GENOME ASSEMBLY

TERMINOLOGY!

READ

- Raw fragment of DNA as it has been “read” by the sequencer
- These are “real” except for any base-calling errors
- Note: different sequencing platforms produce different kinds of errors with different frequencies

LIBRARY

- Set of DNA fragments of a particular size, attached to adapters
    - e.g. a 250bp library, or a 3Kbp library

INSERT SIZE

- Length of your fragment with adapters excluded

*This is an Illumina paired-end HiSeq example

PAIRED-END

- Sequencing from both ends of a particular fragment.

MATE-PAIR

- A kind of library that allows you to have large insert sizes (up to 40 Kbp for Illumina sequencing).

KMER

- A short substring of a particular length (k)
- Before contigs can be built, de Bruijn graph assemblers count occurrences of all such substrings
- kmer distribution can give us an estimate of genome size, and repeats
- JELLYFISH is the best known kmer counting program

LONG READS

- PacBio: reads up to ~150 Kbp
- Oxford Nanopore: reads up to 1Mbp!
- Higher error rate than short read sequencing

CONTIG

- Definition from Celera website:
    - A contig consists of a set of reads, a layout that includes all the reads and leaves no gaps, a multiple sequence alignment of the reads, and a consensus sequence. In practice contigs consist of one or more unitigs. Note the consensus may contain (small) gaps spanned by reads even though the layout includes no (0X) gaps.

SCAFFOLD

- Definition from Celera website:
    - A linear ordering of contigs joined by mate pairs. A scaffold defines the order and orientation (DNA strand) for each component contig. There are two ways to measure scaffold length. "Scaffold bases" is sum of contig lengths. "Scaffold span" is that plus the sum of gap lengths. Celera Assembler uses complex criteria to build scaffolds, but some generalizations apply. Every gap in a scaffold was spanned by at least two mate pairs. A gap with negative length means the sequence data and mate data disagree. Usually, negative gaps are small (20bp) and induced by low-quality sequence at the end of a read. In the FASTA representation of a scaffold, negative gaps are represented by a fixed number (20) of N's.

N50

- The contig length such that using equal or longer contigs produces half the bases of the genome.

NG50

- From a set of sorted scaffold lengths, at what contig or scaffold length do we see a sum length that is greater than half of the genome size?

FINISHED GENOME

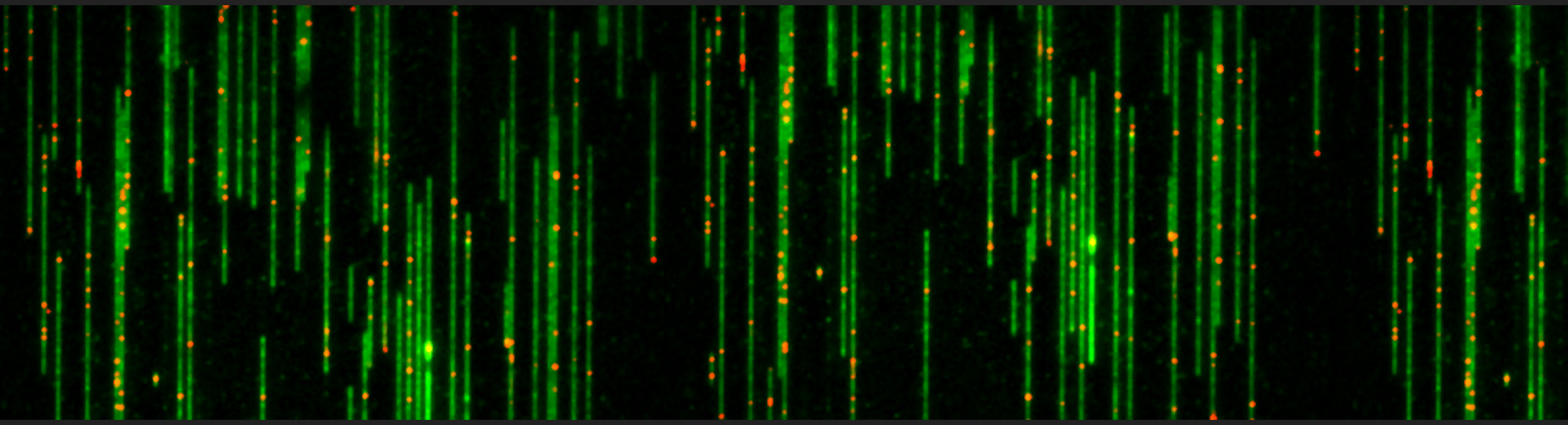
- Assembled to chromosome:
  - Lots of Bacteria and Archaea
  - Arguably no Eukaryotes
    (even the human genome has gaps)

GENOME ASSEMBLY

Platform	Cost	Amount of DNA required	DNA quality required	Assembly output
Illumina short reads	$	•	•	short contigs
Illumina mate pairs	$$$	•	●	short contigs/ longish scaffolds
PacBio	$$$	●	●	long contigs/long scaffolds
Oxford Nanopore	$$	●	●	long contigs/long scaffolds
10X Genomics	$	•	●	short contigs/long scaffolds
Hi-C/Chicago libraries	$$$	●	●	longer scaffolds
BioNano	$$$	●	●	optical map

OPTICAL MAPPING

- BioNano:

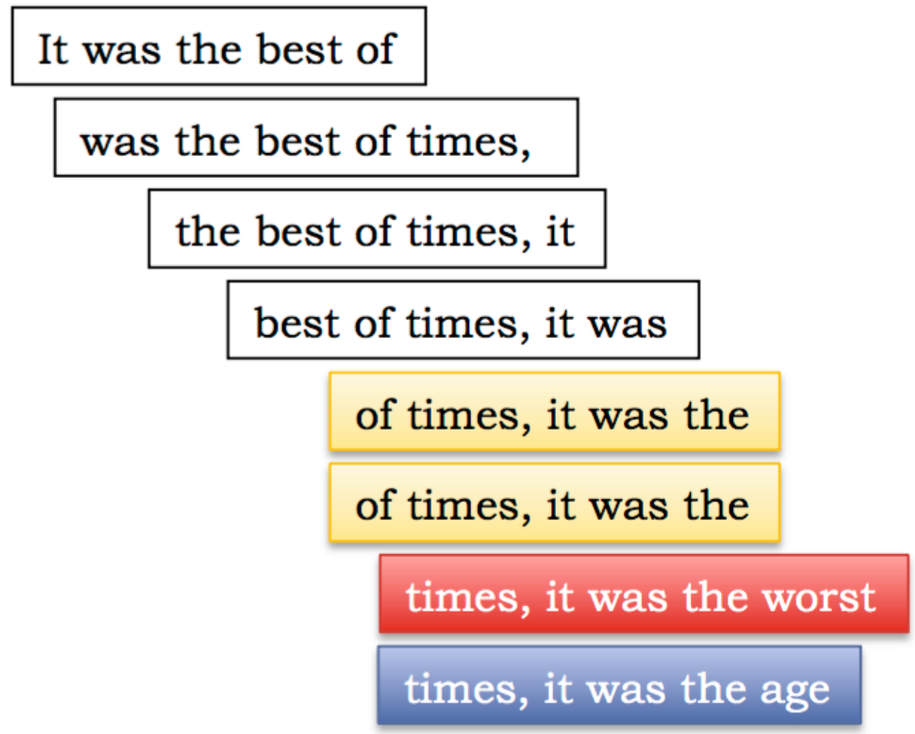


WHY IS ASSEMBLY SUCH A BIG CHALLENGE???

- Mike Schatz (Johns Hopkins) teaches an example using the Charles Dickens novel *A Tale of Two Cities*
- Available on his website: http://schatzlab.cshl.edu/teaching/2014/

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,



The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

http://schatzlab.cshl.edu/teaching/2014/

RAW DATA

↓

TRIMMING

↓

Screen for contaminants??? →

↓

BUILDING CONTIGS

↓

Screen for contaminants??? →

↓

BUILDING SCAFFOLDS

↓

GAP FILLING

↓

ERROR CORRECTION

↓

Screen for contaminants??? →

↓

QUALITY ASSESSMENT

WHICH ASSEMBLER DO I NEED?

- Depends on:
  - Data (sequencing platform, libraries)
  - Genome size
- Compute resources at your disposal

ILLUMINA PAIRED-END ONLY

- DISCOVAR (but only if 2X250bp)
- w2rap-contigger
- SPAdes (but only small genomes: bacteria, archaea, fungi, protists)
- ABySS
- MIRA
- Meraculous
- Velvet

ILLUMINA PAIRED END + MATE-PAIR

- ALLPATHS-LG
- SOAP
- MaSuRCA
- Meraculous
- Platanus

10X GENOMICS: SPECIAL LIBRARY + ILLUMINA READS

- Supernova

MITOCHONDRIAL OR CHLOROPLAST GENOMES

- NovoPlasty
- FastPlast
- MITObim (uses MIRA)
- Velvet
- SPAdes
- ABySS

HIGHLY HETEROZYGOUS GENOMES

- DISCOVAR
- Platanus
- Haplomerger (not actually an assembler, but tries to merge your contigs split apart due to heterozygosity)
- Redundans (similar to Haplomerger)
- OR, GET LONG READS!

PACBIO/NANOPORE ONLY

- Canu
- FALCON (PacBio)
- wtdbg2/Redbean
- STILL NEED SHORT READS FOR POLISHING!!

HYBRID ASSEMBLY

- MaSuRCA
  - lots of specialized error correction plus Celera/FLYE
  - **can include low coverage long reads
- SPAdes

OTHER COMBINATIONS, LEGACY DATA

- MIRA (no PacBio yet, but 454, Sanger, Illumina, Ion Torrent)
- SPAdes (but not for large genomes)

POST-ASSEMBLY PROCESSING

- Scaffolding: SSPACE/SSPACE-longread
- Gap-filling: PBJelly
- Post assembly error correction:
  - Pilon
  - Nanopolish

VISUALIZATION

- JBROWSE
- UCSC genome browser

LET'S GO TO THE TUTORIAL!

- Today, we will visualize read quality with **FastQC**, trim adapters with **TrimGalore!**, count kmers with **Jellyfish**, calculate genome size and heterozygosity with **GenomeScope**, and start a hybrid assembly with **MaSuRCA**.

- The data we are using is from the Red Siskin, an endangered South American bird