Cribari, Long Read Sequencing Team Lead
- Manage PacBio and ONT projects at the Core
  - Sample Preparation
  - Library Preparation
  - Sequencing

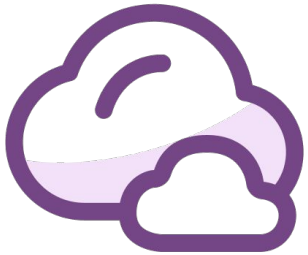
We help FAS researchers with bioinformatic analysis

ONE-ON-ONE CONSULTS

ONGOING COLLABORATIONS

BIOINFORMATICS WORKSHOPS

- Danielle Khost, PhD University of Rochester Biology
- Expertise in genome assembly, long-read sequencing technologies
- Recent project: developing pipelines to identify structural variants from long-read sequencing data



What comes to mind when you think Long Read Sequencing and Analysis?

How much experience do you have with Long Read Sequencing?

Are you interested in general information or project specific questions regarding sample prep/sequencing, or bioinformatics?

Outline

Bauer Core Information

- Submission best practices
- Quality checking your samples
- Bauer Core instrumentation and Submission
- Sequencing options

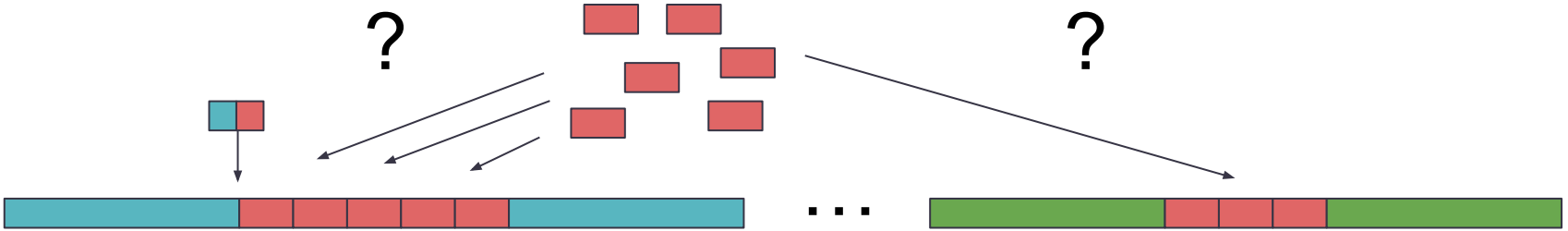
Bioinformatics

- ONT vs PacBio
- Assembly with Hifiasm
- HiC - Arima Lunch and Learn tomorrow Nov. 20th!
- QC Metrics
- How do I pick?
- Real world examples

Sequencing

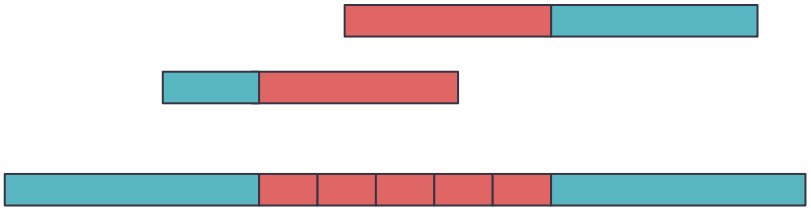
Why do long read sequencing?

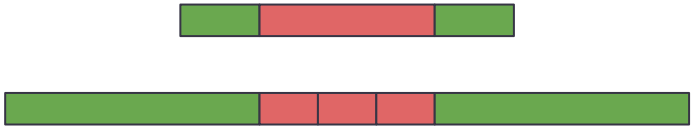
- To sequence a genome: fragment the genome, sequence the pieces (i.e. reads), put everything back together
- *The problem: with short sequencing reads, cannot confidently assign read to position in case of repetitive or structurally complex locus*

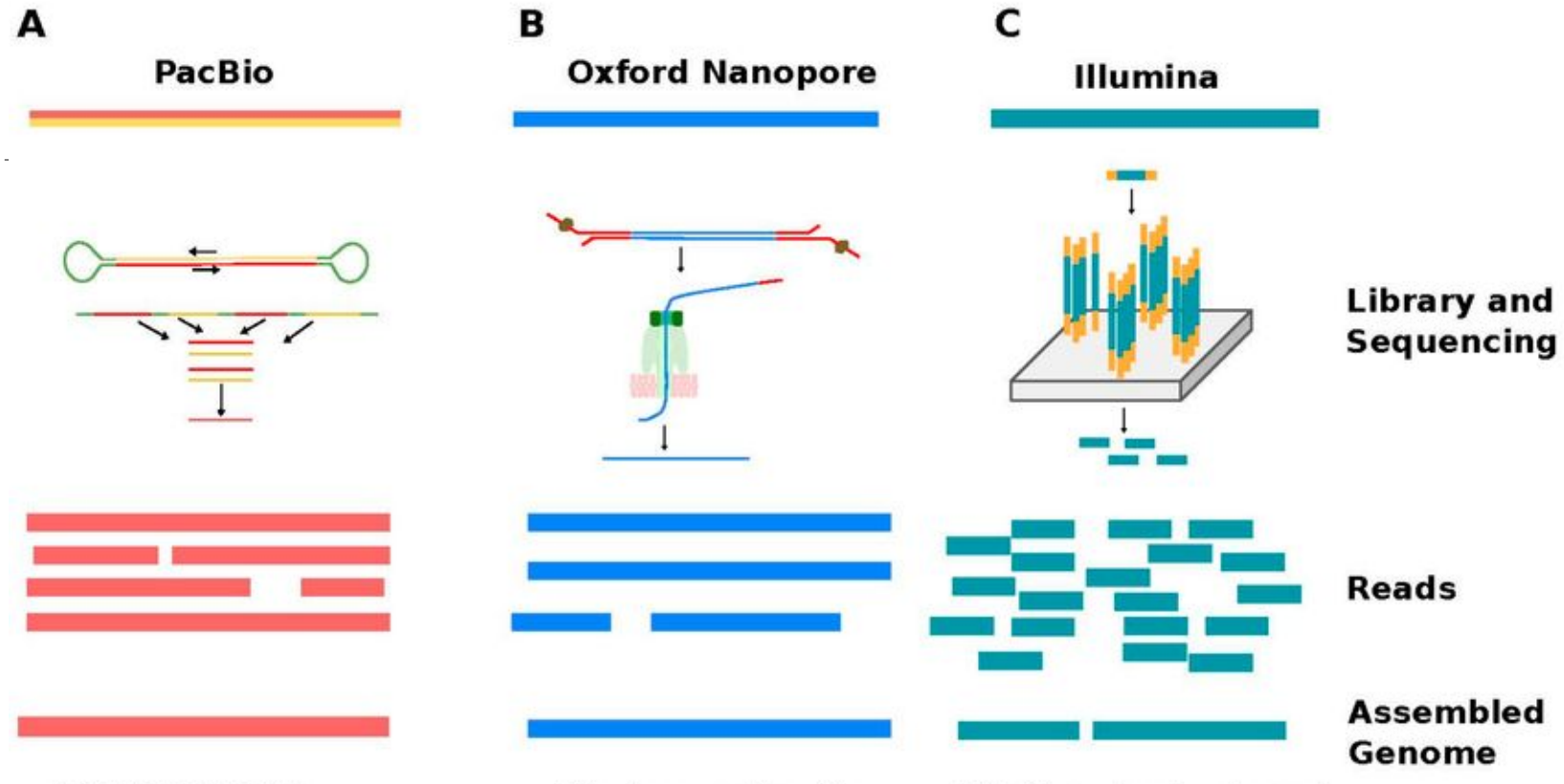


Why do long read sequencing?

- To sequence a genome: fragment the genome, sequence the pieces (i.e. reads), put everything back together
- ***Long reads can bridge repetitive and problematic regions, allowing us to properly assemble them***






A
PacBio
B
Oxford Nanopore
C
Illumina
Library and Sequencing
Reads
Assembled Genome

Definitions

HMW DNA

- High Molecular Weight DNA
- Long, Intact DNA Fragments
- Generally 50 Kb or longer

Kb

- kilobases

Best Practices to Achieve HMW DNA

Good quality HMW DNA starts with sample storage

1. **Blood samples**
    a. **Within week**
2. **Tissue samples**
    a. **Fresh or frozen**
    b. **Avoid freeze- thaws whenever possible.**
3. **Cells**
    a. **Washed and pelleted**

Avoid EtOH storage when possible

Extraction Options

Pick what's best for your organism!

a. **Modifications to the protocol**
    i. **Longer lysis**
    ii. **Long elutions**
    iii. **Proper mixing**

Tips

Elution and Solubilization of the DNA

Gentle pipetting

Gently heating at 37°C

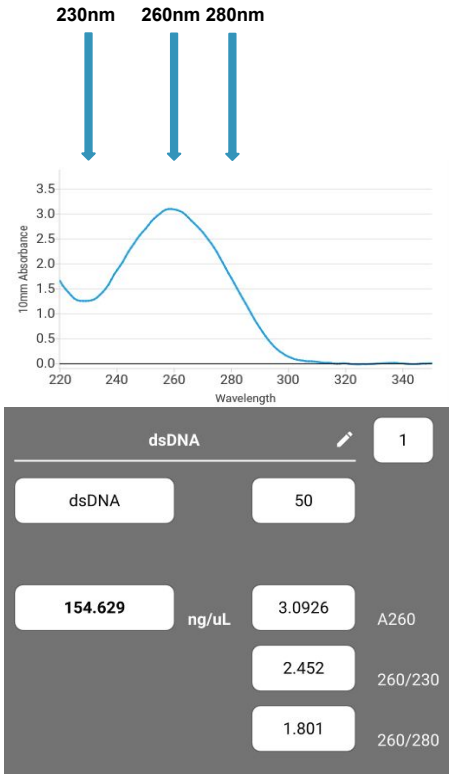
Definitions

Purity Ratios

- Ratios of absorbance at certain wavelengths
- Different molecular are absorbed at different wavelengths
  - 260 nm - DNA
  - 280 nm - Proteins
  - 230 nm - salts, EDTA, carbohydrates, EtOH, etc.

HARVARD
Faculty of Arts and Sciences

Post-Extraction Quality Checks: Purity



Purity

1. 260/280 Ratio
   a. Protein or phenol
   b. 1.8 is considered pure
2. 260/230 Ratio
   a. Salts, phenol, EDTA, or carbohydrates
   b. 2.0-2.2 best

Impurities impact downstream sequencing

Not all impurities show up

Post-Extraction Quality Checks: Concentration

Nanodrop concentrations can be inaccurate

Qubit is best practice

1. **HMW DNA can be difficult**
    a. **Long strands**
    b. **Viscosity**
2. **Take two or three readings for a more accurate quantification**
3. **1ug of DNA needed for most LRS Library Preps**

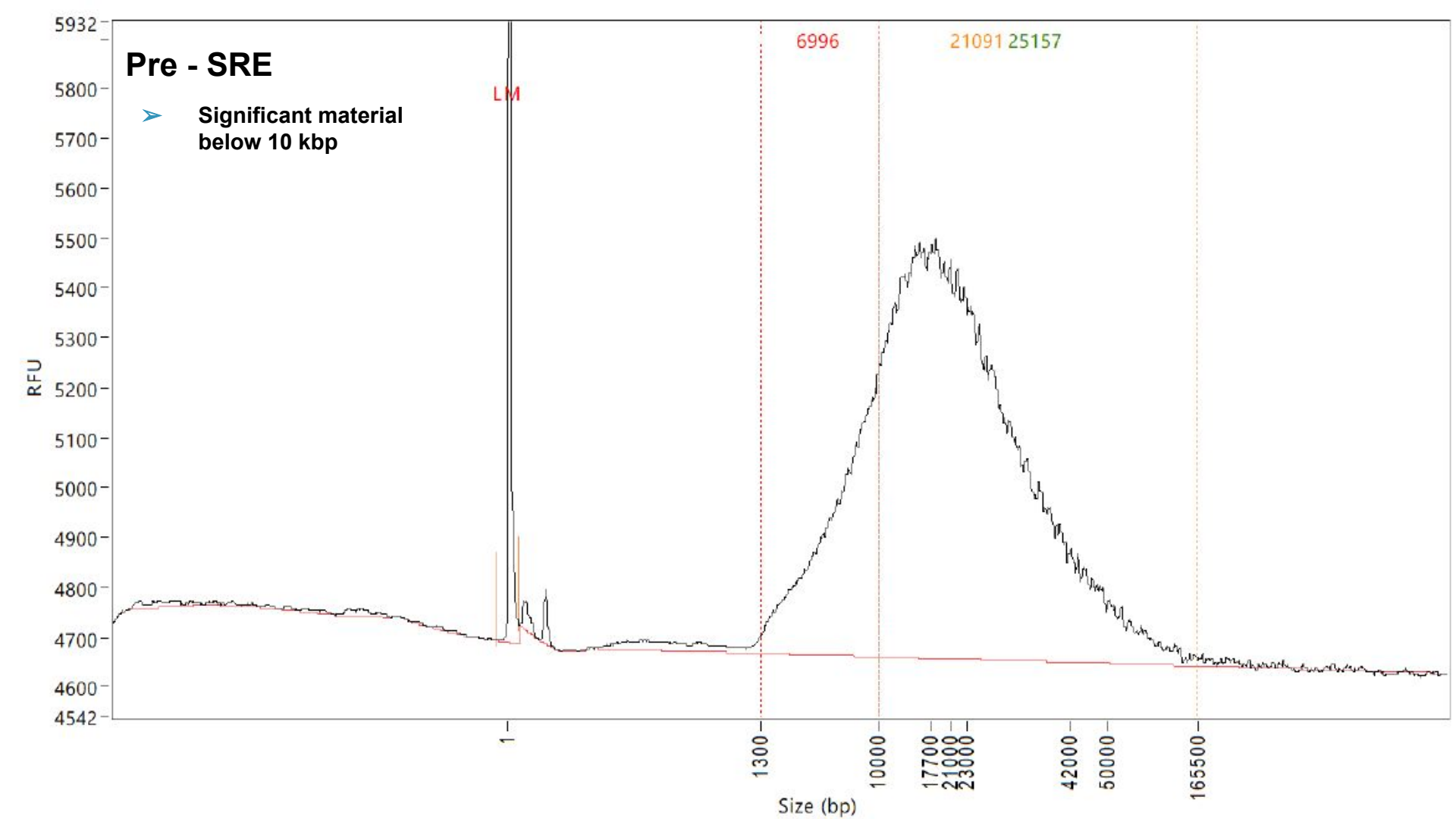
Post-Extraction Quality Checks: Size

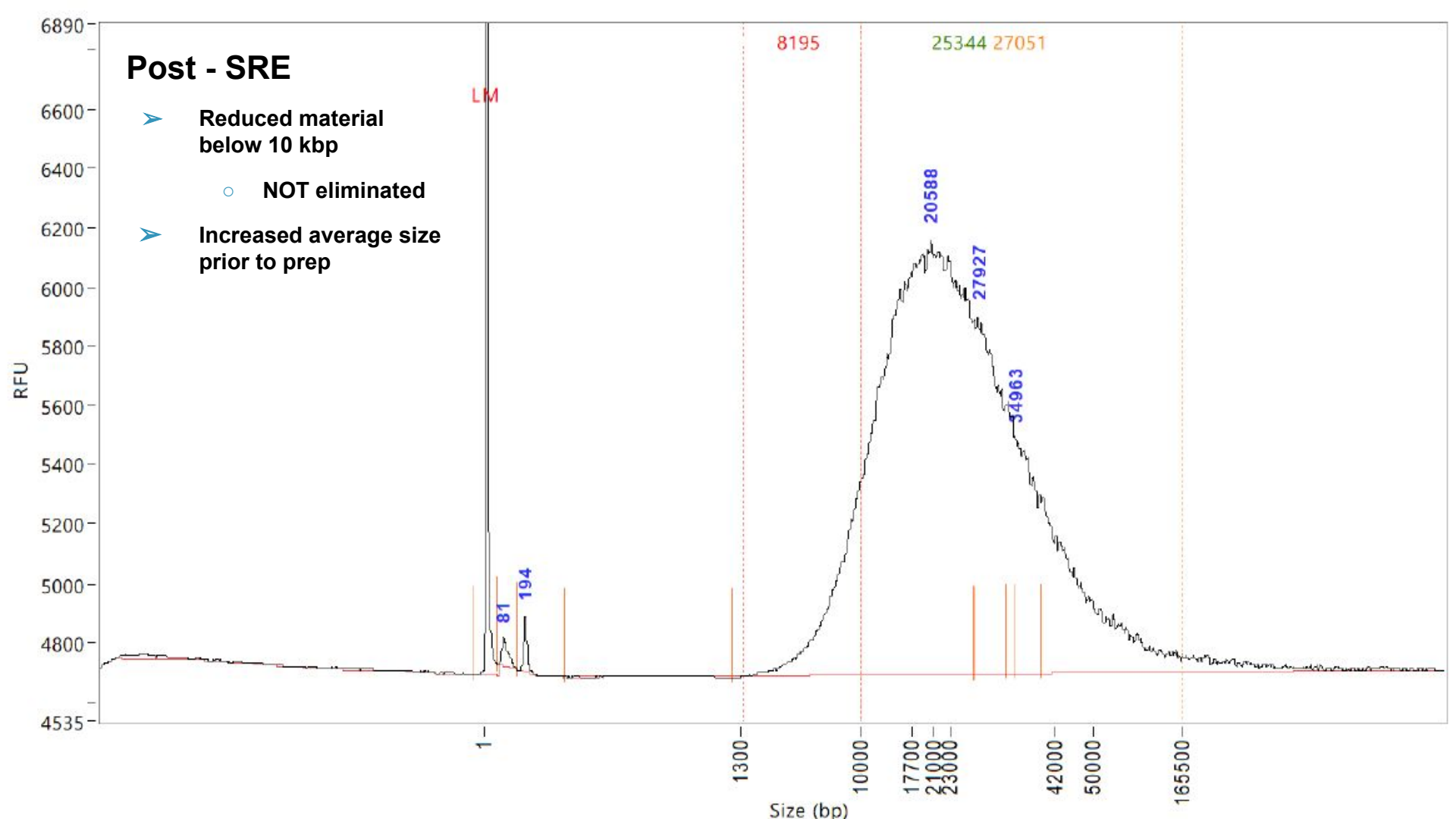
TapeStation

1. Determining success rather than size
2. Look for small fragments below 10kb

Short Read Elimination (SRE)

1. Reduces fragments below 10kb
2. Expect to lose ~50% of the material
3. Plan ahead and provide extra material to account for loss
    a. 500 ng per sample for the PacBio preps after loss
    b. 1000 ng for ONT preps after loss


Pre - SRE
Significant material below 10 kbp
LM
6996
21091 25157
RFU
5932
5800
5700
5600
5500
5400
5300
5200
5100
5000
4900
4800
4700
4600
4542
1
1300
10000
17700
21000
23000
42000
50000
165500
Size (bp)


Post - SRE
Reduced material below 10 kbp
NOT eliminated
Increased average size prior to prep
LM
8195
25344 27051
20588
27927
34963
194
81
RFU
6890
6600
6400
6200
6000
5800
5600
5400
5200
5000
4800
4535
1
1300
10000
17700
21000
23000
42000
50000
165500
Size (bp)

Bauer Core Offerings

Core Provided Instrumentation:

1. NanoDrop
2. Qubit
3. TapeStation
4. Femto Pulse

➢ **Training is available to use the TapeStation!**

Sign up here to secure a spot at an upcoming training

Submission Process

- Switching from the Minilims system to LockBox
  - Currently in the user testing phase
- All in one system for:
  - Sample information
  - Sequencing information
  - Communication
  - Results
  - Billing
- More to come as testing completes

Definitions

Q-Score

- A log scale measurement of the probability an incorrect base is called at a specific position along the DNA sequencing read
  - Q30 = 1 in 1,000 bases are incorrect (0.1%)
  - Q40 = 1 in 10,000 bases are incorrect (0.01%)

Polymerase

- Enzyme that synthesizes long chains of nucleic acids

Nicks

- A break or discontinuous segment in a double stranded DNA molecule

Sequencing Outputs and Coverage

Exact output of a sequencing run will depend on fragment lengths and quality

- **Smaller fragments will usually have higher Q-Scores when sequenced with PacBio**
- **Lower than expected yields may be attributed to:**
  - **Small fragments still present after SRE**
  - **Nicks in DNA**
  - **Unknown contaminants or inhibitors**
- **Most research goals are achieved with 30x coverage**

Sequencing Options

	PacBio Revio	ONT PromethION	ONT MinION
Input Requirements	1-50 ng, 500 ng	1000 ng	1000 ng
Expected Output	7-9M Reads 80-100 Gb	100-120 Gb	10 Gb
Library Prep Options	SMRTbell 3.0 Kinnex Suite AmpliFi	LSK114 NBD114 Rapid Barcoding Ultra Long Read	LSK114 NBD114 Rapid Barcoding Ultra Long Read
QScore	35-40	20	20

Sequencing Options

	PacBio Revio	ONT PromethION	ONT MinION
Library Prep Cost	$647 - $1526	$108 - 1038	$108 - 1038
Flowcell Cost	$1581	$1261	$968
Cost Per GBase (sequencing only)	$19.76	$12.61	$96.80
Multiplexing	Yes	Yes	Yes

Definitions

Methylation

- Process in which a methyl group is added to DNA
- Can turn a gene “off” by preventing activation and therefore protein production

Kinetics

- Timing, rate, and mechanisms of DNA methylation
- Important in replication

HARVARD
Faculty of Arts and Sciences

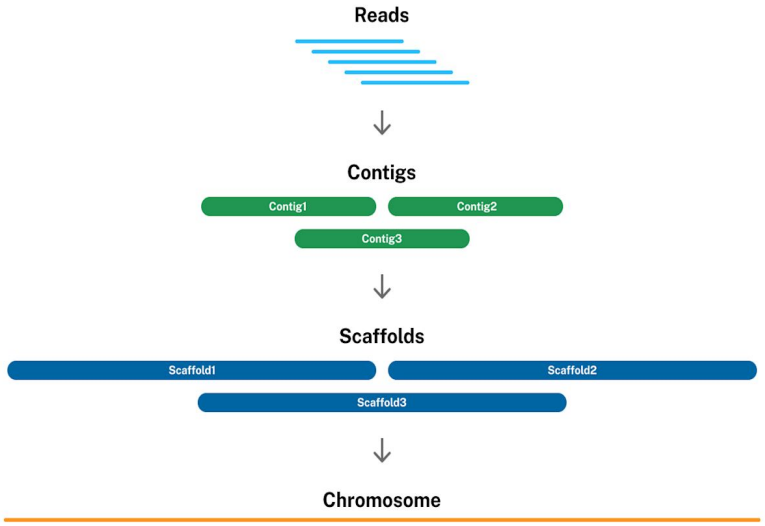
Methylation, Kinetics, and Modifications

- Methylation is a key mechanism that regulates gene expression
- Analysis of these modifications is most useful for Whole Genome Sequencing applications
- Methylation data is not recommended for any applications involving a cDNA step
- Both ONT and PacBio offer sequencing options that allow analysis of methylation

Bioinformatics

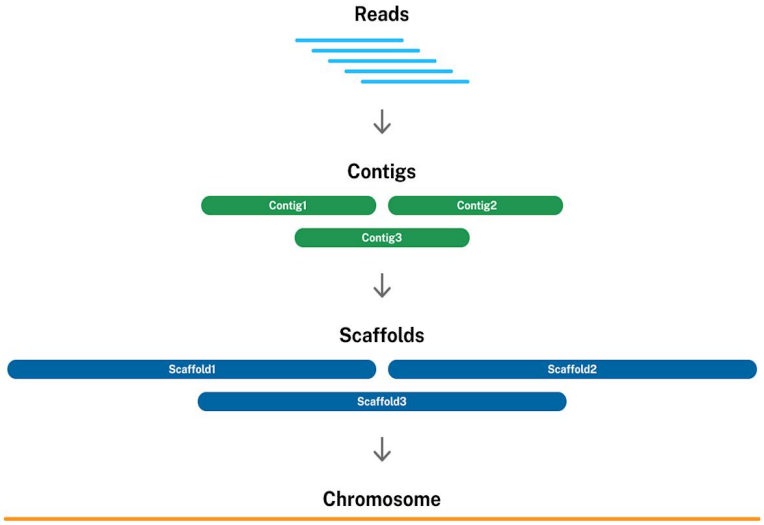
Definitions

- Coverage: proportion of genome/region that is covered by at least one sequencing read
- Depth: redundancy of coverage, i.e. how many reads align to each base of assembly
- Assembler: software used to generate (i.e. assemble) genome from reads. Most likely fragmentary
- Contig: a contiguous sequence of bases generated by an assembler. Contains no gaps or unknown sequence
- Scaffold: multiple contigs that have been stitched together. May contain gaps of known or unknown size



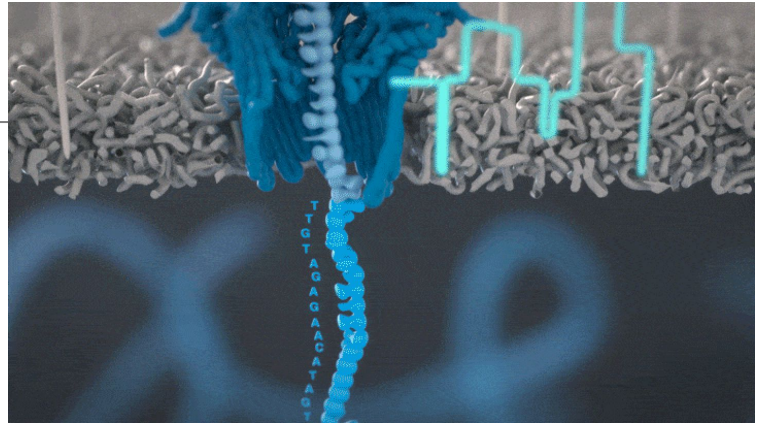
Definitions

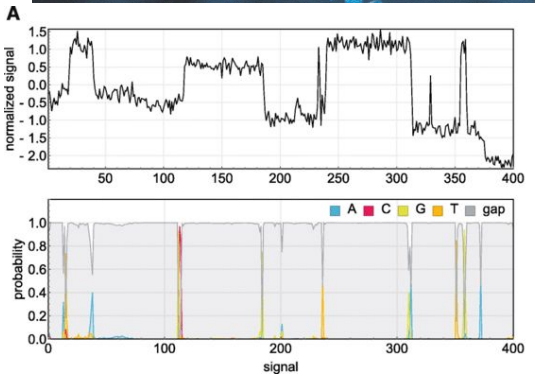
- **Primary assembly: complete assembly representing haploid genome, but unphased**
- **Phased assembly: resolving an assembly into haplotypes. Can be partially or fully phased**
  - **Partial: complete assembly with long stretches of phased blocks, but some switching of parental alleles**
  - **Fully: complete assembly with each parental haplotype correctly separated w/ no switching**
- **HiC: sequencing using chromatin capture to detect interactions between chromatin (3d structure)**
- **Q score: Phred quality score, estimates accuracy of base in read. Log scaled (e.g. Q20 = 99%, Q30 = 99.9%...)**

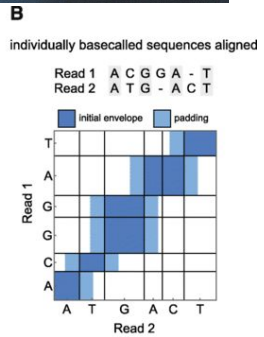


ONT Sequencing

- DNA/RNA read thru pore across voltage differential
- Causes current to fluctuate, generating a “squiggle”
- *Basecalling algorithms* translate squiggle into base pair (kmer)
  - Signal vs noise determines accuracy
  - Struggles with certain sequences (homopolymers) but resolution improving
- Also can detect modified bases (5mC, 5hmC, 6mA)
  - Integrated into basecalling process (need to enable!)
  - Also possible post-calling
- Ultra-long read optimization

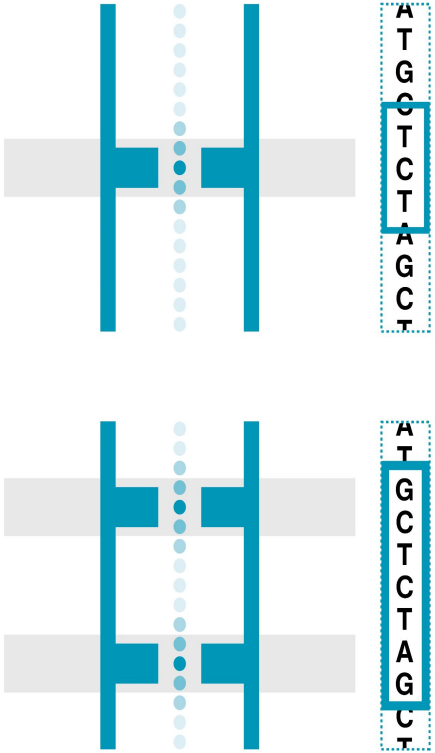






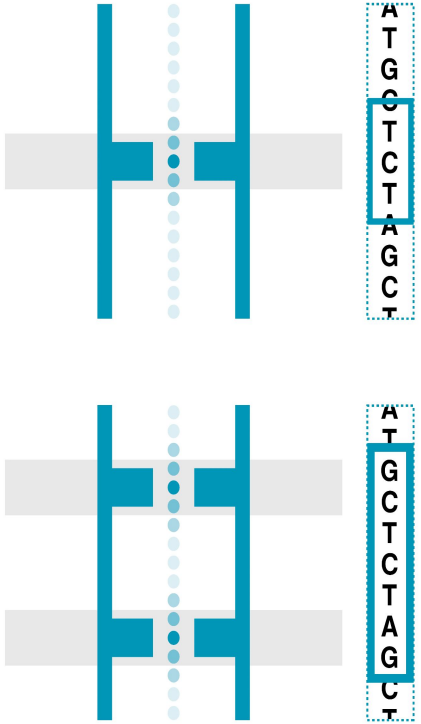
Nanopore and Accuracy

- Older chemistries had higher error rates (10-15% for R9)
  - Necessitated post-assembly polishing (Illumina reads or self-correction)
  - Struggles with homopolymer regions
- R10 chemistry adds second reader in pore to increase accuracy
  - Basecalling models tuned for higher accuracy
  - Error rate 1-2% (10x PacBio, 100x Illumina)
- Pre-correction of reads (HERRO, DeChat, etc.) or as part of assembly process
- <u>TL;DR: post-assembly polishing with short reads not necessary with latest chemistries</u>



Nanopore and Accuracy

- Older chemistries had higher error rates (10-15% for R9)
  - Necessitated post-assembly polishing (Illumina reads or self-correction)
  - Struggles with homopolymer regions
- R10 chemistry adds second reader in pore to increase accuracy
  - Basecalling models tuned for higher accuracy
  - Error rate 1-2%
- Pre-correction of reads (HERRO, DeChat, etc.) or as part of assembly process
- TL;DR: post-assembly polishing with short reads not necessary with latest chemistries
- Duplex reads option: 99.8% accuracy
  - Reads double strand instead of single strand
  - Much less output

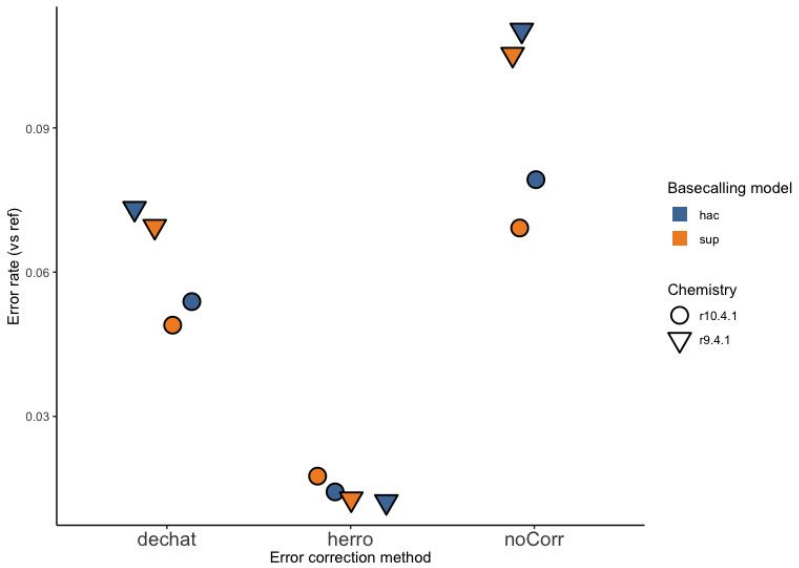


ONT Basecalling

- Dorado basecaller latest from Oxford Nanopore
  - 3rd party callers exist as well
- ML models decode the raw nanopore data
  - ‘Fast’, ‘high accuracy (HAC)’, or ‘super accurate (SUP)’ models
- Basecalling done on machine; don’t need to do yourself (model selected automatically)
  - Models frequently updated
- Option of recalling data yourself with newer models

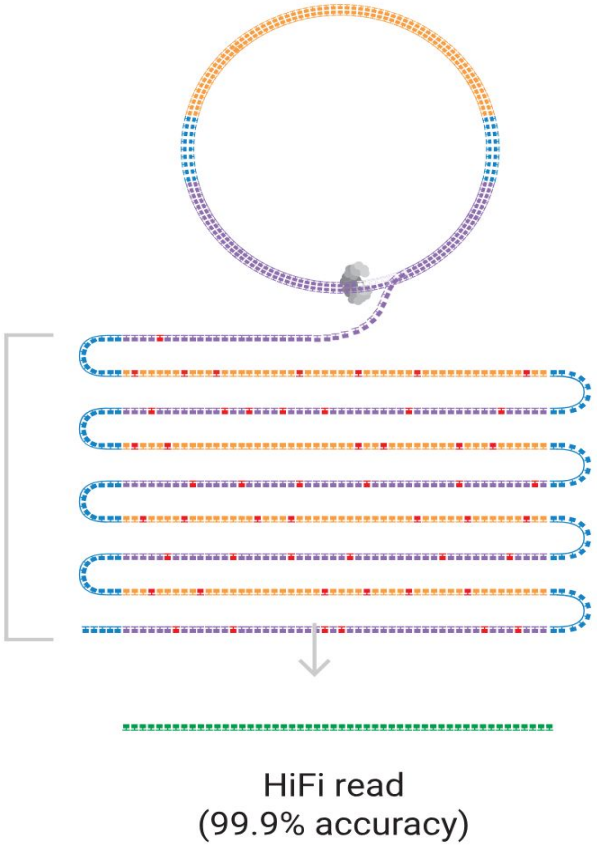
Nanopore and Accuracy

- "Is it worth correcting or re-calling my Nanopore data?"
- *D. melanogaster* ONT data
  - r9 vs r10 chemistry
  - HAC vs SUP basecalling models
  - No read correction vs HERRO vs DeChat
- Aligned vs *D. mel* reference
- Takeaways:
  - Read precorrection has biggest effect on error rate
  - SUP vs HAC effect marginal
  - HERRO more effective than DeChat…
  - BUT much more data loss with HERRO: 50% to 90% reads discarded



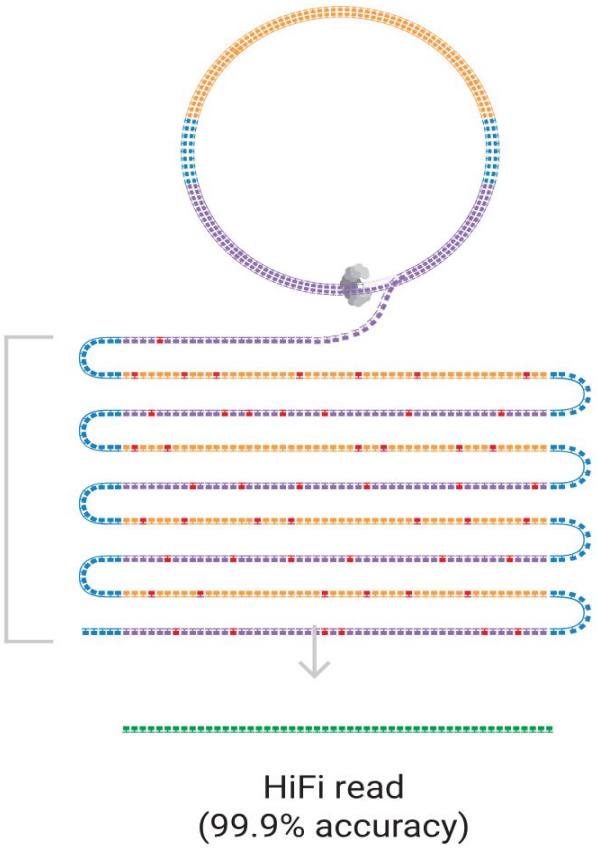
PacBio HiFi Sequencing

- Latest iteration is HiFi single molecule real-time sequencing
- How it works:
  - Circular DNA molecule added to well with polymerase anchored to bottom
  - Polymerase incorporates fluorescent bases; sequencer reads flashes of light
  - Sequence read multiple times; takes the consensus to obtain high accuracy
- Reads returned in BAM format
  - Convert to FASTQ and use for downstream analysis!

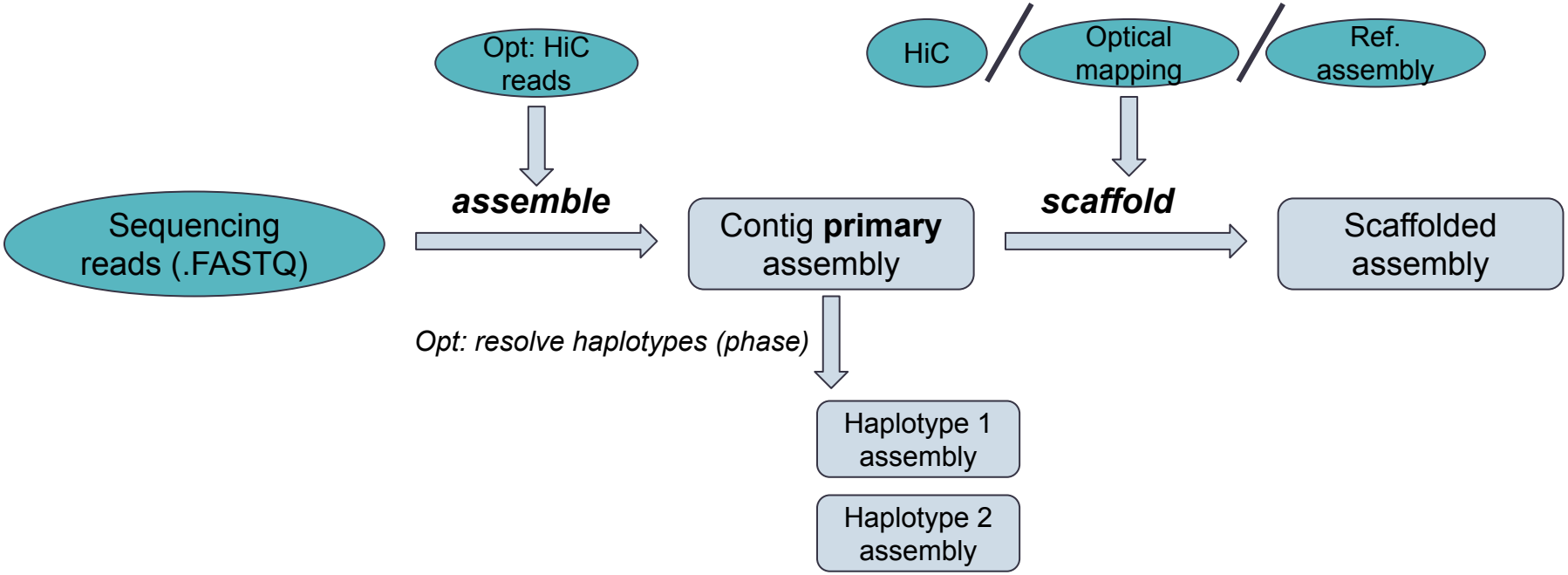


PacBio HiFi Sequencing

- Error rate close to short read technologies
  - ~0.1% vs 1-2% for latest ONT
  - Better for analysis depending on accuracy (assembly phasing, SNP calling, etc.)
- Read length typically around 12-15kb
  - ONT read length N50 ~35kb; tails of distribution >100kb



Assembly workflow



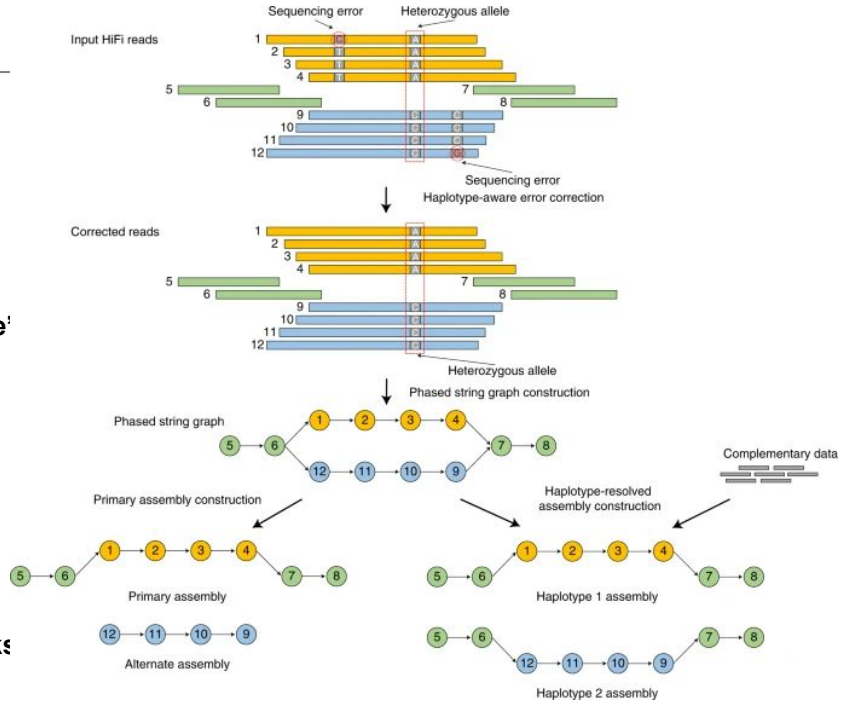
Hifiasm

- **Has become the standard for assembly**
  - **Fast & relatively memory efficient: human-sized genome <1 day, ~140Gb RAM**
  - **Lots of features:**
    - **Integrate HiC data for contiguity or phasing (N.B: NOT scaffolding!)**
    - **Integrate ONT ultra-long reads for scaffolding**
    - **Integrate short read trio data for phasing**
  - **Now supports Hifi *and* ONT data for assembly**

- **The Informatics group has this handy snakemake workflow to automate assembly with hifiasm on the Harvard computing cluster**

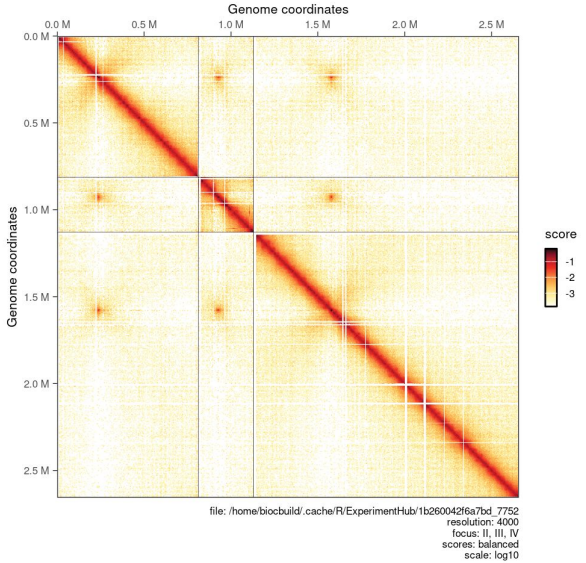
Hifiasm

- **Algorithm overview:**
  - Haplotype-aware error correction of input reads
  - Alignment of corrected reads; produces string graph
  - "Bubble" in graph where alternate haplotypes
  - With Hifi-only assembly, will random select one "bubble" as the primary assembly
  - Purges duplicate contigs
- **Hifi-only assembly output:**
  - Primary assembly
  - Two *partially phased* assemblies (hap1 and hap2)
    - Represents diploid genome: large phased blocks but contains switch errors (i.e. haplotypes switches within contig)



Phasing Assemblies

- If want *fully phased* assembly, need HiC data
  - Uses contact info from HiC to resolve to resolve haplotypes with no switch errors
  - Still struggles phasing across centromeres
- Designed for diploid assembly
  - Note CAN use Hifiasm on polyploid genomes to produce a primary assembly, but won't be able to phase it
- Is NOT the same as using HiC to scaffold; need to run stand-alone tool for that…



Scaffolding Assemblies

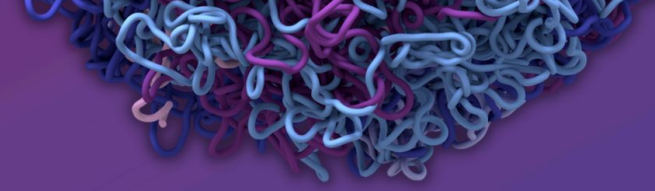
- **Scaffolding with HiC: recommend Yet Another HiC Scaffolder (YAHS)**
  - **“HiC reads” are Illumina short reads done with special prep**
  - **Map the HiC reads to the contig assembly; use the contacts to stitch together into scaffolds**
- **Scaffolding with optical mapping**
  - **Label certain motifs in genome with fluorescent marker using restriction endonucleases**
  - **Linearize the molecules and visualize them, “beads on a string”**
  - **Performed by external company (Bionano) as a service**
- **Reference-based scaffolding**
  - **Align contigs against closely related species, recommend RagTag to stitch contigs**

HiC with Arima

Lunch and Learn tomorrow November 20th

Register Here!

HARVARD
Faculty of Arts and Sciences



ARIMA GENOMICS

HiC: New Frontiers in Genomics

Thursday, November 20, 2025 | 12 pm ET
Room 425, Northwest Labs
In collaboration with the Bauer Core

Food will be provided and a chance for prizes

Identify new biomarkers and potential drug targets

Detect and discover structural variants and gene fusions

Link chromatin conformation to impacts on gene regulation

Guest Speaker:
Mario Aragon, PhD
Arima Genomics

QC Metrics for Assembly

- QUAST metrics
  - Size of assembly, number of contigs, assembly contiguity (N50 and L50)
  - If reference assembly available, base pair & structural differences vs reference
- BUSCO: “benchmarking universal single copy orthologs”
  - Curated database of genes universally present within taxonomic groups
  - Presence, duplication or fragmentation gives estimate of genome completeness
- Kmer estimates
  - Comparing kmer present in reads vs assembly to estimate completeness, levels of duplication
  - MERQURY, KAT, GenomeScope, etc.

So Which Do I Pick?

- Trade-offs more clear-cut in the past…
- Both technologies developing quickly, hard to give recommendation

Before considering your sequencing options, think about:

- Needs for assembly
  - Telomere-to-telomere (T2T)? Or contig-level ok?
  - Used for SNP calling?
  - Haplotype-resolved assembly? Or primary only ok?
- Size of genome
  - Small (<1 Gbase) vs large (>3 Gbase)?
  - Complexity of genome…repeat density? Polyploid?
- How much material available?
  - High quality?

So Which Do I Pick?

Based on Research Questions

1. I want to look at variation across populations
    a. Accuracy important, contiguity less important
2. I want to look at “dark matter” regions
    a. Contiguity more important, more T2T
3. I just want to have a decent draft assembly for this weird animal/plant

So Which Do I Pick?

- PacBio
  - Pros:
    - High accuracy right out the box (nothing below Q20)
    - Lower error rate == easier phasing assemblies, variant detection, etc.
  - Cons:
    - Less output == more expensive
    - Reads shorter than ONT

So Which Do I Pick?

- **ONT**
  - **Pros:**
    - **Longest read length, can get up to 1 Mb long**
    - **More output == cheaper**
  - **Cons:**
    - **Lower accuracy, needs methods to deal with it (improving)**
      - **Q9 cutoff**

So Which Do I Pick?

- **Theoretical “kitchen sink” assembly:**
  - **>=30X PacBio Hifi**
  - **Nanopore (ultra)long for contiguity**
  - **HiC for scaffolding/haplotype resolution**
  - **Could reasonably expect something approaching chromosome-scale!**

So Which Do I Pick?

- **Theoretical “kitchen sink” assembly:**
  - **>=30X PacBio Hifi**
  - **Nanopore (ultra)long for contiguity**
  - **HiC for scaffolding/haplotype resolution**
  - **Could reasonably expect something approaching chromosome-scale!**
- **Our services:**
  - **Snakemake pipeline for assembly with hifiasm**
  - **Schedule a consult for help with common post-assembly analysis (QC, HiC, annotation…)**
  - **Possible longer term collaboration for more specialized help…**

Real World Examples of Assembly

- *Phlox* wildflower species (*P. drummondii*)
  - Difficult genome; ~6.5Gbase and ~90% repetitive DNA
  - Older project; PacBio Hifi prohibitively expensive
  - R9.4 ONT chemistry, normal + ultralong preps, ~35X coverage depth
  - Assembly time >2 months on the cluster using Flye assembler
- Post-assembly:
  - Polished using short reads & self-correction
  - Scaffolding using optical mapping (Bionano) and genetic map (ALLMAPS)
- Final results: highly contiguous assembly (N50 406Mb), >50% genome contained in 7 chromosomal scaffolds

Real World Examples of Assembly

- ***Rhizanthes*** **corpse flower**
  - **Another difficult plant; genome size est. 11Gb, 75% genome simple repeats**
  - **Limited sample availability**
  - **Low coverage (<20X) PacBio Hifi + ~5-10X ONT**
  - **Assembled using Hifiasm**
- **Final assembly size 10.4Gb, contig N50 of 1.5Mb for primary assembly**
  - **Assembly time < 1 week; lower error rate = less costly read overlapping**

Real World Examples of Assembly

- Scrub jay pangenome: examining genetic variation within and between species
  - Graph-based representation of diversity of entire population
  - Hifi sequencing 45 individuals across 4 scrub jay species
  - 1 individual chose as “reference”, additional HiC sequencing
  - Got sequencing data piecemeal; assembled using snakemake Hifiasm pipeline
- Post-assembly:
  - Scaffolded reference individual using HiC, plus reference-based scaffolding vs Hawaiian crow and prior Florida scrub jay assembly
  - Constructed pangenome graph: whole genome alignment of haplotype assemblies (90 assemblies)
- Bird genomes fairly “well-behaved”...assembly straight-forward, graph construction was main challenge

Real World Examples of Assembly

- ***Stentor*** **ciliate genome**
  - **Small genome: estimated ~90Mb "haploid" genome size…**
  - **Unique biology making assembly and evaluation difficult!**
    - **Single-celled but continuously grows**
    - **Micronucleus and macronucleus**
    - **Genome duplicated proportionally to organism size**
    - **Genome duplicated unevenly (e.g. rDNA locus more amplified)**
  - **>1000X Hifi coverage**
  - **Experimented with different assemblers and methods…**
- **Optimal approach:**
  - **Subsetting reads to ~60X coverage, 20 subsets**
  - **Assemble each subset using *hifiasm***
  - **Sequentially merge assemblies using *quickmerge*; remove redundant sequence**

Definitions

Q score: Phred quality score, estimates accuracy of base in read. Log scaled (e.g. Q20 = 99%, Q30 = 99.9%...)

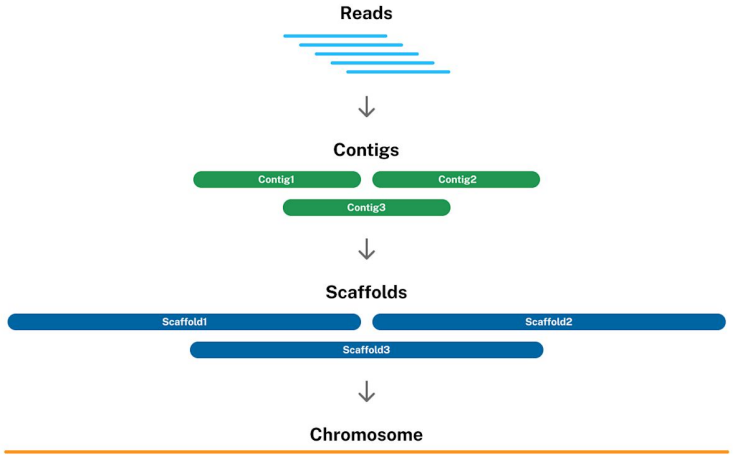
Assembler: software used to generate (i.e. assemble) genome from reads. Most likely fragmentary

Contig: a **contig**uous sequence of bases generated by an

Scaffold: multiple contigs that have been stitched togeth

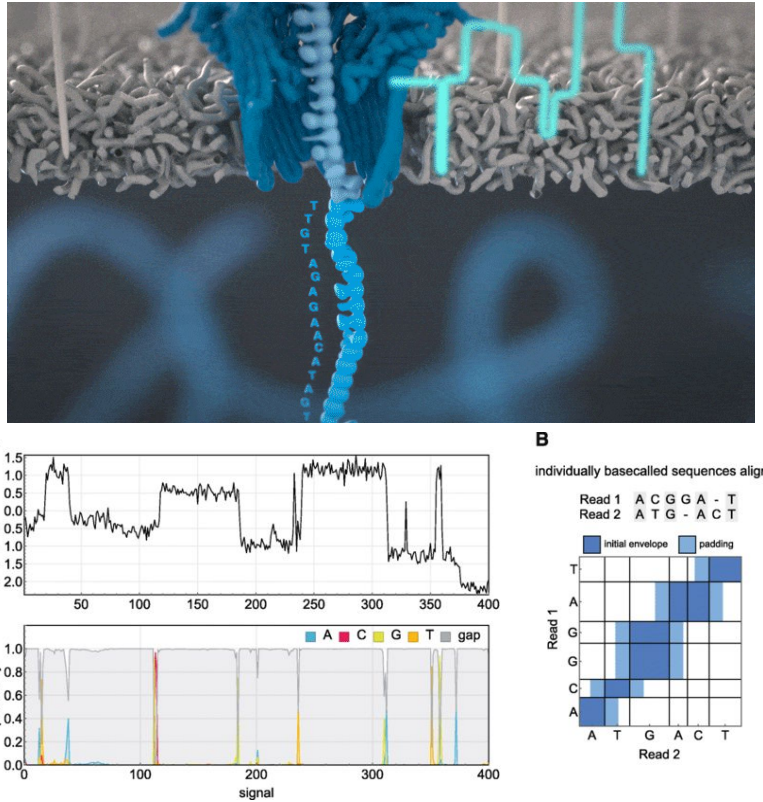
HiC: sequencing using chromatin capture to detect intera

Homopolymers: runs of identical nucleotides > 2 bp in le



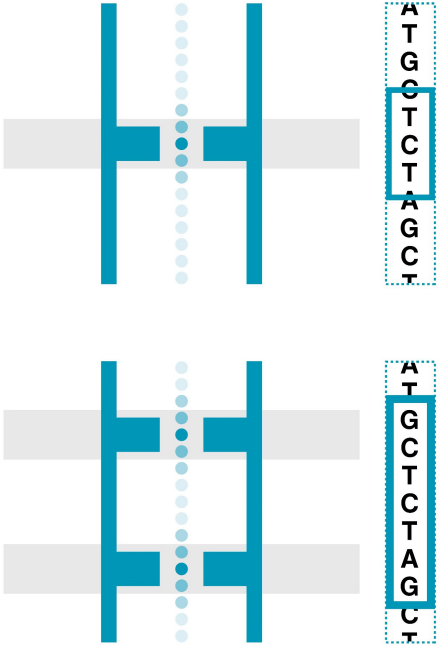
ONT sequencing

- DNA/RNA read thru pore across voltage differential
- Causes current to fluctuate, generating a “squiggle”
- *Basecalling algorithms* translate squiggle into base pair (kmer)
  - Signal vs noise determines accuracy
  - Struggles with certain sequences (homopolymers) but resolution improving
- Also can detect modified bases (5mC, 5hmC, 6mA)
  - Integrated into basecalling process (need to enable!)
  - Also possible post-calling
- Ultra-long read optimization



Nanopore & accuracy

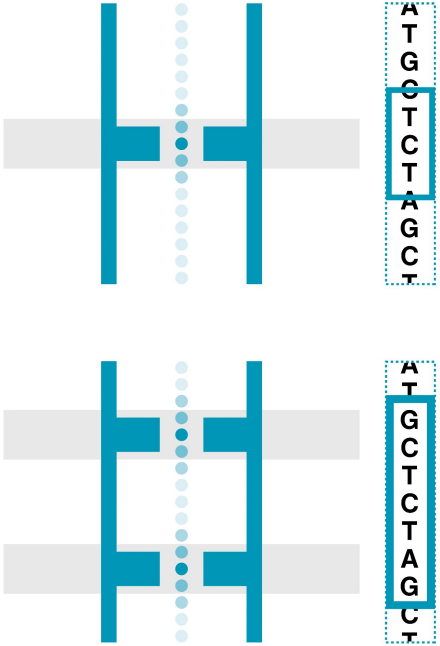
- Older chemistries had higher error rates (10-15% for R9)
  - Necessitated post-assembly polishing (Illumina reads or self-correction)
  - Struggles with homopolymer regions
- R10 chemistry adds second reader in pore to increase accuracy
  - Basecalling models tuned for higher accuracy
  - Error rate 1-2% (10x PacBio, 100x Illumina)
- Pre-correction of reads (HERRO, DeChat, etc.) or as part of assembly process
- **TL;DR: post-assembly polishing with short reads not necessary with latest chemistries**



Nanopore & accuracy

- Older chemistries had higher error rates (10-15% for R9)
  - Necessitated post-assembly polishing (Illumina reads or self-correction)
  - Struggles with homopolymer regions
- R10 chemistry adds second reader in pore to increase accuracy
  - Basecalling models tuned for higher accuracy
  - Error rate 1-2%
- Pre-correction of reads (HERRO, DeChat, etc.) or as part of assembly process
- **TL;DR: post-assembly polishing with short reads not necessary with latest chemistries**

- Duplex reads option: 99.8% accuracy
  - Reads double strand instead of single strand
  - Much less output



ONT basecalling

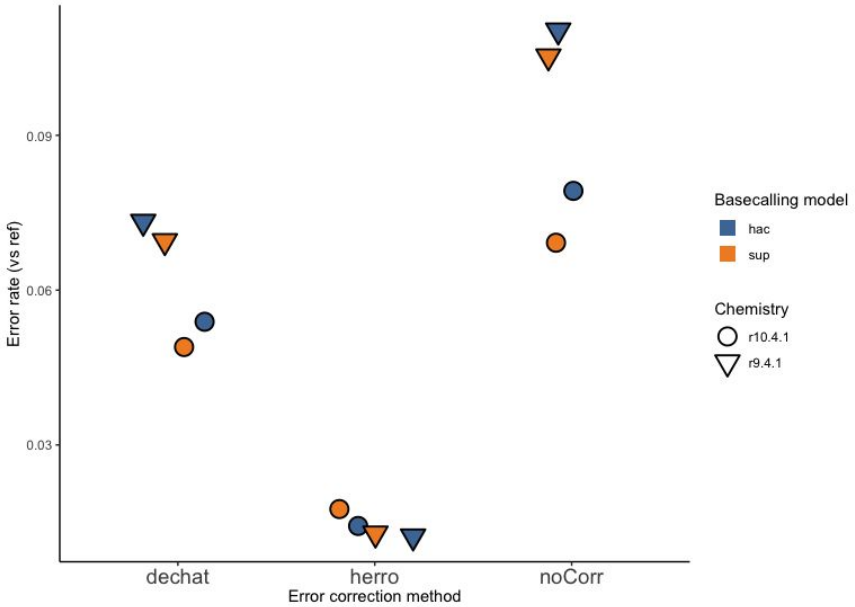
- Dorado basecaller latest from Oxford Nanopore
  - 3rd party callers exist as well
- ML models decode the raw nanopore data
  - ‘Fast”, “high accuracy (HAC)’, or ‘super accurate (SUP)’ models
- Basecalling done on machine; don’t need to do yourself (model selected automatically)
  - Models frequently updated
- Option of recalling data yourself with newer models

Nanopore & accuracy

- "Is it worth correcting or re-calling my Nanopore data?"

- *D. melanogaster* ONT data
  - r9 vs r10 chemistry
  - HAC vs SUP basecalling models
  - No read pre-correction vs HERRO vs DeChat
- Aligned vs *D. mel* reference

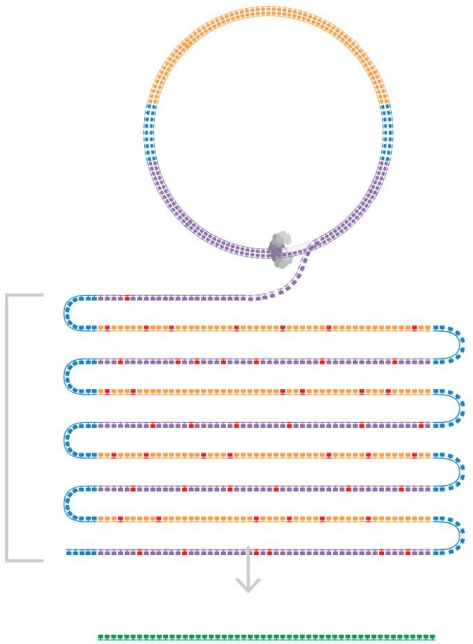
- Takeaways:
  - Read precorrection has biggest effect on error rate
  - SUP vs HAC effect marginal
  - HERRO more effective than DeChat…
  - BUT much more data loss with HERRO: 50% to 90% reads discarded



PacBio HiFi sequencing

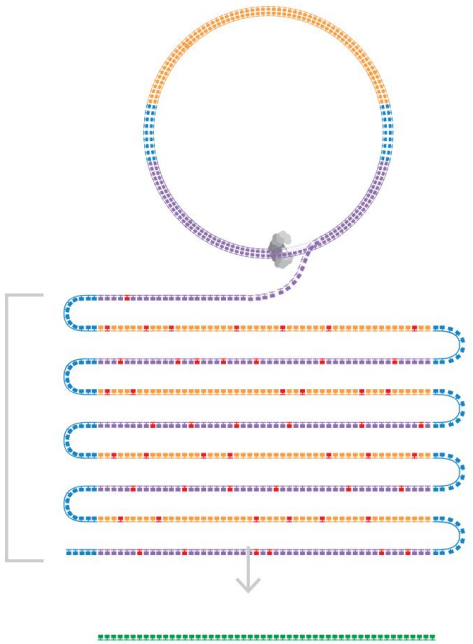
- Latest iteration is HiFi single molecule real-time sequencing
- How it works:
  - Circular DNA molecule added to well with polymerase anchored to bottom
  - Polymerase incorporates fluorescent bases; sequencer reads flashes of light
  - Sequence read multiple times; takes the consensus to obtain high accuracy

- Reads returned in BAM format
  - Convert to FASTQ and use for downstream analysis!



PacBio HiFi sequencing

- Error rate close to short read technologies
  - ~0.1% vs 1-2% for latest ONT
  - Better for analysis depending on accuracy (assembly phasing, SNP calling, etc.)
- Read length typically around 12-15kb
  - ONT read length N50 ~35kb; tails of distribution >100kb



HiFi read
(99.9% accuracy)

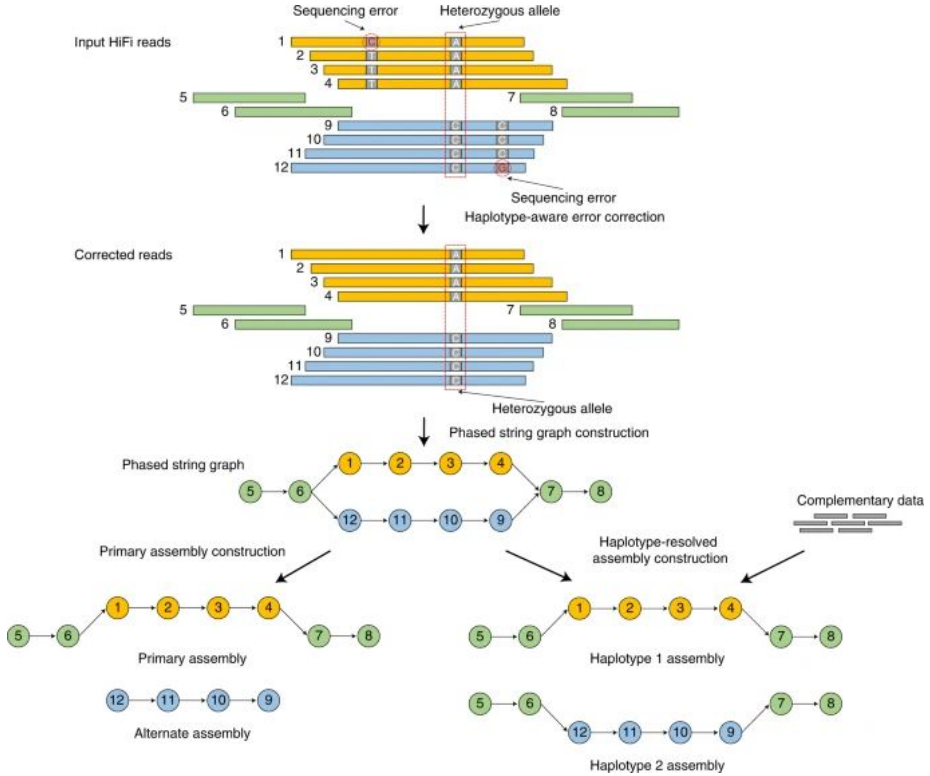
Hifiasm

- Has become the standard for assembly
  - Fast & relatively memory efficient: human-sized genome <1 day, ~140Gb RAM
  - Lots of features:
    - Integrate HiC data for contiguity or phasing (N.B: NOT scaffolding!)
    - Integrate ONT ultra-long reads for scaffolding
    - Integrate short read trio data for phasing
  - Now supports Hifi *and* ONT data for assembly

- The Informatics group has this handy snakemake workflow to automate assembly with hifiasm on the Harvard computing cluster

Hifiasm

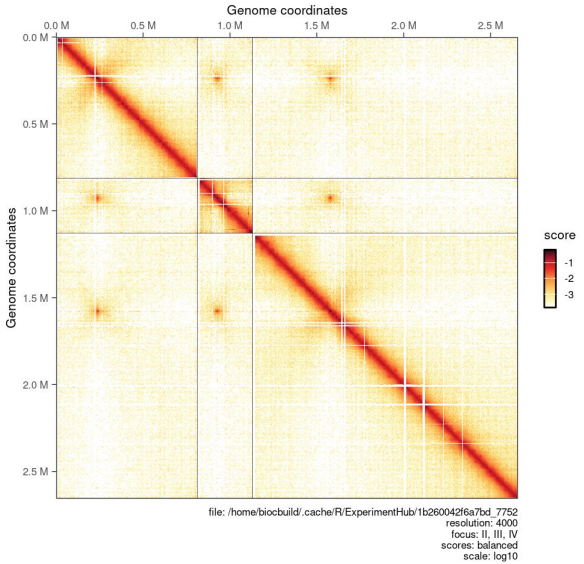
- Algorithm overview:
  - Haplotype-aware error correction of input reads
  - Alignment of corrected reads; produces string graph
  - "Bubble" in graph where alternate haplotypes
  - With Hifi-only assembly, will random select one "bubble" as the primary assembly
  - Purges duplicate contigs
- Hifi-only assembly output:
  - Primary assembly
  - Two *partially phased* assemblies (hap1 and hap2)
    - Represents diploid genome: large phased blocks, but contains switch errors (i.e. haplotypes switches within contig)



Phasing assemblies

- If want *fully phased* assembly, need HiC data
  - Uses contact info from HiC to resolve to resolve haplotypes with no switch errors
  - Still struggles phasing across centromeres
- Designed for diploid assembly
  - Note CAN use Hifiasm on polyploid genomes to produce a primary assembly, but won't be able to phase it

- Is NOT the same as using HiC to scaffold; need to run stand-alone tool for that…



Scaffolding assemblies

- Scaffolding with HiC: recommend Yet Another HiC Scaffolder (YAHS)
  - “HiC reads” are Illumina short reads done with special prep
  - Map the HiC reads to the contig assembly; use the contacts to stitch together into scaffolds
- Scaffolding with optical mapping
  - Label certain motifs in genome with fluorescent marker using restriction endonucleases
  - Linearize the molecules and visualize them, “beads on a string”
  - Performed by external company (Bionano) as a service
- Reference-based scaffolding
  - Align contigs against closely related species, recommend RagTag to stitch contigs

QC metrics for assembly

“So which one do I pick?”

- Trade-offs more clear-cut in the past…
- Both technologies developing quickly, hard to give recommendation

Before considering your sequencing options, think about:

- Needs for assembly
  - Telomere-to-telomere (T2T)? Or contig-level ok?
  - Used for SNP calling?
  - Haplotype-resolved assembly? Or primary only ok?
- Size of genome
  - Small (<1 Gbase) vs large (>3 Gbase)?
  - Complexity of genome…repeat density? Polyploid?
- How much material available?
  - High quality?

“So which one do I pick?” - based on research questions

- I want a pangenome to look at genes across populations
- I want to look at “dark matter” regions
-

“So which one do I pick?”

- PacBio
  - Pros:
    - High accuracy right out the box (nothing below Q20)
    - Lower error rate == easier phasing assemblies, variant detection, etc.
  - Cons:
    - Less output == more expensive
    - Reads shorter than ONT
- ONT
  - Pros:
    - Longest read length, can get up to 1 Mb long
    - More output == cheaper
  - Cons:
    - Lower accuracy, needs methods to deal with it (tho improving)
      - Q9 cutoff

“So which one do I pick?”

- Theoretical “kitchen sink” assembly:
  - >=30X PacBio Hifi
  - Nanopore (ultra)long for contiguity
  - HiC for scaffolding/haplotype resolution
  - Could reasonably expect something approaching chromosome-scale!

“So which one do I pick?”

- Theoretical “kitchen sink” assembly:
  - >=30X PacBio Hifi
  - Nanopore (ultra)long for contiguity
  - HiC for scaffolding/haplotype resolution
  - Could reasonably expect something approaching chromosome-scale!

- Our services:
  - Snakemake pipeline for assembly with hifiasm
  - Schedule a consult for help with common post-assembly analysis (QC, HiC, annotation…)
  - Possible longer term collaboration for more specialized help…

Real-world examples of assembly