GENOME ASSEMBLY

TERMINOLOGY!

READ

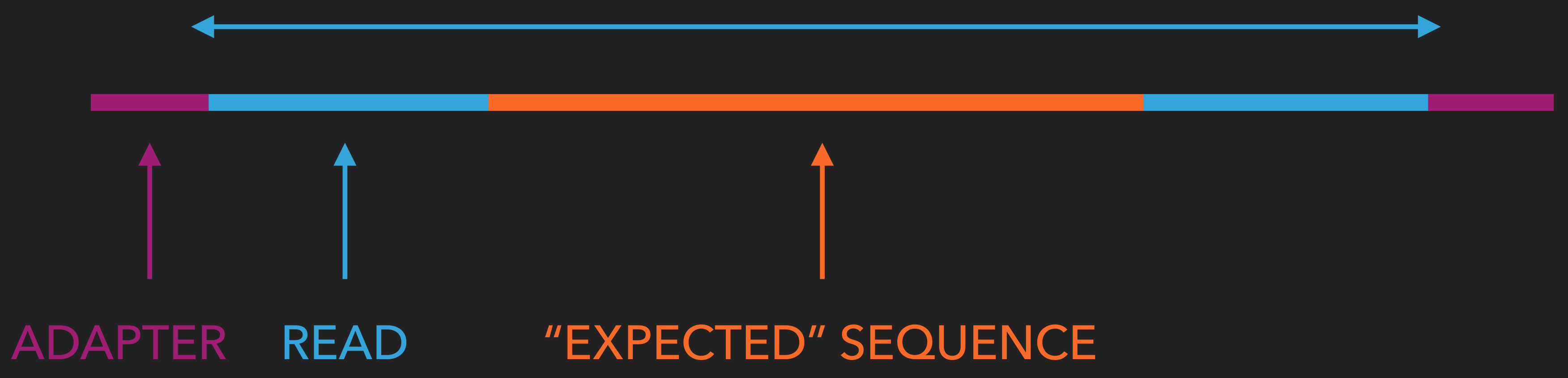
- Raw fragment of DNA as it has been “read” by the sequencer
- These are “real” except for any base-calling errors
    - Note: different sequencing platforms produce different kinds of errors with different frequencies

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.

UNITIG

- Definition from Celera website:
    - A high-confidence contig seed. The end of a unitig is, by definition, a place where the overlap data shows multiple, mutually contradictory, paths. Unitigs are supposed to end at repeats.

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?

L50

- Bradman: the number of sequences evaluated at the point when the sum length exceeds 50% of the assembly size is sometimes referred to as the L50 number. Admittedly, this is somewhat confusing: N50 describes a sequence length whereas L50 describes a number of sequences

FINISHED GENOME

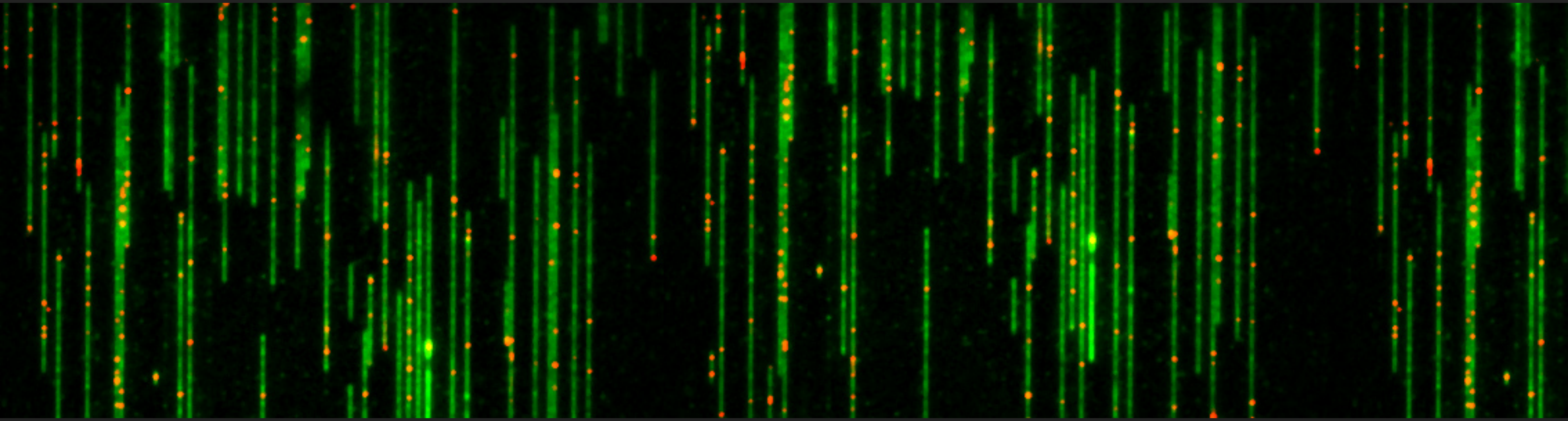
- Assembled to chromosome:
  - Lots of Bacteria and Archaea
  - Arguably no Eukaryotes
    (even the human genome has gaps)

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

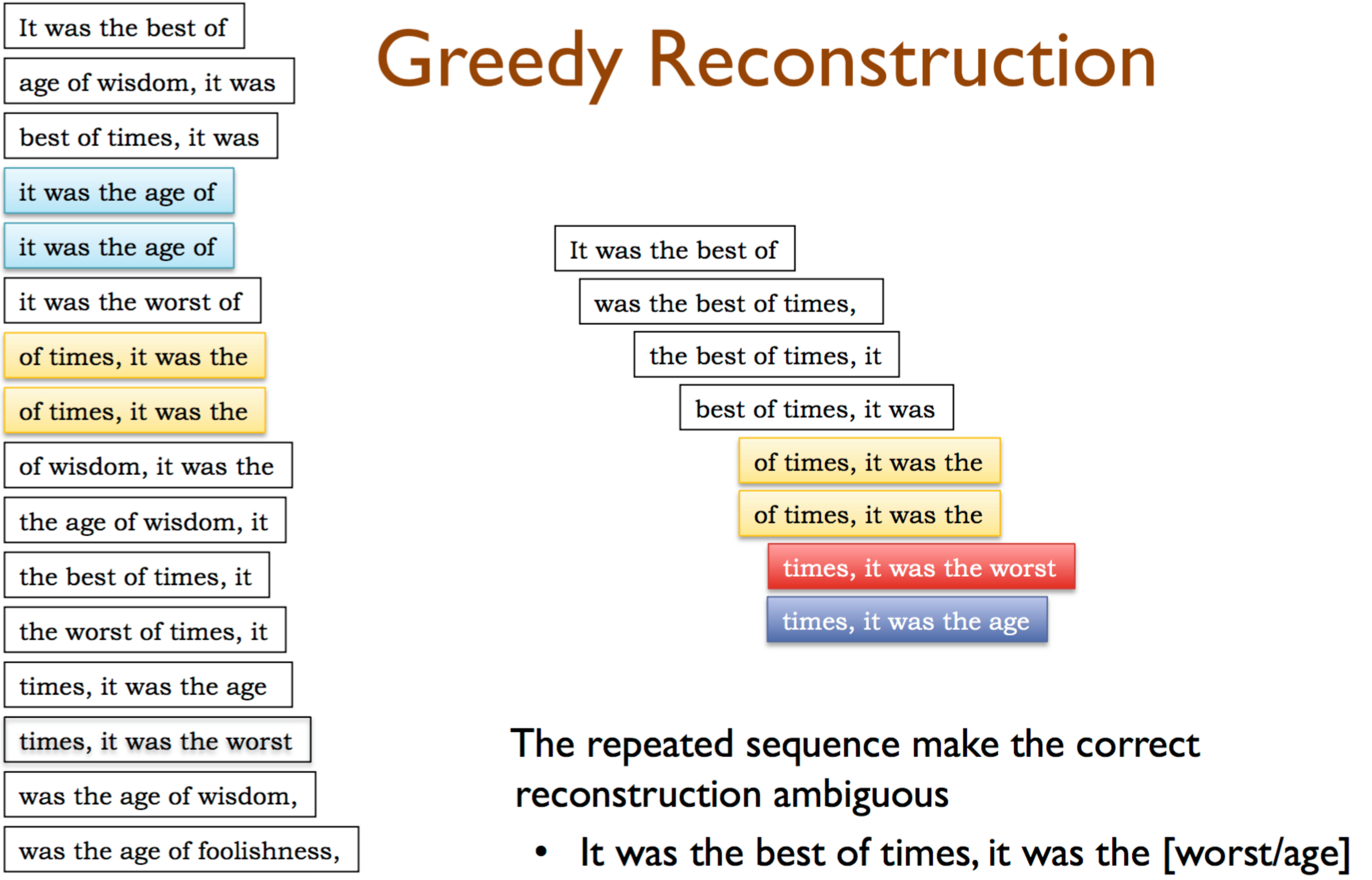
OPTICAL MAPPING

- 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,

It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

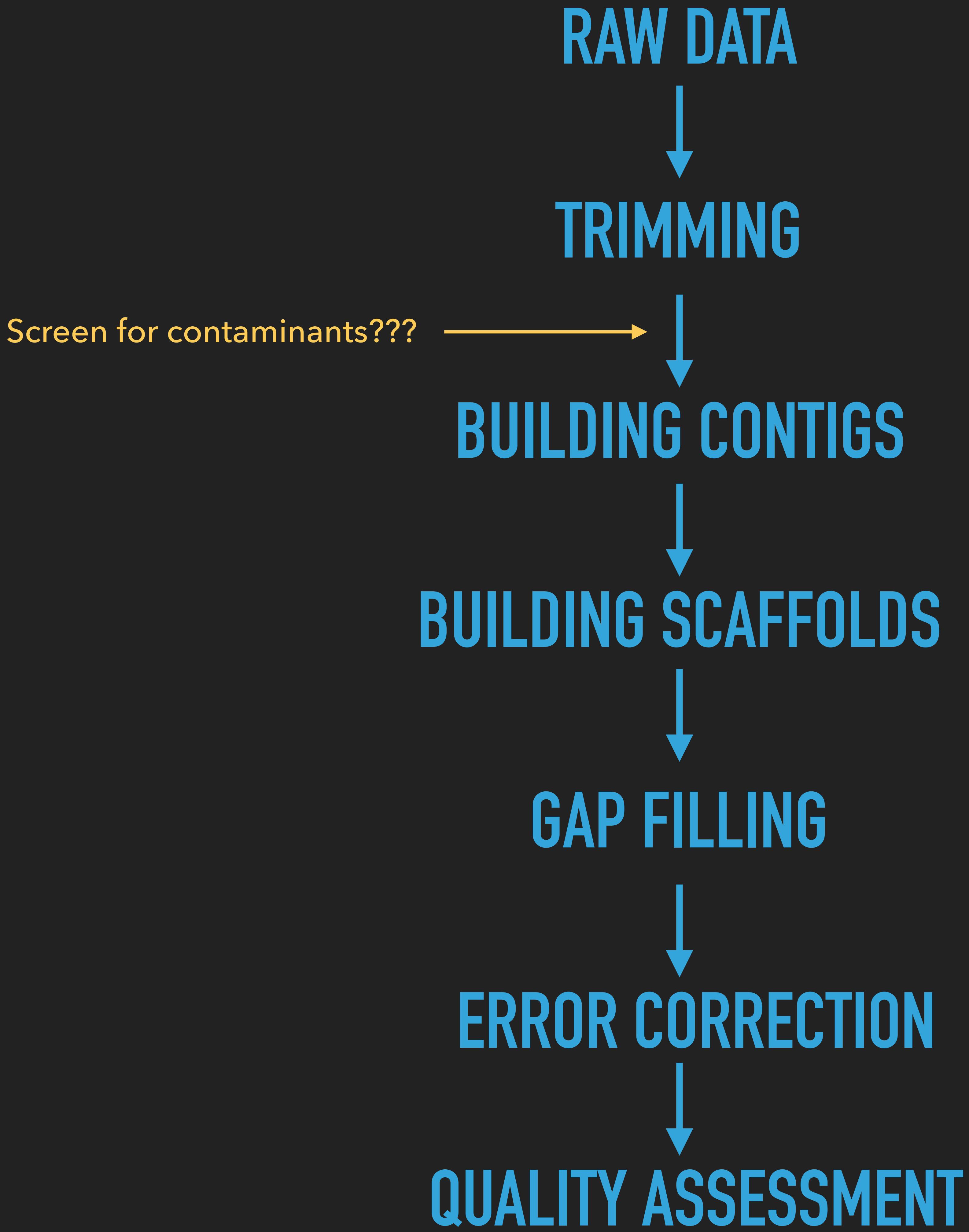
times, it was the worst

times, it was the age

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

BUILDING SCAFFOLDS

GAP FILLING

ERROR CORRECTION

QUALITY ASSESSMENT

OLC ASSEMBLERS VS. DE BRUIJN GRAPH ASSEMBLERS

- Take home message:
  - OLC developed first
  - De Bruijn graph method developed to deal with repetitive bacterial genome
  - De Bruijn methods work better with short fragment data, which took over for a while
  - OLCs are back now, with the infusion of PacBio

OLC

- Overlap, Layout, Consensus
- e.g. Celera (Myers, 2000)
    - Celera most known (and still used) OLC assembler: others are TIGR, Arachne, Newbler, Phrap, PCAP
    - Canu is the fork of Celera assembler still being developed to deal with PacBio and Nanopore data: https://github.com/marbl/canu

DE BRUIJN GRAPH

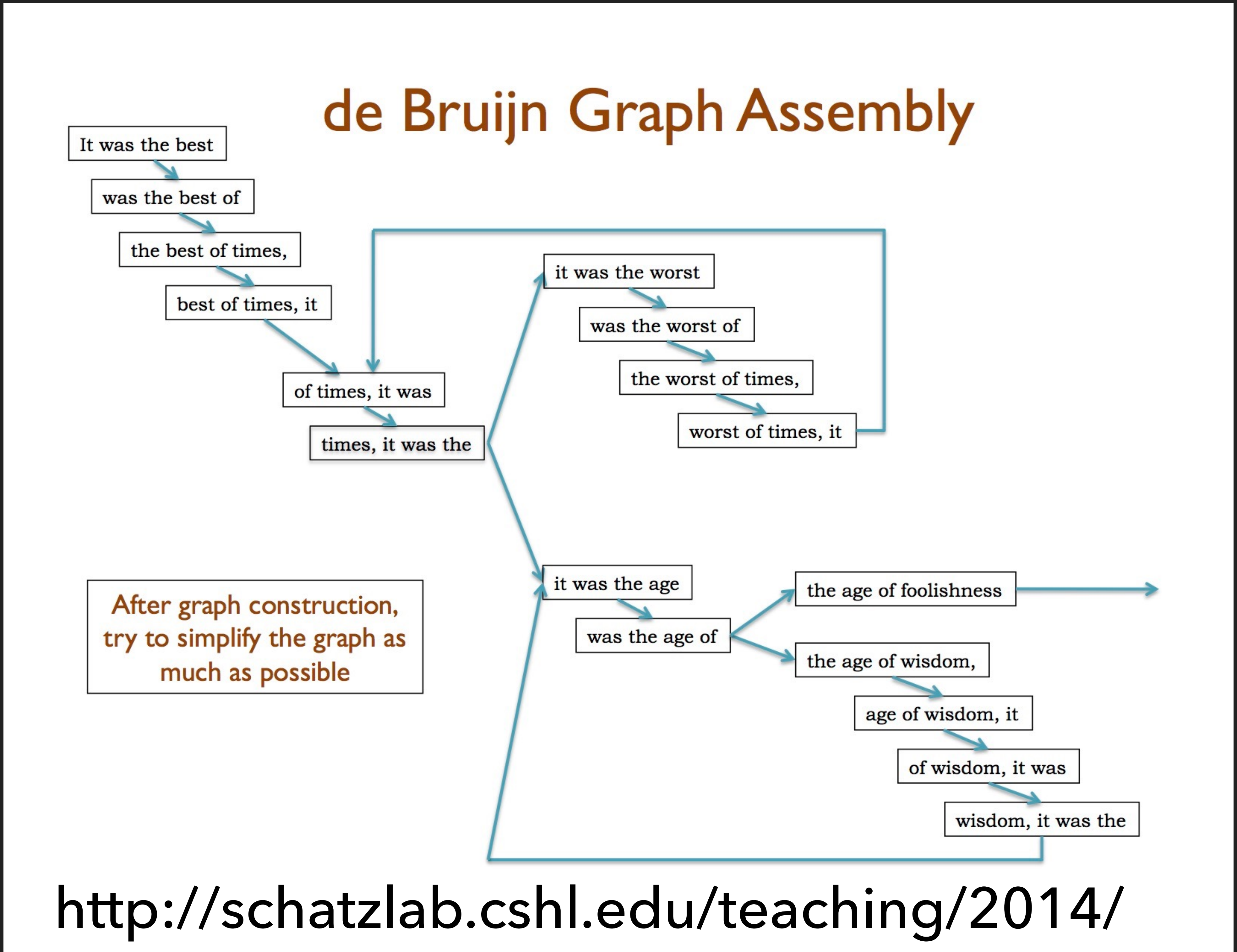
- De Bruijn graph: Pevzner, 2001 (Euler)
  - Most of the assemblers you know use De Bruijn graphs.

OLC VS. DE BRUIJN

- With short reads, overlap consensus assembly suffers from two main problems:
    - short read length means the overlaps must be calculated over a large proportion of the read to retain accuracy
    - the huge number of reads increases the number of links, so that the contig path is difficult to compute.
- The de Bruijn graph approach circumvents the problems of overlap consensus assembly. Rather than using the reads 'as is' and trying to link them, the k-mers (all subsequences of length k within the reads) are computed and the reads are represented as a path through the k-mers. Such a paradigm handles redundancy better than the overlap consensus approach and makes the computation of paths more tractable.

MacLean *et al.* 2009, Nature Reviews Microbiology

DE BRUIJN GRAPH



de Bruijn Graph Assembly

It was the best

was the best of

the best of times,

best of times, it

of times, it was

times, it was the

it was the worst

was the worst of

the worst of times,

worst of times, it

it was the age

was the age of

the age of foolishness

the age of wisdom,

age of wisdom, it

of wisdom, it was

wisdom, it was the

After graph construction, try to simplify the graph as much as possible

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

STRING GRAPH (OLC)

- Myers, 2005

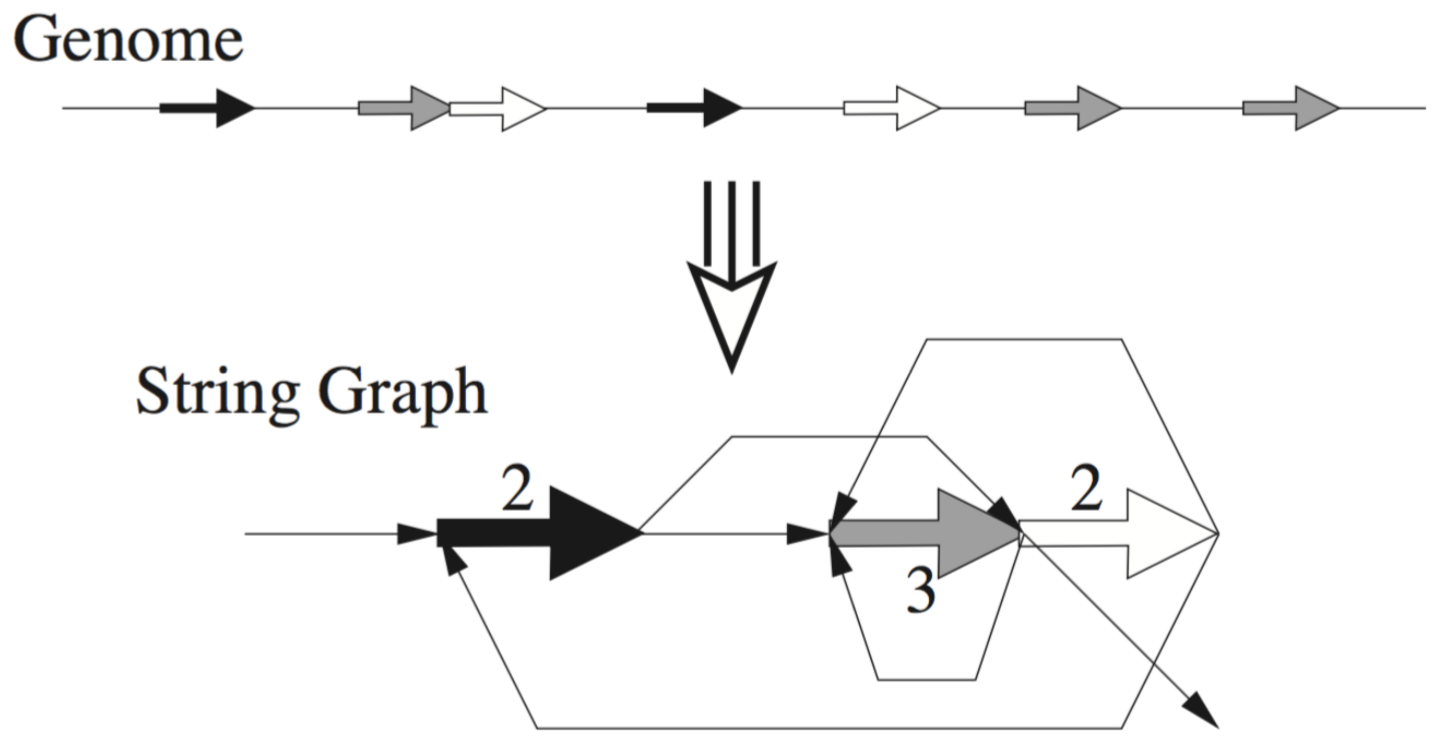


Fig. 1. A genome and its string graph. The thick arrows of the same shade represent identical repetitive sequences. The numbers in the string graph give the number of copies of each repeat inferable by counting entry and exits into the collapsed segment.

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)
- 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

- 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)
- STILL NEED SHORT READS FOR POLISHING!!

HYBRID ASSEMBLY

- MaSuRCA
  - lots of specialized error correction plus Celera (stay tuned for Aleksey Zimin's talk tonight!)
  - **can include low coverage long reads
- SPAdes (not for large genomes)

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

VISUALIZATION

- JBROWSE
- UCSC genome browser

LET’S GO TO THE TUTORIAL!