Genome Assembly and Long Reads

Chad Tomlinson
BGA Group
McDonnell Genome Institute
4444 Forest Park Ave.
St. Louis, MO 63108

Outline

- Definition of long read sequencing and its applications.
- Brief introduction to the two main long read sequencing technologies.
- Explanation of genome assembly and how long reads improve assemblies.
- Examples of commonly used long read genome assembly algorithms.
- Overview of the pb-assembly (Falcon-Unzip) assembly algorithm.

Evolution of DNA Sequencing

First Generation

Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments

Second Generation
(Next Generation Sequencing)

454, Solexa,
Ion Torrent
Illumina

- High throughput from the parallelization of sequencing reactions
- ~50-500 bp fragments

Third Generation

PacBio
Oxford Nanopore

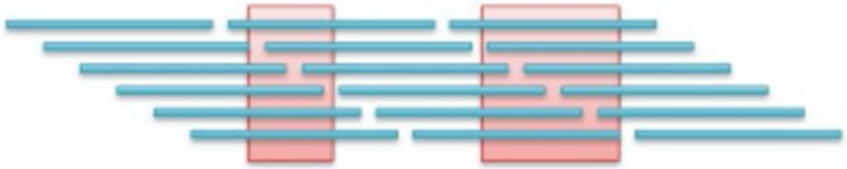
- Sequence native DNA in real time with single-molecule resolution
- Tens of kb fragments, on average

Short-read sequencing

Long-read sequencing

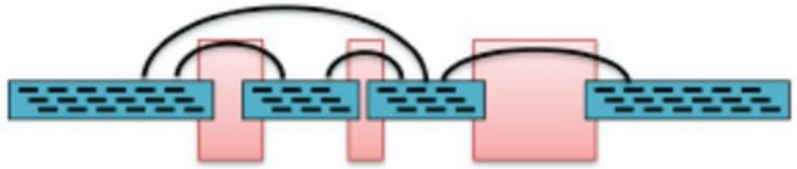
Long Read Analysis Applications

a) De novo Assembly



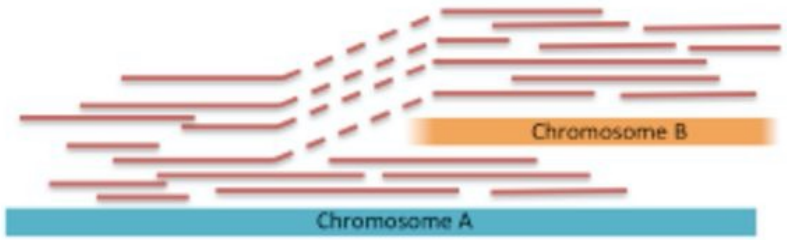
Reconstruct the genome sequence directly from the sequenced reads (blue). Longer reads will span more repetitive elements (red), and produce longer contigs.

b) Chromosome Scaffolding



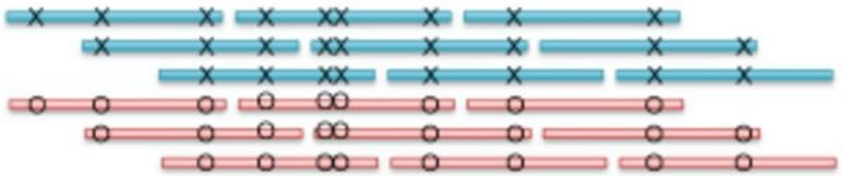
Order and orient contigs (blue) assembled from overlapping reads (black) into longer pseudo-molecules. Longer spans are more likely to connect distantly spaced contigs, especially those separated by long repeats (red).

c) Structural Variation Analysis



Identify reads/spans (red) that map to different chromosomes or discordantly within one. The longer the read/span, the more likely to capture the SV, and will have improved mappability to resolve SVs in repetitive element.

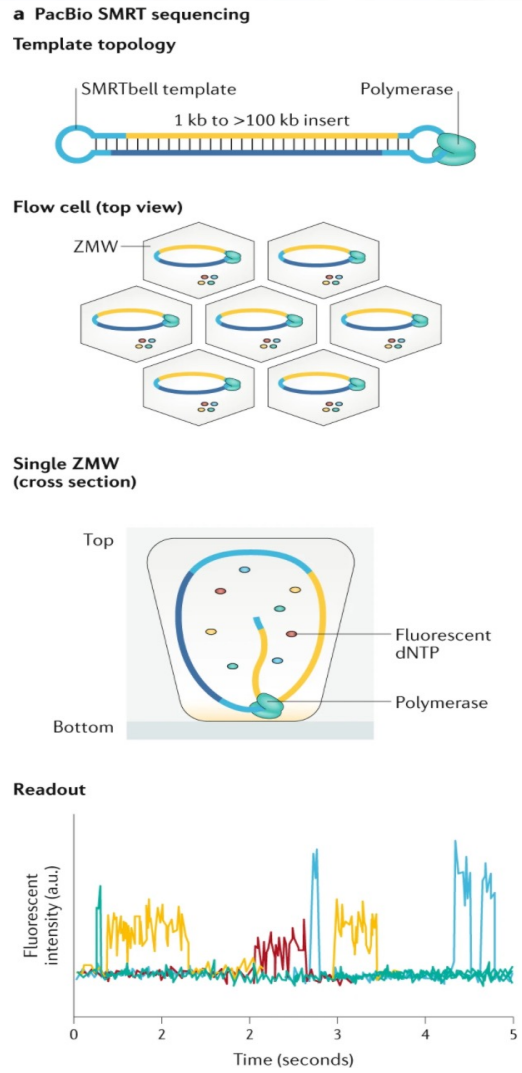
d) Haplotype Phasing



Link heterozygous variants (X/O) into phased sequences representing the original maternal (red) and paternal (blue) chromosomes. Longer reads and longer spans will be able to connect more distantly spaced variants.

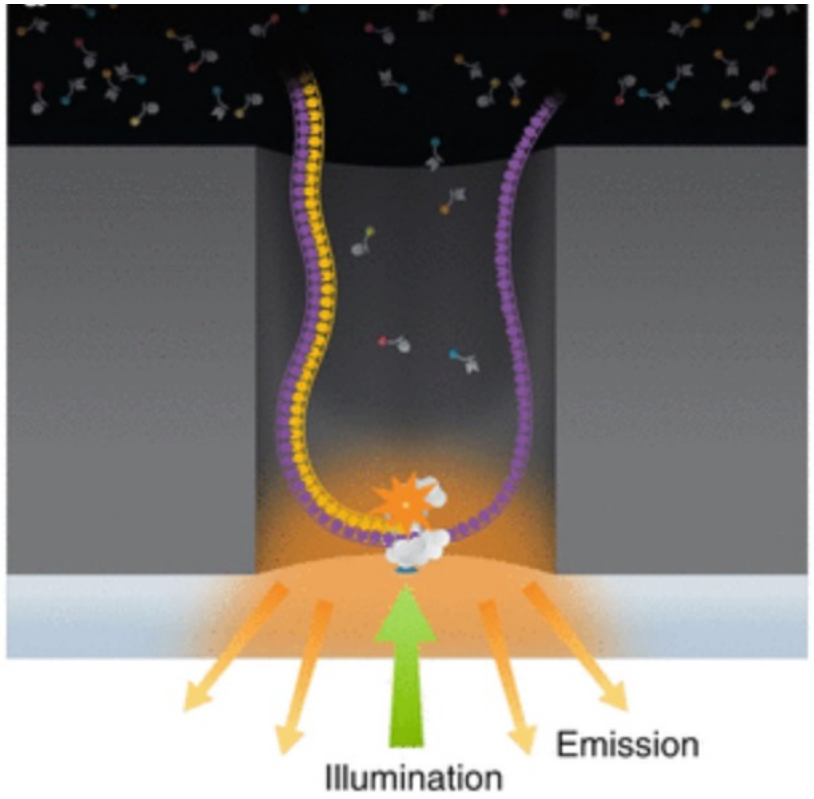
Pacific Biosciences SMRT Sequencing

- Uses a topologically circular DNA molecule template called a SMRTbell, comprised of a double-stranded DNA insert with single-stranded hairpin adapters on either end.
- DNA insert can range from 1 to more than 100 Kb in length.
- Once the SMRTbell has been assembled, it is bound by a DNA polymerase and loaded onto a SMRT cell
- A SMRT cell is a chip that contains 8 million sequencing units called ZMWs – zero mode waveguides.
- ZMW is a nanophotonic confinement structure consisting of a circular hole ~70 nm in diameter and ~100 nm in depth. This provides the smallest available volume for light detection.
- A single SMRTbell diffuses into each ZMW.



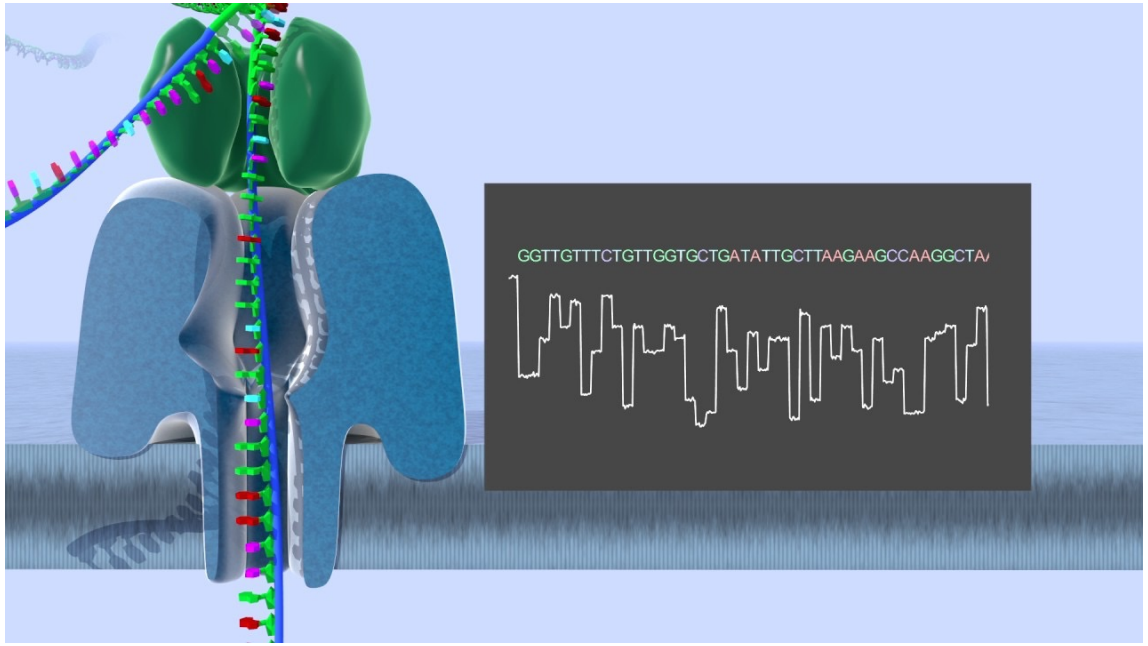
Pacific Biosciences SMRT Sequencing

- Within each ZMW, a single polymerase is immobilized at the bottom. The polymerase can bind to either hairpin of the SMRTbell and start the replication process.
- Four flourescent-labeled nucleotide bases, which generate distinct emission spectrums, are added to the SMRT cell. As each base is held by the polymerase, a light pulse is produced that identifies the base.
- The replication processes in all ZMWs of a SMRT cell are recorded as a “movie” of light pulses, and the pulses corresponding to each ZMW can be interpreted to be a sequence of bases called a continuous long read, CLR.



Oxford Nanopore Sequencing

- Uses linear DNA molecules instead of circular ones. These molecules are usually between one and several hundred Kb in length, but can be up to several megabases long.
- Sequencing begins by attaching a double-stranded DNA molecule to a sequencing adapter that is pre-loaded with a motor protein.
- The DNA mixture is loaded onto a flow cell that contains hundreds to thousands of nanopores embedded in a synthetic membrane.
- The motor protein unwinds the double-stranded DNA and in conjunction with an electric current, drives the negatively charged DNA through the pore at a controlled rate.
- As the DNA passes through the pore, it causes specific disruptions to the current. These disruptions are analyzed in real time to determine the sequence of the bases in the DNA strand.
- Reads >1Mb in length have been generated with the longest reported reads close to 2.3Mb in length.

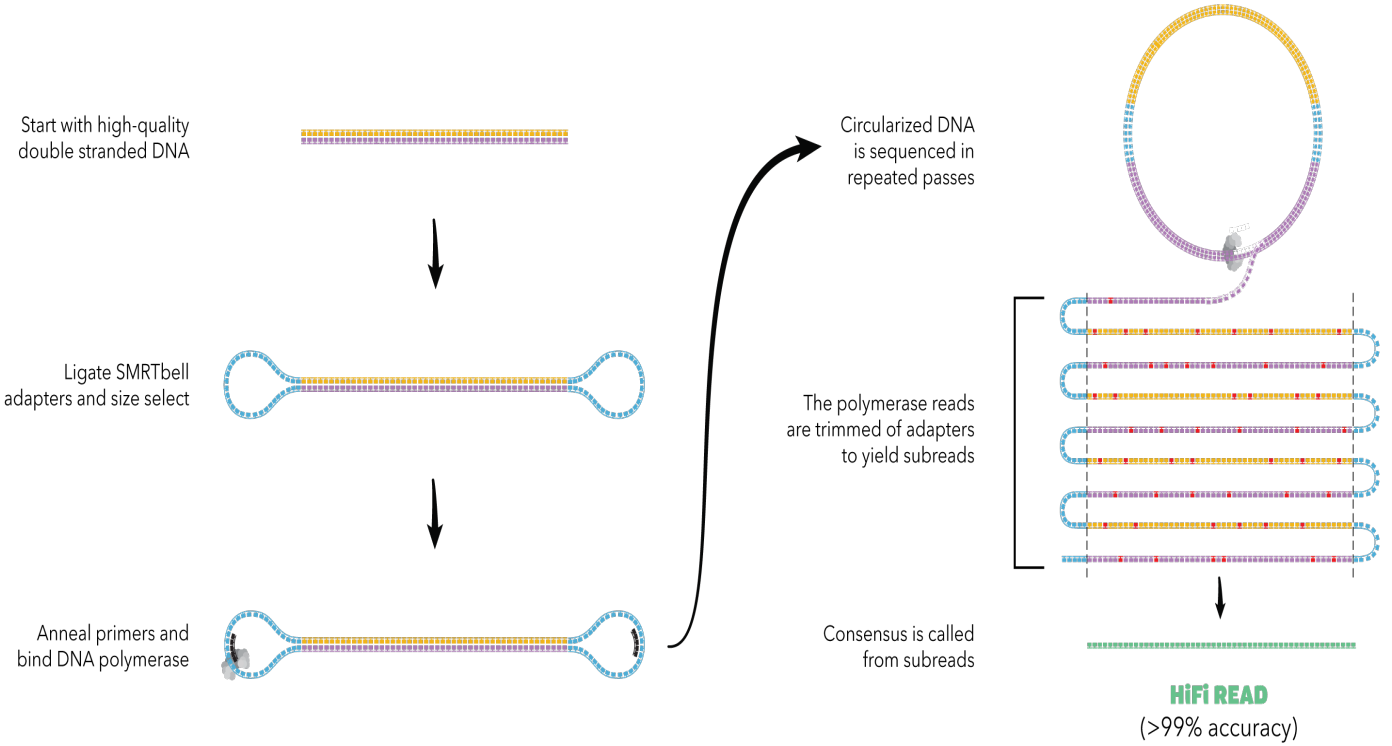


Long Read Sequencing Overview

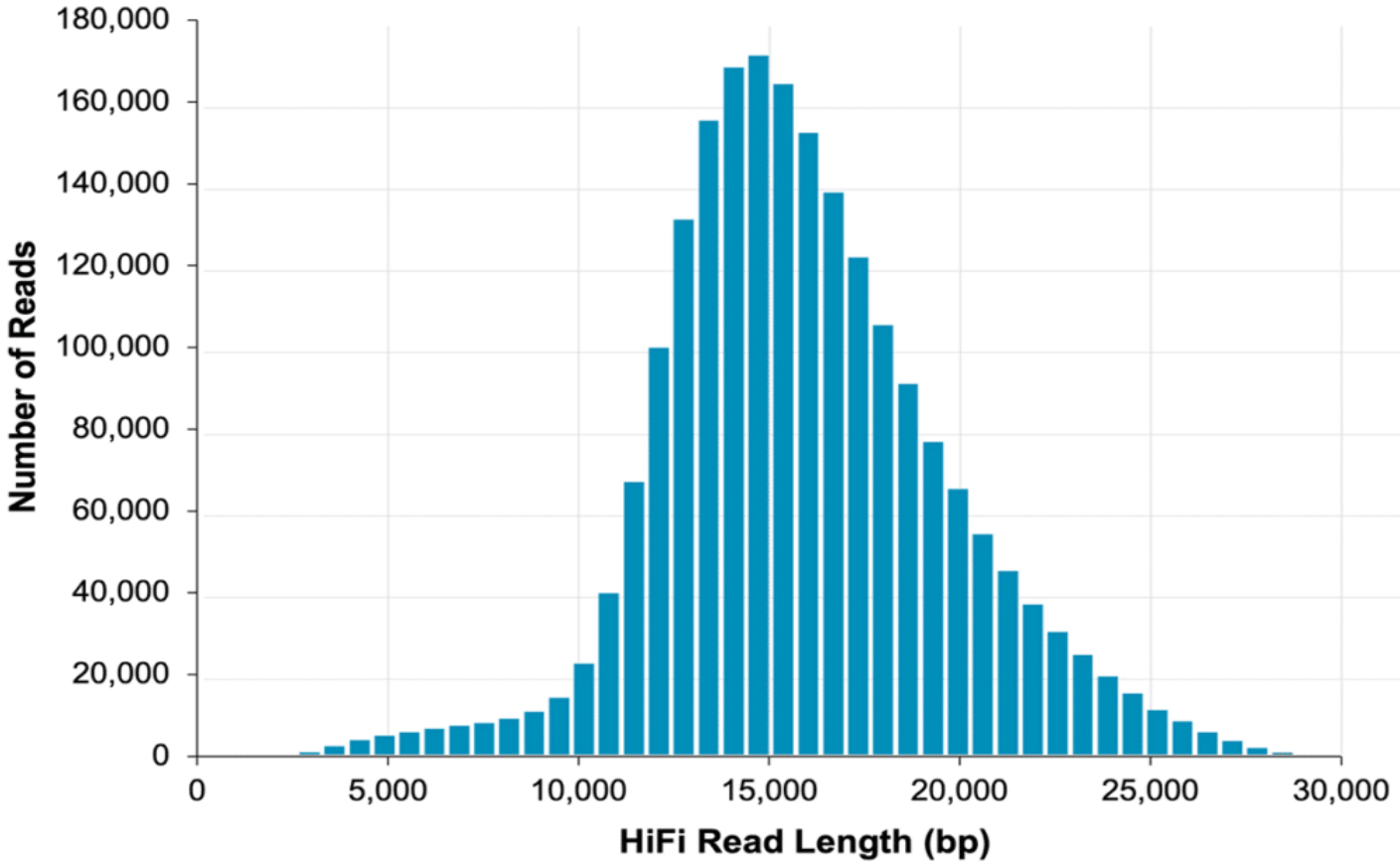
Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60	87–92	0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100		5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/GridION	Long	10–60	>1,000	87–98	2-20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500		0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15	>99.9	16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25		65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

PacBio HiFi (CCS) Sequencing

- Pacbio CLR subread accuracy typically ranges between ~85% to 92% with error distributed randomly.
- Only ~85% of homopolymers at least five bases long are called accurately in CLR reads.
- HiFi (CCS) sequencing involves multiple passes around the SMRTbell.
- Since error is distributed randomly, you can call a more accurate HiFi (CCS) read consensus with 3 to 4 passes around the insert.
- HiFi (CCS) reads have a median accuracy of greater than 99.9%, with over 99.5% of homopolymers at least 5 bp long accurately called.
- More than 50% of the regions previously inaccessible with Illumina short-read sequencing data in GRCh37 (human reference) are now accessible with HiFi reads.



PacBio HiFi Read Length Distribution



Goals of Genome Assembly

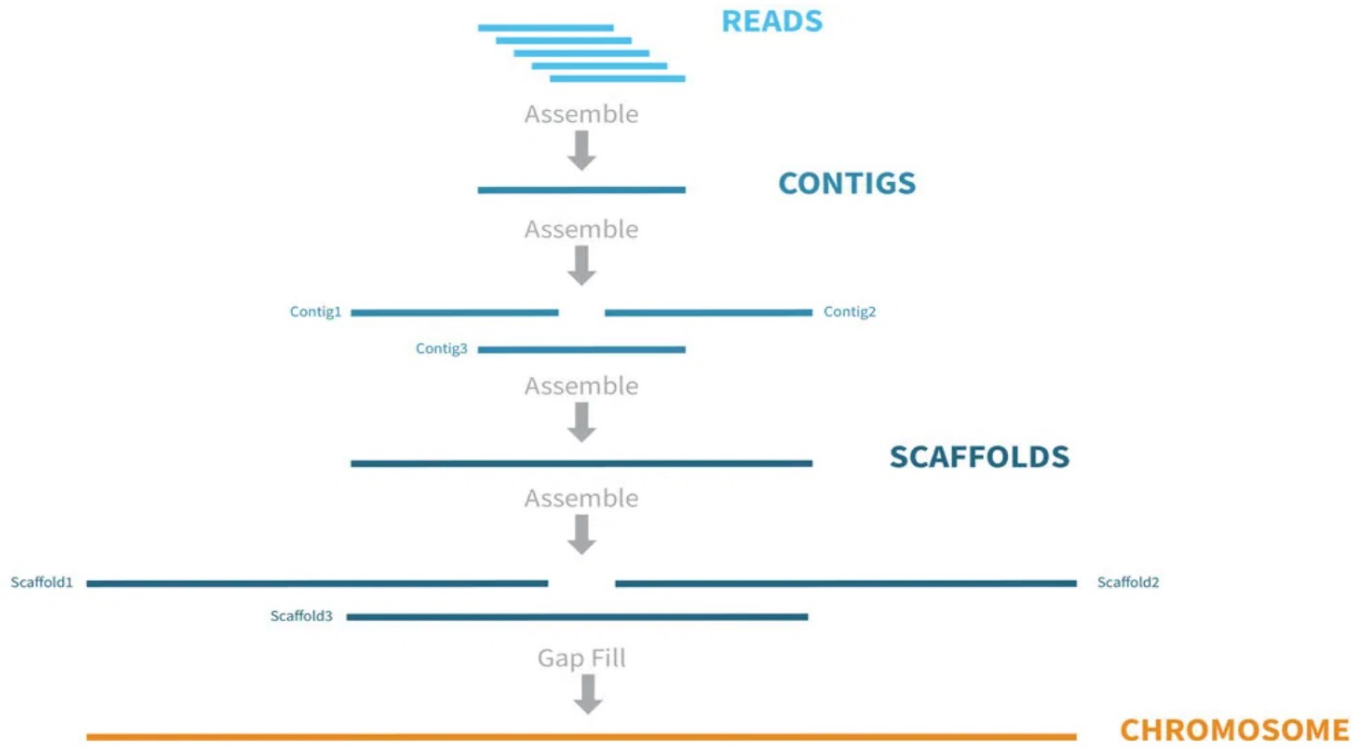
- **Determine the complete genome sequence of an organism.**
  - As few gaps in the sequence as possible
  - As accurate of a sequence as possible

- **Assemble and organize the sequence.**
  - Into contigs
  - Into scaffolds
  - Into chromosomes

- **Annotate the protein-coding gene sequence.**
  (and other genetically important functional features)

De Novo Genome Assembly Process and Outputs

- Algorithms are used to find accurate overlaps between reads. The consensus of a set of overlapping read sequences is called a contig.
- Contigs are segments of the genome that have gaps in between. In most instances, we do not know how the contigs are to be ordered and oriented in respect to one another.
- We can use additional methods to order and orient contigs, fill in gap regions, and chain contigs together into larger sequences called scaffolds.
- Scaffolding can be an additional step performed by the assembly algorithm, but it usually involves a separate process using BioNano, Hi-C, Oxford Nanopore data. or other data types.
- Additional methods can be used to organize scaffolds into more complete chromosomal level assemblies.
- Contigs and scaffolds are typically available in FASTA and/or FASTQ formats.



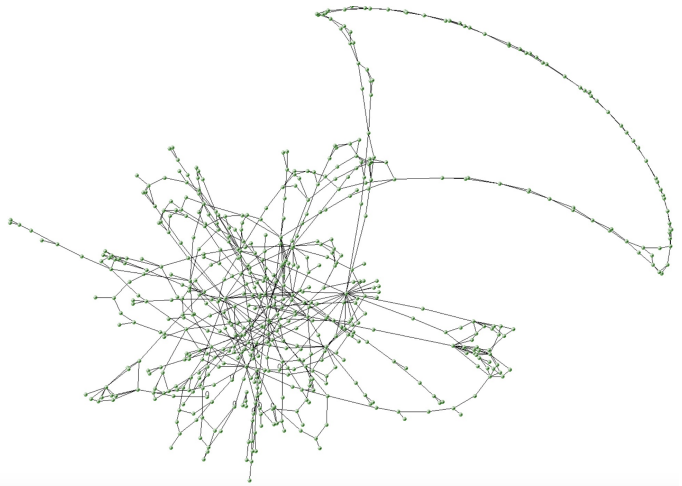
Two main assembly methods

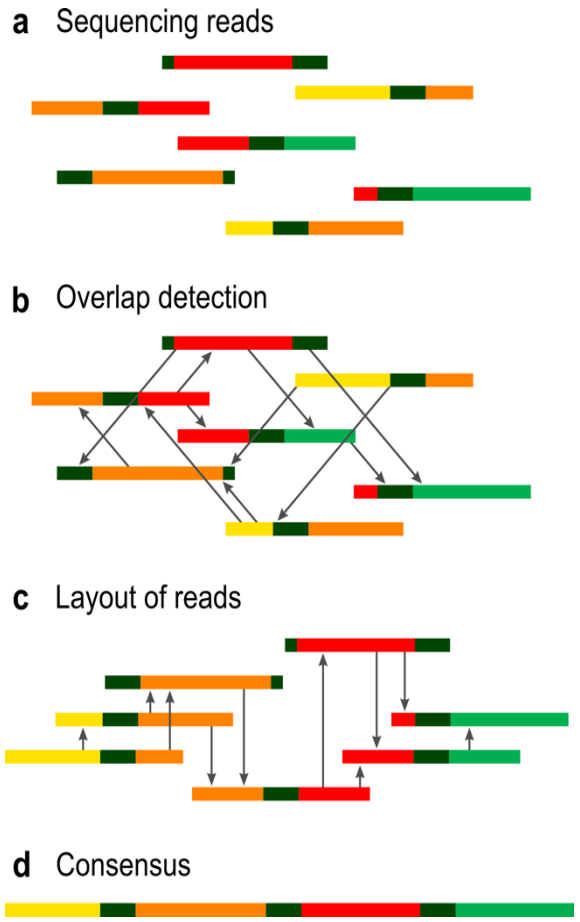
- **Overlap Layout Consensus Assemblers** – Mainly used for long reads
  - Construct overlap graph directly from reads, eliminating redundant reads; trace path for assembly.
  - Examples: pb-assembly (Falcon), Canu, Hifiasm

- **de Bruijn graph based Assemblers** – Mainly used for short reads
  - Construct k-mer graph from the reads; original reads are discarded.
  - Trace a path through the graph to arrive at the assembly.
  - Examples: MEGAHIT, ABySS

Overlap Layout Consensus Graph

- All read vs. all read alignment and identify all possible overlaps between reads.
- The overlap relationship between reads is captured in a large assembly graph.
- The graph is refined to correct errors and simplify.
- Find the best path through the graph and traverse each node in the graph once.
- Output the consensus of the path as the assembly.





Challenges to Genome Assembly

- **Repeats**
  - Repetitive regions can make it more difficult to determine read placement leading to misassemblies, which are errors in the construction of the assembly.
  - Long reads can help this by traversing the longest repeats in the genome.

- **Sequencing Errors or Assembly Errors**
  - Base errors in sequencing reads or consensus errors introduced by assemblers can confound the assembly process.

- **Heterozygosity**: Presence of different alleles at the same loci in homologous chromosomes
  - Alleles from the same locus are more likely to be mistaken as sequences from different loci.
  - The assembler may incorporate two different contigs that actually represent the same regions of the genome.
  - This might be desirable if you are seeking haplotype separated diploid or polyploid assemblies.

- **Contamination**
  - Contamination in the sequencing data can lead to contigs of contamination in your final assembly.
  - Adapter contamination of the read data can impact assembly results.

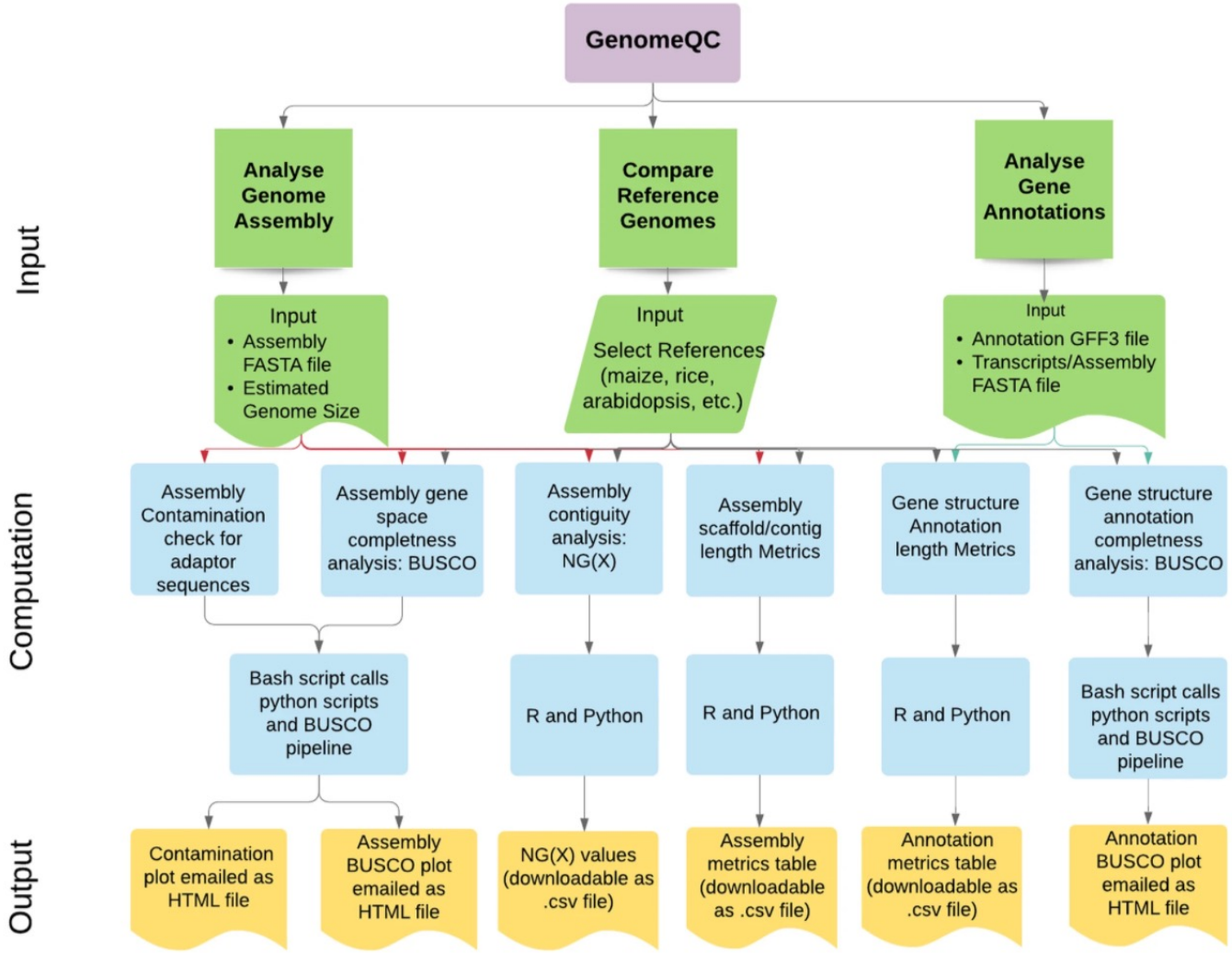
Methods to QC Genome Assemblies

- **Assembly scaffold/contig length metrics**
  -N50 length, avg. length, # contigs/scaffolds

- **Compare assembly to Reference Genome**
  -Identify misassemblies
  -Identify real SNPs and SVs (indels, translocations, duplications)

- **Evaluate against an optical map**: Ordered, genome wide high resolution restriction map
  -Identify misassemblies
  -Scaffold contig assemblies

- **Busco Analysis**: Evaluation of the assembly against a set of single-copy orthologs present in 90% of species of a particular group.
  -What percentage of core genes does your assembly contain?
  -Measure of assembly completeness.



N50 Contig/Scaffold Length

N50 size

Def: 50% of the genome is in contigs larger than N50

Example: 1 Mbp genome

N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:
N50 values are only meaningful to compare when base genome size is the same in all cases

Assemblers for PacBio Data

- **De novo Assemblers**
  - Pb-assembly (Falcon/Falcon-Unzip)
  - Canu/HiCanu

- **Trio Based Assemblers:** Use short reads from parents to partition child reads by maternal/paternal haplotypes prior to assembly.
  - TrioCanu
  - Hifiasm

- **Reference Assisted Assemblers:** Using the reference to assist in building the reads into contigs/scaffolds. This can bias assembly results towards the reference used.
  - RefKA

- **Hybrid Assemblers :** Use short and long reads. More effective for assembly of complex genomes (e.g. Plants)
  - Masurca
  - PBcR

PB Assembly (Falcon/Falcon-Unzip) of HiFi Data

- **Falcon Assembly:**
  -Input is PacBio HiFi data in FASTA format.

  -Output is haplotype-fused assembly in FASTA format and set of associated contigs (SVs from primary contig assembly).

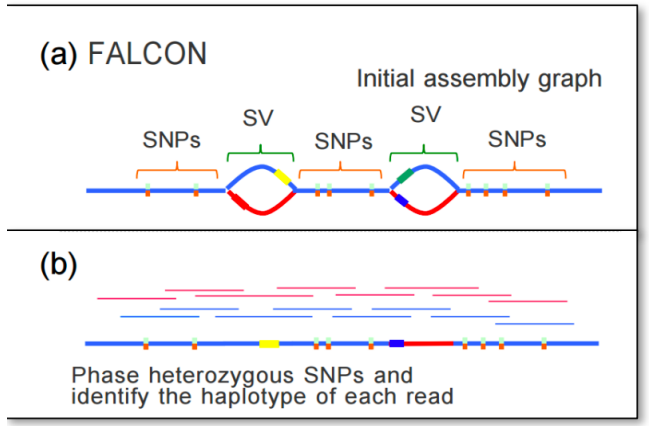
- **Falcon-Unzip:**
  -Align reads to contigs and phases the reads using heterozygous SNPs.

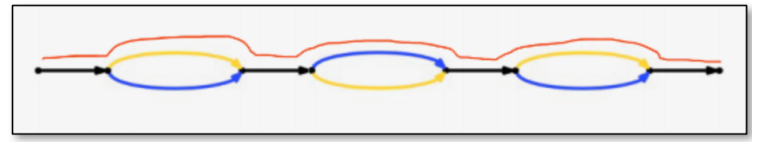
  -SNPs are used to separate the haplotypes into partially phased primary contigs and fully phased haplotigs.

  -There are switch errors in the output between maternal and paternal haplotypes.

  -Output is a FASTA file of primary contigs and a FASTA file of haplotigs.

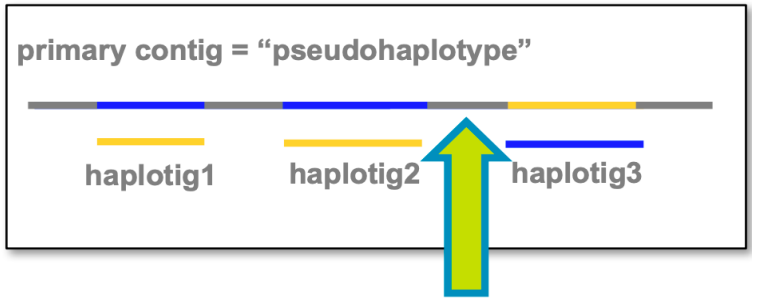
- PacBio data for diploid individual (no trio)
- Phase PacBio reads using SNPs identified in initial assembly graph
- Output phased and collapsed regions in high contiguity contigs





Weisenfeld et al. 2017

PSEUDOHAPLOTYPE AND HAPLOTIGS

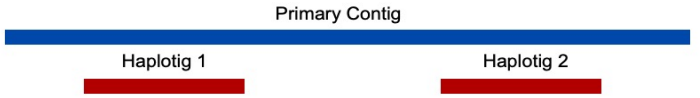


"Phase/Haplotype Switch"

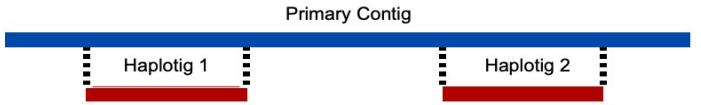
Falcon-Phase

FALCON-PHASE WORKFLOW

Input: FALCON-Unzip assembly & HiC data

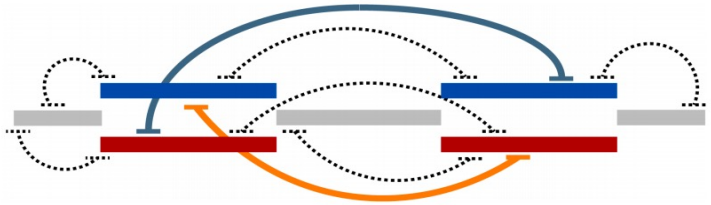


1. Identify **haplotig placement** on primary contigs

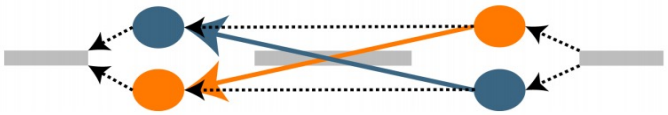


2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**

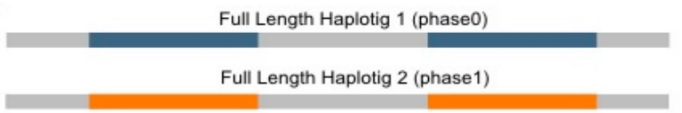
3. **Map** paired HiC reads to minced contigs and generate normalized **contact matrix**



4. **Phase** haplotigs along each primary contig



Output: phased full length haplotigs



Acknowledgements

Long-read human genome sequencing and its applications

Logsdon, G.A., Vollger, M.R. & Eichler, E.E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21,** 597–614 (2020). https://doi.org/10.1038/s41576-020-0236-x

Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome

Wenger, A.M., Peluso, P., Rowell, W.J. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37,** 1155–1162 (2019). https://doi.org/10.1038/s41587-019-0217-9

Long-read sequencing in deciphering human genetics to a greater depth

Midha, M.K., Wu, M. & Chiu, KP. Long-read sequencing in deciphering human genetics to a greater depth. *Hum Genet* **138,** 1201–1215 (2019). https://doi.org/10.1007/s00439-019-02064-y

Introduction to Genome Assembly

Bioinformatics Workbook (Online)

Andrew Severin - Author

https://bioinformaticsworkbook.org/dataAnalysis/GenomeAssembly/Intro_GenomeAssembly.html

Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing

Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13(12):1050-1054. doi:10.1038/nmeth.4035

Extended haplotype phasing of *de novo* genome assemblies with FALCON-Phase

Zev N. Kronenberg, Arang Rhie, Sergey Koren, Gregory T. Concepcion, Paul Peluso, Katherine M. Munson, Stefan Hiendleder, Olivier Fedrigo, Erich D. Jarvis, Adam M. Phillippy, Evan E. Eichler, John L. Williams, Tim P.L. Smith, Richard J. Hall, Shawn T. Sullivan, Sarah B. Kingan
bioRxiv 327064; doi: https://doi.org/10.1101/327064

GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations

Manchanda, N., Portwood, J.L., Woodhouse, M.R. *et al.* GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics* **21,** 193 (2020). https://doi.org/10.1186/s12864-020-6568-2